Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics

Goldhammer, Frank; Martens, Thomas; Lüdke, Oliver

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Goldhammer, F., Martens, T., & Lüdke, 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*, *5*, 1-25. https://doi.org/10.1186/s40536-017-0051-9

Goldhammer *et al. Large-scale Assess Educ (2017) 5:18*
DOI 10.1186/s40536-017-0051-9

Large-scale Assessments in Education

RESEARCH Open Access

Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics

Frank Goldhammer[1*], Thomas Martens[2] and Oliver Lüdtke[3]

*Correspondence: Goldhammer@dipf.de
[1] German Institute for International Educational Research (DIPF)/Centre for International Student Assessment (ZIB), Schloßstr. 29, 60486 Frankfurt/Main, Germany
Full list of author information is available at the end of the article

Abstract

Background: A potential problem of low-stakes large-scale assessments such as the Programme for the International Assessment of Adult Competencies (PIAAC) is low test-taking engagement. The present study pursued two goals in order to better understand conditioning factors of test-taking disengagement: First, a model-based approach was used to investigate whether item indicators of disengagement constitute a continuous latent person variable by domain. Second, the effects of person and item characteristics were jointly tested using explanatory item response models.

Methods: Analyses were based on the Canadian sample of Round 1 of the PIAAC, with $N = 26{,}683$ participants completing test items in the domains of literacy, numeracy, and problem solving. Binary item disengagement indicators were created by means of item response time thresholds.

Results: The results showed that disengagement indicators define a latent dimension by domain. Disengagement increased with lower educational attainment, lower cognitive skills, and when the test language was not the participant's native language. Gender did not exert any effect on disengagement, while age had a positive effect for problem solving only. An item's location in the second of two assessment modules was positively related to disengagement, as was item difficulty. The latter effect was negatively moderated by cognitive skill, suggesting that poor test-takers are especially likely to disengage with more difficult items.

Conclusions: The negative effect of cognitive skill, the positive effect of item difficulty, and their negative interaction effect support the assumption that disengagement is the outcome of individual expectations about success (informed disengagement).

Keywords: Test-taking disengagement, Response time threshold, Explanatory item response modelling, Person effects, Item effects

Background

The validity of inferences based on (average) test scores obtained from large-scale assessments depends heavily on test-takers' engagement when taking the test, that is, the degree to which they were motivated to show what they actually know and can do, in other words, to deliver their maximum performance (Cronbach 1970). However, in

low-stakes assessments such as the Programme for the International Assessment of Adult Competencies (PIAAC) (OECD 2013a), test-takers or groups of test-takers may differ in the effort they exert when taking the test (Wise and DeMars 2005). The negative consequences of this can include, inter alia, the underestimation of respondents' true proficiency levels and the introduction of construct-irrelevant variance (Finn 2015; Haladyna and Downing 2004; Kong et al. 2007; Wise 2015).

Ideally, low test-taking engagement for test instruments administered under low-stakes testing conditions should be avoided. One option is for test administrators to employ strategies that can elicit effort and decrease inattention (Lau et al. 2009). Another option is to give a monetary reward (Braun et al. 2011). However, empirical findings on whether incentives increase test-taking engagement seem to be heterogeneous, dependent on various factors, and also raise ethical issues (Finn 2015).

Alternatively, disengaged responses can be identified after the assessment and taken into account when estimating test scores and population parameters (e.g., Rios et al. 2017). For instance, the effort-moderated IRT model proposed by Wise and DeMars (2006) applies a 3-parameter logistic (PL) IRT model for responses given in the solution behavior mode, while a constant probability model is applied for rapid-guessing behavior. Information on disengagement can also be used to fine-tune the scoring of response behavior. In the PIAAC, fast non-responses that can be understood as disengaged responses were classified as not attempted items, while non-responses taking more than 5 s were considered wrong responses (OECD 2013b).

Regardless of which strategy is chosen—avoiding disengaged responses or dealing with them in the data analysis phase—it is important to understand the process of disengaged responding and related conditioning factors. Therefore, the present study pursued two goals: First, we used a model-based approach to investigate whether behavioral item indicators of disengagement constitute a continuous latent person variable by assessment domain in PIAAC. Second, we tested the joint effects of person and item characteristics on disengagement in PIAAC using explanatory item response models.

Representing differences in test-taking engagement

Previous research has applied approaches other than (continuous) latent variable modelling to represent differences in test-taking engagement. These studies used both model-based and descriptive methods to capture differences in test-taking engagement and took item responses, item response times or both into account.

Schnipke and Scrams (1997) suggested distinguishing between two modes of response behavior: solution behavior, indicating that the test taker is engaged in the task of obtaining a correct response, and rapid-guessing behavior, indicating that the test taker is making quick responses, which can occur because he or she is running out of time, for example. In line with this distinction, the HYBRID model by Yamamoto and Everson (1997) incorporates a mixture of response processes. The regular response process is captured by an IRT model of a particular form, and the random response strategy by an alternative response model in which the (constant) probability of success is independent of ability. Solution behavior is not assumed to be known, but the switching-point from solution behavior to rapid guessing, which may differ across test-takers, is estimated as part of the model. In contrast, the effort-moderated IRT model proposed by

Wise and DeMars (2006) incorporates a variable derived from response time to indicate solution behavior and whether or not the regular IRT model holds for a particular item-person combination. To identify different response modes, Schnipke and Scrams (1997) proposed a log-normal mixture model of item response time assuming two types of response-time distributions, one for rapid guessing and the other for solution behavior (expressed as a bimodal empirical response time distribution). The model has been used to, inter alia, investigate whether the proportion of guessing behavior increases with item position. Meyer (2010) combined the log-normal mixture model with a Rasch mixture model to identify the mode of response behavior using both item response times and item responses. These (mixture) item response models have proven to be beneficial for estimating model parameters accurately in the context of rapid guessing behavior.

Another line of research has directly addressed the degree to which test-takers exert effort when proceeding through a test. To detect low effort in low-stakes testing, Wise and Kong (2005) developed a continuous measure of test taking effort called response time effort (RTE) as the proportion of items completed with solution behavior. Wise and colleague used the effort measure to filter test taker data from the data set (motivation filtering), and investigated beneficial effects on test score reliability and convergent validity.

The approach in the present study expands upon previous work by using a model-based method to define a continuous latent variable of test-taking engagement. Specifically, (item response) measurement models are used to investigate whether binary indicator variables representing (non-)solution behavior for a person and an item constitute a common continuous latent variable. The concept of RTE proposes that test-takers' engagement when proceeding through a test differs continuously. In line with this, Setzer et al. (2013) analyzed binary solution-behavior indicators by means of a hierarchical generalized linear model including random intercepts for person and institution (but without random item intercepts or explanatory person and item variables). Thus, our first goal was to test whether there are actually systematic person differences in disengagement across test items that can be captured by a latent variable. Providing evidence that a measurement model can be established would also justify summing across indicator variables, as is done when computing the RTE measure. If the uni-dimensional 1-parameter logistic (1PL or Rasch) model holds, the sum score accurately represents the 1PL person parameter (Rost 2004). A model-based approach is also beneficial for complex test designs (e.g., multi-matrix design, adaptive test design) such as PIAAC, where different test-takers complete different item sets within a domain, and summing across different sets of engagement indicator variables may not provide comparable measures.

Behavioral indicators of test-taking engagement

Self-report effort measures completed after finishing the test are sometimes used to assess test-taking engagement; however, such measures may have accuracy and validity problems (Wise and DeMars 2005; Wise and Kong 2005). An alternative approach is to infer test-taking engagement directly from test-taking behavior (see Fig. 1). Specifically, engagement as the willingness to deliver maximum performance can be derived from the amount of time taken to complete a task, as the investment of time is a necessary (although not sufficient) condition of completing a task successfully. Note that the relation between task completion time and success can be curvilinear (cf. Fig. 3). Thus, although a minimum amount

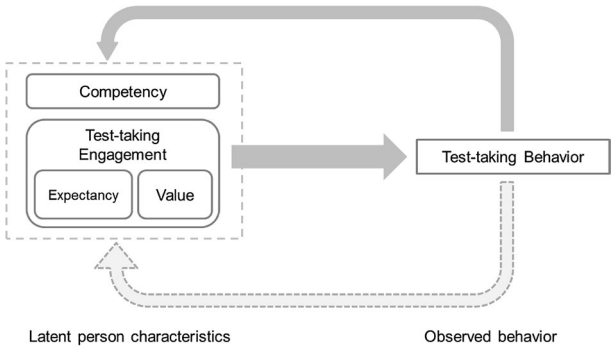

Fig. 1 Test-taking behavior is influenced by both the to-be assessed competency and individual test-taking engagement. Test-taking behavior is used to draw inferences about competency (response data) and can also be used to judge test-taking engagement (response time data). The expectancy of solving an item successfully and the personal value of taking the test are considered antecedents of test-taking engagement

of time is needed to obtain a correct solution, taking much more time can be indicative of failure and quicker responses of greater success (Goldhammer et al. 2014).

