Kuang and Sahin
Large-scale Assessments in Education (2023) 11:4
https://doi.org/10.1186/s40536-023-00152-0

Large-scale Assessments in Education

RESEARCH Open Access

Comparison of disengagement levels and the impact of disengagement on item parameters between PISA 2015 and PISA 2018 in the United States

Huan Kuang[1] and Fusun Sahin[2*]

*Correspondence: fsahin@cainc.com

[1] School of Human Development and Organizational Studies in Education, University of Florida, Gainesville, FL 32611, USA
[2] Curriculum Associates, 153 Rangeway Road, North Billerica, MA 01862, USA

Abstract

Background: Examinees may not make enough effort when responding to test items if the assessment has no consequence for them. These disengaged responses can be problematic in low-stakes, large-scale assessments because they can bias item parameter estimates. However, the amount of bias, and whether this bias is similar across administrations, is unknown. This study compares the degree of disengagement (i.e., fast and non-effortful responses) and the impact of disengagement on item parameter estimates in the Programme for International Student Assessment (PISA) across the 2015 and 2018 administrations.

Method: We detected disengaged responses at the item level based on response times and response behaviors. We used data from the United States and analyzed 51 computer-based mathematics items administered in both PISA 2015 and PISA 2018. We compared the percentage of disengaged responses and the average scores of the disengaged responses for the 51 common items. We filtered disengaged responses at the response- and examinee-levels and compared item difficulty ($P+$ and b) and item discrimination (a) before and after filtering.

Results: Our findings suggested that there were only slight differences in the amount of disengagement in the U.S. results for PISA 2015 and PISA 2018. In both years, the amount of disengagement was less than 5.2%, and the average scores of disengaged responses were lower than the average scores of engaged responses. We did not find any serious impact of disengagement on item parameter estimates when we applied response-level filtering; however, we found some bias, particularly on item difficulty, when we applied examinee-level filtering.

Conclusions: This study highlights differences in the amount of disengagement in PISA 2015 and PISA 2018 as well as the implications of the decisions made for handling disengaged responses on item difficulty and discrimination. The results of this study provide important information for reporting trends across years.

Keywords: Disengaged responses, Disengagement, Process data, Item parameter estimates, Large-scale assessment, PISA

Introduction

Disengagement, which is defined as providing or omitting responses to test items without making an adequate effort, is a problem for many tests. Examinees' true ability cannot be understood from scores if they do not exert sufficient effort to solve the items. Consequently, the interpretation of the test scores may be inappropriate, and the validity of the inferences based on these scores would deteriorate (Wise, 2017). The issue of disengagement can be particularly problematic in low-stakes assessments where scores do not have any consequence on the examinees. Examinees may provide disengaged responses due to a lack of motivation (Wise, 2015). As such, disengagement has been recognized as a problem for many low-stakes, large-scale assessments, such as the Programme for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP).

Disengagement has been found to bias the estimation of examinees' ability (Wise, 2015; Wise & Kingsbury, 2016) and item parameters (Wise & DeMars, 2006; Yamamoto, 1995). Although studies have examined differences in disengagement between students, schools, and countries (Debeer et al., 2014; Rios & Guo, 2020), they have not done so across years. This is an important gap, because many large-scale assessments, such as PISA and NAEP, report score trends between years across countries and subgroups of students. If the degree of disengagement differs across administrations, the impact of disengagement on item parameter estimates can also vary across administrations, and this would lead to an issue of comparability of scores and, therefore, of the score trends reported for countries and subgroups. This study addresses this gap in the literature by investigating the level of disengagement (i.e., the percentage of responses or examinees detected as disengaged for each item) in items administered in both PISA 2015 and PISA 2018 and the impact of disengagement on item parameter estimates.

Most of the literature on disengagement has focused on one type of disengagement, namely rapid guessing, to multiple-choice single-select (MCSS) items. Recently, Sahin and Colvin (2020) broadened the conceptualization of disengagement to include rapid guesses to MCSS items, as well as rapid-omit, rapid-irrelevant responses to constructed-response items, coining the term *rapid disengagement* to represent this broader concept, which covers different item types (e.g., constructed-response items) and response decisions (e.g., no response). Because there are multiple item types in PISA, we followed the broader conceptualization of disengagement in this study and used the term "disengagement" in the same sense as "rapid disengagement."

Literature

In this section, we first discuss the relationship between disengagement and item parameters and then review approaches for detecting disengagement.

Disengagement and item parameters

Recent research has found that disengagement differs by item. For example, Schnipke and Scrams (1997) claimed that the rapid guessing was essentially the same across items. However, Goldhammer et al. (2017) indicated that disengaged responses were provided more commonly in response to difficult items than to easy items in the Programme for the International Assessment of Adult Competencies (PIAAC). This suggests that the bias stemming from disengaged responses can differ from item to item.

Many studies have examined the impact of disengagement on bias in the estimation of item parameters. In a meta-analysis study, Rios and Deng (2021) investigated 53 studies that used different criteria for classifying examinees as disengaged and produced effect sizes on how much bias that disengagement introduced to item parameters. Rios and Deng made three key observations: (a) studies typically investigated the impact of disengagement on item difficulty, leaving out item discrimination, (b) the different methods used to detect disengagement resulted in differences in the number of disengaged examinees detected, and (c) these differences were not associated with statistically significant differences in average item difficulty after applying a process called *motivation filtering*, a term suggested by Sundre and Wise (2003), which suggests removing disengaged responses.

As for specific examples of research on the impact of disengagement, Yamamoto (1995) found that 30% of examinees omitted and rapid-guessed on one-third of the items in a simulation study, resulting in changes in both item discrimination and item difficulty parameters when they were estimated using a two-parameter logistic (2PL) item response theory (IRT) model. However, Yamamoto (1995) did not observe a clear pattern in how the omitted and rapid-guessed responses influenced the item parameters. Item discrimination and item difficulty parameters increased dramatically in some items but decreased in others. Similarly, Wise and DeMars (2006) compared the original and estimated values of item parameters when rapid-guessing was present in 2.3%, 6.7%, and 11.3% of the responses. They found that rapid-guessing led to overestimation of both item difficulty and item discrimination when a three-parameter logistic (3PL) IRT model was used. Wise et al. (2006) labeled approximately 11–53% of examinees as disengaged in five different assessments. They found that mean test scores increased but that standard deviations decreased after filtering out the disengaged examinees. Bovaird and Embretson (2006) found that item discrimination decreased significantly, but that item difficulty increased significantly, in a 2PL IRT model after applying motivation filtering. While these studies highlight the possible relationship between disengaged responses and item parameters, they do not indicate how much item disengagement is sufficient to cause significant bias in item parameters. The stability of these differences across administrations is also unknown.

Approaches to identifying disengagement

A few statistical approaches have been developed for identifying disengagement. Most approaches use variables, derived from process data, which refer to the cumulation of records of examinees' clicks and keystrokes while they are taking computer-based tests. Identifying disengagement at the item level (i.e., item-level disengagement) typically requires establishing thresholds on variables such as response time (i.e., total time spent on an item). If the response time associated with a response is at or below the threshold for a specific item, the response is labeled as disengaged; if it is above the threshold, the response is labeled as engaged.

In a meta-analysis, Rios and Deng (2021) found that the choice of the response time threshold was associated with nonnegligible differences in the number of responses and examinees identified as disengaged. Therefore, we believe it is useful to outline some of the most common ways to set response time thresholds and how they are used with other variables. The item-level detection approaches described in the literature fall into one of the three categories: (a) response time only, (b) response time and accuracy, and (c) response time and response behaviors.

Response time only This approach requires defining a response time threshold, which corresponds to the minimum response time that is needed for an examinee to provide an engaged response (e.g., Wise & Kong, 2005). Kong et al. (2007) outlined four ways to specify a threshold: (a) the Common Threshold Method, which proposes a constant threshold (e.g., three seconds) for all of the items on the test; (b) the Reading Time Method, which estimates reading time with item surface features, such as the number of characters, and ancillary reading; (c) the Visual Spike Method, which inspects the response time distribution visually and sets the threshold at the endpoint of an early spike in a bimodal response time distribution; and (d) the Mixture Model-Based Method, which fits the response time distribution of an item to a finite mixture model and sets the threshold based on the best-fitting model. Wise and Ma (2012) introduced a fifth method, namely the Normative Threshold Method, which defines the threshold as a certain percentage of the average item response time of all examinees. Wise and Ma (2012) found that a threshold as 10% of the average response time, with a maximum value of 10 seconds, best distinguished rapid guessing from solution behavior compared to other percentages studied. One caveat of response time only methods is that they can misclassify fast-thinking test takers as disengaged (Wise, 2017). To overcome this shortcoming, researchers have proposed methods to detect disengagement that use response times in conjunction with other variables.

Response time and accuracy Using both response time and response accuracy is an alternative approach to setting the response time threshold (Guo et al., 2016; Lee & Jia, 2014; Ma et al., 2011). The first step is to compute the *proportion correct conditional on response time* for each item (Ma et al., 2011). Then, the threshold is set at the first response time corresponding to a proportion correct that is greater than the random chance level (i.e., 25% for an MCSS item with four options). One caveat of using this method is that the response accuracy associated with rapid guessing can be significantly different from random chance (Wise, 2017; Wise & Ma, 2012). Sahin and Colvin (2020) reported that the probability of a correct response is zero for a rapid response to constructed-response items. Similarly, the probability of a correct response is zero for a rapid omit to any kind of item (Sahin & Colvin, 2020). Thus, response accuracy is inapplicable to specify rapid guessing at a random chance level to item types other than MCSS and to rapid-omit behaviors.

Response time and response behaviors The use of response time and response behaviors, what Sahin and Colvin (2020) referred to as the “enhanced” method, has been shown to be a better approach to detecting examinees who display disengagement

(Sahin & Colvin, 2020). Specifically, this method detects disengagement using the *number and type of response behaviors* (e.g., keypresses, clicks, and clicking interactive tools), which are derived from the process data, in addition to response time. To apply this approach, two sets of thresholds are set jointly for an item: one for response time and one for the number of response behaviors. The threshold for the number of response behaviors specifies the maximum number of response behaviors that exhibit no or minimum engagement. If an examinee responds to an item faster than (or equal to) the response time threshold and performs fewer actions than (or equal to) the number of response behaviors threshold for that item, that response is flagged as disengaged. Sahin and Colvin (2020) used a constant value as the threshold for the number of actions for all of the items under investigation and suggested that the distribution of the number of actions can be used to set the threshold in future research.

