Organisation for Economic Co-operation and Development

EDU/PISA/GB(2023)5

For Official Use **English text only**

27 February 2023

DIRECTORATE FOR EDUCATION AND SKILLS
PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT

Governing Board

ITEM CHARACTERISTICS AND TEST-TAKER DISENGAGEMENT IN PISA

55th meeting of the PISA Governing Board

21-23 March 2023
Rome, Italy

The PGB is invited to:

- **NOTE** and **COMMENT** on the working paper

Andreas Schleicher, Director for Education and Skills and Special Advisor on Education Policy to OECD's Secretary-General (andreas.schleicher@oecd.org).

JT03513137

Item Characteristics and Test-Taker Disengagement in PISA

Abstract

If test-takers do not engage with the assessment, the reliability of test scores and the validity of inferences about their proficiency may suffer. Test-taker disengagement is particularly likely in low-stakes assessments and for certain types of students. Prior research also suggests that levels of engagement are related to aspects that test developers might manipulate to reduce disengaged response behaviour. This paper investigates which item characteristics are associated with two indicators of test-taker disengagement, rapid guessing and breakoffs, in the 2018 Programme for International Student Assessment (PISA) assessment of reading. Analyses of data from almost 500 000 students from 67 countries or economies showed that rapid guessing was observed mainly on simple multiple-choice questions and in the presence of long and complex texts. Such texts also increased the likelihood of breakoffs, which were most likely to coincide with idiosyncratic selected-response formats, such as hot spot or matching tasks.

Table of contents

FIGURES

1. Introduction

1. If test-takers do not engage fully with the assessment tasks, the reliability of test scores and the validity of inferences about the test-takers' proficiency may suffer (Guo et al., 2022[1]). These considerations are particularly meaningful in the context of low-stakes assessments, which include a large number of educational assessments. In these, some test-takers may lack the motivation to engage with the test and put their best effort forward because their performance has very few, if any at all, personal consequences (Wise, Pastor and Kong, 2009[2]). Disengaged behaviours like rapid guessing or failing to complete the assessment in the absence of time pressures are indeed frequently documented in the context of educational low-stakes assessments (Buchholz, Cignetti and Piacentini, 2022[3]; Wise and DeMars, 2005[4]). This working paper aims to contribute to previous research by exploring how different item characteristics relate to test-taker disengagement in the context of the Programme for International Student Assessment (PISA), and to offer recommendations for test developers to mitigate its occurrence. PISA presents a number of desirable characteristics for exploring the topic of test-taker engagement: it is based on representative samples of 15-year-old students and is administered in more than 60 countries, thus allowing to explore the robustness of findings and supporting claims which go beyond the typical convenience samples used in other studies; and, because it reports results only at aggregate levels, it has low stakes for individual students who sit the test.

1.1. Relevance of test-taker disengagement in PISA

2. Researchers have addressed test-taker disengagement and its implications using a variety of names: test-taker engagement (Goldhammer, Martens and Lüdtke, 2017[3]), test-taker disengagement (Wise, Soland and Dupray, 2021[4]), test-taker effort (Wise, 2006[5]), or test-taker motivation (Wise and DeMars, 2005[6]). Regardless of the concept employed, researchers have highlighted that test scores reflect not only the cognitive capacity of the test-takers, but also their effort, motivation, and non-cognitive skills (Zamarro, Hitt and Mendez, 2019[7]; Borghans et al., 2016[8]).

3. The accumulation of research findings that underline the role of non-cognitive skills in test performance has attracted renewed scrutiny into the implications drawn from test scores about the quality of education. This applies to teacher value-added modelling (Petek and Pope, 2022[10]), or to system-level assessments such as PISA (Lee and Stankov, 2018[11]).

4. The implications of PISA, for example, extend much beyond the statistical analysis of student performance on a test. Over the years, PISA has become a central source of information for many countries on aggregate national levels performance (i.e. reading, math, and science), and for some countries or economies PISA is one of the *few sources* of this kind of information (Zamarro, Hitt and Mendez, 2019[7]). It is not uncommon, therefore, that policy makers around the world use PISA results to analyse the strengths and weaknesses of their national educational policies (Santos and Centeno, 2021[11]), and to direct reform efforts toward improving instruction in the areas in which the country's results were weakest. For example, Germany implemented educational reforms following a disappointing performance in the PISA 2000 assessment (Ertl, 2006[13]). Likewise, Denmark (Egelund, 2008[14]), Japan (Takayama, 2008[15]), Spain (Engel, 2015[16]), and Australia (Gorur and Wu, 2014[17]) have engaged in comparable actions following low (or declining) performance of students in PISA. Yet, despite these well-intended efforts, not all implemented changes have resulted in meaningful improvements in national performance.

5. In the presence of low or declining levels of test engagement, policy changes that aim at improving students' skills may have limited effects on test scores; conversely, performance in low-stakes assessments may be improved simply by offering students material rewards contingent on performance (Duckworth et al., 2011[18]) students' "will" without affecting their skill. Ensuring that students who sit the PISA test engage with the assessment and questionnaire items, even in the absence of direct incentives to do so, is central to the value of PISA for policy makers.

1.2. Adjusting for test-taker disengagement

6. Several adjustments that aim to resolve the negative consequences of test-taker disengagement on test-score reliability and validity have been proposed. More commonly, such approaches either (a) filter out disengaged responses from the dataset, or (b) correct test scores to account for the observed levels of (dis)engagement (Kuhfeld and Soland, 2019[19]). In both instances, these adjustments are applied *post hoc* (i.e. after data have been collected). Buchholz, Cignetti and Piacentini (Buchholz, Cignetti and Piacentini, 2022[3])discuss in more detail the usefulness of *post hoc* remedies to account for test-taker disengagement in the context of low-stakes assessments.

7. Filtering implies the removal of observations that are believed to be disengaged from the dataset. This can either be done at the level of a single response or at the level of the respondent (Rios et al., 2016[21]). However, some of the concerns about this approach are associated with the binary decision that is required for classifying an observation as engaged or disengaged, as well as with the fact that removing observations reduces the sample size and is likely to introduce sampling bias in inferences to broader populations.

Adjusting test scores to account for disengaged responses in the data can be accomplished in several ways. For example, simple sum-scores can be corrected *post hoc* using regression by controlling for one or multiple measures of disengagement. Alternatively, disengagement can be controlled in the construction of test scores derived from item-response-theory (IRT) models, for example by including disengagement in the scaling model. The effort-moderated IRT model (Wise and DeMars, 2006[22]) and the speed-accuracy + engagement model (Pohl, Ulitzsch and von Davier, 2021[12]; Ulitzsch, Davier and Pohl, 2019[51]) are examples for the latter approach.

8. Both approaches require the selection of one or multiple indicators that reflect test-taker disengagement validly and reliably. In the context of low-stakes assessments such as PISA, Computer-Based Assessment (CBA) is slowly becoming customary (Kuhfeld and Soland, 2019[19]). CBA has expanded the range of indicators for detecting and describing disengaged response behaviour, as it allows to keep track of response processes in the form of log-file data associated with test items. For example, student response times have been used to develop indicators for rapid guessing, which may reveal a lack of time, interest or knowledge (Wise, 2017[22]). Meanwhile, records of the responses provided to the test items continue to be used both in paper- (Borgonovi and Biecek, 2016[23]) and computer-based assessments (Zamarro, Hitt and Mendez, 2019[8]) to compute indicators such as completion or non-response rates, which may relate to fatigue or exhaustion.

9. Rapid-guessing indicators and indicators based on missing-response patterns are available for most computer-based assessments, irrespective of the test-takers' characteristics or test content. Their use as indicators of disengagement is justified by the fact that responses that are lacking because the test-taker did not meaningfully interact with the stimulus, and very rapid responses, do not provide valid indications of students' capacity to solve the task.

10. However, the choice of a particular measure or set of measures to mitigate the consequences of disengagement on test scores through *post hoc* adjustments is likely to change the result of the adjustment, as the correlations between them is low (Buchholz, Cignetti and Piacentini, 2022[3]). This introduces an undesirable level of arbitrariness in the construction of test scores.

1.3. Revisiting item characteristics to reduce test-taker disengagement

11. Prevention represents an alternative way to address student disengagement: it consists in limiting its occurrence by administering tests and items that are less disengaging. A first step, in order to do so, is understanding how disengaged response behaviours are related to test and item characteristics.

12. Some studies have inferred the effect of item and test characteristics on engagement only indirectly, by studying how they relate to performance. As an example of this line of research, Wolf, Smith and Birnbaum (1995[24]) addressed how item *position* is related to achievement. More specifically, this study compared mathematics performance between high-school sophomores and juniors in an exam that was of consequence only for the first group, and was thus low-stakes for the second group. Results showed that item position negatively affected the performance of junior students only (those for which the exam had low stakes). Similarly, DeMars (2000[25]) provided evidence of differential item functioning depending on *response formats* across groups defined by test stakes. In this study, high school students were assigned to high-stakes and low-stakes groups, and results showed that students performed better under high stakes for both constructed response and multiple-choice formats, but that the difference in performance was larger for constructed response items (DeMars, 2000[26]).

13. More recently, research has focused on understanding the association of test and item characteristics with specific forms of disengagement, which can be observed at the item level: item-level measures allow to exploit the within-student variation in test engagement levels in order to identify how it relates to differences in item characteristics. The following sections review these studies, and group them in terms of the indicators of disengagement used (i.e. rapid guessing, non-response, and breakoffs).

1.3.1. Rapid guessing and response time

14. Rapid guessing has been an increasingly common measure of test-taker disengagement since the widespread adoption of computer-based assessment (CBA).

15. Item position is repeatedly found to relate to rates of rapid guessing. For instance, Wise (2006[7]) investigated the effect of item features on rapid guessing during a low-stakes computer-based test among university students, and found that *position* was one of the strongest predictors of rapid guessing: items that were placed later in the test had higher rates of rapid guessing (Wise, 2006[7]). In a similar study conducted with university students, Wise, Pastor and Kong (2009[2]) found that items occurring later in the test had significantly higher rates of rapid guessing. Setzer et al. (2013[26]) also showed that rapid guessing was strongly positively associated with item position in the context of large-scale low-stakes assessments of university students. Similarly, Goldhammer, Martens and Lüdtke (2017[5]) found that respondents spent less time on items located in the second of two assessment modules in the Programme for the International Assessment of Adult Competencies (PIAAC), an assessment of adults between the ages of 16 and 65 administered outside of traditional educational institutions; and Bowling et al. (2020[27]) found that university students spent less time on questionnaire items as they progressed further into a questionnaire, and for some students this remained true even when warned

about carelessness potentially having negative consequences for them (i.e. forfeit of participation credits).