Wise and Kong (2005) proposed using item response times to distinguish between solution behavior and rapid guessing behavior (Wise 2017). Following this notion, we assume that engaged item completion (i.e., solution behavior) involves taking at least a certain minimum amount of time required to read and understand the test instructions, process the stimulus' content, and finally give a response, whereas disengaged test-taking behavior means taking less time or guessing rapidly.

Response time thresholds distinguishing between engaged and disengaged responses can be identified in various ways. The three-second rule is commonly used as a constant threshold (Kong et al. 2007; Lee and Jia 2014). The idea of item-specific thresholds relates to the assumption that engaged test-taking behavior is associated with taking a minimum amount of time to be able to respond correctly, and that this amount of time can be assumed to differ across items (Goldhammer et al. 2016). One approach to determine item-specific thresholds is to inspect the response time distribution visually (Kong et al. 2007). The goal is to identify the threshold as the response time at what is judged to be the end point of the short time spike in a bimodal response time distribution. Wise and Ma (2012) proposed an automated way to determine the threshold. Their normative threshold method defines a certain percentage (e.g., 10%) of the average item response time as the threshold and assumes a maximum threshold value of, for instance, 10 s. Lee and Jia (2014) applied another method, previously considered by Ma et al. (2011), to multiple-choice (MC) items. First, the proportion correct conditional on response time was computed for each item. The threshold was defined as the first response time which is clearly associated with a proportion correct greater than the chance level for success (e.g., 25% for a MC item with four response options).

Similarly, the present study obtains item-specific response-time thresholds by conditioning proportion correct on response time. We consider all response behavior with a response time below the threshold as disengaged, that is, rapid responses and rapid non-responses (omissions), while response behavior above the threshold was considered to be engaged, that is, slow responses and slow non-responses (Goldhammer et al. 2016; Wise and Gao 2017). Note that slow (non-)responses are not necessarily engaged responses (see "Discussion" section).

Explaining differences in test-taking engagement

This section outlines how disengaged responses can occur for various reasons at the person level, the item level, or the interaction of both (Finn 2015). The term test-taking engagement already suggests that performance on a test depends not only on ability but also on motivational and emotional aspects (Asseburg and Frey 2013). Differences in test-taking motivation among test-takers influence the degree to which test scores truly reflect individual differences in competence or ability. Some test-takers may not reveal their true competence level simply because they are not motivated to comply with the instructions and, for instance, rush through the test or skip items. Theories of motivated behavior provide a conceptual framework for identifying sources of test-taking engagement. The most prominent of these is expectancy-value theory (Eccles (Parsons) et al. 1983). Thus, a basic model of test-taking motivation and engagement should include the "expectancy" of solving the test item as well as the "value" that the test-taker attaches to solving the test item (see Fig. 1). Note that expectancy and value may be positively correlated, but still additively predict performance, as might the interaction term (see Trautwein et al. 2012).

Person level

Assuming individual differences in test-taking disengagement suggests that there is an individual disposition to be more or less engaged when proceeding through a test. That is, some test-takers consistently give more disengaged answers than others. Identifying the person as a source of variation in disengagement is a descriptive step, and a precondition for explaining individual differences by way of person-level variables, as discussed in this section.

Large-scale assessments are typically experienced as a low-stakes situation (Asseburg and Frey 2013; Sundre and Kitsantas 2004; Wise 2009) in that test-taking behavior does not have consequences for the test taker. Thus, from the expectancy-value theory perspective, the "value" component should be similarly low across test-takers and therefore not related to individual differences in test-taking engagement. However, the perceived expectancy of being capable of solving an item may vary considerably across test-takers (Asseburg and Frey 2013; Cole et al. 2008). A major factor determining expectancy is ability self-concept, that is, one's perceived competence in performing specific tasks. Thus, it can be assumed that test-takers with a more positive self-concept will have more positive expectations and thus higher test-taking engagement than those with a negative self-concept.

In a recent review, Finn (2015) discussed several person-level predictors of low test-taking motivation. Test-takers who were less compliant, that is, less motivated, to take the test tended to show higher levels of reactance (Brown and Finney 2011). Another line of research has shown that boredom negatively affects test-taking effort (e.g., Asseburg and Frey 2013). With regard to gender differences, previous studies suggest that male students tend to exhibit lower levels of test-taking engagement than female students (e.g., DeMars et al. 2013). Similarly, Setzer et al. (2013) demonstrated that females exhibited greater response time effort than males, and those whose primary language is English exhibited greater response time effort than speakers of other languages. Personality measures of agreeableness and conscientiousness have also proven to be positively

related to test-taking effort (DeMars et al. 2013). In a study by Penk et al. (2014), invested effort, that is, the self-reported willingness to engage with test items, was explained by individual differences in task-irrelevant cognition, specifically distraction.

It seems worthwhile to also consider findings from the field of missing data since disengagement is defined in the present study as including both rapid responding and skipping items rapidly. In fact, omitted responses are often at least partially due to a lack of test-taking motivation (Jakewerth et al. 1999; Wise and DeMars 2005). Latent variable modelling of omission propensity has revealed a negative relation with ability, that is, stronger test-takers omit fewer items (Holman and Glas 2005; Pohl et al. 2013). Köhler et al. (2015) provided some evidence that people without a migration background and with higher levels of education exhibit a lower omission propensity.

Item level

According to expectancy-value theory, a major determinant of the expectancy component at the item level is the test-taker's estimate of item difficulty (Eccles (Parsons) et al. 1983). More specifically, if perceived item difficulty is high relative to one's competence, test-taking engagement will be negatively affected. Wolf et al. (1995) demonstrated that performance differences between more highly and less motivated students can be specifically explained by item difficulty (p value), the degree to which an item is mentally taxing (expert rating), and fatigue (item position). Thus, these findings suggest that differences in test-taking engagement have a particularly strong negative effect on performance for difficult and mentally taxing items presented late in the test. Relatedly, Asseburg and Frey (2013) showed that test-taking effort was higher when items had only moderate difficulty relative to ability. In the study by Penk et al. (2014), self-reported effort invested into completing test items was accounted for by the test's perceived attractiveness, reflecting how much fun one had when taking the test, and perceived usefulness.

As suggested by the study by Wolf et al. (1995), an item's position can be assumed to be another determinant of differences in test-taking engagement. Research on item position effects has shown that an item presented at a later position in the test is more difficult than when presented at the beginning (e.g., Debeer et al. 2014). This common phenomenon is usually explained by a decrease in test-taking motivation and/or an increase in fatigue. Increased item difficulty may also be due to more and more rapid guessing towards the end of a timed test (Schnipke and Scrams 1997). Setzer et al. (2013) showed that items presented later in the test and including more text as well as ancillary reading material are more strongly associated with test-taking disengagement. Other potential item-level predictors of disengagement are suggested by research on response omissions. For instance, there is evidence that item difficulty increases the probability of omitting an item (Stocking et al. 1988).

Person and item level

Following expectancy-value theory, the present study particularly focusses on interactions between person- and item-level factors that may influence expectations about completing an item successfully. Specifically, we assume that disengaged responding depends on the test taker's ability self-concept and perceived item difficulty, and is

thus the outcome of an informed decision process. Given the positive relation between self-concept and corresponding ability (Eccles (Parsons) et al. 1983; Marsh and Craven 2006) as well as between perceived item difficulty and actual item difficulty (e.g., Wolf et al. 1995), we predict that individual differences in cognitive ability and differences in item difficulty explain test-taking engagement. Moreover, we assume that the positive effect of item difficulty on test-taking disengagement is weaker for more able test-takers because the relative item difficulty is lower for them.

Research goals and hypotheses

Overall, the present study aimed to investigate the conditioning factors of disengaged responses observed in the PIAAC domains of literacy, numeracy, and problem solving. Thereby, we hope to shed some light on whether the process of disengagement is erratic or instead systematic or even strategic (i.e., informed disengagement). This was done by pursuing two related goals.

