Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 15, December 2023 ISSN 1531-7714

Measuring Test-Taking Effort on Constructed-Response Items with Item Response Time and Number of Actions

Militsa G. Ivanova, *University of Cyprus*
Michalis P. Michaelides, *University of Cyprus*

Research on methods for measuring examinee engagement with constructed-response items is limited. The present study used data from the PISA 2018 Reading domain to construct and compare indicators of test-taking effort on constructed-response items: response time, number of actions, the union (combining effortless responses detected by either response time or number of actions measures or both), and the intersection of response time and number of actions (responses identified as effortless by both response-time and number of actions measures). A 10% normative threshold identification method was used for both response time and number of actions. Pre-defined validation criteria were used to explore the validity of each of the four indicators. Number of actions yielded a similar number of effortless responses as the union measure. Response time and intersection measures also had similar results and were related to lower disengagement than the number of actions and union indicators. With the normative threshold identification method, number of actions and the union of the two process data may result in a higher level of response misclassifications on some constructed-response items than response time and the intersection measures. Response time appears to be a more valid indicator of test-taking effort on constructed-response items than number of actions.

Keywords: test-taking effort measures, response time, number of actions, constructed-response items, PISA test-taking effort measures, response time, number of actions, constructed-response items, PISA

Introduction

International assessment programs, such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA), use tests to conduct international comparative studies and evaluate educational systems by assessing student academic performance and skills. Studies using data from such programs may serve as scientific evidence for essential policy decisions (Schleicher, 2019), therefore it is crucial to have valid estimates of student proficiency levels in the examined content area. Yet, test-takers may not be willing to invest enough effort when participating in low-stakes international large-scale assessments, which usually pose no consequences to examinees for low performance.

Previous studies have revealed that the effort examinees are willing to invest in a low-stakes assessment situation affects their performance (Eklöf, 2010b; Eklöf et al., 2014) and the lack of effort adds construct-irrelevant variance to test scores. A metanalysis by Wise and DeMars (2005) has shown that examinee effort is positively related to test scores. Depending on the effort measure used, moderate to strong positive correlations between effort and performance were observed in a recent metanalytic study as well (Silm et al., 2020).

Examinee disengagement with test items negatively impacted the psychometric properties of low-stakes assessments: it reduced the test construct validity (Wise, 2009), inflated internal reliability (Sundre &Wise, 2003; Wise & DeMars, 2009), and led to differential item functioning (DeMars & Wise, 2010). Consequently, monitoring test-taking effort in low-stakes achievement tests and controlling for its possible effects on test results would be beneficial for the assessment process.

The Measurement of Test-Taking Effort

Verbalization methods, such as self-report scales and interviews, have been used as indicators of test-taking effort for years. However, they are vulnerable to response bias, test-takers difficulty understanding the questions (Johnson, 2005), or examinee reluctance to invest effort on a self-report scale (Eklöf, 2010a). Besides, verbalizations represent global measures of overall effort invested in a test (Wise, 2015), therefore they do not allow tracking examinee engagement with different types of items.

Behavioral measures are less susceptible to such drawbacks since they capture overt test-taking behavior. The digitalization of low-stakes assessment programs allows for the collection of a large amount of log-file data (i.e., all data recorded in a digital-based assessment) which may reflect examinee process behavior while working with an item (Provasnik, 2021). Such process data have been used in previous literature to detect disengagement in learning (Gobert et al., 2015) and game persistence (Ventura & Shute, 2013). Data on response time and number of actions examinees undertake while interacting with a test item can also be used in the assessment context to describe examinees' effort.

Examinees' item responses can be divided into rapid-guesses and solution behavior based on the amount of time taken by an examinee to answer an item. Responses occurring within a very short time interval, insufficient to meaningfully engage with an item, are supposed to reflect a lack of engagement with the item solution and are classified as rapid guesses. The rest of the responses are supposed to indicate engagement with the test items, and they are recorded as solution behavior[1]. A threshold identification is essential to distinguish the two types of behavior (Wise & Kong, 2005).

In the literature on item response times, a number of threshold identification techniques have been proposed. A common threshold can be applied to all items (Wise et al., 2010) or a threshold can be set based on item length (Wise & Kong, 2005). Information about examinee item response time-frequency distribution (Setzer et al., 2013), a percentage of the average examinee response time on a particular item (normative threshold; Wise & Ma, 2012), or combinations of response time with item accuracy (cumulative proportion threshold; Guo et al., 2016) and item accuracy with test performance (change in information threshold; Wise, 2019) have also been used to identify time thresholds. Threshold comparison studies are inconclusive about a preferred threshold identification method (Kong et al., 2007; Wise, 2019). Each method leads to a certain amount of misclassification, which should be acknowledged when selecting a threshold identification method (Wise, 2017).