16. The relationship between *difficulty* and rapid guessing has also been examined. For example, Goldhammer, Martens and Lüdtke (Goldhammer, Martens and Lüdtke, 2017[5]) showed that item difficulty was positively related to test-taker disengagement in the context of PIAAC. In this study, the higher the difficulty of the item, the lower the response time. However, it is worth noting that these associations were moderated by test-takers' cognitive skills (i.e. lower performing students were more likely to disengage on difficult items than higher performing test-takers. Rios and Guo (2020[28]) also found that rapid guessing was positively related to the perceived difficulty of the test in a study that collected data from university students in four different countries. In contrast, Wise (2006[7]) found no statistically significant relationship between item difficulty and rapid guessing.

17. Wise (2006[7]) also investigated the relationship between *item length* and effort, and found that item length (i.e. how much reading or scanning was required) was among the strongest predictors of rapid guessing: the longer the item text, the more rapid guessing was observed among test-takers (Wise, 2006[5]). Subsequently, Wise, Pastor and Kong (2009[2]) verified in a similar study that for items with more text there was a significant increase in the rates of rapid guessing. Setzer et al. (2013[26]) found that response times were negatively associated with how much test-takers had to read and/or scan.

18. Current lines of inquiry are also trying to identify item characteristics, such as the *inclusion of multimedia*, that are positively associated with test-taker engagement and that could therefore be added to items to decrease some forms of disengagement. For example, Wise, Pastor and Kong (2009[2]) showed that the inclusion of graphics in items decreased rapid guessing, and moderated the effect of item position on rapid guessing. Similarly, Lindner et al. (2017[29]) examined the presence of representational pictures in text-based items, in a computer-based test for elementary school children; results showed that the presence of representational pictures significantly reduced participants' rapid guessing behaviour (Lindner et al., 2017[31]). Rios and Soland (2022[30]) studied different correlates of rapid guessing in the context of the PISA 2018 science assessment, and also showed that the presence of multimedia item content was significantly related to decreases in rapid guessing (Rios and Soland, 2022[32]).

19. In a similar vein, (Wise, Soland and Dupray, 2021[4]) examined the rates of rapid guessing in *technology-enhanced*, multiple-choice, and multiple select items, and showed that technology-enhanced items (i.e. selecting options from one area of the item display and move them to other areas, and/or selecting a response from within a piece of text or information table) received higher levels of test-taker engagement as the rates of rapid guessing were consistently lower than the other two item types examined in their study.

1.3.2. Non-response

20. Borgers and Hox (2001[31]) investigated the effect of *position* on non-response among elementary school children and high-school youth. This study found that the items that were shown later in the opinion and attitude questionnaires had higher proportions of missing responses (Borgers and Hox, 2001[33]). More recently, (Zamarro, Hitt and Mendez, 2019[7]) also explored test-taker engagement over the course of the PISA 2009 test (i.e. performance) and questionnaire (i.e. non-response and careless response rates). The results showed that, on average, as test takers progress in the test and questionnaires, their engagement tends to decline (Zamarro, Hitt and Mendez, 2019[7]).

21. Another feature that has received attention when examining associations with test-taker disengagement is *response format*. Guo et al. (2022[1]) collected data from an assessment of university students and examined how non-traditional multiple-choice items were associated with test-taker disengagement (i.e. non-response and rapid guessing). This study showed that while multiple-choice items with many options are harder than items with fewer options, neither variant increased the non-response or rapid guessing rate (Guo et al., 2022[1]). In contrast, Borgers and Hox (2001[31]) utilized data from elementary and secondary school children to investigate how the number of response options affected item non-response in opinion and attitude questionnaires. The results in this study showed a negative effect, meaning that the greater number of response options in the items, the higher the proportion of non-response (Borgers and Hox, 2001[33]).

22. Adding to these findings, Borgers and Hox (2001[31]) also examined the relation between item non-response and the *length* of ancillary reading. Contrary to previously cited results on rapid guessing, this study showed that the length of the introductory text had a positive effect on item response (Borgers and Hox, 2001[31]). In the same way, the inclusion of technology and/or of *technology enhancements* has been associated with decreasing proportions of disengagement. Namely, Zehner et al. (2020[32]) analysed participants' responses to PISA reading questions assigned to computer, paper, or both modalities, and showed that, while some items were more affected than others, generally test-takers assigned to the computer modality incorporated longer texts into their constructed responses.

1.3.3. Breakoffs

23. Breakoff behaviour refers to respondents who start a test or a questionnaire but fail to complete it in the absence of time pressure. Although such behaviour can be observed in interviewer-administered surveys, higher rates occur in online, self-administered surveys (Peytchev, 2009[33]; Steinbrecher, Roßmann and Blumenstiel, 2014[34]).

24. In general, research in this field expands on the findings of non-response studies to better understand correlates of declines in completion. For example, Peytchev (2009[33]) conducted one of the first studies that tried to identify the correlates to survey breakoffs in a sample of adult online volunteers in the US. In this study, higher breakoff rates were associated to perceived difficulty (e.g. grids of multiple questions, or multiple open-ended questions), section introductions (i.e. logical break and commitment to start a new part of the survey), sensitive questions (e.g. alcohol consumption), novel respondent tasks (i.e. different from multiple choice or text entry), and technical issues (Peytchev, 2009[35]). In the same line of research, Steinbrecher, Roßmann and Blumenstiel (2014[34]) found that task difficulty, as well as technical issues were the strongest predictors of breakoff in a sample of adult volunteers in Germany. That is, the more complex the response format, and the more technical issues, the greater the likelihood of a breakoff. The authors thus recommend creating surveys that are as simple as possible, reducing survey length, and providing incentives to resume participation (Steinbrecher, Roßmann and Blumenstiel, 2014[36]). Similarly, a more recent meta-analysis of breakoff rates in mobile Web surveys (Mavletova and Couper, 2015[37]) found that shorter surveys, reminders, and less complex designs decreased breakoff rates. However, this study did not find a significant effect of incentives or allowing participants to skip questions (Mavletova and Couper, 2015[37]).

25. Complementary research has aimed to describe the demographic correlates of breakoff. In this regard, (Mittereder, 2019[38]) sampled both students and faculty in a US-based university and found that male respondents were more likely to break off at the beginning of the questionnaire, while female respondents tended to quit toward the end of the questionnaire. This study also found that item non-response, as well as rapid

guessing, were positively associated with greater chances of breaking off (Mittereder, 2019[38]). Similarly, Fortunato, Hibbing and Provins (2022[37]) collected data from a representative sample of US adults, and found that although females were more likely to break off than males, their responses were on average of a substantial higher quality (e.g. no straight-lining or shirking). This study's main recommendation is to employ attention checks throughout the questionnaires (Fortunato, Hibbing and Provins, 2022[39]).

26. McGonagle (2013[38]) collected data through a computer-assisted telephone interview among US families and found that the risk of first breakoff was related to the duration of each section. That is, the longer it took for a section to be completed, the greater the likelihood of breakoff. Results also highlighted that introductory sections and/or first questions on the screen increased the risk of breakoff (McGonagle, 2013[40]).

27. Although most studies on breakoff behaviour have been conducted in the context of self-administered web surveys, this phenomenon is also relevant in a group-administered assessment. In this setting, breakoff behaviour holds the potential to disrupt the testing situation for the whole group. However, not enough research has been conducted to examine the determinants of breakoffs in that context.

1.4. The purpose of this working paper

28. Taken together, these findings suggest that certain item and test characteristics hold the potential to reduce test-taker disengagement. Carefully considering these when designing tests and questionnaires is especially crucial in the context of low-stakes assessments such as PISA, where test-taker performance has very few, if any at all, personal consequences (Wise, Pastor and Kong, 2009[2]). Therefore, the main motivation of this working paper is to investigate whether test-taker disengaged response behaviour can be reduced through changes in assessment design, particularly through better item design. More specifically, the main research question guiding this working paper is: what item characteristics are associated with indicators of test-taker disengagement?

2. Methods

2.1. Data

29. PISA is a low-stakes assessment administered every three years by the Organization for Economic Co-operation and Development (OECD) that measures 15-year-old's proficiency in reading, mathematics, and science, as well as in an innovative domain. All our analyses are based on the PISA 2018 reading test. After completing a test focusing on two or three of these cognitive domains, students responded to questionnaires covering a diversity of topics (e.g. socio-economic background, social emotional characteristics, reading-related attitudes and behaviours). Participation takes approximately three hours, including a two-hour test, the questionnaire, time for tutorials, and breaks (including one between the first and second hour of testing).

30. In 2018, the focus of the assessment was on reading, and an hour-long reading test was administered to all PISA participants; half of the participants took the reading test in the first hour, while the remaining participants took it in the second hour, after the break. A multi-stage adaptive test design with two routing points was implemented in the computer-based versions of the reading test (in use in the vast majority of countries): after a first, core block of items, participants were routed to test versions of varying difficulty, depending on their performance in preceding blocks. In total, the item pool comprised 245 reading items, partitioned into three sets and assembled, within each set,

into non-exclusive testlets. Each test-taker responded to three testlets (one from each set), comprising between 33 and 40 items in total. The testlets assigned to each student were chosen in part at random, in part based on prior performance.

31. PISA tests were administered to more than 600 000 15-year-old students attending educational institutions in grades seven and higher across more than 70 countries and economies. Only data from countries that administered the PISA test on computers are used in the present study. Indeed, response times, which are required for the definition of rapid guessing behaviour, are only available in this mode. As a result, the present study is based on data from 67 countries and economies. PISA samples are drawn according to a two-stage stratified design. In each country, at least 150 schools were selected in the first stage (or all schools, in countries/economies in which there were fewer than 150 schools in which 15-year-old could be enrolled). In the second stage, students were sampled within the selected schools; the typical within-school sample size is 42 (and all 15-year-old students were selected if fewer than 42 were enrolled), but countries/economies could implement a smaller target cluster size (and increase the number of sampled schools) if preferred. The final target sample size was therefore 6 300 students for each participating jurisdiction. Students with a permanent physical, cognitive, behavioural, or emotional disability preventing them to participate and/or limited proficiency in the language of assessment could be excluded from the within-school sampling (up to a limit of 5% of exclusions overall) or assigned to a shorter, adapted version of the test; students in the latter case were excluded from this analysis. The final sample for this study therefore consisted of about 17 million item responses, given by 499 387 students from 67 participating countries or economies to the items included in their PISA reading test.