We first addressed the question of whether disengaged responses across items and by domain can be explained by a single latent person variable, which would suggest that each individual test-taker is engaged or disengaged to a consistent degree when taking a test. If a continuous latent person variable can be defined using a uni-dimensional measurement model for each domain, the next step is to investigate whether these individual differences are similar across domains. Therefore, we also explored the correlational structure of disengagement across the domains of literacy, numeracy, and problem solving.

Second, we investigated the joint effect of person and item characteristics on disengagement using explanatory item response models. At the person level, we tested the effects of educational attainment, language, gender, age, and cognitive skill on individual test-taking disengagement in literacy, numeracy, and problem solving. Given previous research, we expected that educational attainment, fluency in the test language, being female and cognitive skill would be negatively related to test-taking disengagement. We had no hypothesis with regard to age. We further investigated how item characteristics affect disengagement. Given previous findings, we expected items completed in the second part of the PIAAC assessment to be associated with significantly greater disengagement than the same items completed in the first part. Furthermore, in accordance the informed disengagement hypothesis presented above, we assumed that disengagement would increase with item difficulty across all three domains. This effect was hypothesized to be moderated by individual cognitive skill such that the negative effect of item difficulty on disengagement would be smaller for stronger participants. Put simply, strong test-takers stay engaged when they encounter difficult items, whereas poor test-takers give up quickly.

Several previous studies have addressed how to represent differences in test-taking disengagement and how these are related to person and item characteristics. The present study adds to them by focusing on the adult population as assessed in the Canadian sample of the PIAAC. Furthermore, unlike previous studies, we propose a model-based approach assuming both (random) person and item effects on test-taking disengagement, and incorporate explanatory variables at both the person and item levels as well as their interaction to test the hypothesis of informed disengagement.

Methods

Sample

The target population for PIAAC 2012 (Round 1) consisted of all non-institutionalized adults between the ages of 16 and 65 (inclusive) residing in each country (meaning that their usual place of residency is in that country) at the time of data collection. To address our research goals, we selected the largest PIAAC sample from Round 1, which was Canada with $N = 26{,}683$ participants. The public use file (PUF) was downloaded from the OECD webpage on 24 October 2015. Canada published age information in bands of 10 years: 17.30% of the participants were 24 years old or younger, 17.10% were 25–34 years old, 20.10% were 35–44 years old, 23.30% were 45–54 years old, and 22.10% were 55 years old or older. Among all Canadian participants, 46.60% were male and 53.40% female. Only the Canadian subsample that completed the computer-based assessment ($N = 20{,}923$) was included in the present analysis because item response times were not available for the paper-based assessment. Table 1 shows the distributions of the person-level variables used in the explanatory models for this subsample. Since the present study did not seek to describe features of the population of Canada or compare populations (see Goldhammer et al. 2016), PIAAC sampling weights were not included in the analyses.

PIAAC test design

The PIAAC test design (OECD 2013b) assumed 60 min of testing time for the cognitive assessment. However, no time constraint was imposed; that is, some participants

Table 1 Distribution of person characteristics in the Canadian subsample completing the computer-based assessment (N = 20,923)

Predictor	*N*	*%*
Educational attainment		
Less than high school	2719	13.40
High school	4693	23.13
Above high school	11,212	55.25
NA	2353	11.60
Test language same as native language		
Yes	16,804	82.81
No	4118	20.29
NA	1	0.00
Score cognitive pre-test		
3	495	2.44
4	1581	7.79
5	5721	28.19
6	13,126	64.68
Gender		
Male	9573	47.17
Female	11,350	55.93
Age group		
24 or less	4101	20.21
25–34	4033	19.87
35–54	9101	44.85
55 or more	3688	18.17

NA not available

were expected to take longer. Participants first completed the background questionnaire (BQ), which asked, inter alia, about their computer experience, which was crucial to route test-takers to either the paper-and-pencil or computer-based assessment (CBA; see Fig. 2). Participants with no computer experience were given the paper-based assessment, as were participants who refused to take the assessment on the computer.

Participants in the computer-based condition had to pass two short tests taking about 5 min each (CBA Core Stages 1 and 2). Participants who failed these test assessing basic Information and Communication Technology (ICT) skills (CBA Core Stage 1) were rerouted to the paper-based core section. Participants who succeeded in the first task but failed the following cognitive pre-test (CBA Core Stage 2) with three literacy and three numeracy items subsequently took only the paper-based reading components. Participants who successfully completed both pre-tests were randomly assigned to one of three possible types of computer-based cognitive assessments, each consisting of two modules (see grey boxes in Fig. 2) that took about 50 min in total: (i) 50% took a random combination of the literacy (Lit) and numeracy (Num) items (Lit-Num or Num-Lit), (ii) 33% were assigned randomly to either the literacy or numeracy items plus one of the two sets of problem solving (PS) items (Lit-PS2, Num-PS2, PS1-Lit or PS1-Num), and (iii) 17% completed only the two sets of problem solving items (PS1-PS2). Only those participants who took the CBA modules were included in the present study.

Literacy and numeracy were assessed using a two-stage adaptive test design. That is, each module included two stages, each of which consisted of various testlets differing in difficulty (three testlets at Stage 1, four testlets at Stage 2). The selection of the testlet for Stage 1 depended on participants' scores in the short cognitive pre-test (three literacy and three numeracy items), language, and educational attainment; for Stage 2, the score obtained in Stage 1 was used as an additional selection criterion.

Overall, 49 literacy items, 49 numeracy items and 14 problem solving items were administered in the assessment. A test-taker completing the adaptive literacy and numeracy modules was required to respond to 20 items (9 items in Stage 1 and 11 items in Stage 2). Each of the two problem solving modules (PS1, PS2) consisted of 7 items. Thus, test-takers completed a total of 40 items (Lit, Num), 27 items (PS1 or PS2 combined with Lit or Num), or 14 items (PS1, PS2) over the course of the cognitive assessment. In the present study, all possible combinations of item sets across domains within the CBA were included in the analysis.

It follows from the test design that the literacy and numeracy items were administered at two positions (Module 1 and Module 2) in balanced order, with the order of items fixed within each module; for problem solving, the order of content was also fixed across modules. Thus, random equivalent groups completed the literacy items and numeracy items in the first or the second part of the cognitive assessment, respectively, while there was only one order for the two sets of problem solving items.

Indicator of test-taking disengagement P +>0%

Definition

In line with previous research (e.g., Lee and Jia 2014; Wise 2006), we used item-specific response time thresholds to distinguish between engaged and disengaged responses. The proportion correct greater than zero (P +>0%) method was applied to obtain the

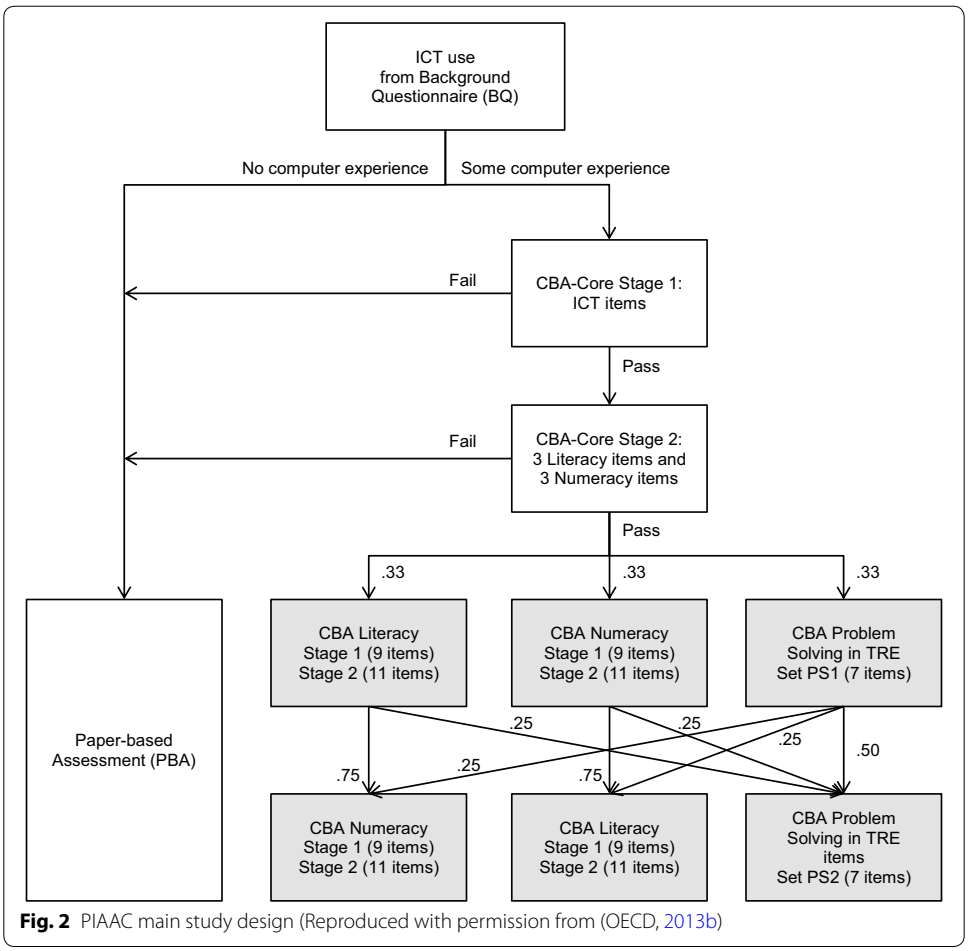

Fig. 2 PIAAC main study design (Reproduced with permission from (OECD, 2013b)