Research questions

The aim of this study is to compare the degree of disengagement and the impact of disengagement on item parameter estimates in low-stakes, large-scale assessments across administrations. To achieve this goal, we investigated differences between the prevalence and impact of disengagement in the 2015 and 2018 administrations of PISA. The research questions are:

1. How much does the percentage of disengagement differ between the items common to PISA 2015 and PISA 2018?
2. How much do the scores of disengaged responses differ between the items common to PISA 2015 and PISA 2018?
3. How much do estimates of item difficulty and item discrimination, with and without disengagement, change between 2015 and 2018?

Method

In this section, we will first introduce the data used in this study and then discuss the analyses conducted, step by step, from detecting disengagement to comparing the percentage of disengagement in PISA 2015 and PISA 2018; comparing scores for disengaged responses in PISA 2015 and PISA 2018; and comparing weighted item parameter estimates with and without disengagement in PISA 2015 and PISA 2018.

Data

The Programme for International Student Assessment (PISA) was developed by the Organisation for Economic Co-operation and Development (OECD) to monitor student performance and provide comparative indicators of education systems across the world (OECD, 2000). It is administered every three years to 15-year-old students in more than 70 countries in reading, mathematics, and science. In each cycle, PISA focuses on one of these subjects, and the other two subjects are administered as minor assessment areas for trend purposes. In this study, we analyzed mathematics data from U.S. students in the two most recent PISA administrations, conducted in 2015 and 2018. Specifically, we analyzed 51 items common to both the 2015 and 2018 administrations. These items were distributed in various blocks due to matrix sampling.

The total number of U.S. students who participated in the mathematics assessment was 5712 in 2015 and 4838 in 2018, including 2854 (50%) female and 2858 (50%) male students in 2015, and 2376 (49.1%) female and 2462 (50.9%) male students in 2018. The majority of the students were in grade 10 (73.7% in 2015 and 74.4% in 2018), about 10% were 9th graders or lower (9.5% in 2015 and 8.5% in 2018), and about 15% were 11th graders or higher (16.8% in 2015 and 17.2% in 2018). In both years, 90% of the students did not repeat a grade, and the rest 10% repeated a grade. Because randomly equivalent students receive each block of test items, it is reasonable to assume that the overall size, demographic composition, and ability level of the analytical sample are similar for each item. The sample size for each item ranged from 643 to 736 in 2015 and from 767 to 833 in 2018. For analyses of the weighted $P+$ and item parameter estimates, we used the final student weight ("W_FSTUWT") variable, which is available in the public-use datasets.

A computer-based assessment (CBA) was the main mode of assessment in both PISA 2015 and 2018. In this study, we used three variables that are available in the PISA public-use datasets in both 2015 and 2018: *Total Time, Number of Actions,* and *Scored Response. Total Time* is a continuous variable derived from the process data for each item that specifies the total amount of time that each student spent on the items. *Number of Actions* is another continuous variable derived from the process data for each item that specifies the number of steps each student took before giving their final responses (OECD, 2017). *Scored Response* is a categorical variable with six categories: 0 = No Credit, 1 = Full Credit, 6 = Not Reached, 7 = Not Applicable, 8 = Invalid, and 9 = No Response (OECD, 2017, p. 198). None of the responses for the 51 items included in this study were in category 7 (Not Applicable) or category 8 (Invalid). In order to analyze the average score, we recoded category 6 (Not Reached) responses to missing values. Omitted responses (category 9, No Response) were recoded to 0, following the same coding procedure for missing scores that PISA uses in its own methodology (OECD, 2017, p. 149).

Analysis

We followed the enhanced method (Sahin & Colvin, 2020) to detect disengagement due to the limitations of the other methods, as discussed above. The enhanced method detects responses that are more likely to represent disengagement and covers all of the types of disengagement that are likely to be present in the PISA items: rapid guessing to MCSS items, rapid omitting, and rapid-irrelevant responses to constructed-response items. To apply this method, we set thresholds for *Total Time* and for *Number of Actions*. The remainder of this section provides detailed information on the steps that we took: (1) establishing *Total Time* thresholds for each item, (2) establishing *Number of Actions* thresholds for each item, (3) comparing the percentage of disengagement between 2015 and 2018, (4) comparing the scores of disengaged responses between 2015 and 2018, (5) removing the responses identified as disengaged from the analytical sample (i.e., motivation filtering), and (6) comparing the weighted item parameters ($P+$, a, and b) before and after applying motivation filtering in 2015 and 2018.

Establishing total time thresholds

Among the various threshold-setting methods suggested in the literature, we utilized the *Normative Threshold Method* (Wise & Ma, 2012). It suggests setting the response time threshold for an item as 10% of the average response time, with a maximum value of 10 seconds. For each item in this study, average response time was computed based on the *Total Time* variable. Any value larger than 10 minutes for one item was considered an outlier and removed from computing the average response time since examinees were expected to complete the test within 60 minutes.

For comparison purposes, we set a common threshold value for the same item in 2015 and 2018. To achieve this, we first checked and confirmed that the *Total Time* variables for each item in 2015 and 2018 had the same distribution. Specifically, we inspected whether the distributions were similarly based on minimum, maximum, mean, and mode values, which are also related to the location of the peak and the overall shape. We did not observe a binomial distribution in many items, which is another reason for our decision to use the *Normative Threshold Method*. Then, we merged the *Total Time* variables in 2015 and 2018 for each item and computed 10% of the average response time, with a maximum of 10 seconds, as the common threshold.

Establishing number of actions thresholds

Because the minimum number of interactions needed to provide an effortful response can vary substantially across the items in our data, we set the *Number of Actions* thresholds by adapting the *Visual Spike* idea developed for response time (Wise & Kong, 2005) to the number of actions. When we inspected the distribution of the *Number of Actions* variables, we observed a unimodal distribution. Therefore, we specified the threshold at the beginning of the spike, assuming that this value represents the minimum number of actions that effortful students take. Given that each character entry or click is counted as an action, 200 actions represented a relatively lengthy response (e.g., 50 words) to constructed-response items, and any number of actions greater than 200 was considered an outlier and not included in plotting the distribution.

For comparison purposes, we set a common *Number of Actions* threshold value for the same item in 2015 and 2018. The first step in this procedure was to inspect the distribution of the *Number of Action*s variable for each item in 2015 and 2018. While the shape of the distribution is similar in both years, the values for the *Number of Action*s variable in 2018 were one less than the corresponding values in 2015. Confirming our observation, we learned that clicking the “next” button to move to the next item in a unit or to the next unit in the test was counted in computing the total number of actions in 2015 but not in 2018 (M. Ikeda, personal communication, August 20, 2020). To obtain the same scale for this variable, we added the value of “one” to the *Number of Action*s variable in 2018 and then merged the variables in 2015 and 2018 for each item. We then plotted the variables based on the merged data for each item and set the value at the beginning of the spike in the distribution as the threshold.

Comparing the percentage of disengagement

To answer the first research question, we detected disengaged responses based on the *Total Time* threshold and *Number of Actions* threshold for each item. Then we compared the percentage of disengagement for each item between administration year.

Comparing the scores of disengaged responses

To answer the second research question, we first compared the scores for disengaged responses and engaged responses separately in each year. Then we compared the differences between the scores for the engaged and disengaged responses across years.

Motivation filtering

Removing disengagement from the data is termed *motivation filtering*, and it requires treating the data points associated with disengagement as missing. However, there is no consensus on how motivation filtering should be applied. While some studies remove only the responses that are detected as disengaged, other studies remove all of an examinee's responses if any response from that examinee is associated with disengagement. Rios et al. (2017) coined the term *response-level filtering* to refer to the first type of motivation filtering and *examinee-level filtering to refer* to the latter. Wise (2009) used the term *rapid-response filtering* to refer to response-level filtering.

In this study, we used both response-level and examinee-level filtering to understand the role of the filtering method in examining the impact of disengagement on item parameters. When applying examinee-level filtering in this study, all of the responses from students who provided a disengaged response to at least one item were removed. In total, 145 students, or 3% of the sample, were removed in 2015 when examinee-level filtering was applied. Similarly, 242 students, or 4.23% of the sample, were removed in 2018 when examinee-level filtering was applied.

Comparing the impact of disengagement on weighted item parameters

To answer the third research question, we compared item parameter estimates computed under both classical test theory (CTT) and IRT. First, we took the subset of students' scores to all 51 items common to both the 2015 and 2018 administrations of PISA. We then computed item difficulty (i.e., the proportion of correct responses, $P+$) following the same coding procedure used for missing scores in PISA, where omitted responses are scored as incorrect. We compared the value for $P+$ before and after applying response- and examinee-level filtering.

Next, we conducted the national item calibration by computing the IRT item parameter estimates for the items administered in the United States in 2015 and 2018. Consistent with the technical procedures followed in PISA, a 2PL model was used for the binary items and a generalized partial credit model (GPCM) for the polytomous item responses (OECD, 2017, 2022). The latent trait (θ) was assumed to be normally distributed, and the mean was set to 0 and the variances to 1 to identify the models. The 51 common mathematic items were scaled separately for each of six conditions: (1) before any filtering in 2015, (2) after applying response-level filtering in 2015, (3) after applying examinee-level filtering in 2015, (4) before any filtering in 2018, (5) after applying response-level filtering in 2018, and (6) after applying examinee-level filtering in 2018. We used the "*mirt*" (Chalmers, 2012) package in the R (R core team, 2020) environment (Version 4.0.2) to estimate the weighted IRT model parameters. We then examined the impact of disengagement on the estimated item difficulty (b) and item discrimination (a). We compared a and b before filtering and after applying response- and examinee-level filtering, following the same procedure used to compare $P+$.

Results

Percentage of disengagement

We examined the percentage of disengagement for each CBA mathematics item in 2015 and 2018 to answer the first research question: *How much does the percentage of disengagement differ between the items common to PISA 2015 and PISA 2018* (see Fig. 1; Table 2). At the item level, the percentage of disengagement was slightly higher in PISA 2018 than in PISA 2015. The percentage of disengagement ranged from 0% to 2.86% in 2015 and from 0.13% to 5.20% in 2018, with an average of 0.79% in 2015 and 1.37% in 2018. Item CM992Q02 was associated with the highest percentage of disengagement in 2015 and in 2018. In both years, the level of disengagement detected was below 1% for most items (36 items in 2015 and 22 items in 2018). No disengagement was detected for two items, CM423Q01 and CM919Q01, in 2015 (see Table 3).