2.2. Measures of test-taker disengagement, models, and explanatory variables

2.2.1. Indicators of test-taker disengagement

32. Two indicators of test-taker disengagement were examined in this working paper:

- *Rapid guessing* was calculated at the item level for each test-taker by considering the total response time (across all item visits) when responses were not missing. A cut-off of ≤5 seconds was applied to code non-missing responses as rapidly guessed responses. Missing responses (omitted and non-reached items) as well as items for which no valid response time was available were coded as missing and excluded from the analyses of rapid-guessing behaviour.
- *Breakoffs* were also calculated at the item level for each test-taker. First, non-reached items were defined as sequences of at least two items with no recorded response at the end of each test session. The first item in this sequence was then coded as a "breakoff event" if it was reached within the first 45 minutes of the one-hour long test sessions, i.e. if it could be assumed that the onset of sequential non-response occurred in the absence of time pressures. The student was then considered in "breakoff" state for the rest of the test.

2.2.2. Models

33. To examine the association of item characteristics with rapid guessing, three-level logistic regression models were estimated using Stata's (version 17) melogit command. In these models, the logarithm of the odds of observing a rapid guess is described as a linear function of item, student, and school characteristics; random intercepts for students and schools are included to account for the unobserved student and school factors that may influence rapid guessing.

34. To examine the association of item characteristics with breakoff events, Weibull proportional hazard survival models were estimated using Stata's (version 17) streg command. These models assume that the baseline hazard rate of observing a breakoff event increases exponentially with the position of the item, and that concomitant item and student characteristics included in the model have a multiplicative effect on the baseline hazard. Standard errors are clustered at the school-level to account for the multi-level structure of the data.

35. Separate models were estimated on each national dataset, and average results across all 67 samples are presented in the results section. Results by national sample are available as supplementary material on the PISA Sharepoint Site. For both models, exponentiated coefficients, corresponding to odds ratios (OR) or hazard ratios (HR) are reported in the remainder of this paper; standard errors are transformed using the delta method. No weights are applied: results therefore describe, in the first place, how the data yield and quality in the 2018 assessment were affected by disengaged response behaviours that could be related to item (and student) characteristics. Any generalisation beyond the sample of students selected for PISA, either at national or international level, relies on the validity of the regression models themselves; consistent with this, we report model-based standard errors in the analysis.

2.2.3. Explanatory variables

36. Among the variables that explain test-taker disengagement, the main focus of this paper is on surface characteristics of the items, which are not strongly related to the target construct or to the measurement properties of the items. All of these were coded as categorical variables and included in models as (sets of) dummy variables to facilitate the interpretation of coefficients as odds ratios. The following item characteristics were included in models to explain differences in test-taker disengagement:

- *Response format* includes four categories:
    - open ended response (25% of the items): corresponding to items in which students had to respond by writing in open text format (see Annex A for an example);
    - hot spot and match questions (4% of the items): students respond by marking certain areas of an image or by pairing elements from one column to another (see Annex B for an example);
    - matrix and complex multiple-selections (17% of the items): students answer a set of questions (displayed in rows) with the same answer options (displayed in columns) by making one selection per row, or select multiple options (e.g. "all that apply") for a single question (see Annex C for an example);
    - simple multiple-choice (54% of the items): students select a single response among a number of options provided (typically four) (see Annex D).
- *Inclusion of images/figures* can take three values ("none", "one", "two or more"). Of the 233 items in this study,[1] 24% items had no image and/or figures, 49% had one, and 27% had two or more images and/or figures. For example, all items

[1] Of the 245 reading items, a few items were excluded because they are presented on the same screen as others; only one item per screen is included in the analysis.

presented in the Annex (A-D) belong to the same unit which includes multiple figures.

- *Interactive Material* is a binary indicator capturing whether the reading stimuli required students to navigate and/or interact with different pages, tabs, websites, etc. About half (48%) of the items had such interactive material. The items presented in the Annex (A-D) are examples of items that required students to interact with tabs.
- *Length* reflects the number of words in reading stimuli and can take three values ("≤200 words", "201-500 words", "≥501 words"). Out of all items, 24% had ancillary text with 200 or less words, 64% text between 201 and 500 words, and 12% text with 501 or more words.
- *Position* is a binary indicator capturing whether an item is the first in its unit (21% of items), thus describing a situation in which new stimulus material (i.e. a new text, thematically unrelated to previous texts) is presented.

37. A number of additional variables (control variables), at the item and student level, were also included in models. Their main role is to ensure that effects of the surface characteristics listed above are not confounded by factors that have a direct effect on test engagement, and that are, within the specific item set considered here and due to the adaptive routing design used, correlated with the former.

38. First, models control for the *exact item position* within the test; this constitutes the "time" variable in the survival models, while a quadratic polynomial was included in the logistic regression among the predictors of rapid guessing. In the next session, we provide descriptive graphical evidence of the association of item position with disengagement indicators to justify the choice of these functional forms. Position is related to item characteristics because of the multi-stage adaptive design: for example, open-ended items were deliberately placed mostly towards the end of the test and avoided in the core testlets, where they could not be used to inform the adaptive routing of students.

39. *Item Difficulty* was also accounted for; difficulty is measured on the PISA described proficiency scale, with the simplest tasks in the test corresponding to Level 1. Level 2, Level 3, Level 4, and Level 5/6 correspond to increasingly difficult tasks. Of the 233 items included, 23% were in Level 1, 28% in Level 2, 20% Level 3, 14% Level 4, 14% Level 5/6. Item difficulty is related to a number of characteristics of interest; for example, a larger proportion of difficult items have an open response format, compared to easier items. Because of the adaptive assignment of testlets to students, it is also related to the student profile: more proficient students are, on average, exposed to a greater proportion of difficult items.

40. Because of the adaptive test design, which introduces a dependency of item characteristics on student performance, and of prior evidence showing that disengaged behaviours are related to the test-takers' skill or ability level, it is important to also control for students' expected performance in the test in order to interpret the associations of surface characteristics of items with disengagement (captured in regression coefficients) as reflecting the causal effects of these characteristics. We take advantage of the fact that all students who sat the PISA reading test also completed tests in one or two additional domains (mathematics, science, and/or global competence, the innovative domain), before or after completing the reading test. Because not all students took the same tests, for each domain we include two variables: the first "plausible value" (PV1) for the respective domain, corresponding to an imputed performance score (a single draw from the posterior likelihood distribution, given test responses and background characteristics), and a binary

variable indicating whether test scores were imputed based solely on background characteristics (i.e. whether the student did not actually sit the corresponding test). The imputed score (PV1) was recoded to a constant for students for whom no test responses were available to compute the posterior likelihood: in this way, variation in PV1 reflects true variation in performance for students who sat the corresponding test, and the "imputation dummies" reflect the random assignment of students to the three domains.

41. Finally, we introduce control variables for the *test session* – a dichotomous variable indicating whether the student took the reading test in the first or the second hour – and for *sex* (male, female).

3. Results

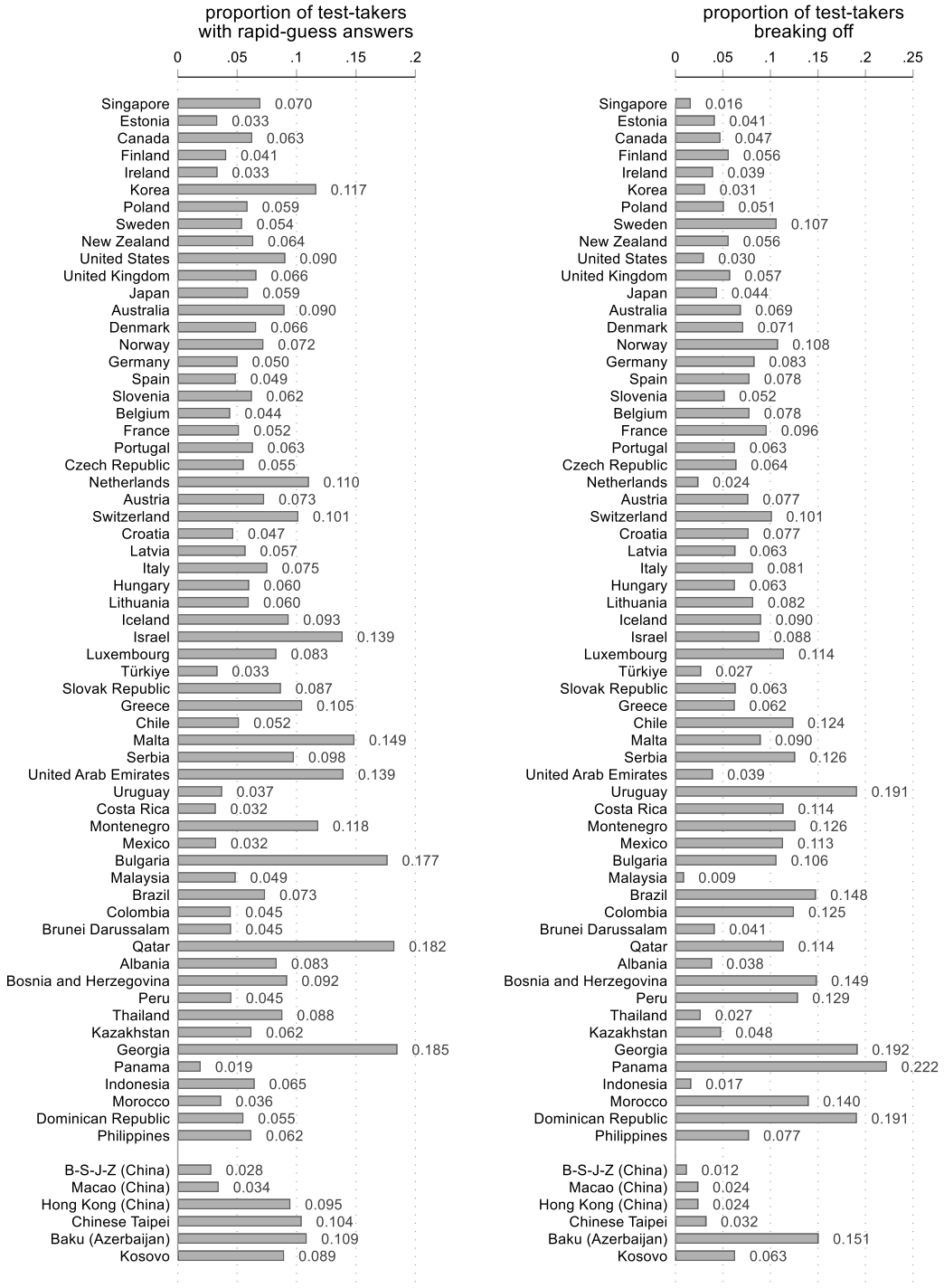
42. This section describes the variation in test-taker disengagement across countries and their incidence during the reading test. It then presents findings from the models described above, which aim at isolating the effect of (surface) item characteristics on the probability of rapid guessing and of breakoffs.