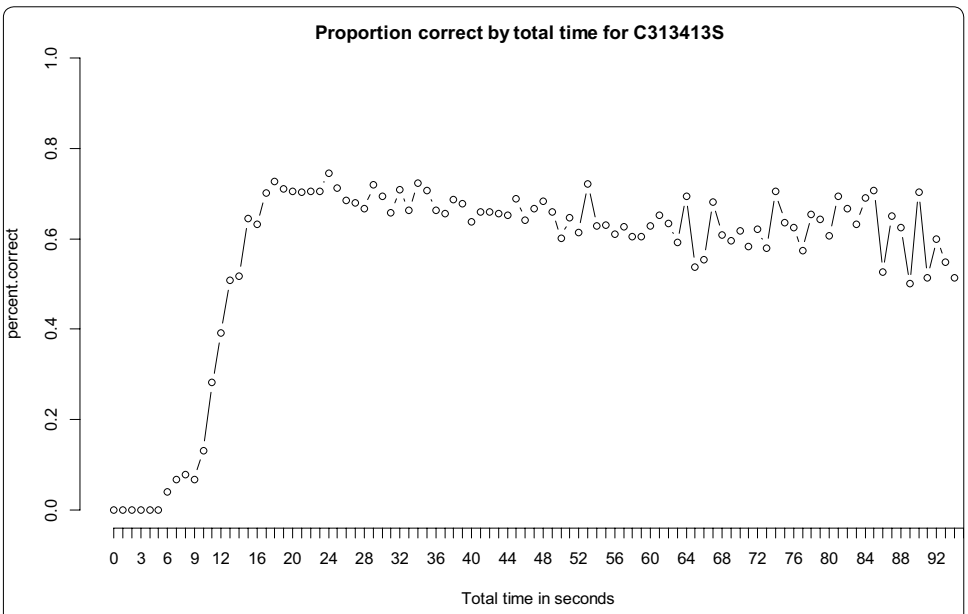

Fig. 3 Proportion correct conditional on response time (total time on task, truncated at the 95th percentile) for the sample literacy item C313413, completed in Module 2

thresholds (Goldhammer et al. 2016). This is an adapted version of Lee and Jia's (2014) method conditioning the proportion correct (P +) on response time to determine the threshold as the shortest response time where conditional P + first exceeds chance level. For the PIAAC items, we assumed that the chance level of obtaining a correct response is zero. This seems justifiable because almost none of the response formats allow for rapid correct guesses. There are only five MC-like numeracy items; all other items require participants to enter numbers or interact with the stimulus, for instance, by highlighting text or clicking on a graphical element. The assumption that the rate of rapid correct guesses is negligible is supported by Goldhammer et al. (2016), who found that the average proportion correct for response behavior taking less than 3 s was 1% for literacy, 4% for numeracy, and 0% for problem solving.

To determine the P +>0% threshold, the proportion correct conditional on the response time was computed item by item at one second intervals. The threshold was identified as the shortest response time associated with a proportion correct of greater than zero. Figure 3 shows the proportion correct conditional on response time for a sample item. When the response time hits the threshold of 6 s, the probability of success becomes greater than zero (P +>0 threshold). Thus, response behavior taking 6 s or longer was classified as engaged, while response behavior taking less than 6 s was regarded as disengaged.

To classify response behavior as engaged or disengaged, all items visited by the test-taker and for which response times were available were considered. Whether a response was given or not (omission) was not relevant for the classification; instead, we sought to determine how engaged the test-taker was as reflected by the time spent on the item. Thus, omissions with a response time below the threshold were classified as disengaged, while those with a response time above the threshold were classified as engaged.

Empirical properties

Since item position (Module 1 vs. Module 2) could have an impact on the location of the response time threshold (e.g., due to fatigue effects), the response time thresholds for literacy and numeracy items were determined by module (Goldhammer et al. 2016). While there were some differences between the two modules, there was high consistency overall, as indicated by the cross-module correlation of $r = 0.92$ ($p < 0.001$) for literacy and $r = 0.63$ ($p < 0.001$) for numeracy. The difference in average response time thresholds was small (Module 1 vs. Module 2: 6.51 s vs. 6.98 s for literacy, 2.27 s vs. 2.16 s for numeracy). For literacy, thresholds varied between 1 and 26 s for Module 1 and between 1 and 33 s for Module 2; for numeracy, they varied between 1 and 8 s for Module 1 and 0 and 8 s for Module 2; for problem solving, between 3 and 76 s.

As shown by Goldhammer et al. (2016), the proportion correct for some items was greater than 0% for all empirical response time intervals. In this case, all responses were considered to be engaged responses. This concerned several numeracy items (Module 1: 28.57%, Module 2: 24.49%), a few literacy items (Module 1: 4.17%, Module 2: 2.04%) and no problem solving items.

Goldhammer et al. (2016) investigated whether the P +>0% indicator of disengaged response behavior can be considered valid, which requires the indicator to identify responses with no chance for success (e.g., due to rapid guessing or because the time

spent on the item was below the minimum allowing for success above chance level). Following the procedures described by Lee and Jia (2014), they determined the average proportion correct for engaged versus disengaged response behavior across items (and additionally by item) for each construct, as well as the correlations between score group and proportion correct for engaged and disengaged response behavior by item. One example of their findings is that the proportion correct for engaged responses in literacy, numeracy, and problem solving were 0.56, 0.63 and 0.43, respectively, in comparison to 0 for disengaged responses by definition according to the P +>0% method. Compared to other methods specifying constant thresholds of 5000 or 3000 ms, or item-specific thresholds obtained by inspecting visually the (bimodal) response time distribution, the P +>0% method resulted in the greatest difference in proportion correct for engaged versus disengaged responses, suggesting that the method separates disengaged and engaged responses very well. Taken together, these validity checks indicate that the P +>0% indicator can validly be interpreted as a measure of test-taking disengagement.

Explanatory variables

Differences in test-taking engagement were explained by the following person-level variables: gender ("male" and "female"; PUF variable GENDER_R); age group ("Aged 24 or less", "Aged 25–34", "Aged 35–54", and "Aged 55 or more"; PUF variable AGEG10LFS, collapsing the age groups "Aged 35–44" and "Aged 45–54" into "Aged 35–54"); educational attainment ("Less than high school", "High school", and "Above high school"; PUF variable B_Q01a_T); native language ("Test language same as native language", and "Test language not same as native language"; PUF variable NATIVELANG); as well as score on the cognitive pre-test (PUF variable CBA_CORE_STAGE2_SCORE) as an indicator of cognitive skill. Furthermore, to investigate a potential position effect on test taking engagement, we included a variable indicating whether literacy and numeracy items were completed in Module 2 ("LIT", and "NUM", PUF Variable CBAMOD2). Finally, we used RP67 difficulties (i.e., items are located on the scale where they have a 67% probability of being completed successfully in the target population) as provided in the PIAAC technical report (OECD 2013b) as an item variable. Item difficulties were rescaled (divided by 100) to facilitate model estimation. The (interacting) variables cognitive pre-test score and item difficulty were centered when testing the explanatory item response models to ease the interpretation of effects.