Response time is not prone to response bias as self-report scales (Wise & Kingsbury, 2016). Additionally, it allows taking into consideration possible changes in effort across the items of a test, which is important since examinees have been shown to invest less effort on items appearing toward the end of the test (Setzer et al., 2013; Wise et al., 2009) or on different item types (Michaelides & Ivanova, 2022). Response time has been shown to be a reliable and valid indicator of test-taking effort, however, it has only been validated for selected-response items (Wise & Gao, 2017; Wise & Kong, 2005); therefore, its applicability to constructed-response (CR) items, e.g., items that necessitate students to generate their own response (OECD, 2019a), remains unclear.

While the use of response time is a widely recognized indicator of effort, research examining the relationship between number of actions performed on an item and test-taking effort is limited. The frequency of active interactions with an item was supposed to reflect examinees' goal-directed behavior on that item (Greiff et al., 2016). A quadratic relationship was

[1] Note that, even though the terms test-taking effort, engagement and (non-)rapid guessing/solution behaviors can be defined conceptually in slightly different ways, they were used in the current study as synonyms.

observed between the number of actions and task success, indicating that a moderate frequency of interactions with an item was associated with higher examinee engagement than low item interaction frequency (Goldhammer, Naumann, et al., 2017). Task demands moderated this quadratic relationship; the association between the number of actions and task success was strong for items that required a long navigation path and weak for less demanding items. Using items from one science booklet from PISA 2015, Yavuz (2019) found a moderate relationship between the frequency of interactions and item accuracy level. When only CR items from a cluster of science items were considered, the mean number of actions in PISA 2015 was moderately associated with self-reported effort and strongly related to students' performance on the examined items (Ivanova et al., 2020).

Measuring test-taking effort on constructed-response items. CR items have been associated with lower levels of motivation and effort than multiple-choice (MC) items (DeMars, 2000; Eklöf & Knekta, 2017; Michaelides & Ivanova, 2022), but research on using behavioral indicators and process data to measure effort on CR items is scarce. Liu and Hau (2020) suggested that missing responses, such as omitted items or invalid responses, indicate a lack of test-taking effort. Although, missingness may address the disadvantages of the response time approach, such as identifying students who spent a substantial amount of time on an item but did not engage in its solution, it could also reflect low ability (Liu & Hau, 2020).

The number of characters written by test-takers as a response to an item has been used to indicate examinee effort on CR items. Examinees may give very brief, often nonsensical, answers to CR items in a very short time. Rapid perfunctory responses contain a small number of characters provided in a time period, which is insufficient to meaningfully engage with an item (Wise & Gao, 2017). However, with rapid perfunctory responses as an indicator of low effort, some effortful behaviors, such as deleting and correcting responses, are ignored and examinees engaging in such behaviors may be misclassified as disengaged. Additionally, in international large-scale assessment programs, where item pools are reused and details about specific test-taker responses on items are often not publicly available, the detection of rapid perfunctory responses may be impossible.

A recent study by Ivanova et al. (2020) has shown that number of actions on CR items in PISA 2015 mediated the relationship between self-reported effort and performance on CR items, and it was positively associated with both constructs when controlling for cluster (i.e., a group of items) position in the assessment. The authors suggested that the frequency of active interactions with CR items may be used as an indicator of engagement. Sahin and Colvin (2020) used a small number of items (3 MC and 4 CR) to show that fewer disengaged responses were identified when using a combination of response time and number of actions than when using response time alone as an indicator of effort.

When proposing a new measure of test-taking effort, to demonstrate its usefulness as an indicator of the construct assessed, the suggested measure should be validated. Studies recommended the use of the correlation between the new measure and supplementary indicators of test engagement, such as self-reported effort scales (Wise & Gao, 2017; Wise & Kong, 2005) or achievement motivation scale scores (Liu & Hau, 2020) as evidence of convergent validity. Other research explored the informativeness of the effortless responses on MC items: it was hypothesized that effortless responses would be less informative about examinee's overall test performance than effortful ones, meaning that a low relationship was anticipated between disengaged responses' accuracy rate and test results (Wise, 2019). Regarding item accuracy, non-effortful responses were expected to be accurate at a rate close to the chance level for MC items and at a lower rate than the solution behaviors (Michaelides et al. 2020; Wise & Gao, 2017; Wise & Kong, 2005). However, estimating the chance level accuracy of CR items is challenging. It is probably reasonable to expect very low, close to zero percent accuracy of effortless responses on CR items, as shown by Wise and Gao (2017).