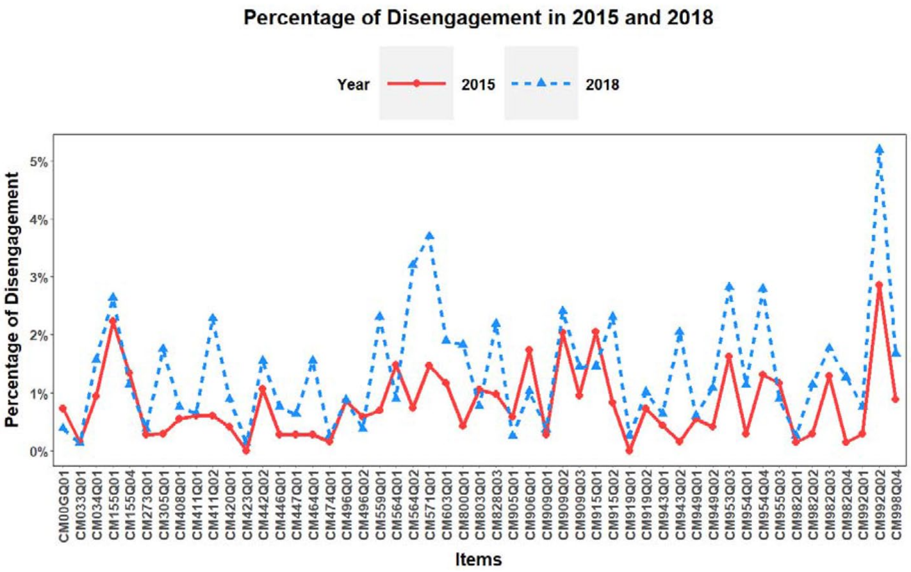

Fig. 1 Percentage of disengagement in 2015 and 2018. *Source:* Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

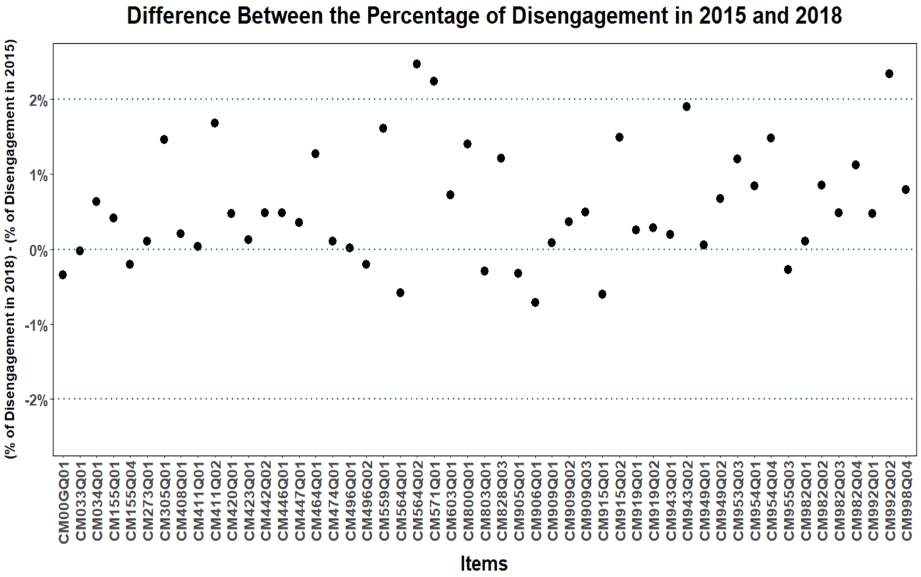

Fig. 2 Difference between the percentage of disengagement in 2015 and 2018. The difference is the percentage of disengagement in 2015 subtracted from the percentage in 2018. *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

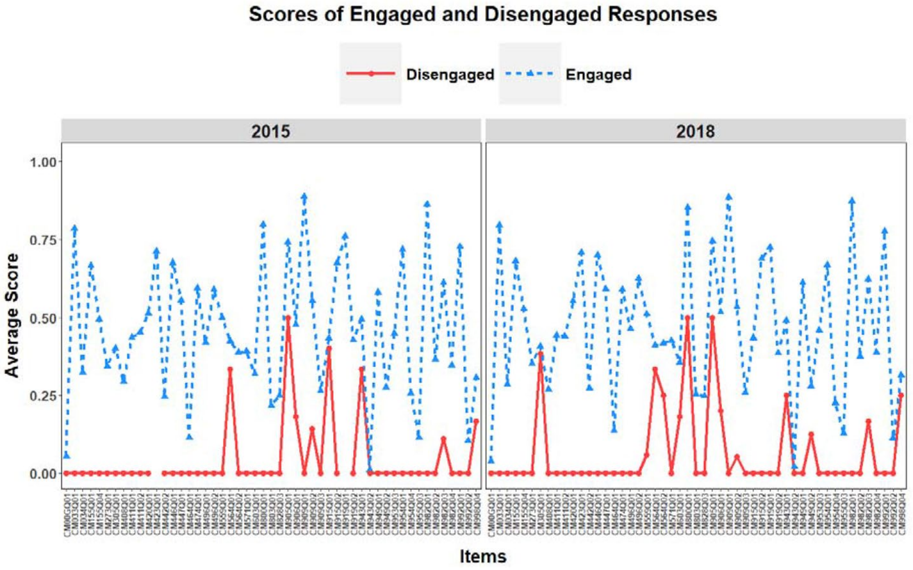

Fig. 3 Scores of engaged and disengaged responses in 2015 and 2018. SOURCE: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

Figure 2 shows the difference between the percentage of disengagement in 2015 and 2018 for each of the 51 items, computed as the percentage of disengagement for an item in 2015 subtracted from the percentage of disengagement for that item in 2018. Each dot represents an item, and many of the dots are located around the horizontal line $y = 0$, thus representing very small differences. The largest difference between the percentage of disengagement in 2015 and 2018 was 2.47% on item CM564Q02, and the smallest difference was 0.01% on item CM496Q01. The differences were less than 1% in 37 (or 72.5%) of the items, less than 2% but more than 1% in 11 (or 21.6%) of the items, and less than 3% but more than 2% in 3 (or 5.9%) of the items.

Scores of disengaged responses

The results for the second research question—*How much do the scores of disengaged responses differ between the items common to PISA 2015 and PISA 2018*—suggest that disengaged responses received lower scores than engaged responses in both 2015 and 2018 (see Fig. 3). This finding is consistent with the expectation that disengaged responses are less likely to be correct than are engaged responses.

The average scores of disengaged responses ranged from 0 to 0.5 in both 2015 and 2018, while the average scores of engaged responses ranged from 0.01 to 0.89 in 2015 and from 0.02 to 0.89 in 2018 (see Fig. 3; Table 4). In both 2015 and 2018, the average scores of engaged responses were similar to the average scale scores reported for the population (OECD, 2017), with the largest difference equal to 0.07. The average scores of disengaged responses were 0 for most of the 51 items (41 items in 2015 and 38 items in 2018). Given that omitted responses were also scored as incorrect, and represented with a score of 0, most of the disengaged responses were either incorrect or represented no response in most of the items. The average scores of disengaged responses were not applicable (N/A) for two items in 2015 (CM423Q01 and CM919Q01), because no disengagement was detected for these items.

Among the 51 items examined, the average scores of the disengaged responses did not change between years for 37 items; in 35 of these items, the average score was zero. For 14 items, the average scores under disengagement differed slightly between 2015 and 2018 without a clear pattern being observed. In nine of the items, the average score under disengagement was greater in 2018, but in three items, the average score was greater in 2015. In two of the items, no examinees were detected as disengaged in 2015; therefore, the average could not be compared between years.

Changes in item parameters

To answer the third research question—*How much do estimates of item difficulty and item discrimination, with and without disengagement, change between 2015 and 2018*—we computed and compared item difficulty ($P+$ and b) as well as item discrimination (a).

Comparison of P+

The overall pattern of differences in $P+$ with disengagement (i.e., before filtering) and without disengagement (i.e., after filtering) was the same in 2015 and 2018. $P+$ increased for most items (about 40) in both years after applying either response-level or

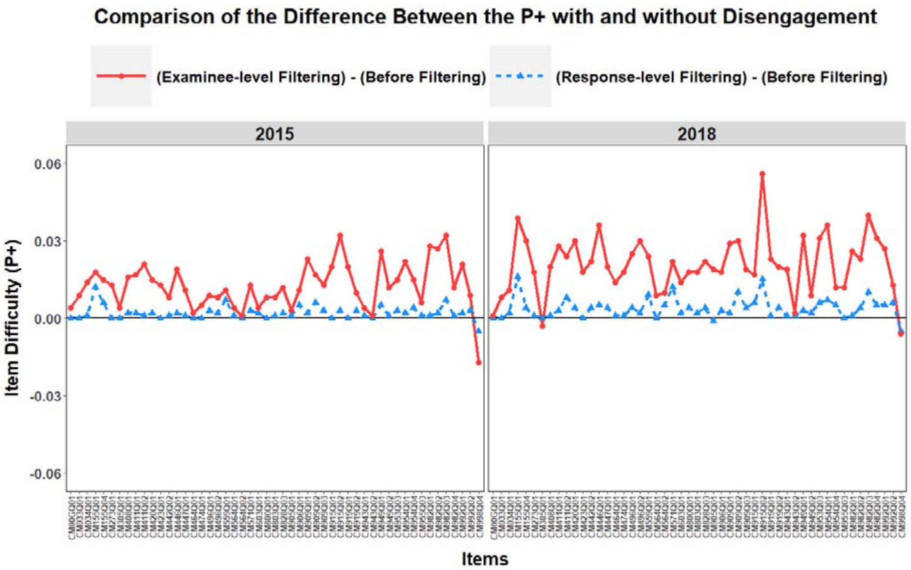

Fig. 4 Comparison of the difference between the $P+$ with disengagement and without disengagement. The blue line refers to the difference between $P+$ with disengagement and after applying response-level filtering; and the red line refers to the difference between $P+$ with disengagement and after applying examinee-level filtering. *Source*: Organisation for Economic Cooperation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

examinee-level filtering (see Fig. 4). In other words, most items became slightly easier after motivation filtering was applied. In both years and for most items, applying examinee-level filtering resulted in larger differences in $P+$ than response-level filtering (Fig. 4; Table 5).