Data analysis

To address the first research goal, we tested a 1-parameter logistic (1PL) item response model for each construct with dichotomous disengagement indicators (0 = engaged, 1 = disengaged) as item response variables. To judge the item fit, we inspected information-weighted (Infit) and unweighted (Outfit) mean squared residual-based item fit statistics. As a rule of thumb, an Infit and Outfit between 0.5 and 1.5 can be considered acceptable (de Ayala 2009; Wright and Linacre 1994). The Infit is sensitive to unexpected responses in items located close to the person parameter, while the Outfit is sensitive to unexpected responses in items located away from the person parameter (i.e., very difficult or easy items for a person). Items with a value smaller than the lower bound of 0.5 are typically not excluded since this indicates overfit (i.e., observations can be better

predicted by the model than expected). In addition, we visually inspected whether the model-expected item characteristic curve fit the (non-parametric) observed item characteristic curve. Specifically, we checked whether the observed curves show humps, non-monotonicity or an unexpected asymptote (Douglas and Cohen 2001).

For literacy and numeracy, a latent regression model for the person parameter was incorporated to take the adaptive two-stage test design into account. Including the background variables that served as selection criteria in the adaptive design (i.e., educational attainment, language, score in the cognitive pre-test) makes the assumption that the not-administered items were missing at random (MAR) more justifiable. If the propensity for disengaged responses would be related to one of the selection criteria, but the latter were not included in the model, the MAR assumption would be violated and the parameter estimates biased. For literacy and numeracy, the Stage 1 score was additionally used to select the testlet at Stage 2. However, we did not include Stage 1 score in the model because it was highly correlated with the other selection criteria and we wanted to keep the background model constant across the three domains.

Furthermore, we also sought to test a three-dimensional 1PL model with between-item multidimensionality to explore the correlational structure of construct-specific latent engagement variables for literacy, numeracy, and problem solving. However, this model exhibited estimation problems and did not converge using numerical integration or quasi Monte Carlo integration, probably due to its complexity and the low proportion of disengaged responses for many items. To recover the latent disengagement correlations, we used plausible values from the three uni-dimensional models. In a first step, a uni-dimensional model was tested for each domain, and expected-a-posteriori (EAP) estimates were obtained as person parameter estimates. In a second step, a uni-dimensional model was tested for each domain with the EAP estimates of the other two domains as predictors in the latent regression model. Ten plausible values were drawn for each person on the basis of these domain-specific measurement models. Afterwards, correlations among domains were computed for each of the ten plausible values, and these were averaged in a final step.

To test the joint effects of person and item characteristics on disengagement, we applied explanatory item response models using the generalized linear mixed modelling (GLMM) framework (De Boeck et al. 2011; Doran et al. 2007). The model explains the logit for the probability of making a disengaged response for person p and item i with the effects of K person covariates and H item covariates as well as their interaction:

$$logit\left(P\left(Y_{pi}=1\right)\right)=\beta_0+\sum_{k=1}^{K}\gamma_k X_{p,k}+b_{0p}+\sum_{h=1}^{H}\gamma_h Z_{i,h}+b_{0i}+\omega Z_{i,1}X_{p,1} \qquad (1)$$

where β_0 denotes the fixed intercept, γ_k the fixed effect of person covariate $X_{p,k}$, γ_h the fixed effect of item covariate $Z_{i,h}$, ω the fixed interaction effect of an item covariate $Z_{i,1}$ (i.e., item difficulty) and a person covariate $X_{p,1}$ (i.e., cognitive skill), b_{0p} the (residual) random person intercept (person disengagement), and b_{0i} the (residual) random item intercept (item easiness with regard to disengagement). A normal distribution was assumed for the random item and person intercepts, with a mean of zero, $b_{person}\sim N\left(0, Var\left(b_{0p}\right)\right)$, and $b_{item}\sim N(0, Var(b_{0i}))$.

To address the second research aim, we first tested Model 1 with person characteristics $X_{p,k}$ as predictors, that is, educational attainment, language, gender, age group, and cognitive skill. We then tested Model 2, which included only item characteristics $Z_{i,h}$ as predictors, that is, module position (for numeracy and literacy only) and item difficulty. Finally, the full Model 3 was tested as given in Eq. (1) with all person and item characteristics and the interaction of item difficulty with cognitive skill. We also tested a baseline model, Model 0, to investigate the extent to which the predictor variables reduce the variances of the random person and item intercepts in order to determine their explanatory power (effect size). Note that only Model 1 and the final Model 3 account for the adaptive two-stage test design by including the adaptive selection criteria.

All models were estimated in the R environment (R Core Team 2016). The TAM package (Kiefer et al. 2016) was used for scaling and dimensionality analysis; it performs a marginal maximum likelihood estimation of the model parameters. The 1PL models tested assume uni-dimensionality and equal discriminations across items (by constraining them to one). The glmer function from the lme4 package (Bates et al. 2015) was used to test explanatory item response models (GLMMs). The maximum likelihood estimation method in lme4 utilizes a Laplace approximation.

Results

Measurement models for test-taking engagement

Before testing measurement models, we computed the proportion of disengaged responses by item. The following analyses only include those items that exhibited variation in disengagement. Forty-eight of 49 items for literacy were included, 37 of 49 items for numeracy, and 14 of 14 items for problem solving. The proportion of disengaged responses in the remaining items varied between 0.12 and 18.67% for literacy, between 0.04 and 5.51% for numeracy, and between 0.13 and 27.77% for problem solving.

Testing a 1PL item response model for the literacy disengagement indicators revealed an acceptable item fit for almost all items. The Infit statistic was between 0.68 and 1.47 for all items, while all Outfit values were between 0.04 and 1.40, with the exception of one item at 10.02. This item was dropped for the following analyses. Comparing the expected and observed item characteristic curves showed that the model fit the data quite well (see sample items in Fig. 4, top).

For numeracy, the Infit statistic obtained from the unidimensional 1PL model was between 0.69 and 1.37 for all items. The Outfit statistic was between 0.01 and 1.29 for all items except for two with values of 2.23 and 14.67. These two items were excluded from the following analyses. Comparing the expected and observed item characteristic curves indicated that the model fit the data acceptably (see sample items in Fig. 4, middle). Observations were only available for the lower part of the item characteristic curve for most items, indicating that items were very difficult and were completed by most participants in the mode of engagement.

Finally, the 1PL item response model for the disengagement indicators for problem solving had also an acceptable fit. The Infit statistic varied between 0.76 and 1.29, and the Outfit statistic between 0.07 and 1.53. No items were dropped. Comparing expected and observed item characteristic curves also suggested that the model fit the data (see sample items in Fig. 4, bottom).