Purpose of Study

The current study aimed to categorize test-taking behavior on CR items in PISA as effortless or effortful using item response time only, number of actions on an item only, or combinations of the two (i.e., the union and the intersection of response time and number of actions). A conservative, item-specific threshold identification method (i.e., 10 % normative threshold- NT10) was used to distinguish effortful from effortless responses with each effort measure, to minimize the possibility of falsely identifying effortful behavior as effortless (i.e., reducing the Type I error). Such conservative thresholds can be practically useful in research where the point of interest is on individual results, rather than aggregated scores (Wise, 2019). Differences in proportions of effortless behavior across effort measures were presented and discussed. The various effort measures were compared in terms of four different validation criteria: a) comparison of accuracy levels across response behaviors (i.e., effortful and effortless); b) accuracy rate of effortless responses; c) effortless response informativeness about examinee's overall test performance; and d) relationship of the response behavior with PISA user-defined missingness. It was hypothesized that response time and number of actions separately would be weaker indicators of test-taking effort than an indicator based on a combination of both. Validity results were described separately for low-difficulty and high-difficulty testlets; no difference in the validity of effort indicators across testlet difficulty levels was expected.

Effort was measured separately for the first and second PISA test sessions (i.e., for the two halves of the assessment) since item position was previously found to have a significant effect on test-taking behavior reflecting effort (Wise et al., 2009). It was hypothesized that the proposed test-taking effort measure would be valid for both test sessions, however, the magnitude of effort was expected to decrease when the items were presented in the second session of the assessment.

PISA has measured student test-taking effort, using the "effort thermometer" self-report scale, since 2003 (Butler & Adams, 2007; Organization for Economic Co-operation and Development [OECD], 2010). The importance of estimating and evaluating the magnitude of examinees' effort in the assessment, however, has been acknowledged almost a decade later when the assessment program started paying special attention to reporting and discussing the construct (OECD, 2015; OECD, 2019b). Still, valid behavioral indicators of examinee effort on CR items have been significantly understudied, even though CR items have been associated with lower levels of engagement (DeMars, 2000). The current study aims to address this gap by contributing to the literature through the exploration and validation of item-level measures of test-taking effort on CR items. It also represents the first empirical exploration of a normative approach for the number of actions as an alternative or in combination with response time to measure test-taking effort on CR items.

Method

Sample

Publicly available data from the PISA 2018 test were used in the study. The PISA target population included 15-year-old students enrolled in educational institutions at seventh grade or higher. Within countries, students were selected via a two-stage stratified sampling design. In the first stage, at least 150 schools were chosen to participate in each country, while in the second stage about 42 students were selected from each school (OECD, 2019b).

For the purpose of the current study, data from the Spanish administration of PISA have been used due to its large sample size (n= 35943). Certain data from Spain, such as the Reading plausible values (PVs), timing, and response pattern data were initially masked when international PISA datasets were released because of some Spanish students' negative disposition towards the assessment (OECD, 2020). Further analysis of Spanish data revealed that data from a minority of students attending only a small number of schools seemed to be associated with low test-taking effort and the initially missing Spanish data were included in the publicly available datasets[2]. The higher-than-usual degree of students' disengagement in PISA 2018 is unlikely to affect the validity of the proposed effort measure examined in this paper. On the contrary, a larger number of rapid guessers may yield more stable estimates of their accuracy level and informativeness, which have been used as validation criteria for all examined effort measures.

[2] More information about the PISA Spanish administration and data can be seen in: Organization for Economic Co-operation and Development (2019). Annex A9. A note about Spain in PISA 2018: Further analysis of Spain's data by testing date (updated on 23 July 2020) in *PISA 2018 Results. What students know and can do [Volume I]*, PISA, OECD Publishing. 5f07c754-en.pdf (oecd-ilibrary.org)

Instruments

PISA test design. PISA is a large-scale international comparative study focusing on three main school subjects: reading, mathematics, and science, with reading being the major domain and the focus in the current study. The two-hour PISA assessment consisted of two testing sessions (OECD, 2019b). Each test session consisted of a combination of units of items, which included multiple-choice (i.e., simple or complex) and constructed-response (i.e., computer-scored or human-coded) items. The one-hour reading assessment was administered during the first or the second session (hour) of the test. The rest of the test consisted of two half-hour clusters of items assessing one or two minor domains (OECD, 2020).