3.1. Cross-country variation in test-taker disengagement

43. On average across student samples from 67 countries and economies, rapid guessing and breakoffs were observed for 7.4% and 8.0% of students who sat the PISA reading test in 2018, respectively (some students exhibited both behaviours). In most cases, only a minority of their responses were affected: viewed in terms of item responses, about 1.6% of them were identified as rapid guesses, and another 1.6% were missing because of breakoff.

44. There is considerable variation in the prevalence of rapid guessing and breakoffs across countries: Figure 3.1 shows the respective proportion of test-takers (students) in each national sample included in the present study.

Figure 3.1. Prevalence of rapid guessing and breakoff patterns



Notes: Countries/economies are ranked in descending order of their mean reading performance in 2018. For Spain, the rank reflects the country's performance in 2015.
Source: OECD, PISA 2018 Dataset.

45. The highest prevalence of rapid guessing is observed in Bulgaria, Georgia, Qatar and Bulgaria, where more than one in six students who sat the PISA test had rapid-guess responses, and more than 4% of responses overall were classified as rapid guesses in this study. In contrast, rapid guessing is rarely observed among students who sat the PISA test in Costa Rica, Mexico and Panama, but also in Beijing-Shanghai-Jiangsu-Zhejiang (China): fewer than one in two hundred (0.5%) responses, at most, were considered rapid guesses in the samples from these countries and economies. At the level of countries/economies, the prevalence of rapid guessing behaviour is unrelated to performance in PISA: the linear correlation coefficient between the rate of rapid guessing and performance on the PISA reading test is comprised between -0.2 and 0.2. Among high-performing countries, for example, high rates of rapid guessing are observed in Chinese Taipei, Hong Kong (China) and Korea.

46. In contrast, there is a higher tendency to breakoff in low-performing countries (the linear correlation coefficient between the rate of breakoff among students in the PISA sample of a country/economy, and mean performance on the PISA reading test, is about -.5). The highest prevalence of test breakoff is observed in Latin American countries: for example, 22% of students in Panama, and 19% of students in the Dominican Republic and Uruguay (as well as in Georgia), did leave unanswered questions at the end of the reading test even though they had at least 15 more minutes left to complete the test. On the other hand, students in Asian countries were the least likely to break off: fewer than 2% of students in Beijing-Shanghai-Jiangsu-Zhejiang (China), Indonesia, Malaysia and Singapore did so. Among high-performing countries, breakoffs were frequently observed in Sweden and Norway.

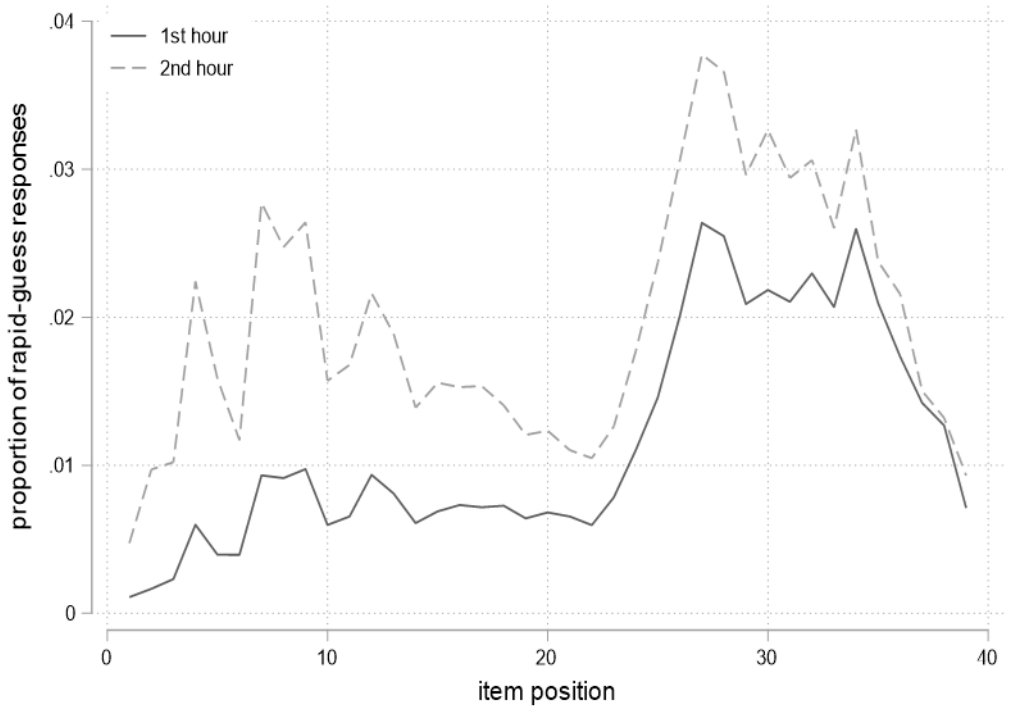
47. The correlation of rapid guessing rates and breakoff rates across countries/economies is close to 0. This suggests that test-taker disengagement tends to manifest itself in different forms across countries.

48. In the PISA scaling procedures, missing answers due to breakoffs, like any non-reached items (irrespective of their cause) are treated as "not administered" and do not affect a country's mean score. A high proportion of students with breakoff patterns may still cast questions on the interpretation of mean scores, as students who voluntarily did not complete the assessment may have failed to do their best even in the parts of the assessment that they completed. Breakoffs are also a strong threat to the unbiased recovery of item parameters, as their estimation has to rely on smaller and potentially biased samples in the presence of breakoffs. In contrast, valid answers which are classified, based on response-time information, as rapid guesses, are treated as either correct or incorrect and used in scoring. Their effect on scores may be positive (if students were likely to give an incorrect answer, or to omit the question, had they not guessed rapidly) or negative. Rapid guessing also poses a significant threat to the unbiased recovery of item parameters.

3.2. Incidence of test-taker disengagement indicators

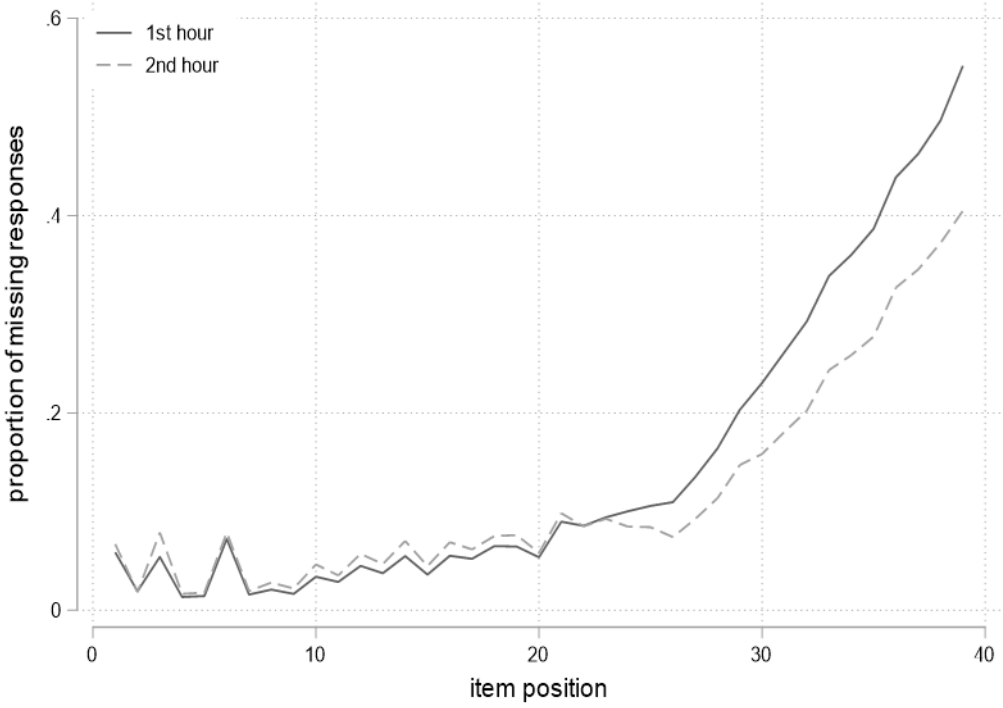
49. By definition, the breakoff indicator is not independent across items: once a student decides to "break off" (or is forced to break off due, for example, to a glitch in the delivery system), all subsequent items are affected. Rapid guessing also appears to be related to the position of items in the test.

Figure 3.2. Relationship Between Item Position and Average Rapid Guessing by Testing Hour



50. Figure 3.2 illustrates the relationship between item position and rapid-guessing rates, by testing hour. The figure shows that both in the first and in the second testing hour, the occurrence of rapid guessing increases as students advance in the test and the time limit approaches. It is also interesting to note that the incidence of rapid guessing appears to be higher at the beginning of the second testing hour, compared to the beginning of the first hour. Finally, the rate of rapid guessing does not increase (or decrease) smoothly, but exhibits peaks at particular positions in the test; because of the specific test design in PISA, certain items (and item types) are more likely to be found in particular positions and it is likely, therefore, that these peaks reflect the influence of such item characteristics rather than the effect of a particular item position. We will turn to the effect of item characteristics on rapid guessing towards the end of this section.