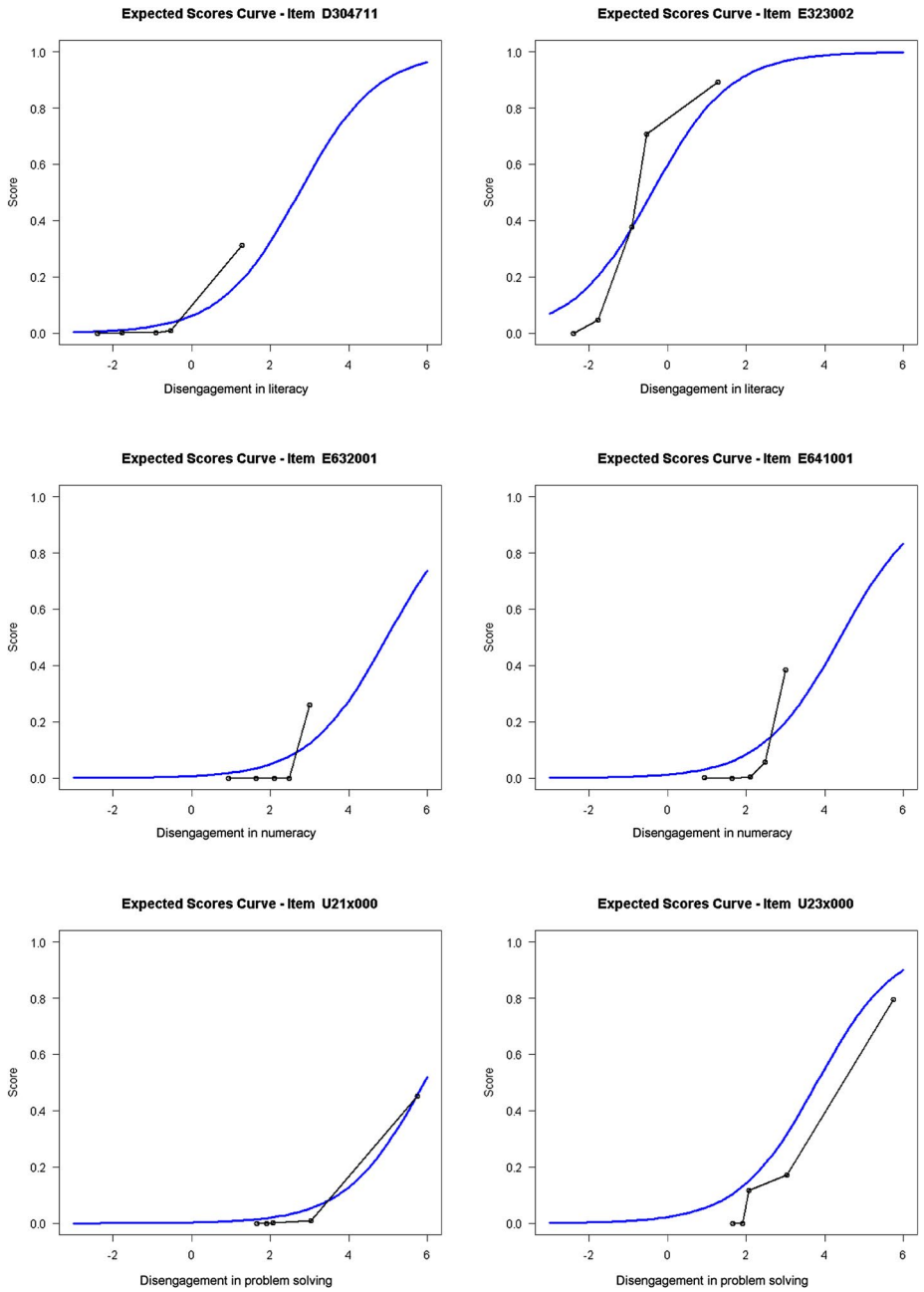

Fig. 4 Model-expected and observed item characteristic curves of two selected items by domain (top: literacy, middle: numeracy, bottom: problem solving)

Using plausible values, the average correlations of literacy disengagement with numeracy and problem solving disengagement were 0.43 and 0.50, respectively, whereas the correlation between numeracy and problem solving disengagement was 0.46. These medium-sized correlations suggested that test-taking disengagement represents a domain-specific construct.

Explaining differences in test-taking disengagement

Literacy disengagement

The results for Model 1, testing the effects of person variables on test-taking disengagement in literacy, are presented in Table 2. Higher educational attainment was associated with lower disengagement. In addition, the main effect of cognitive skill was negative and significant, suggesting that stronger test-takers exhibit a lower level of disengagement. Disengagement was higher among test-takers whose native language was a language other than the test language. There were no significant effects of gender or age group. Model 2 reveals the relation between item properties and disengagement. Items completed later, that is, in Module 2, were associated with higher disengagement. Item difficulty also had a positive effect on disengagement. The full Model 3 reproduces the effects of the person and item variables. Contrary to expectations, the effect of item difficulty was not significantly attenuated among strong test-takers, as indicated by the nonsignificant negative interaction between item difficulty and cognitive skill.

The decrease in variance for the random person intercept, $Var\left(b_{0p}\right)$, from Model 0 to Model 3 was 21.32%, while it was 31.11% for the random item intercept, $Var(b_{0i})$. Thus, the predictors explained a substantial portion of the variance in literacy disengagement.

Numeracy disengagement

Table 3 presents the results for Model 1, testing the relation between person characteristics and test-taking disengagement in numeracy. The highest level of educational attainment was associated with lower disengagement, while this was not the case for a medium educational level. Disengagement was greater among participants for whom the test

Table 2 Explanation of test-taking disengagement in literacy (N = 14,039)

Predictor	Model 0		Model 1		Model 2		Model 3	
	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Intercept	− 14.82***	(0.15)	− 14.03***	(0.26)	− 14.90***	(0.17)	− 14.11***	(0.27)
Educational attainment								
High school			− 0.70**	(0.21)			− 0.77***	(0.22)
Above high school			− 1.35***	(0.23)			− 1.48***	(0.23)
Language (test language not same as native language)			0.73***	(0.18)			0.83***	(0.18)
Score cognitive pre-test			− 0.80***	(0.09)			− 0.86***	0.09)
Gender (female)			− 0.01	(0.14)			0.02	(0.14)
Age group								
25–34			0.29	(0.24)			0.34	(0.24)
35–54			0.27	(0.20)			0.30	(0.21)
55 or more			0.22	(0.24)			0.24	(0.24)
Position (module 2)					1.27***	(0.14)	1.43***	(0.15)
Item difficulty					2.98***	(0.20)	3.00***	(0.20)
Item difficulty × score cognitive pre-test							− 0.07	(0.05)
$Var\left(b_{0p}\right)$	132.33		121.37		116.43		104.11	
$Var(b_{0i})$	0.45		0.44		0.33		0.31	

SE standard error

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3 Explanation of test-taking disengagement in numeracy (N = 13,947)

Predictor	Model 0		Model 1		Model 2		Model 3	
	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Intercept	− 15.91***	(0.26)	− 15.71***	(0.42)	− 16.03***	(0.29)	− 15.79***	(0.43)
Educational attainment								
High school			− 0.46	(0.33)			− 0.48	(0.33)
Above high school			− 0.78*	(0.35)			− 0.84*	(0.35)
Language (test language not same as native language)			0.65*	(0.26)			0.69**	(0.26)
Score cognitive pre-test			− 0.52***	(0.13)			− 0.43**	(0.14)
Gender (female)			0.02	(0.22)			0.00	(0.22)
Age group								
25–34			0.33	(0.37)			0.34	(0.37)
35–54			0.33	(0.32)			0.32	(0.32)
55 or more			0.32	(0.38)			0.32	(0.38)
Position (module 2)					0.84***	(0.23)	0.92***	(0.23)
Item difficulty					3.42***	(0.39)	3.34***	(0.38)
Item difficulty × score cognitive pre-test							− 0.43**	(0.14)
$Var(b_{0p})$	147.47		141.71		137.12		130.94	
$Var(b_{0i})$	0.92		0.91		0.69		0.68	

SE standard error

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

language was not their native language. Cognitive skill exhibited a significant negative effect. Once again, there were no significant effects of gender or age group on disengagement. Model 2 provided insights into the relation between item properties and disengagement. When items were completed in Module 2, disengagement was significantly higher than when they were completed in Module 1. Item difficulty had a significant positive effect on disengagement. Finally, Model 3 exhibited person and item variable effects very similar to those from Model 1 and Model 2. Most importantly, Model 3 revealed that item difficulty interacted significantly with cognitive skill, with the effect of item difficulty on disengagement smaller among strong test-takers than poor test-takers.

The variance in the random person intercept, $Var(b_{0p})$, decreased by 11.21% from Model 0 to Model 3, while the variance in the random item intercept, $Var(b_{0i})$, decreased by 26.09%. Thus, the amount of variance explained by person variables was only about half as much for numeracy compared to literacy.

Problem solving disengagement

The results of Model 1 (see Table 4) revealed that participants with higher educational attainment showed lower test-taking disengagement in problem solving. Test-takers with higher scores on the cognitive pre-test were less disengaged when completing the problem solving test. Disengagement was higher among participants whose native language was not the same as the test language and who were more than 24 years old. There was a particularly strong increase for the oldest group, participants aged 55 or above. However, there was no significant effect of gender. Model 2 showed that item difficulty had a positive effect on disengagement. These findings were reflected again in the full Model 3. This model additionally revealed that the positive effect of item difficulty was

Table 4 Explanation of test-taking disengagement in problem solving (N = 10,367)

Predictor	Model 0		Model 1		Model 2		Model 3	
	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Intercept	− 6.30***	(0.61)	− 6.66***	(0.60)	− 6.29***	(0.48)	− 6.63***	(0.50)
Educational attainment								
High school			− 1.06***	(0.14)			− 1.07***	(0.14)
Above high school			− 2.16***	(0.15)			− 2.16***	(0.15)
Language (test language not same as native language)			0.97***	(0.12)			0.96***	(0.12)
Score cognitive pre-test			− 1.26***	(0.06)			− 1.19***	(0.06)
Gender (female)			− 0.05	(0.09)			− 0.05	(0.09)
Age group								
25–34			1.30***	(0.15)			1.30***	(0.15)
35–54			1.55***	(0.13)			1.55***	(0.13)
55 or more			2.25***	(0.15)			2.24***	(0.15)
Item difficulty					4.51***	(0.96)	4.54***	(1.24)
Item difficulty × score cognitive pre-test							− 0.40***	(0.08)
$Var(b_{0p})$	10.88		11.02		10.89		11.04	
$Var(b_{0i})$	4.80		5.06		3.08		3.24	

SE standard error

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

smaller among strong test-takers as indicated by the significant negative interaction between item difficulty and cognitive skill.