After applying response-level filtering, $P+$ increased slightly for 38 items in 2015 and 50 items in 2018, and it remained the same for 12 items in 2015. For two items, CM998Q04 in both 2015 and 2018, as well as CM905Q01 in 2015, $P+$ decreased slightly (< 0.005). Item CM155Q01 showed the largest difference in $P+$, increasing by 0.012 in 2015 and by 0.016 in 2018. The absolute values of the differences for all 51 items in both 2015 and 2018 were less than 0.02.

After applying examinee-level filtering, $P+$ increased slightly for 43 items in 2015 and 49 items in 2018, and it remained the same for 6 items in 2015. For two items, CM998Q04 in both 2015 and 2018, as well as CM305Q01 in 2018, $P+$ decreased slightly (< 0.017). 12 items in 2015 and 29 items in 2018 had nonignorable differences (absolute values of differences ≥ 0.02) in $P+$ after applying examinee-level filtering. Among these items, CM915Q02 had the largest difference in $P+$, with increases of 0.032 in 2015 and 0.056 in 2018.

Moreover, of the 15 items in 2015 and 29 items in 2018 that had a higher degree of disengagement (i.e., 1% or higher), the majority (10 in 2015 and 26 in 2018) also had a larger increase in $P+$ (about 0.01 or higher) after applying examinee-level filtering. However, we did not observe this pattern after applying response-level filtering. There were exceptions in which $P+$ did not increase very much after applying examinee-level filtering

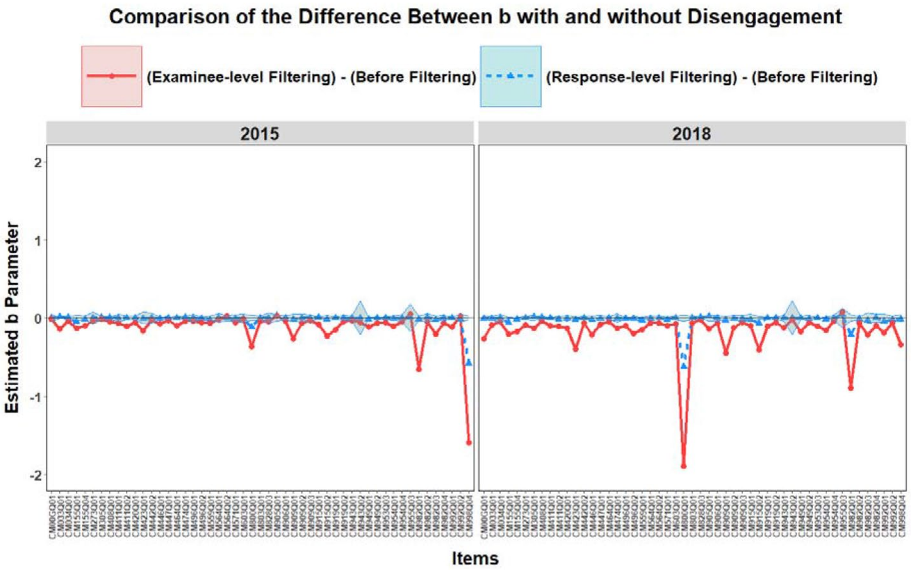

Fig. 5 Comparison of the difference between *b* with disengagement and without disengagement. The blue line refers to the difference between *b* with disengagement and after applying response-level filtering; the red line refers to the difference between *b* with disengagement and after applying examinee-level filtering; and the light green represents the range of the SE of the *b* parameter estimates with disengagement (i.e., before filtering). *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

even though a relatively high percentage of the responses were disengaged. For example, item CM943Q02, where disengagement was detected in 2.05% of the responses in 2018, $P+$ increased to only 0.002 after examinee-level filtering was applied.

Comparison of item difficulty (b)

The overall pattern of differences in b parameter estimates with disengagement (i.e., before filtering) and without disengagement (i.e., after filtering) was the same in 2015 and 2018. The b parameter estimates decreased for most items in both years after applying either response-level filtering (35 in 2015 and 41 in 2018) or examinee-level filtering (47 in 2015 and 50 in 2018, see Fig. 5; Table 6). Consistent with the findings for $P+$, most items became slightly easier after motivation filtering. In both years and for all 51 items, the differences in the b parameter estimates were larger after applying examinee-level filtering than response-level filtering.

After applying response-level filtering, the change in the b estimates ranged from -0.583 to 0.028 in 2015 and from -0.621 to 0.044 in 2018. For two items in 2015 and two items in 2018, the absolute values of the differences in the b parameter estimates before and after applying response-level filtering were larger than 0.1. To observe changes in individual items, we also took the standard error (SE) of the b parameter estimates into account. We concluded that the b values were different if the absolute difference between the b parameter estimates before and after applying filtering was larger than the SE of the b parameter estimates before filtering. Based on this comparison, the b values were found to be different for 8 items in 2015 and 16 items in 2018 after applying response-level filtering.

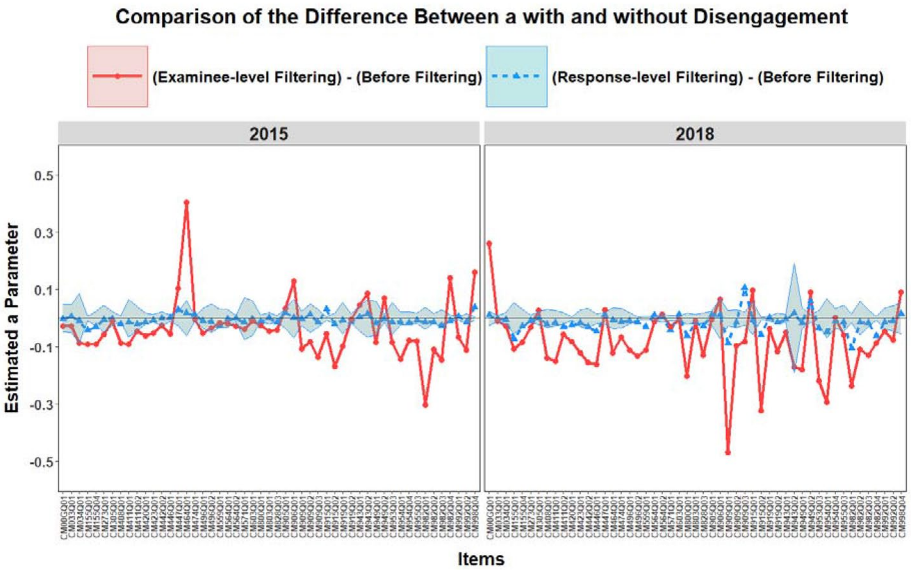

Fig. 6 Comparison of the difference between *a* with disengagement and without disengagement. The blue line refers to the difference between *a* with disengagement and after applying response-level filtering; the red line refers to the difference between *a* with disengagement and after applying examinee-level filtering; and the light green represents the range of the SE of the *a* parameter estimates with disengagement (i.e., before filtering). *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

After applying examinee-level filtering, the changes in the *b* values ranged from −1.595 to 0.05 in 2015 and from −1.891 to 0.078 in 2018. For ten items in 2015 and 25 items in 2018, the absolute values of the difference in the *b* parameters before and after applying examinee-level filtering were larger than 0.1. Applying examinee-level filtering resulted in slightly larger differences in the *b* values than applying response-level filtering. Considering the SE, the *b* values were different for 36 items in 2015 and for 48 items in 2018 after applying examinee-level filtering.

Most items that had a relatively higher degree of disengagement (i.e., 1% or higher) also had a larger decrease in *b* estimates after applying response-level (about 0.2 or larger) or examinee-level (about 1 or larger) filtering. In particular, *b* estimates with extreme values tended to decrease more after applying examinee-level filtering: for example, CM998Q04 ($b = 4.198$ before filtering; decreased by 1.595 after applying examinee-level filtering in 2015), CM800Q01 ($b = -3.849$ before filtering; decreased by 1.89 after applying examinee-level filtering in 2018), and CM982Q01 ($b = -2.244$ before filtering; decreased by 0.90 after applying examinee-level filtering in 2018).

Comparison of item discrimination (a)

The overall pattern of differences in item discrimination (*a*) with disengagement (i.e., before filtering) and without disengagement (i.e., after filtering) was the same in 2015 and 2018. The *a* parameter estimates decreased for most items in both years after applying either response-level filtering (35 items in 2015 and 38 items in 2018) or examinee-level filtering (42 items in 2015 and 42 items in 2018, see Fig. 6). In both years, the differences in *a* parameter estimates were larger for almost all items (47 items in both 2015 and 2018) after applying examinee-level filtering than response-level filtering.

After applying response-level filtering, the change in *a* values ranged from −0.042 to 0.039 in 2015 and from −0.105 to 0.106 in 2018. The *a* parameter estimates decreased less than 0.1 for 35 items in 2015 and 37 items in 2018, and decreased more than 0.1 but less than 0.2 for none of the items in 2015 and only one item in 2018. Similar to our comparison of the *b* values, we took the SE of the *a* parameters into consideration to observe changes in individual items. We concluded that the *a* values were different if the absolute difference between the *a* parameter estimate before and after applying filtering was larger than the SE of the *a* parameter estimate before filtering. With respect to the SE, the *a* values were different for 11 items in 2015 and 18 items in 2018 after applying response-level filtering.

After applying examinee-level filtering, the change in *a* values ranged from −0.303 to 0.406 in 2015 and from −0.468 to 0.262 in 2018. The *a* parameter estimates decreased less than 0.1 for 34 items in 2015 and 20 items in 2018, decreased more than 0.1 but less than 0.2 for 7 items in 2015 and 16 items in 2018, and decreased more than 0.2 for 1 item in 2015 and 6 items in 2018. With respect to the SE, the *a* values were different for 39 items in 2015 and 44 items in 2018 after applying examinee-level filtering. Applying examinee-level filtering resulted in slightly larger differences in the *a* values than did applying response-level filtering.