Figure 3.3. Relationship Between Item Position and Average of Missing Items by Testing Hour

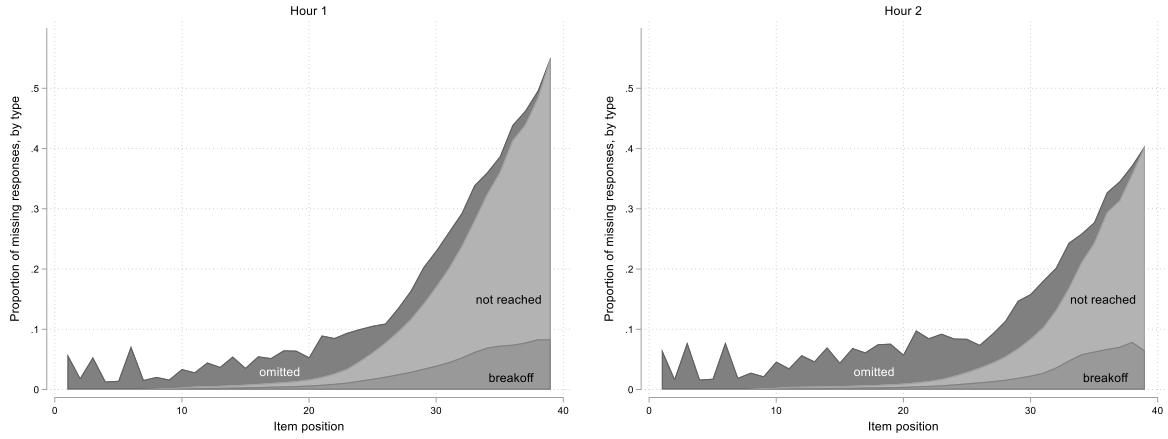


51. Figure 3.3 illustrates the relationship between item position and the rate of missing responses, by testing hour. Even more than the previous figure regarding rapid guessing, this figure shows that as students approach the end of the session, the incidence of missing responses increases, in both the first and second testing hour. In contrast to rapid guessing, however, there is a larger proportion of missing responses at the end of the first testing hour, compared to the end of the second hour. Just like for rapid guessing, certain positions appear to be associated with peaks in the incidence of missing responses, most likely because of the influence of the characteristics of items which appear in these positions.

52. Missing responses may come in distinct patterns which reflect a number of different behaviours. A first distinction is between omitted items (missing responses which are followed by non-missing responses) and non-reached items (sequences of missing responses which extend through the end of the test). For the purpose of producing test scores, the former are treated as valid, but "incorrect" responses in PISA: indeed, even highly engaged respondents may omit certain responses, either after a quick scan, and as a strategic time-allocation decision, or after trying hard, but without success, to understand and solve a task. A second distinction is among non-reached items, depending on whether these are simply the result of students "running out of time" before they reach the end of the test, or of proper breakoff (i.e. where the onset of non-response is observed at least 15 minutes before the time limit).

53. Figure 3.4 shows this breakdown of missing responses by type.

Figure 3.4. Type of Missing Responses during the First (Left) and Second (Right) Testing Hour



54. The largest share of missing responses appears to be driven by non-reached items. That is, in both the first and second testing hour (but particularly so in the first testing hour), a significant fraction of students appears to run out of time before completing the reading test. Perhaps in order to avoid this (and to reach the end of the test faster), a larger share of students seems to omit certain items in the second hour, compared to the first hour. The figures also show that a significant portion of non-reached items can be attributed to students who would have had at least 15 minutes of time left to engage with test items. These students – and the items that trigger the onset of such breakoffs – are the focus of the analysis that follows.

3.3. Associations between item characteristics and rapid guessing

55. Table 3.1 shows the average results of the multi-level logistic regression models, which estimates the extent to which item characteristics are associated with the probability of rapid guessing. Two specifications are presented: the baseline model, including all explanatory and control variables described in Section 2, and an auto-regressive model which includes, in addition, an indicator for "previous rapid guessing": the coefficient on this indicator signals to what extent the likelihood of rapid guessing increases (or decreases) after a student engages for the first time in rapid guessing. In both models, all item characteristics of interest were statistically significant predictors of rapid guessing ($p < .05$).

56. Findings show that *response format* is one of the strongest predictors of rapid guessing, with simple multiple-choice items being more than 20 times more likely to trigger rapid guessing behaviour compared to open-response items, and more than 13 times more likely than complex multiple-choice items. The latter, including matrix formats and multiple-selection formats, were slightly more likely to trigger rapid guessing (odds ratios between 1.6 and 1.7, depending on the model) compared to open-response formats.

57. The inclusion of multimedia showed differential associations with rapid guessing. On the one hand, the *inclusion of images and/or figures* increased the likelihood of disengaged behaviour. Compared to items with no images or figures, rapid guessing was more likely to be observed on items with multiple images and/or figures (i.e. 1.4 times higher odds in both the main and in the auto-regressive models); as well as on items with

one image or figure (i.e. 1.3 times in both models). On the other hand, the *inclusion of interactive material* appeared to decrease the likelihood of rapid guessing in both models (in the main model, for example, the odds were only 0.79 times as large when interactions were present).

Table 3.1. Item-Level and Student-Level Predictors of Rapid Guessing

	Baseline Model	Lagged Model
Response Format: Open Response	ref.	ref.
Hot Spot/Match	1.204* (0.057)	1.123* (0.053)
Matrix/Multiple Selections	1.673* (0.033)	1.643* (0.032)
Simple Selection	23.100* (0.348)	21.702* (0.322)
Images/Figures: No	ref.	ref.
Single	1.259* (0.011)	1.289* (0.011)
Multiple	1.360* (0.014)	1.396* (0.014)
Interactive Material: No	ref.	ref.
Yes	0.785* (0.005)	0.805* (0.006)
Reading Stimuli Length: 1-200 words	ref.	ref.
201-500 words	1.706* (0.015)	1.755* (0.015)
501+ words	2.028* (0.027)	2.129* (0.028)
First In Unit: No	ref.	ref.
Yes	0.386* (0.004)	0.405* (0.004)
Second Test Hour: No	ref.	ref.
Yes	2.459* (0.034)	1.717* (0.017)
Previous Rapid Guessing		6.417* (0.081)
Difficulty: Level 1	ref.	ref.
Level 2	1.374* (0.011)	1.304* (0.011)
Level 3	1.439* (0.013)	1.415* (0.013)
Level 4	2.304* (0.031)	2.142* (0.028)
Level 5/6	3.061* (0.045)	2.890* (0.041)
Sex: Male	ref.	ref.
Female	0.302* (0.004)	0.461* (0.005)
Position	Yes	Yes
Performance	Yes	Yes
School r.effects	Yes	Yes
Student r.effects	Yes	Yes
Schools	18345	18345
Students	498687	498687
Observations	16046941	16046941

Note: exponentiated coefficients, corresponding to odds ratios, are reported, along with standard errors in Parentheses. *: $p < .05$.

58. Examining their *length*, results also showed that the longer the ancillary content the more likely items were to receive rapid guessing (i.e. 1.706 and 2.028 times in the main model, 1.755 and 2.129 times in the lagged model for texts of moderate and large length, respectively).

59. In addition, when examining the associations of certain control variables, the more *difficult* an item the more likely it was to obtain a rapid guessing response. For instance, in both models, items increased from slightly more likely to be rapid guessed in Level 2 and Level 3, to more than twice as likely in Level 4, and more than three times as likely in Level 5/6.

60. All estimates control for *position* of the item within the test; in addition, a few indicator variables included in the models are also related to item position or order. In particular, students who answered the reading test in the second testing hour were about 2.5 times more likely to provide rapid guess answers to an item in comparison to students who answered the reading test in the first testing hour (odds ratio of 2.459 in the main model). The lagged model also considered *previous rapid guessing* response as a potential explanatory variable of this disengaged behaviour. The coefficient on this indicator implies that previous rapid guessing increases the likelihood of repeating the same behaviour again by more than six times. In turn, in this lagged model, the ratio between the likelihood of rapid guessing in the second hour and the likelihood of rapid guessing in the first hour is reduced to about 1.7, suggesting that part of the increase in rapid guessing in the second hour operates through the fact that students begin engaging in rapid guessing earlier in the test, during the second hour. Finally, order effects also seem to play a role in both models: being the first item in the unit decreased the chance of rapid guessing about 2.5 times (oddsratios of about 0.4, or 1:2.5).

61. When examining test-takers *sex*, females were less likely to give a rapid guessing response compared to males (i.e. odds ratios of 0.302 in the main model, and 0.461 in the lagged model).

3.4. Associations Between Item Characteristics and Breakoffs

62. Survival models were estimated to analyse the associations between item characteristics and breakoffs. Table 3.2 shows the average results of the survival model across 67 countries or economies. Similarly, all item characteristics were statistically significant predictors of breakoffs in all models ($p < .05$).

63. The results of these models are reported in terms of hazard ratios. Hazard ratios indicate how the risk (or "hazard") of observing a certain event (breakoff, in the current analysis) increases, or decreases, when a particular factor (measured by the independent variables included in the model) is observed. Similar to odds ratios, hazard ratios are multiplicative factors applied to the baseline hazard; ratios above 1 indicate factors that increase the risk of breakoff, while ratios between 0 and 1 indicate factors that decrease the risk of breakoff. In our models, the baseline hazard is a function of item position only.