The variance of the random person intercept, $Var(b_{0p})$, was much lower for problem solving than for literacy and numeracy. Unexpectedly, this variance component did not decrease from Model 0 to Model 3 even though the person level predictors expected significant effects. However, as expected, the variance of the random item intercept, $Var(b_{0i})$, fell substantially, by 32.50%.

Discussion

The overall goal of the present study was to investigate the conditioning factors of disengaged responses as observed in the PIAAC domains of literacy, numeracy, and problem solving. For this purpose, binary item disengagement indicators were defined for the Canadian sample by means of response time thresholds and subjected to an item response analysis. The results showed that disengagement indicators define a latent dimension by domain. Furthermore, individual and item differences could be explained substantially by the test-taker's educational attainment, language, and cognitive skill level, and by the item's difficulty and position.

Gender did not exhibit any effect on disengagement. Previous studies reporting gender effects are based on more homogenous and younger samples than the PIAAC, such as university students (Setzer et al. 2013). Thus, it would be interesting to explore whether the gender effect depends on other variables such as age or educational level. Interestingly, age showed differential effects on disengagement, that is, it had a significant effect for problem solving, but not for literacy or numeracy, supporting the assumption that disengagement represents a domain-specific construct. Increasing disengagement by

elderly test-takers in items assessing problem solving in technology-rich environments may be related to their lower levels of ICT experience and skills (OECD 2013a).

Applying item response models to the item disengagement indicators was challenging because the response variation was very low for many items (i.e., very low rate of disengagement and very high item difficulty). As a result, model estimations including all three domains simultaneously had serious problems and did not converge. This points to the need for an alternative measurement model for this kind of data. One option would be to use a (multi-dimensional) latent class model (Bartolucci 2007) distinguishing between engaged and disengaged respondents by means of a categorical latent person variable measured by binary disengagement indicators.

Notably, the variance in the random person and item intercepts varied across domains (see Tables. 2, 3, 4). There was a much greater variation in the person intercept (individual disengagement) for literacy and numeracy than for problem solving, while the pattern was reversed for the variance in the item intercepts (item difficulty regarding disengagement). The huge variance in individual disengagement for literacy and numeracy may suggest that the latent disengagement variable represents two groups of test-takers, those who are mostly engaged (by far the majority of the sample) and those who are mostly disengaged. The lower variance in individual disengagement for problem solving suggests that test-takers were less extreme in being engaged or disengaged. How this relates to characteristics of the problems solving assessment requires further investigation; for instance, more diverse and less familiar kinds of items such as multiple simulated software applications might re-engage unmotivated test-takers.

Following expectancy-value theory, we assumed that test-takers would make disengaged responses depending on their (perceived) cognitive skill and (perceived) item difficulty, which together determine the expected task success. The obtained findings support this hypothesis of informed disengagement. Specifically, the negative interaction between cognitive skill and item difficulty suggests that relative item difficulty helps determine disengagement. That is, poor test-takers encountering more difficult items are more likely to become disengaged than strong test-takers. Interpreting the results in this way requires test-takers to be able to evaluate the difficulty of an item relatively quickly, that is, below the response time threshold. This raises the question of whether the interpretation of informed disengagement is compatible with defining disengaged responses as relatively quick (non-) responses. For sure, empirically there may be a portion of test-takers who simply rush through the test without any strategic reflection about their probability of success. However, we define the threshold as the shortest response time where the conditional probability of success first exceeds chance level. This means that it may take (almost) as much time to try to solve an item before finally judging it to be too difficult as it does to complete it correctly. Further research is needed to justify the interpretation of informed disengagement. For instance, future studies could conduct cognitive interviews after administering a test on a (sub)sample of test-takers to learn more about the decision processes resulting in disengagement. From a test-taker's perspective, informed disengagement can be regarded as an efficient test-taking strategy as long as the test-taker's perceptions of his or her ability and item difficulty are correct. Modelling approaches for intentional omissions (Mislevy and Wu 1996), that is, models of the joint distribution of item response and disengagement, can be applied to investigate the relations between the model-predicted correctness

of an item response and disengaged responding. If the relationship is strong, disengaged responses carry information about the to be estimated competence level.

The proposed basic model of test-taking engagement (see Fig. 1) assumes that expectancy and value are determinants of test-taking engagement. However, as already pointed out by Eccles and Wigfield (2002), the expectancy-value approach needs to be extended by integrating concepts of action regulation, particularly volition, which describes action execution more comprehensively by assuming action phases and related volitional processes (Gollwitzer 1996). Specifically, a mind-set that supports efficient means of self-control might be helpful to prevent other intentions from distracting one from the task at hand (Kuhl 2000). From this, it can be inferred that engaged test-taking requires not only high expectancy but also a high level of self-control, while informed disengagement is the result of low expectancy and high level of self-control. Low levels of self-control are associated with aberrant test-taking behavior.

A potential limitation of using response time thresholds to identify disengaged responses is that disengaged responses may also be associated with long response times, for instance, if the test-taker pretends to be engaged with task completion in the test situation without making any real effort. Such disengaged responses cannot be discovered using the current approach. Therefore, an interesting further development would be to extend the item-level measure of test-taking disengagement to a sequential measure considering sequences of items. Test-takers pretending to take the test seriously would not rapidly respond to items but may distribute response time erratically across items without regard for the items' difficulty or time intensity. Thus, a pattern of repeated deviations of the observed response time from the expected response time given the person's test-taking speed and the items' time intensity (see van der Linden and Guo 2008) could indicate disengaged responding. Implementing this approach would require estimating individual latent speeds using observed response times as indicators and item time intensity parameters obtained from a sample of engaged test-takers.

A more fundamental concern about binary disengagement indicators is that continuous response time information is transformed into categories (engaged vs. disengaged) even though conceptually disengagement for an item might best be regarded as a continuous phenomenon. For instance, a person is probably less engaged when their response time is only slightly above the threshold than when their response time is clearly above it. Thus, an interesting future research direction would be to incorporate observed response times into the modelling of disengagement and define a continuous indicator for engagement rather than a single cut-off. For instance, Fox and Marianti (2017) recently proposed person-fit statistics for joint models for item responses and response times to detect aberrant response behavior (e.g., guessing).

Another potential limitation of our study refers to the explanatory models. We implicitly assumed measurement invariance in the levels of grouping variables (e.g., male vs. female). If this assumption were to be violated because of differential item functioning, the group comparisons could be biased. However, we decided against testing this assumption given the low rate of disengagement, which would be even smaller if the data set were to be split into multiple groups.

An important future research direction is to investigate the impact of considering disengaged responses when modelling individual and group differences. For instance, response

behavior classified as disengaged in PIAAC could be coded as not attempted/not reached regardless of whether there was a response or a non-response (omission). Criteria of interest are the reliability of the proficiency scale and the validity of test score interpretation. Regarding the latter, investigating the impact of dealing with disengagement on inferences about country differences in competencies would be of utmost interest for international large-scale assessments such as PIAAC. This would shed light on the question of whether or not the amount of observed disengagement is a severe problem given the intended use of test scores.

It should be emphasized that the design of the PIAAC study does not allow for causal inferences, particularly in terms of person-level predictors. For instance, with regard to the effect of the cognitive skill variable, it may be that worse cognitive skill causes higher disengagement, since test-takers expect that they will not be able to successfully complete the item anyway. However, higher levels of disengagement could also give rise to lower scores on the test for cognitive skills. The findings for item-level predictors are more conclusive, and this is particularly so for position, as this property was varied experimentally by the random assignment of test-takers to modules.

Conclusions

The present study used a model-based approach to provide empirical evidence that disengaged responding in the large-scale assessment PIAAC can be explained by individual and item differences. Thus, whether test-takers in the Canadian sample were more or less disengaged could be explained by educational attainment, native language, and cognitive skill level, as well as by age for problem solving only. In the same vein, items are more or less associated with disengaged responses depending on item difficulty and the position of the module in which the item can be found. The negative effect of cognitive skill, the positive effect of item difficulty, and their negative interaction effect support the assumption that disengagement is the outcome of individual expectations about success (informed disengagement).

Authors' contributions
FG originated the idea for the study, conducted the analyses and wrote the most of the manuscript. OL contributed to the development of the data analysis strategy; he also revised and reworked the manuscript. TM contributed to the conceptual framing and reworked the manuscript. All authors read and approved the final manuscript.