Summary

This study provides insight into the comparability of the percentage of disengagement and of the average scores of disengaged responses in PISA 2015 and PISA 2018, as well as of the impact of disengagement on item parameter estimates.

As to the first research question—*How much does the percentage of disengagement differ between the items common to PISA 2015 and PISA 2018*—the results suggest that in the U.S. sample there was only small differences between PISA 2015 and PISA 2018. Less than 5.2% of the responses were associated with disengagement in individual CBA math PISA items in both years, and 3% of the examinees in 2015 and 4.23% in 2018 were associated with disengagement in at least one item.

For the second research question—*How much do the scores of disengaged responses differ between the items common to PISA 2015 and PISA 2018*—we found that the average scores of disengaged responses were less than 0.5 and that they were lower than the average scores of engaged responses. This pattern was the same in PISA 2015 and PISA 2018.

For the third research question—*How much do estimates of item difficulty and item discrimination, with and without disengagement, change between 2015 and 2018*—the results show that the overall pattern of differences in item parameters with disengagement and after applying response-level and examinee-level filtering was similar across years. A summary of the results of the changes in item parameters is provided in Table 1. Applying response-level filtering resulted in small differences in the *P*+, *a*, and *b* values.

Applying examinee-level filtering resulted in relatively large differences in the $P+$, a, and b values, even when a small percentage of disengaged responses was detected for an item, introducing some bias. A similar pattern of results was obtained in Rios et al. (2017) in that examinee-level filtering biased the mean scores in their study.

Discussion and conclusions

Reporting score trends across years is crucial for large-scale assessments, such as PISA and NAEP. Thus, the results of this study on differences in disengagement across years, which pertain to trend reporting, provide important information for such assessments.

There are many factors that affect the impact of disengagement on item parameters. One of the main factors, and one that is the focus of this study, is the percentage of disengaged responses or examinees. Another factor is the method used to detect disengagement. In this study, we used an enhanced item-level disengagement method that is based on both response time and response actions. A third related factor is how the disengaged responses are handled. To illustrate this, we used both response-level and examinee-level filtering. The results showed that different patterns were observed in changes in the a and b parameter estimates when response-level and examinee-level filtering were applied.

To elaborate on our method for detecting disengaged responses, we chose a conservative approach that detected only the responses that had the highest likelihood of displaying disengagement based on their response time and number of actions. Wise (2017) noted that taking a more conservative approach (i.e., failing to detect some potentially disengaged examinees) is preferable to one that falsely detects examinees as disengaged, given that both errors cannot be minimized simultaneously. We examined two ways in which we could relax our constraints: (1) removing the constraint of a 10-second maximum for the response time threshold; and (2) removing the constraint on the minimum number of actions. We tested the impact of these modifications and found that removing the first constraint did not make much difference in the classification of responses as disengaged and removing the second constraint led to unreliable results. The highest value for 10% of the average total time variable was 15 seconds in our sample, which did not introduce much change to the threshold. The scores for the examinees who were labeled as disengaged based on only their response time were close to the average scores of engaged responses, suggesting that some examinees could be falsely labeled as disengaged.

In this study, we used the same thresholds in both 2015 and 2018 for comparison purposes. We were able to form common thresholds by merging the datasets from both years. Our goal was to detect disengagement based on the same criteria in both years. In some cases, the threshold value was not a perfect fit for either the 2015 or 2018 dataset. In these cases, we tried to specify a threshold value that would be equally imperfect for both years so that any potential error in classification would impact the two datasets equally.

As to the filtering method, we used both response-level and examinee-level filtering to illustrate the impact of filtering on the item parameters. With examinee-level filtering, the students removed were the same for each item; for response-level filtering, the students removed were different for each item. Wise (2009) suggested that response-level filtering is appropriate when examinees are not being compared based on their raw scores. The advantage of response-level filtering is that it retains as much data as possible by keeping valid responses that are not impacted by disengagement.

Overall, we did not find any serious impact of response-level filtering on the average percent correct of items, $P+$. Our results are in accordance with previous studies suggesting that response-level filtering does not impact mean test scores (Kong et al., 2007; Wise, 2006). We observed nonignorable changes in item parameters for only a few items after applying examinee-level filtering. Hauser and Kingsbury (2009) found that proficiency estimates were not impacted by examinee-level filtering if the percentage of disengagement did not exceed 20%. In our study, the percentage of disengaged examinees was low, which may explain why we observed changes regarding item parameter estimates with disengagement and without disengagement only in a few items.

Researchers should consider ability distribution of disengaged examinees when they are deciding whether examinee- or response-level motivation filtering would be more appropriate. Specifically, they should be wary of the assumption that disengaged responses are independent of both item (i.e., item difficulty) and test taker (i.e., the latent trait, θ) characteristics, when applying motivation filtering (Rios et al., 2017). Wise's (2009) assumption about examinee-level filtering was that student effort is unrelated to true proficiency. Thus, removing examinees would result in either an underestimation or overestimation of item parameters depending on whether the examinees removed were of high or low ability, if this assumption was violated. In our study, we assumed that item and examinee characteristics are independent of the disengaged behaviors. We observed that the items with the highest percentage of disengagement were not the most difficult items, which echoed the findings from previous studies that item difficulty is not significantly related to examinee effort (Rios et al., 2017; Wise, 2006; Wise & Kingsbury, 2016). This also supported our assumption that there was not a linear relationship between the disengagement levels we observed and the difficulty of the items we examined. As for test taker ability, our analysis focused on the 51 items common to both assessments; thus, we had limited knowledge with which to test our assumption about students' abilities.

We did not find any patterns in the disengaged responses based on examinee demographics. For each item in both 2015 and 2018, male and female examinees equally provided disengaged responses. In both years, 3% of the examinees in each grade were found to be disengaged. Moreover, the level of disengagement was the same among students who repeated grades as among those who did not repeat grades.

Generally speaking, researchers should consider the possibility that test takers use rapid guessing as a test-taking strategy. Test takers can skip items that are difficult for them to allocate more time and energy to the items that they have a higher probability of answering correctly with reasonable effort. Such calculated behavior may result in observing more disengagement in difficult items and some lower-ability students displaying disengaged behavior more frequently. In such cases, response-level filtering would reduce the amount of information gained on the difficult items that were subject to higher levels of disengagement, and examinee-level filtering would eliminate a distinct subgroup of examinees. Therefore, researchers should examine the patterns in the disengagement before deciding to apply response-level or examinee-level filtering.

Table 1. Summary of changes of item parameters

	P+ **2015**	***P+*** **2018**	***b*** **2015**	***b*** **2018**	***a*** **2015**	***a*** **2018**
Response-level filtering	39 items changed (38 increased; 1 decreased)	51 items changed (50 increased; 1 decreased)	8 items changed (2 increased; 6 decreased)	16 items changed (1 increased; 15 decreased)	11 items changed (5 increased; 6 decreased)	18 items changed (14 decreased; 4 increased)
Examinee-level filtering	45 items changed (43 increased; 2 decreased)	51 items changed (49 increased; 2 decreased)	36 items changed (36 decreased)	48 items changed (48 decreased)	39 items changed (8 increased; 31 decreased)	44 items changed (7 increased; 37 decreased)

Finally, researchers should bear in mind that differences in scores between disengaged and engaged examinees are not informative for gauging the impact of disengagement on item parameters. Researchers have used higher average scores from engaged examinees compared to disengaged examinees as evidence of correctly detecting examinees. Consequently, $P+$ of the items is expected to increase after eliminating disengaged responses. However, higher score differences between engaged and disengaged students do not necessarily correlate with higher differences in $P+$. For example, Rios et al. (2017) reported that examinee-level filtering artificially inflated the true mean score when ability was related to disengaged responding. In particular, when examinee-level filtering is applied (or item parameters are estimated with a model-based approach such as IRT), the impact of disengagement on item parameters is not obvious because the item parameter estimates from one item will be influenced by the disengaged responses from the other items.

Overall, this study provides an example of how to detect and handle disengagement in a large-scale assessment administered in two years in the United States based on PISA data. The study highlights the differences in disengagement in both years as well as the implications of the decisions made for handling scores received under disengagement on item difficulty and discrimination. Since the study uses data from one country, researchers should be cautioned against generalizing the results to student populations in other countries and regions. It would, of course, be desirable for future work to compare disengagement across years with data from other countries.

Future research should also consider the potential effects of examinee-level filtering more carefully; for instance, by suggesting cut-offs for the percentage of disengaged examinees to be used in the filtering. Finally, we would like to note that this study was the first to follow a data-driven approach to set the threshold for the number of actions. We adapted the visual approach to specify the threshold at the beginning of a spike, but future studies may apply other methods for setting the threshold.

Appendix

See Tables 2, 3, 4, 5, 6, and 7.