Table 3.2. Item-Level and Student-Level Predictors of Breakoffs

	Hazard Ratios
Response Format: Open Response	ref.
Hot Spot/Match	1.282* (0.044)
Matrix/Multiple Selections	0.268* (0.011)
Simple Selection	0.425* (0.009)
Images/Figures: No	ref.
Single	1.179* (0.026)
Multiple	0.891* (0.023)
Interactive Material: No	ref.
Yes	1.104* (0.021)
Reading Stimuli Length: 1-200 words	ref.
201-500 words	1.417* (0.037)
501+ words	1.797* (0.057)
First In Unit: No	ref.
Yes	2.192* (0.044)
Second Test Hour: No	ref.
Yes	0.878* (0.016)
Difficulty: Level 1	ref.
Level 2	1.372* (0.043)
Level 3	2.385* (0.069)
Level 4	1.767* (0.062)
Level 5/6	1.413* (0.052)
Sex: Male	ref.
Female	0.902* (0.015)
Item position (*p*)	4.783* (0.063)
Performance	Yes
Schools	18346
Students	499387
Obs. Breakoffs	30030
Observations	17017825

Note: exponentiated coefficients, corresponding to hazard ratios, are reported, along with standard errors that account for clustering at the school level (in parentheses). *: $p < .05$. Item position is treated as the discrete time variable t in the survival model; the reported coefficient (p) corresponds to the Weibull parameter for the baseline hazard function, $h_0(t) = pt^{p-1}$. Values of p above 1 indicate an exponentially increasing hazard rate.

64. *Response format* showed different associations with breakoffs than those observed with rapid guessing. For instance, while hot spot and matching questions increased the likelihood of breakoffs by about 1.3 times (compared to open-response formats), simple multiple-choice along with matrix and complex multiple-choice items were between two and four times less likely to trigger breakoffs (hazard ratios, HR, equal 0.425 and 0.268, respectively).

65. *Inclusion of multimedia* exhibited weak associations with breakoffs, with no clear direction. That is, items with one image and/or figure were 1.179 times more likely, and items with ancillary material interactions were 1.104 times more likely to outset breakoffs. Contrarily, items with multiple images and/or figures were less likely (HR = 0.891) to start breakoff behaviour.

66. Regarding item *length*, results showed that the longer the reading stimuli the more likely items were to lead to breakoffs. This finding mirrors a similar finding for rapid-guessing behaviour. For instance, items with longer ancillary content were about

1.8 times more likely to onset breakoffs, while shorter 200-500-word ancillary-content items were 1.4 times more likely.

67. Examining the associations of control variables, the likelihood of breakoffs increased with *difficulty* only up to Level 3; items at Level 4 or Level 5/6 were less likely to result in breakoff, compared to items at Level 3, when presented to students (and after controlling for students' proficiency). This pattern too differs from the one observed for rapid guessing behaviour.

68. Regarding item *position*, being the first item in the unit increased the chance of breakoffs by more than 2 times: students were more likely to breakoff on the first item in a new unit, than on subsequent items within the same unit. Breakoffs were also slightly less likely in the second testing hour (HR = 0.878), perhaps because some breakoffs in the first hour are related to technical difficulties which could either be solved by the beginning of the second hour, or which caused the corresponding students to abandon the test entirely.

69. In line with the findings for rapid guessing, *females* were less likely to breakoff compared to males.

4. Discussion

70. The main purpose of this working paper was to investigate which item characteristics were associated with different types of test-taker disengagement indicators, in order to explore the potential of reducing disengaged response behaviour through changes in item and/or test design.

71. Previous research has identified multiple manifestations of disengaged response behaviour; however, studies that computed multiple measures found that the correlation among these measures is low (Buchholz, Cignetti and Piacentini, 2022[3]), suggesting that they correspond to distinct behaviours, rather than multiple manifestations of the same underlying causes.

72. In this paper, two measures of disengagement were examined in greater detail: rapid guessing and breakoffs. The previous section presented both descriptive evidence about the relation between test length and student engagement, based on these two indicators, and explanatory models which related rapid-guessing and breakoff behaviour in the reading test to specific characteristics of the items presented to students.

73. Regarding test length and item position, the descriptive evidence suggests that rapid-guessing behaviour is, in part, a response to time pressures: students are more likely to provide rapid-guess answers if they feel pressured by time (towards the end of the testing session), and if they failed to reach the end of the test in a previous test session: while non-reached items decrease in the second test hour, rapid-guess answers increase. As noted earlier, in PISA students are not penalised for not reaching the end of the test, while they are penalised for wrong answers (including wrong rapid-guess answers). Therefore, if students substitute rapid-guess answers (at the beginning of the test) for missing responses (at the end of the test, because of non-reached items), their performance suffers.

74. Regarding more specific determinants of rapid guessing, a number of findings from the explanatory models confirmed those reported in prior literature. In particular, rapid guessing in PISA is observed mainly on simple multiple-choice questions, where students have to tick only one among a limited number of answer options to provide a valid (though possibly incorrect) answer. While rapid guessing appeared to be encouraged by simple response formats, it also was found to relate to the complexity and length of the stimulus material: longer texts and more difficult questions were associated with higher rates of

rapid guessing. An aspect on which the results in the previous section do not concur with prior literature is the effect of figures and images: the findings suggest that the presence of visuals also increases the likelihood of rapid guessing, perhaps because it contributes to the complexity of the stimulus material in a reading test.

75. Before drawing implications for item design from the analysis of rapid-guessing alone, it is important to verify that efforts to suppress this disengaged response behaviour through careful item design do not backfire and increase the occurrence of another undesirable behaviour. Inspired by the literature on self-report surveys, the present study also examined whether certain item characteristics triggered disengaged response behaviour in the form of breakoffs, i.e. strings of missing responses which correspond to students who fail to complete the assessment in the absence of time pressures. Breakoffs are of particular relevance in the context of group-administered assessments: students who break off early in the test hold the potential to disrupt the testing situation for fellow test-takers.

76. The models that relate breakoff behaviour to item (and person) characteristics suggest that a number of item characteristics do not only raise the likelihood of rapid guessing, but also the risk of breakoff. Such is the case, for example, of text length and, although to a lesser extent, of item difficulty: longer reading stimuli and more difficult items seem to increase the likelihood of both rapid guessing and breakoffs.

77. In contrast, the associations with response formats were unique to each behaviour. While rapid-guess answers were strongly associated with simple response formats, breakoffs were particularly likely after open-ended text responses and, even more so, after response formats which required complex selections (e.g. hot spot items). The latter correspond often to innovative item formats, and it is possible that this effect results from the unfamiliarity of students with such formats.

78. Regarding student characteristics, only their sex and their proficiency (proxied by their scores in the remaining domains) was included in the model; the associations of these characteristics is consistent across models, with girls being less likely to exhibit either kind of disengaged behaviour, and less proficient students being more likely to do so.

79. The scope and design of this investigation was limited, in ways that affect the implications that can be drawn from its results. In particular, the analysis focused only on PISA 2018 reading items – results may not hold for other domains where characteristics such as the inclusion of images and of interactive navigation features, or the length of stimulus text, may play a different role. In addition, certain characteristics were present only in a limited number of items (e.g. hot spot and match questions), and the results associated with these characteristics may therefore, in part, reflect more idiosyncratic features of these items.

4.1. Implications for PISA test and item design

80. Taken together, these findings provide meaningful *practical* insights to potentially mitigate test-taker disengagement through purposeful item and test design. In particular, item developers and test designers could consider reducing item characteristics that increase the likelihood of a disengaged response behaviour, particularly if they are associated with multiple types of disengaged response behaviour.

81. Increasing the proportion of easy items, for example, appeared to increase the likelihood that most students engage with the test as intended, and that their response behaviour reflects their true proficiency. While this might deteriorate test targeting and therefore measurement precision, in theory, the cost in terms of reliability might be more

than offset by a gain in validity – i.e. a greater assurance that test scores reflect not only how students responded, but also that students' responses reflect what they know and can do.

82. Even though specific associations with response formats were not consistent for breakoffs and rapid guessing, limiting the variation in response formats within the PISA test to only a few, familiar formats, might help students to demonstrate their best proficiency. Indeed, response formats that are unfamiliar or the subject of insufficient practice during tutorials risk disrupting students and are associated with higher risks of breakoff, with only limited benefits for reducing rapid guessing behaviour.

83. The strongest predictor of rapid guessing are items that have simple, multiple-choice response format. In turn, the strongest predictor of breakoff behaviour are items that introduce new stimulus material. Both characteristics, however, are almost inevitable in PISA tests: simple multiple-choice format have desirable psychometric properties and can be administered and scored economically, which is an advantage as the constructs assessed in PISA require that a variety of stimulus situations are presented to every student. But because each characteristic has opposite effects on the other behaviour (simple multiple-choice formats reduce the likelihood of breakoff, and the first item in a unit is the least likely to be rapidly guessed), one way of reducing the risk of disengagement might be to combine these two characteristics, and to ensure that the first items in a sequence of thematically related items present the simplest possible response format.

84. More generally, regarding those characteristics – such as complex response formats or interactive stimulus materials – which appeared to have contradictory effects on the likelihood of rapid guessing and breakoffs, the baseline likelihood of each behaviour might guide the choices of test developers in a national context. While on average, a similar amount of test-takers exhibited either type of behaviour, in some countries/economies the priority is more clearly to address one type of behaviour: for example, in many Latin American countries, breakoffs are widespread, but few students engage in rapid guessing. In contrast, in Hong Kong (China), the Netherlands and Korea, breakoffs are very rare, but more than one in ten students engaged in rapid guessing during the reading test.

85. The findings on test length and item position suggest that in order to ensure that most students engage in the test, in a way that reflects their actual proficiency, it is important to help students' time management, by including specific guidelines in the test-administrator manual and tutorials and more scaffolding during the test. Before they begin the test, for example, students should be reassured that they incur no penalty for not reaching the end of the test. Including additional breaks and shortening the test sessions may also help students' time management. For example, in light of the decision to introduce greater balance across the three core domains of PISA, and in consideration of the positive effect on test engagement observed at the beginning of the second test session (compared to the end of the first session), a possibility would be to organise the test around three 40-minutes sessions, rather than two 60-minute sessions. To avoid situations in which students break off because they are stuck on an item that is too hard for them, or where they do not understand the response format, soft reminders about the possibility of skipping items might also be introduced: for instance, such reminders might appear on the screen after an extended period of inactivity, or after a fixed period of time. The effect of such changes to the interface or test design on students behaviour and, ultimately, data quality, should ideally be investigated thoroughly prior to their introduction in the main study, through pilot studies and/or field-trial experiments.