Author details
[1] German Institute for International Educational Research (DIPF)/Centre for International Student Assessment (ZIB), Schloßstr. 29, 60486 Frankfurt/Main, Germany. [2] Medical School Hamburg, Am Kaiserkai 1, 20457 Hamburg, Germany. [3] IPN-Leibniz Institute for Science and Mathematics Education/Centre for International Student Assessment (ZIB), Olshausenstraße 62, 24118 Kiel, Germany.

Acknowledgements
We are grateful to three anonymous reviewers for their helpful remarks on an earlier version of this paper.

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

Availability of data and materials
The dataset analyzed in this paper is based on the public use file for Canada from the first round of the Programme for the International Assessment of Adult Competencies (PIAAC). The public use file is available at the OECD website http://www.oecd.org/site/piaac/publicdataandanalysis.htm.

Ethics approval and consent to participate
We use data from the PIAAC Survey of Adult Skills which adheres to ethics standards stated by the OECD; see PIAAC Technical Standards and Guidelines (June 2014), http://www.oecd.org/skills/piaac/PIAAC-NPM(2014_06)PIAAC_Technical_Standards_and_Guidelines.pdf.

Funding
We thank the Centre for International Student Assessment (ZIB) for financial support.

Appendix

R syntax for computing the binary indicator of test-taking disengagement P+>0%

```
## Variables (for a given item as included in the PIAAC public use file)
# item.r = vector of item responses
# item.rt = vector of item response times in ms

percentile.rt = .95
bin.width = 1000 #in ms

breaks <- seq(0, quantile(item.rt, c(percentile.rt),na.rm = TRUE),by =
bin.width)
percent.correct <- matrix(NA, length(breaks),1)

for (i in c(1:length(breaks)-1)) {
  filter <- item.rt > breaks[i] & item.rt <= breaks[i+1]
  percent.correct[i] <- table(item.r[filter])[2]/sum(filter, na.rm = T)
   }

min.time.correct <- min( c(1:length(breaks))[percent.correct>0],  na.rm = T)
min.time.correct <- min.time.correct-1 # lower bound of the 1 second interval

## Variables
# value = binary disengagement indicator (engaged vs. disengaged)
# variable = item intercept of disengagement
# subject = person intercept of disengagement
# pos = position (module 1 vs. module 2)
# difficultyrp67 = item difficulty
# CBA_CORE_STAGE2_SCORE = score in the cognitive pre-test
# Age = test-taker's age
# Gender = test-taker's gender (male vs. female)
# NATIVESPEAKER = native language (test language same vs. not same as native
language)
# EdLevel3 = educational attainment (less than high school vs. high school vs.
above high school)
```

R syntax for estimating the generalized linear mixed models (GLMM)

```
## Models for Literacy, Numeracy, and Problem solving

## Note, for Problem solving, the predictor pos was not included


library(lme4)


#M0 - Model 0 (baseline model)

Output.M0 <- glmer(value ~ 1 + (1|variable) + (1|subject),

family=binomial("logit"), data = data)


#M1 - Model 1 with person covariates

Output.M1 <- glmer(value ~ 1 + (1|variable) + (1|subject) + NATIVESPEAKER +

EdLevel3 + CBA_CORE_STAGE2_SCORE + Age + Gender, family=binomial("logit"),

data = data)


#M2 - Model 2 with item covariates

Output.M2 <- glmer(value ~ 1 + (1|variable) + (1|subject) + pos +

difficultyrp67, family=binomial("logit"), data = data)


#M3 - Model 3 with person and item covariates

Output.M3 <- glmer(value ~ 1 + (1|variable) + (1|subject) + pos +

CBA_CORE_STAGE2_SCORE*difficultyrp67 + Age + Gender + NATIVESPEAKER +

EdLevel3, family=binomial("logit"), data = data)
```

Publisher's Note

Received: 6 December 2016 Accepted: 24 October 2017
Published online: 20 November 2017

References

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92–104.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*(2), 141. https://doi.org/10.1007/s11336-005-1376-9.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

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*(11), 2309–2344.

Brown, A. R., & Finney, S. J. (2011). Low-stakes testing and psychological reactance: Using the hong psychological reactance scale to better understand compliant and non-compliant examinees. *International Journal of Testing, 11*(3), 248–270. https://doi.org/10.1080/15305058.2011.570884.

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*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002.
Cronbach, L. J. (1970). *Essentials of psychological testing*. New York: Harper & Row.
de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.
De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–28. https://doi.org/10.18637/jss.v039.i12.
Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502–523. https://doi.org/10.3102/1076998614558485.
DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practice in Assessment, 8,* 69–82.
Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel rasch model: With the lme4 package. *Journal of Statistical Software, 20,* 1–18. https://doi.org/10.18637/jss.v020.i02.
Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*(3), 234–243. https://doi.org/10.1177/01466210122032046.
Eccles (Parsons), J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., et al. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). San Francisco: W. H. Freeman.
Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153.
Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series, 2015*(2), 1–17. http://doi.org/10.1002/ets2.12067.
Fox, J.-P., & Marianti, S. (2017). Person-Fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement, 54*(2), 243–262. https://doi.org/10.1111/jedm.12143.
Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. Vol. 133. *In: OECD Education Working Papers.* Paris: OECD Publishing.
Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106,* 608–626. https://doi.org/10.1037/a0034716.
Gollwitzer, P. M. (1996). The Volitional Benefits of Planning. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action. Linking cognition and motivation to behavior* (pp. 287-312). New York, London: The Guilford Press.
Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x.
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), 1–17. https://doi.org/10.1348/000711005x47168.
Jakewerth, P. M., Stancavage, B. S., & Reed, E. D. (1999). *An investigation of why students do not respond to questions*. CA: Palo Alto.
Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test analysis modules.* R package version 1.99–6. Retrieved from http://CRAN.R-project.org/package=TAM.
Köhler, C., Pohl, S., & Carstensen, C. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling, 57*(4), 499–522.
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*(4), 606–619. https://doi.org/10.1177/0013164406294779.
Kuhl, J. (2000). Chapter 5—A functional-design approach to motivation and self-regulation: The dynamics of personality systems interactions A2—Boekaerts, Monique. In P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 111–169). San Diego: Academic Press.
Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58,* 196–217. https://doi.org/10.1353/jge.0.0045.
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, 2*(1), 1–24. https://doi.org/10.1186/s40536-014-0008-1.
Ma, L., Wise, S. L., Thum, Y. M., & Kingsbury, G. (2011). *Detecting response time threshold under the computer adaptive testing environment.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.
Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x.
Meyer, J. P. (2010). A mixture rasch model with item response time components. *Applied Psychological Measurement, 34*(7), 521–538. https://doi.org/10.1177/0146621609355451.
Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Vol. RR96-30). Princeton: Educational Testing Service.
OECD. (2013a). *OECD skills outlook 2013: First results from the survey of adult skills*. Paris: OECD Publishing.
OECD. (2013b). *Technical report of the survey of adult skills (PIAAC)*. Paris: OECD Publishing.
Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences. *Large-scale Assessments in Education, 2*(1), 5. https://doi.org/10.1186/s40536-014-0005-4.
Pohl, S., Gräfe, L., & Rose, N. (2013). 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*. https://doi.org/10.1177/0013164413504926.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: to filter unmotivated examinees or not? *International Journal of Testing*, 17, 74–104. http://doi.org/10.1080/15305058.2016.1231193.

Rost, J. (2004). *Lehrbuch Testtheorie—Testkonstruktion [Textbook Test theory—Test construction]* (2nd ed.). Bern: Huber.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453.

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures. *ETS Research Report Series, 1988*(2), i–71. http://doi.org/10.1002/j.2330-8516.1988.tb00297.x.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6–26. https://doi.org/10.1016/S0361-476X(02)00063-2.

Team, R. C. (2016). *R: A language and environment for statistical computing* (Version 3.1.3). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy—value theory: A latent interaction modeling study. *Journal of Educational Psychology, 104*(3), 763. https://doi.org/10.1037/a0027470.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8.

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*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*(3), 152–166.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237–252. doi:10.1080/08957347.2015.1042155.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12165.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1–17. https://doi.org/10.1207/s15326977ea1001_1.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x.

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. http://doi.org/10.1080/08957347.2017.1353992.

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2.

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education, 8*(4), 341–351. https://doi.org/10.1207/s15324818ame0804_4.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Münster: Waxman.