Table 2 Number and percentage of disengaged responses in 2015 and 2018. *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

No.	Item code	Total number of responses		Number of disengaged responses		Percentage of disengaged responses (%)	
		2015	2018	2015	2018	2015	2018
1	CM00GQ01	691	785	5	3	0.724	0.382
2	CM033Q01	668	797	1	1	0.150	0.125
3	CM034Q01	643	766	6	12	0.933	1.567
4	CM155Q01	674	796	15	21	2.226	2.638
5	CM155Q04	671	792	9	9	1.341	1.136
6	CM273Q01	730	795	2	3	0.274	0.377
7	CM305Q01	697	799	2	14	0.287	1.752
8	CM408Q01	728	793	4	6	0.549	0.757
9	CM411Q01	668	788	4	5	0.599	0.635
10	CM411Q02	665	789	4	18	0.602	2.281
11	CM420Q01	730	788	3	7	0.411	0.888
12	CM423Q01	697	797	**0**	1	0.000	0.125
13	CM442Q02	660	776	7	12	1.061	1.546
14	CM446Q01	729	790	2	6	0.274	0.759
15	CM447Q01	731	797	2	5	0.274	0.627
16	CM464Q01	718	775	2	12	0.279	1.548
17	CM474Q01	676	795	1	2	0.148	0.252
18	CM496Q01	698	802	6	7	0.860	0.873
19	CM496Q02	697	799	4	3	0.574	0.375
20	CM559Q01	723	780	5	18	0.692	2.308
21	CM564Q01	679	784	10	7	1.473	0.893
22	CM564Q02	677	780	5	25	0.739	3.205
23	CM571Q01	683	784	10	29	1.464	3.699
24	CM603Q01	685	791	8	15	1.168	1.896
25	CM800Q01	712	767	3	14	0.421	1.825
26	CM803Q01	663	784	7	6	1.056	0.765
27	CM828Q03	719	778	7	17	0.974	2.185
28	CM905Q01	694	793	4	2	0.576	0.252
29	CM906Q01	692	783	12	8	1.734	1.022
30	CM909Q01	736	832	2	3	0.272	0.361
31	CM909Q02	736	832	15	20	2.038	2.404
32	CM909Q03	734	829	7	12	0.954	1.448
33	CM915Q01	730	824	15	12	2.055	1.456
34	CM915Q02	729	822	6	19	0.823	2.311
35	CM919Q01	693	792	**0**	2	0.000	0.253
36	CM919Q02	691	793	5	8	0.724	1.009
37	CM943Q01	685	788	3	5	0.438	0.635
38	CM943Q02	680	782	1	16	0.147	2.046
39	CM949Q01	733	829	4	5	0.546	0.603

Table 2 (continued)

No.	Item code	Total number of responses		Number of disengaged responses		Percentage of disengaged responses (%)	
		2015	**2018**	**2015**	**2018**	**2015**	**2018**
40	CM949Q02	732	831	3	9	0.410	1.083
41	CM953Q03	677	779	11	22	1.625	2.824
42	CM954Q01	688	790	2	9	0.291	1.139
43	CM954Q04	686	788	9	22	1.312	2.792
44	CM955Q03	687	782	8	7	1.164	0.895
45	CM982Q01	702	792	1	2	0.142	0.253
46	CM982Q02	702	792	2	9	0.285	1.136
47	CM982Q03	702	792	9	14	1.282	1.768
48	CM982Q04	702	791	1	10	0.142	1.264
49	CM992Q01	701	790	2	6	0.285	0.759
50	CM992Q02	699	789	20	41	2.861	5.196
51	CM998Q04	681	775	6	13	0.881	1.677

Table 3 Numbers of items by disengagement levels. *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

Percentage of disengagement (%)	Number of items in 2015	Number of items in 2018
0%	2	0
0.01–1%	34	22
1.01–2%	11	17
2.01–3%	4	9
3.01–4%	0	2
4.01–5%	0	0
5.01–6%	0	1

Table 4 Comparison of the scores of disengaged and engaged responses in 2015 and 2018. *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

No	Item code	Average scores of engaged responses		Average scores of disengaged responses	
		2015	**2018**	**2015**	**2018**
1	CM00GQ01	0.054	0.038	0	0
2	CM033Q01	0.785	0.794	0	0
3	CM034Q01	0.322	0.283	0	0
4	CM155Q01	0.653	0.663	0	0
5	CM155Q04	0.490	0.523	0	0
6	CM273Q01	0.344	0.351	0	0
7	CM305Q01	0.400	0.406	0	0.385
8	CM408Q01	0.292	0.266	0	0
9	CM411Q01	0.435	0.439	0	0
10	CM411Q02	0.450	0.432	0	0
11	CM420Q01	0.512	0.548	0	0

Table 4 (continued)

No	Item code	Average scores of engaged responses		Average scores of disengaged responses	
		2015	2018	2015	2018
12	CM423Q01	0.712	0.707	N/A[a]	0
13	CM442Q02	0.244	0.270	0	0
14	CM446Q01	0.673	0.694	0	0
15	CM447Q01	0.553	0.587	0	0
16	CM464Q01	0.115	0.137	0	0
17	CM474Q01	0.593	0.587	0	0
18	CM496Q01	0.415	0.459	0	0
19	CM496Q02	0.585	0.622	0	0
20	CM559Q01	0.496	0.502	0	0.058
21	CM564Q01	0.423	0.410	0.333	0.333
22	CM564Q02	0.387	0.413	0	0.25
23	CM571Q01	0.388	0.412	0	0
24	CM603Q01	0.317	0.352	0	0.181
25	CM800Q01	0.795	0.848	0	0.50
26	CM803Q01	0.216	0.252	0	0
27	CM828Q03	0.248	0.244	0	0
28	CM905Q01	0.740	0.746	0.50	0.50
29	CM906Q01	0.472	0.516	0.182	0.20
30	CM909Q01	0.886	0.883	0	0
31	CM909Q02	0.546	0.522	0.143	0.053
32	CM909Q03	0.262	0.255	0	0
33	CM915Q01	0.433	0.430	0.40	0
34	CM915Q02	0.670	0.674	0	0
35	CM919Q01	0.759	0.723	N/A	0
36	CM919Q02	0.424	0.383	0	0
37	CM943Q01	0.494	0.488	0.333	0.25
38	CM943Q02	0.009	0.019	0	0
39	CM949Q01	0.577	0.607	0	0
40	CM949Q02	0.274	0.278	0	0.125
41	CM953Q03	0.445	0.453	0	0
42	CM954Q01	0.715	0.660	0	0
43	CM954Q04	0.253	0.220	0	0
44	CM955Q03	0.113	0.127	0	0
45	CM982Q01	0.860	0.872	0	0
46	CM982Q02	0.365	0.371	0	0
47	CM982Q03	0.604	0.614	0.111	0.167
48	CM982Q04	0.346	0.383	0	0
49	CM992Q01	0.723	0.774	0	0
50	CM992Q02	0.100	0.107	0	0
51	CM998Q04	0.307	0.319	0.169	0.25

[a] N/A indicates no disengagement for that item.

Table 5 Comparison of $P+$in 2015 and 2018. *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

No	Item code	$P+$					
		2015			**2018**		
		Before filtering	**Applying response–level filtering**	**Applying examinee-level filtering**	**Before filtering**	**Applying response-level filtering**	**Applying examinee-level filtering**
1	CM00GQ01	0.054	0.054	0.058	0.043	0.043	0.044
2	CM033Q01	0.776	0.776	0.785	0.793	0.793	0.801
3	CM034Q01	0.326	0.327	0.34	0.3	0.302	0.311
4	CM155Q01	0.654	0.666	0.672	0.662	0.678	0.701
5	CM155Q04	0.49	0.496	0.505	0.534	0.538	0.564
6	CM273Q01	0.344	0.344	0.357	0.365	0.366	0.383
7	CM305Q01	0.393	0.393	0.397	0.406	0.406	0.403
8	CM408Q01	0.289	0.291	0.305	0.274	0.275	0.294
9	CM411Q01	0.433	0.435	0.45	0.452	0.455	0.48
10	CM411Q02	0.457	0.458	0.478	0.437	0.445	0.461
11	CM420Q01	0.509	0.511	0.524	0.558	0.562	0.588
12	CM423Q01	0.712	0.712	0.725	0.721	0.721	0.739
13	CM442Q02	0.242	0.243	0.25	0.281	0.285	0.303
14	CM446Q01	0.665	0.667	0.684	0.703	0.708	0.739
15	CM447Q01	0.55	0.551	0.561	0.595	0.599	0.615
16	CM464Q01	0.116	0.116	0.118	0.155	0.156	0.169
17	CM474Q01	0.584	0.584	0.589	0.594	0.595	0.612
18	CM496Q01	0.418	0.421	0.427	0.471	0.475	0.496
19	CM496Q02	0.586	0.588	0.594	0.631	0.633	0.661
20	CM559Q01	0.501	0.508	0.512	0.509	0.518	0.533
21	CM564Q01	0.42	0.421	0.424	0.426	0.426	0.435
22	CM564Q02	0.388	0.388	0.389	0.433	0.438	0.443
23	CM571Q01	0.382	0.385	0.395	0.431	0.443	0.453
24	CM603Q01	0.308	0.31	0.312	0.369	0.371	0.383
25	CM800Q01	0.791	0.791	0.799	0.848	0.852	0.866
26	CM803Q01	0.216	0.217	0.224	0.267	0.269	0.285
27	CM828Q03	0.243	0.245	0.255	0.248	0.252	0.27
28	CM905Q01	0.738	0.739	0.741	0.759	0.758	0.778
29	CM906Q01	0.466	0.471	0.477	0.535	0.538	0.553
30	CM909Q01	0.88	0.882	0.903	0.891	0.893	0.92
31	CM909Q02	0.551	0.557	0.568	0.533	0.543	0.563
32	CM909Q03	0.255	0.258	0.268	0.273	0.277	0.292
33	CM915Q01	0.441	0.441	0.461	0.428	0.434	0.445
34	CM915Q02	0.662	0.665	0.694	0.68	0.695	0.736
35	CM919Q01	0.763	0.763	0.783	0.729	0.73	0.752
36	CM919Q02	0.433	0.436	0.443	0.394	0.398	0.414
37	CM943Q01	0.493	0.494	0.497	0.491	0.492	0.51
38	CM943Q02	0.009	0.009	0.01	0.022	0.023	0.024
39	CM949Q01	0.58	0.585	0.606	0.631	0.634	0.663
40	CM949Q02	0.269	0.27	0.281	0.295	0.297	0.304
41	CM953Q03	0.448	0.451	0.463	0.474	0.48	0.505
42	CM954Q01	0.722	0.724	0.744	0.681	0.688	0.717
43	CM954Q04	0.261	0.265	0.276	0.231	0.236	0.243

Table 5 (continued)

No	Item code	*P*+					
		2015			2018		
		Before filtering	Applying response–level filtering	Applying examinee-level filtering	Before filtering	Applying response-level filtering	Applying examinee-level filtering
44	CM955Q03	0.111	0.112	0.117	0.135	0.135	0.147
45	CM982Q01	0.854	0.855	0.882	0.875	0.876	0.901
46	CM982Q02	0.361	0.363	0.388	0.374	0.378	0.397
47	CM982Q03	0.586	0.593	0.618	0.621	0.631	0.661
48	CM982Q04	0.343	0.344	0.355	0.394	0.399	0.425
49	CM992Q01	0.716	0.718	0.737	0.769	0.774	0.796
50	CM992Q02	0.102	0.105	0.111	0.11	0.116	0.123
51	CM998Q04	0.31	0.305	0.293	0.324	0.319	0.318

Table 6 Comparison of *b* in 2015 and 2018. *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