References

Borgers, N. and J. Hox (2001), "Item nonresponse in questionnaire research with children", *Journal of Official Statistics*, Vol. 17/2, pp. 321-335. [31]

Borghans, L. et al. (2016), "What grades and achievement tests measure", *Proceedings of the National Academy of Sciences*, Vol. 113/47, pp. 13354-13359, https://doi.org/10.1073/pnas.1601135113. [9]

Borgonovi, F. and P. Biecek (2016), "An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test", *Learning and Individual Differences*, Vol. 49, pp. 128-137, https://doi.org/10.1016/j.lindif.2016.06.001. [23]

Bowling, N. et al. (2020), "Will the Questions Ever End? Person-Level Increases in Careless Responding During Questionnaire Completion", *Organizational Research Methods*, Vol. 24/4, pp. 718-738, https://doi.org/10.1177/1094428120947794. [27]

Buchholz, J., M. Cignetti and M. Piacentini (2022), "Developing measures of engagement in PISA", *OECD Education Working Papers*, No. 279, OECD Publishing, Paris, https://doi.org/10.1787/2d9a73ca-en. [3]

DeMars, C. (2000), "Test Stakes and Item Format Interactions", *Applied Measurement in Education*, Vol. 13/1, pp. 55-77, https://doi.org/10.1207/s15324818ame1301_3. [25]

Duckworth, A. et al. (2011), "Role of test motivation in intelligence testing", *Proceedings of the National Academy of Sciences*, Vol. 108/19, pp. 7716-7720, https://doi.org/10.1073/pnas.1018601108. [18]

Egelund, N. (2008), "The value of international comparative studies of achievement – a Danish perspective", *Assessment in Education: Principles, Policy & Practice*, Vol. 15/3, pp. 245-251, https://doi.org/10.1080/09695940802417400. [14]

Engel, L. (2015), "Steering the National: Exploring the Education Policy Uses of PISA in Spain", *European Education*, Vol. 47/2, pp. 100-116, https://doi.org/10.1080/10564934.2015.1033913. [16]

Ertl, H. (2006), "Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany", *Oxford Review of Education*, Vol. 32/5, pp. 619-634, https://doi.org/10.1080/03054980600976320. [13]

Fortunato, D., M. Hibbing and T. Provins (2022), "Hurdles to inference: The demographic correlates of survey breakoff and shirking", *Social Science Quarterly*, https://doi.org/10.1111/ssqu.13128. [37]

Gneezy, U. et al. (2019), "Measuring Success in Education: The Role of Effort on the Test Itself", *American Economic Review: Insights*, Vol. 1/3, pp. 291-308, https://doi.org/10.1257/aeri.20180633. [41]

Goldhammer, F., T. Martens and O. Lüdtke (2017), "Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics", *Large-scale Assessments in Education*, Vol. 5/1, https://doi.org/10.1186/s40536-017-0051-9. [5]

Gorur, R. and M. Wu (2014), "Leaning too far? PISA, policy and Australia's 'top five' ambitions", *Discourse: Studies in the Cultural Politics of Education*, Vol. 36/5, pp. 647-664, https://doi.org/10.1080/01596306.2014.930020. [17]

Guo, H. et al. (2022), "Influence of <scp>Selected-Response</scp> Format Variants on Test Characteristics and <scp>Test-Taking</scp> Effort: An Empirical Study", *ETS Research Report Series*, https://doi.org/10.1002/ets2.12345. [1]

Haladyna, T., M. Rodriguez and C. Stevens (2019), "Are Multiple-choice Items Too Fat?", *Applied Measurement in Education*, Vol. 32/4, pp. 350-364, https://doi.org/10.1080/08957347.2019.1660348. [40]

Hanushek, E. and L. Woessmann (2021), "The Political Economy of ILSAs in Education: The Role of Knowledge Capital in Economic Growth", in *International Handbook of Comparative Large-Scale Studies in Education, Springer International Handbooks of Education*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-38298-8_4-1. [39]

Kuhfeld, M. and J. Soland (2019), "Using Assessment Metadata to Quantify the Impact of Test Disengagement on Estimates of Educational Effectiveness", *Journal of Research on Educational Effectiveness*, Vol. 13/1, pp. 147-175, https://doi.org/10.1080/19345747.2019.1636437. [19]

Lee, J. and L. Stankov (2018), "Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA", *Learning and Individual Differences*, Vol. 65, pp. 50-64, https://doi.org/10.1016/j.lindif.2018.05.009. [11]

Lindner, M. et al. (2017), "The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis", *Contemporary Educational Psychology*, Vol. 51, pp. 482-492, https://doi.org/10.1016/j.cedpsych.2017.09.009. [29]

Mavletova, A. and M. Couper (2015), "A Meta-Analysis of Breakoff Rates in Mobile Web Surveys", in *Mobile Research Methods: Opportunities and challenges of mobile research methodologies*, Ubiquity Press, https://doi.org/10.5334/bar.f. [35]

McGonagle, K. (2013), "Survey Breakoffs in a Computer-Assisted Telephone Interview", *Survey Research Methods*, Vol. 7/2, pp. 79-90. [38]

Mittereder, F. (2019), *Predicting and Preventing Breakoff in Web Surveys*, University of Michigan. [36]

Petek, N. and N. Pope (2022), "The Multidimensional Impact of Teachers on Students", *Journal of Political Economy*, https://doi.org/10.1086/722227. [10]

Peytchev, A. (2009), "Survey Breakoff", *Public Opinion Quarterly*, Vol. 73/1, pp. 74-97, https://doi.org/10.1093/poq/nfp014. [33]

Rios, J. and H. Guo (2020), "Can Culture Be a Salient Predictor of Test-Taking Engagement? An Analysis of Differential Noneffortful Responding on an International College-Level Assessment of Critical Thinking", *Applied Measurement in Education*, Vol. 33/4, pp. 263-279, https://doi.org/10.1080/08957347.2020.1789141. [28]

Rios, J. et al. (2016), "Evaluating the Impact of Careless Responding on Aggregated-Scores: To Filter Unmotivated Examinees or Not?", *International Journal of Testing*, Vol. 17/1, pp. 74-104, https://doi.org/10.1080/15305058.2016.1231193. [20]

Rios, J. and J. Soland (2022), "An investigation of item, examinee, and country correlates of rapid guessing in PISA", *International Journal of Testing*, Vol. 22/2, pp. 154-184, https://doi.org/10.1080/15305058.2022.2036161. [30]

Santos, Í. and V. Centeno (2021), "Inspirations from abroad: the impact of PISA on countries' choice of reference societies in education", *Compare: A Journal of Comparative and International Education*, pp. 1-18, https://doi.org/10.1080/03057925.2021.1906206. [12]

Setzer, J. et al. (2013), "An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment", *Applied Measurement in Education*, Vol. 26/1, pp. 34-49, https://doi.org/10.1080/08957347.2013.739453. [26]

Steinbrecher, M., J. Roßmann and J. Blumenstiel (2014), "Why Do Respondents Break Off Web Surveys and Does It Matter? Results From Four Follow-up Surveys", *International Journal of Public Opinion Research*, Vol. 27/2, pp. 289-302, https://doi.org/10.1093/ijpor/edu025. [34]

Takayama, K. (2008), "The politics of international league tables: PISA in Japan's achievement crisis debate", *Comparative Education*, Vol. 44/4, pp. 387-407, https://doi.org/10.1080/03050060802481413. [15]

Wise, S. (2017), "Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications", *Educational Measurement: Issues and Practice*, Vol. 36/4, pp. 52-61, https://doi.org/10.1111/emip.12165. [22]

Wise, S. (2006), "An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test", *Applied Measurement in Education*, Vol. 19/2, pp. 95-114, https://doi.org/10.1207/s15324818ame1902_2. [7]

Wise, S. and C. DeMars (2006), "An Application of Item Response Time: The Effort-Moderated IRT Model", *Journal of Educational Measurement*, Vol. 43/1, pp. 19-38, https://doi.org/10.1111/j.1745-3984.2006.00002.x. [21]

Wise, S. and C. DeMars (2005), "Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions", *Educational Assessment*, Vol. 10/1, pp. 1-17, https://doi.org/10.1207/s15326977ea1001_1. [4]

Wise, S., D. Pastor and X. Kong (2009), "Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice", *Applied Measurement in Education*, Vol. 22/2, pp. 185-205, https://doi.org/10.10805/08957340902754650. [2]

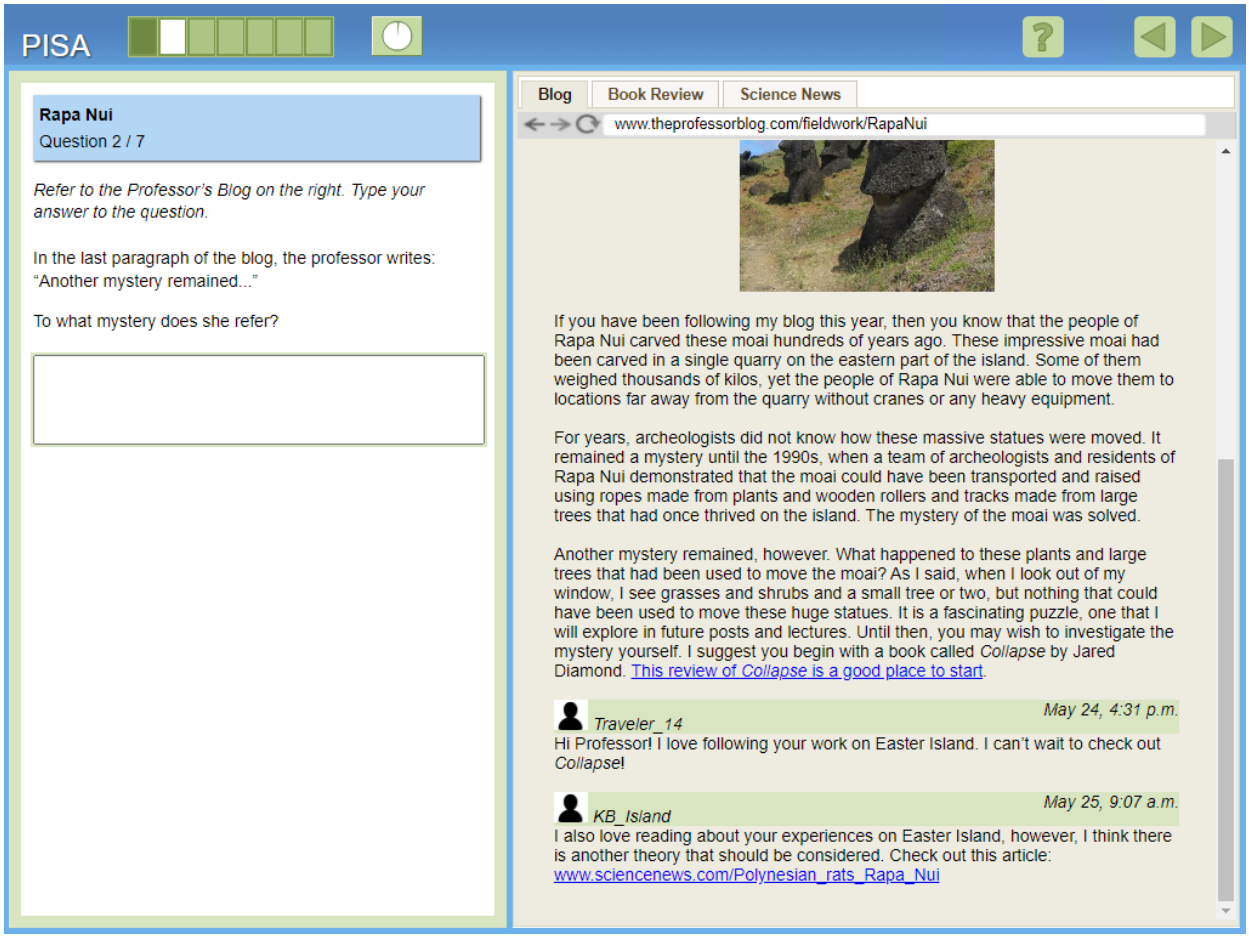
Wise, S., J. Soland and L. Dupray (2021), "The impact of technology-enhanced items on test-taker disengagement", *Journal of Applied Testing Technology*, Vol. 22/1, pp. 28-36. [6]