No	Item code	Model	2015						2018					
			Before filtering		Applying response-level filtering		Applying examinee-level filtering		Before filtering		Applying response-level filtering		Applying examinee-level filtering	
			b	SE[a]	*b*	SE	*b*	SE	*b*	SE	*b*	SE	*b*	SE
1	CM00GQ01	2PLM	2.976	0.048	2.977	0.072	2.959	0.042	2.788	0.026	2.779	v0.04	2.524	0.025
2	CM033Q01	2PLM	−1.808	0.005	−1.793	0.045	−1.946	0.028	−1.94	0.021	−1.947	0.056	−2.032	0.009
3	CM034Q01	2PLM	0.721	0.016	0.725	0.021	0.679	0.006	0.792	0.053	0.791	0.04	0.744	0.016
4	CM155Q01	2PLM	−0.634	0.034	−0.681	0.04	−0.762	0.016	−0.676	0.008	−0.734	0.011	−0.882	0.034
5	CM155Q04	2PLM	0.051	0.026	0.032	0.057	−0.048	0.006	−0.209	0.005	−0.223	0.036	−0.378	0.023
6	CM273Q01	2PLM	0.838	0.08	0.839	0.015	0.802	0.007	0.68	0.025	0.68	0.003	0.592	0.033
7	CM305Q01	2PLM	1.494	0.027	1.501	0.014	1.481	0.018	1.666	0.033	1.681	0.041	1.536	0.023
8	CM408Q01	2PLM	0.843	0.011	0.846	0.038	0.794	0.035	0.941	0.019	0.945	0.027	0.898	0.06
9	CM411Q01	2PLM	0.237	0.049	0.238	0.053	0.171	0.018	0.142	0.007	0.139	0.007	0.044	0.014
10	CM411Q02	2PLM	0.209	0.025	0.209	0.023	0.102	0.013	0.268	0.016	0.247	0.032	0.163	0.052
11	CM420Q01	2PLM	−0.032	0.027	−0.037	0.006	−0.091	0.015	−0.231	0.036	−0.242	0.042	−0.363	0.011
12	CM423Q01	2PLM	−1.215	0.074	−1.224	0.029	−1.382	0.022	−1.416	0.01	−1.443	0.067	−1.814	0.034
13	CM442Q02	2PLM	0.893	0.065	0.892	0.024	0.858	0.058	0.722	0.017	0.721	0.024	0.655	0.061
14	CM446Q01	2PLM	−0.527	0.012	−0.533	0.013	−0.602	0.049	−0.768	0.025	−0.793	0.044	−0.98	0.008
15	CM447Q01	2PLM	−0.187	0.011	−0.188	0.028	−0.221	0.007	−0.368	0.031	−0.376	0.015	−0.452	0.035
16	CM464Q01	2PLM	1.517	0.021	1.512	0.025	1.42	0.04	1.246	0.016	1.247	0.03	1.195	0.017
17	CM474Q01	2PLM	−0.441	0.019	−0.436	0.068	−0.482	0.024	−0.542	0.057	−0.551	0.022	−0.677	0.014
18	CM496Q01	2PLM	0.348	0.053	0.345	0.062	0.305	0.04	0.13	0.005	0.124	0.007	0.036	0.036
19	CM496Q02	2PLM	−0.375	0.006	−0.384	0.07	−0.43	0.026	−0.55	0.016	−0.562	0.032	−0.745	0.013
20	CM559Q01	2PLM	0	0.024	−0.027	0.012	−0.067	0.03	−0.052	0.006	−0.085	0.029	−0.202	0.009
21	CM564Q01	2PLM	0.483	0.056	0.484	0.005	0.46	0.047	0.514	0.027	0.506	0.004	0.45	0.05

Table 6 (continued)

No	Item code	Model	2015						2018					
			Before filtering		Applying response-level filtering		Applying examinee-level filtering		Before filtering		Applying response-level filtering		Applying examinee-level filtering	
			b	SE[a]	*b*	SE	*b*	SE	*b*	SE	*b*	SE	*b*	SE
22	CM564Q02	2PLM	0.969	0.039	0.965	0.008	0.995	0.041	0.381	0.024	0.379	0.023	0.318	0.042
23	CM571Q01	2PLM	0.49	0.045	0.482	0.025	0.432	0.046	0.319	0.012	0.296	0.022	0.223	0.004
24	CM603Q01	2PLM	0.99	0.055	0.983	0.025	0.972	0.036	0.643	0.023	0.641	0.06	0.569	0.027
25	CM800Q01	2PLM	−3.486	0.032	−3.602	0.016	−3.854	0.03	−3.849	0.039	−4.47	0.024	−5.739	0.005
26	CM803Q01	2PLM	0.955	0.027	0.954	0.023	0.919	0.041	0.776	0.023	0.777	0.026	0.709	0.029
27	CM828Q03	2PLM	1.055	0.066	1.054	0.055	1.008	0.066	1.233	0.006	1.241	0.076	1.206	0.052
28	CM905Q01	2PLM	−1.415	0.042	−1.387	0.07	−1.388	0.044	−1.582	0.038	−1.562	0.042	−1.721	0.016
29	CM906Q01	2PLM	0.207	0.026	0.204	0.058	0.164	0.035	−0.115	0.039	−0.118	0.055	−0.181	0.005
30	CM909Q01	2PLM	−1.709	0.049	−1.728	0.038	−1.97	0.054	−1.594	0.015	−1.63	0.055	−2.041	0.02
31	CM909Q02	2PLM	−0.167	0.049	−0.165	0.033	−0.235	0.03	−0.128	0.035	−0.138	0.02	−0.252	0.019
32	CM909Q03	2PLM	0.759	0.045	0.757	0.034	0.725	0.044	0.751	0.039	0.745	0.033	0.693	0.072
33	CM915Q01	2PLM	0.354	0.006	0.365	0.066	0.269	0.036	0.399	0.048	0.375	0.011	0.3	0.009
34	CM915Q02	2PLM	−0.701	0.031	−0.718	0.06	−0.932	0.025	−0.702	0.018	−0.769	0.024	−1.11	0.008
35	CM919Q01	2PLM	−1.062	0.012	−1.064	0.008	−1.208	0.068	−0.839	0.009	−0.841	0.017	−0.947	0.027
36	CM919Q02	2PLM	0.353	0.034	0.348	0.016	0.301	0.011	0.472	0.012	0.469	0.028	0.417	0.046
37	CM943Q01	2PLM	0.086	0.029	0.083	0.046	0.054	0.034	0.07	0.038	0.065	0.009	−0.052	0.031
38	CM943Q02	2PLM	2.893	0.086	2.888	0.086	2.836	0.124	2.106	0.386	2.106	0.383	2.079	0.344
39	CM949Q01	2PLM	−0.299	0.009	−0.315	0.024	−0.415	0.004	−0.491	0.026	−0.494	0.023	−0.66	0.004
40	CM949Q02	2PLM	0.935	0.009	0.932	0.008	0.866	0.011	0.825	0.031	0.812	0.012	0.766	0.007
41	CM953Q03	2PLM	0.218	0.012	0.217	0.012	0.164	0.021	0.099	0.004	0.094	0.03	−0.005	0.05
42	CM954Q01	2PLM	−0.718	0.026	−0.726	0.062	−0.823	0.024	−0.575	0.008	−0.591	0.023	−0.733	0.047
43	CM954Q04	2PLM	0.866	0.045	0.864	0.024	0.819	0.042	0.901	0.034	0.9	0.015	0.863	0.013

Table 6 (continued)

No	Item code	Model	2015						2018					
			Before filtering		Applying response-level filtering		Applying examinee-level filtering		Before filtering		Applying response-level filtering		Applying examinee-level filtering	
			b	SE[a]	*b*	SE	*b*	SE	*b*	SE	*b*	SE	*b*	SE
44	CM955Q03	GPCM	2.844	0.176	2.851	0.201	2.894	0.175	3.903	0.124	3.947	0.136	3.981	0.112
45	CM982Q01	2PLM	−1.78	0.03	−1.798	0.019	−2.435	0.064	−2.244	0.041	−2.465	0.011	−3.142	0.041
46	CM982Q02	2PLM	0.939	0.058	0.946	0.004	0.891	0.03	0.759	0.02	0.749	0.057	0.697	0.023
47	CM982Q03	2PLM	−0.385	0.024	−0.408	0.019	−0.591	0.02	−0.466	0.069	−0.5	0.013	−0.677	0.011
48	CM982Q04	2PLM	0.687	0.039	0.688	0.052	0.621	0.008	0.414	0.053	0.417	0.005	0.314	0.035
49	CM992Q01	2PLM	−0.931	0.007	−0.936	0.024	−1.044	0.044	−1.23	0.054	−1.272	0.028	−1.419	0.019
50	CM992Q02	2PLM	2.001	0.052	1.99	0.026	2.026	0.045	1.675	0.024	1.65	0.044	1.614	0.018
51	CM998Q04	2PLM	4.198	0.026	3.615	0.009	2.603	0.011	2.132	0.046	2.117	0.041	1.793	0.022

[a] SE indicates the standard error.

Table 7 Comparison of *a* in 2015 and 2018. *Source*: Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA), 2015 and 2018 Mathematics Assessment

No	Item code	Model	2015						2018					
			Before filtering		Applying response-level filtering		Applying examinee-level filtering		Before filtering		Applying response-level filtering		Applying examinee-level filtering	
			a	SE[a]	*a*	SE	*a*	SE	*a*	SE	*a*	SE	*a*	SE
1	CM00GQ01	2PLM	1.166	0.048	1.163	0.011	1.14	0.031	1.423	0.025	1.433	0.016	1.685	0.042
2	CM033Q01	2PLM	0.776	0.049	0.783	0.044	0.749	0.03	0.797	0.01	0.796	0.031	0.788	0.033
3	CM034Q01	2PLM	1.319	0.086	1.31	0.021	1.233	0.01	1.414	0.02	1.41	0.056	1.388	0.03
4	CM155Q01	2PLM	1.324	0.008	1.282	0.049	1.234	0.025	1.474	0.054	1.402	0.052	1.368	0.013
5	CM155Q04	2PLM	0.808	0.026	0.779	0.027	0.718	0.009	0.96	0.032	0.932	0.013	0.876	0.005
6	CM273Q01	2PLM	0.907	0.047	0.902	0.019	0.85	0.044	0.964	0.011	0.955	0.045	0.932	0.017
7	CM305Q01	2PLM	0.3	0.024	0.298	0.028	0.29	0.037	0.237	0.023	0.237	0.028	0.265	0.015
8	CM408Q01	2PLM	1.511	0.01	1.491	0.013	1.426	0.02	1.397	0.032	1.376	0.01	1.258	0.037
9	CM411Q01	2PLM	1.73	0.066	1.717	0.008	1.64	0.016	1.738	0.027	1.722	0.023	1.588	0.043
10	CM411Q02	2PLM	0.989	0.042	0.969	0.011	0.944	0.012	1.015	0.021	0.985	0.006	0.958	0.015
11	CM420Q01	2PLM	1.477	0.019	1.462	0.045	1.415	0.038	1.432	0.006	1.408	0.052	1.351	0.022
12	CM423Q01	2PLM	0.847	0.014	0.839	0.039	0.795	0.024	0.728	0.03	0.712	0.012	0.609	0.037
13	CM442Q02	2PLM	2.14	0.024	2.14	0.05	2.115	0.025	2.18	0.034	2.15	0.079	2.025	0.023
14	CM446Q01	2PLM	2.33	0.012	2.332	0.048	2.275	0.041	1.623	0.017	1.578	0.045	1.461	0.025
15	CM447Q01	2PLM	1.515	0.024	1.544	0.049	1.621	0.011	1.459	0.021	1.469	0.023	1.488	0.007
16	CM464Q01	2PLM	2.261	0.062	2.28	0.061	2.667	0.042	2.454	0.036	2.447	0.04	2.333	0.042
17	CM474Q01	2PLM	0.907	0.008	0.917	0.019	0.902	0.028	0.877	0.034	0.866	0.037	0.812	0.042
18	CM496Q01	2PLM	1.29	0.034	1.281	0.013	1.238	0.025	1.582	0.02	1.573	0.066	1.47	0.024
19	CM496Q02	2PLM	1.138	0.05	1.127	0.034	1.105	0.035	1.197	0.006	1.184	0.04	1.066	0.013
20	CM559Q01	2PLM	0.799	0.032	0.773	0.02	0.783	0.028	0.854	0.007	0.825	0.011	0.742	0.026
21	CM564Q01	2PLM	0.772	0.032	0.769	0.027	0.754	0.026	0.679	0.007	0.691	0.009	0.669	0.006

Table 7 (continued)

No	Item code	Model	2015						2018					
			Before filtering		Applying response-level filtering		Applying examinee-level filtering		Before filtering		Applying response-level filtering		Applying examinee-level filtering	
			a	SE[a]	*a*	SE	*a*	SE	*a*	SE	*a*	SE	*a*	SE
22	CM564Q02	2PLM	0.501	0.01	0.501	0.065	0.475	0.013	0.891	0.014	0.895	0.026	0.905	0.006
23	CM571Q01	2PLM	1.341	0.074	1.328	0.041	1.302	0.026	1.279	0.009	1.237	0.011	1.251	0.047
24	CM603Q01	2PLM	0.98	0.061	0.978	0.019	0.972	0.065	1.069	0.008	1.082	0.024	1.066	0.012
25	CM800Q01	2PLM	0.394	0.011	0.382	0.009	0.369	0.03	0.458	0.019	0.397	0.025	0.255	0.017
26	CM803Q01	2PLM	2.463	0.021	2.464	0.023	2.419	0.013	2.163	0.015	2.148	0.048	2.154	0.045
27	CM828Q03	2PLM	1.519	0.01	1.506	0.039	1.478	0.031	1.112	0.018	1.085	0.029	0.983	0.038
28	CM905Q01	2PLM	0.819	0.035	0.84	0.021	0.853	0.009	0.817	0.025	0.828	0.011	0.813	0.02
29	CM906Q01	2PLM	1.212	0.067	1.215	0.022	1.341	0.024	1.207	0.069	1.219	0.012	1.273	0.044
30	CM909Q01	2PLM	1.615	0.025	1.612	0.019	1.509	0.035	2.115	0.027	2.028	0.064	1.647	0.044
31	CM909Q02	2PLM	1.637	0.05	1.651	0.053	1.555	0.011	1.437	0.031	1.424	0.007	1.341	0.042
32	CM909Q03	2PLM	3.477	0.032	3.462	0.033	3.34	0.041	2.392	0.024	2.498	0.014	2.31	0.052
33	CM915Q01	2PLM	0.777	0.006	0.808	0.028	0.723	0.009	0.907	0.057	0.899	0.021	1.005	0.034
34	CM915Q02	2PLM	1.208	0.019	1.187	0.016	1.04	0.04	1.423	0.006	1.366	0.008	1.1	0.021
35	CM919Q01	2PLM	1.495	0.054	1.487	0.014	1.398	0.034	1.769	0.008	1.772	0.014	1.734	0.024
36	CM919Q02	2PLM	1.039	0.016	1.027	0.039	1.034	0.007	1.176	0.018	1.161	0.02	1.061	0.008
37	CM943Q01	2PLM	0.526	0.044	0.531	0.057	0.571	0.007	0.684	0.013	0.681	0.089	0.634	0.031
38	CM943Q02	2PLM	2.676	0.065	2.692	0.059	2.763	0.101	7.054	0.191	7.071	0.183	6.884	0.201
39	CM949Q01	2PLM	1.389	0.059	1.373	0.069	1.306	0.024	1.563	0.037	1.547	0.027	1.383	0.019
40	CM949Q02	2PLM	1.544	0.014	1.541	0.072	1.616	0.079	1.461	0.013	1.519	0.041	1.552	0.017
41	CM953Q03	2PLM	2.104	0.056	2.09	0.018	2.021	0.02	1.999	0.027	1.964	0.054	1.782	0.049
42	CM954Q01	2PLM	2.282	0.023	2.269	0.032	2.139	0.03	2.36	0.068	2.311	0.017	2.068	0.021
43	CM954Q04	2PLM	2.118	0.023	2.101	0.044	2.041	0.017	2.551	0.034	2.541	0.027	2.554	0.023

Table 7 (continued)

No	Item code	Model	2015						2018					
			Before filtering		Applying response-level filtering		Applying examinee-level filtering		Before filtering		Applying response-level filtering		Applying examinee-level filtering	
			a	SE[a]	*a*	SE	*a*	SE	*a*	SE	*a*	SE	*a*	SE
44	CM955Q03	GPCM	1.62	0.018	1.612	0.039	1.54	0.044	1.102	0.046	1.087	0.026	1.044	0.022
45	CM982Q01	2PLM	1.231	0.038	1.218	0.01	0.928	0.102	0.994	0.016	0.889	0.062	0.758	0.013
46	CM982Q02	2PLM	0.707	0.016	0.694	0.02	0.599	0.039	0.811	0.059	0.796	0.006	0.703	0.007
47	CM982Q03	2PLM	1.001	0.031	0.976	0.05	0.856	0.008	1.318	0.029	1.302	0.029	1.189	0.02
48	CM982Q04	2PLM	1.383	0.012	1.373	0.009	1.524	0.058	1.681	0.007	1.62	0.056	1.594	0.026
49	CM992Q01	2PLM	1.246	0.006	1.253	0.064	1.18	0.056	1.197	0.036	1.183	0.013	1.153	0.036
50	CM992Q02	2PLM	1.486	0.017	1.473	0.016	1.375	0.012	1.947	0.044	1.938	0.034	1.873	0.042
51	CM998Q04	2PLM	0.196	0.006	0.235	0.018	0.358	0.006	0.36	0.055	0.376	0.021	0.451	0.017

[a] SE indicates the standard error.

Abbreviations

2PL	Two-parameter logistic
3PL	Three-parameter logistic
CBA	Computer-based assessment
CTT	Classical Test Theory
GPCM	Generalized Partial Credit Model
IRT	Item Response Theory
MCSS	Multiple-choice-single-select
NAEP	National Assessment of Educational Progress
OECD	Organisation for Economic Co-operation and Development
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
SE	Standard error
U.S.	United States

Acknowledgements

The authors thank Markus Broer, Young Yee Kim, and Xiaying "James" Zheng for their review of initial findings and methods. The authors also thank Amy Rathbun for their support in publishing this manuscript.

Author contributions

Fusun Sahin conceived the idea and analytical framework and conducted the literature review. Huan Kuang conducted the analyses and reported results. Both authors worked in writing the manuscript.

Funding

The concept and preliminary results of this paper were developed and conducted during the 2020 NAEP doctoral internship program administered by AIR and funded by NCES under Contract No. ED-IES-12-D-0002/0004. The authors did not receive funding for finalizing the results and preparing this manuscript. The views, thoughts, and opinions expressed in the paper belong solely to the authors and do not reflect NCES position or endorsement.

Availability of data and materials

All results generated for this study are included in this published article. The datasets analyzed during the current study are available publicly in https://www.oecd.org/pisa/data/.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 5 April 2022 Accepted: 12 January 2023

Published online: 08 February 2023

References

Bovaird, J., & Embretson, E. (2006, August). *Using response time to increase the construct validity of trait estimates.* Paper presented at the 114th Annual Meeting of the American Psychological Association, New Orleans, LA.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Debeer, D., Bucholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics, 39*, 502–523.

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education, 5*(18), 1–25.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173–183.

Hauser, C., & Kingsbury, G. G. (2009, April). Individual *score validity in a modest-stakes adaptive educational testing setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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.

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), 8–41.

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, LA.

Organisation for Economic Co-operation and Development (OECD). (2000). *Knowledge and skills for life: First results from PISA 2000*. OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2017). *PISA 2015 technical report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf

Organisation for Economic Co-operation and Development (OECD). (2022). Chapter 9: Scaling PISA data. In *PISA 2018 technical report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/Ch.09-Scaling-PISA-Data.pdf

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/index.html

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education, 9*(18), 1–25.

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential non-effortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education, 33*(4), 263–279.

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*(1), 74–104.

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education, 8*(5), 1–24.

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.

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

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.

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.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61.

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.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343–354.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*(1), 86–105.

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.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*(1), 65–83.

Wise, S. L., & Ma, L. (2012, April). *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.

Yamamoto, K. (1995). *Estimating the effect of test length and text time on parameter estimation using the HYBRID model (ETS Research Report RR-95-02)*. Educational Testing Service.

Publisher's Note