Wolf, L., J. Smith and M. Birnbaum (1995), "Consequence of Performance, Test, Motivation, and Mentally Taxing Items", *Applied Measurement in Education*, Vol. 8/4, pp. 341-351, https://doi.org/10.1207/s15324818ame0804_4. [24]

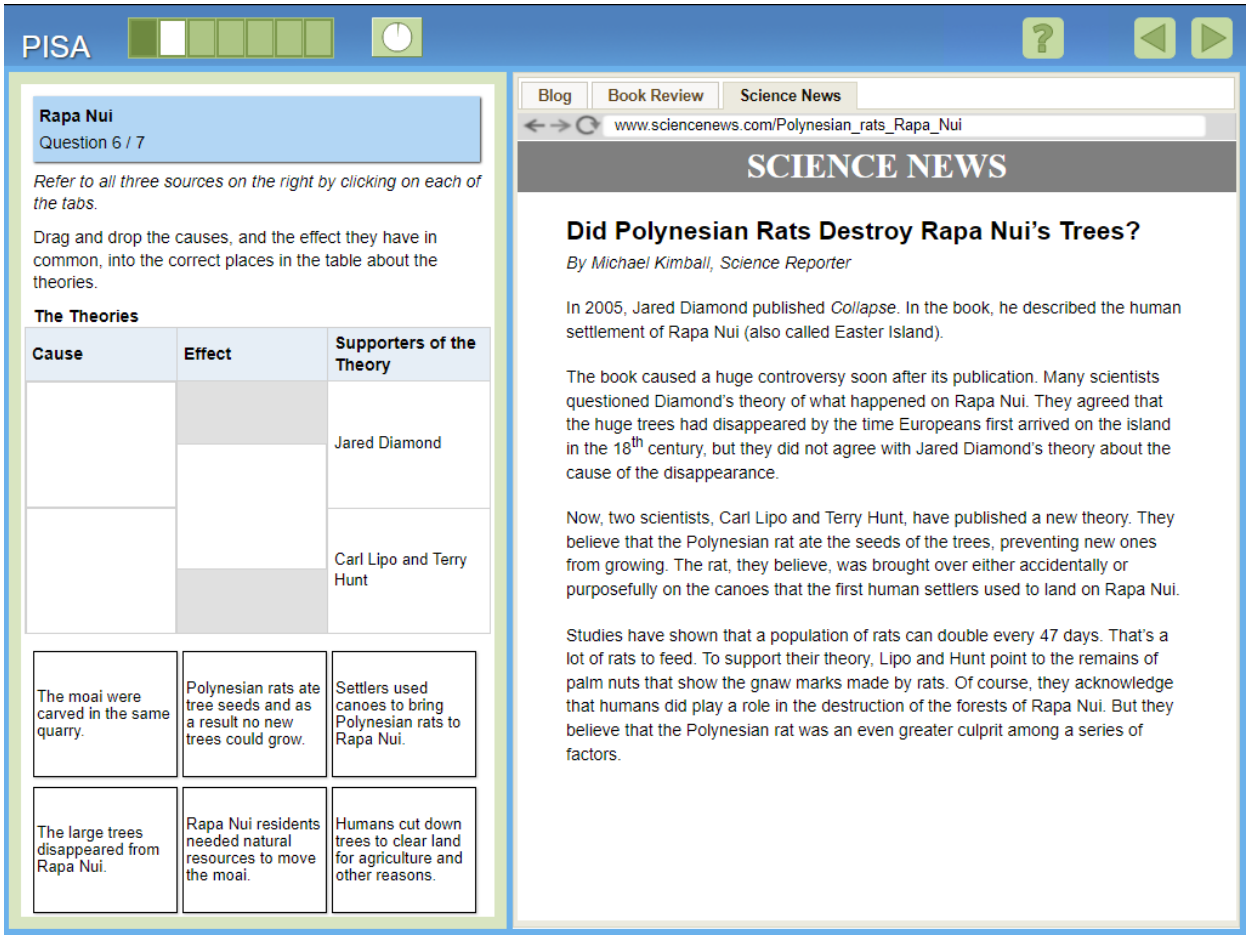
Zamarro, G., C. Hitt and I. Mendez (2019), "When Students Don't Care: Reexamining International Differences in Achievement and Student Effort", *Journal of Human Capital*, Vol. 13/4, pp. 519-552, https://doi.org/10.1086/705799. [8]

Zehner, F. et al. (2020), "PISA reading: Mode effects unveil in short text responses", *Psychological Test and Assessment Modeling*, Vol. 62/1, pp. 85-105. [32]

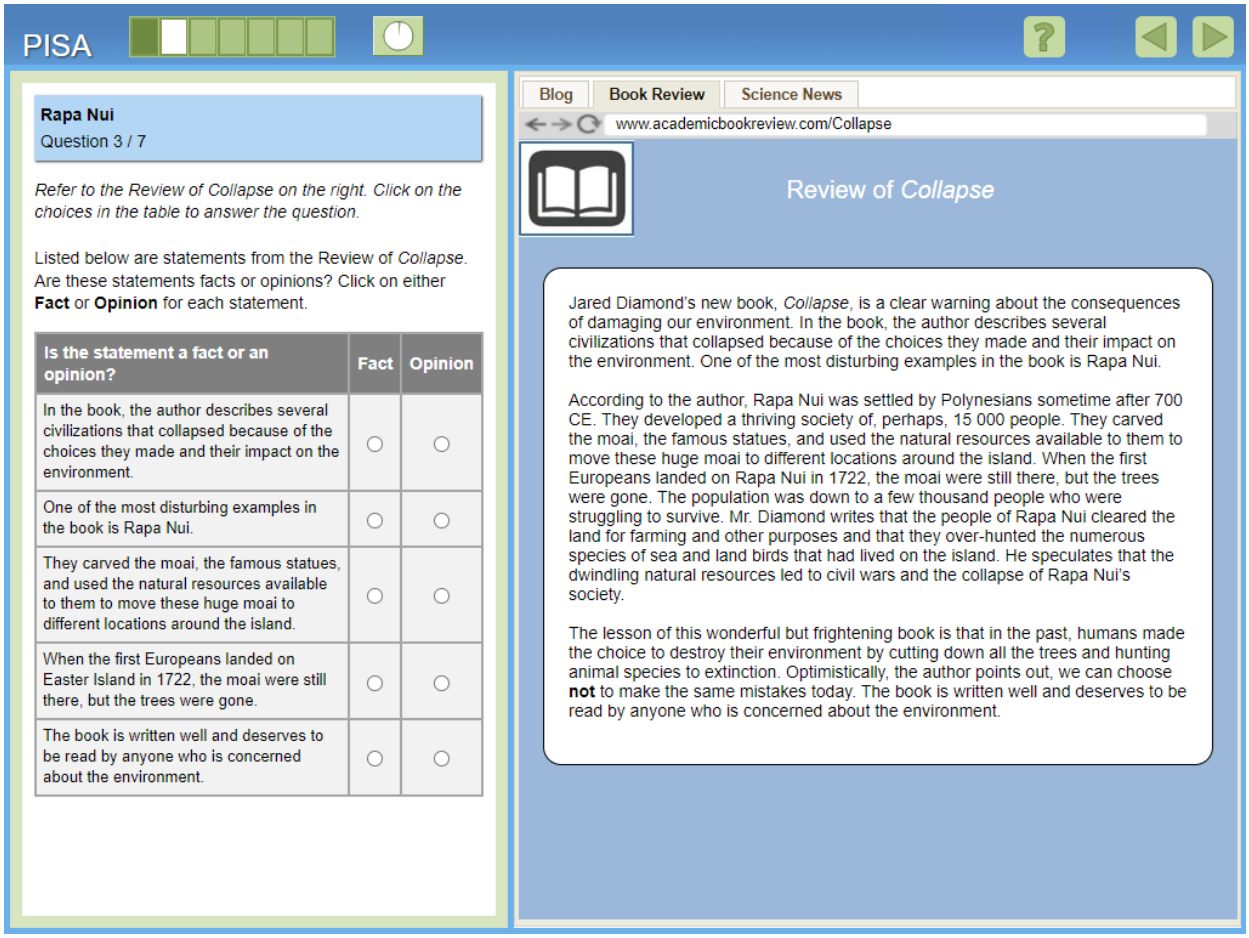
Annex A. Open-Ended Response Item Example



Annex B. Matching Response Item Example



Annex C. Matrix Multiple-Choice Item Example



Annex D. Simple Multiple-Choice Item Example