The current study used data from the PISA 2018 reading test. Since reading was the major domain examined, it comprised a more extensive number of items, which had been administered to a larger sample of students participating in the survey. The reading literacy domain was assessed with a multistage adaptive testing design and consisted of three stages: Core, Stage I, and Stage II. Two different test designs were followed, determining the order of the stage administration. The core stage was always administered first, followed by Stage I and Stage II in test design A, or by Stage II and then Stage I in test design B (OECD, 2020). Students following design A were included in the present study due to the larger sample size in that condition, about 75% of the sample.

Items analyzed. In PISA 2018 reading, the students were assigned to one cluster of items, called a "testlet", at each of the three stages. Testlet assignment to students was done randomly in the Core Stage. In the next two stages students were assigned to a low or a high difficulty testlet based on the testlet assignment in the previous stage, the student's performance on computer-scored items in previous stages, and a probability layer matrix (OECD, 2020). A separate analysis will be conducted for low and high difficulty testlets. For more information about the testlet distribution in reading test Design A, see OECD (2020, Ch. 2, Figures and Tables, p.28).

Core stage testing material, consisting of 3 CR items (1 computer-scored- CR-CS and 2 human-coded- CR-HC), was included in a pilot analysis. Stage I items were used in the main analysis since Stage I[3] had the largest item pool of 46 CR. On two occasions, pairs of items were presented on the same screen and had common process data, but two different scores; their scores were combined, and each pair was handled as a single item. Thus, a total of 44 CR items were used in the Stage I analysis.

Process data. Two types of process data were collected separately for each item: response time and number of actions. Response time was defined as the total time a student spends on an item[4], whereas number of actions stands for all actions (i.e., clicks, double-clicks, key presses, and drag and drop events) a student performs while interacting with an item (OECD, 2017).

Missingness. Beyond system missing (i.e., items not presented to students and not counted in the user-defined missingness calculation) PISA records five types of user-defined missingness: a) "no response/omits" were defined as items presented to students but skipped by them; b) "invalid responses" were student responses that did not match the item response form; c) "not applicable" were items on which students gave a response even though an

[3] Using Stage II instead of Stage I items in test design A could hinder the analysis because high difficulty testlets which can be administered only after success in a high difficulty testlet in Stage I (or low-difficulty testlets administered only after a failure in a low-difficulty testlet in Stage I) may have a very small sample size that may not allow obtaining stable estimates of accuracy and informativeness level of effortless responses.

[4] Item timing variables reported in PISA cognitive database represented the time spent on an item during the students' last visit of the particular item. Since students were allowed to move back and forth among the items within a unit, the recorded timing variables did not reflect the total time spent on items for all students. In 2020, a new international database was made publicly available, reporting the total time each student has spent on an item and the number of visits done on each item, which were used for the time indicator in the current study. For more information about the timing variables see the "Annex K: Uses and Reporting of Process Data".
The variable "number of actions" was not updated in the new international database. Preliminary analysis of the variables (see Table A1 in Appendix A) has shown that students who have revisited an item perform on average more numbers of actions on this item than those who did not revisit, as would be expected. A similar pattern of results was also observed for the total time variables but not for the initial timing variables available in the original international cognitive database. So, it was supposed that the number of action variables were not affected by the glitch in the initial PISA 2018 cognitive database.

instruction had been provided earlier to skip the item; d) "valid skips" were items that were not answered because students had been instructed to skip them; and e) "not reached" were defined as items at the end of the test session which were probably not seen by the examinee. "Omits" and "invalid" responses were used for the item statistics and plausible values estimations, but the "not reached" responses were not included in the student's scores (OECD, 2020). Liu and Hau (2020) showed that user-defined missing values, as no response and invalid responses, can be successfully used as an indication of a lack of test-taking effort. In the current study, the specific (No response and Invalid) user-defined missingness is expected to be related to the proposed ways of measuring test-taking effort as a sign of concurrent validity.

Performance. The ten reading plausible values (PVs), available for each student (OECD, 2020), were used in the study as an indicator of the overall student reading performance.

Validation Criteria

Proposed ways of measuring test-taking effort on CR items in PISA were compared based on four pre-defined validation criteria:

a) accuracy of effortless behavior should be close to the accuracy rate expected for random responses, which is about 0% for CR items;

b) effortful behavior responses should present a higher accuracy rate than the effortless responses;

c) effortless responses should be less informative than the effortful ones (i.e., the correlations between their accuracy rate and the overall PISA Reading scores should be lower), assuming there is an adequate number of effortless responses to estimate their informativeness; and

d) higher proportions of user-defined missingness were expected to be observed on the effortless than on the effortful responses. The relationship between missingness (i.e., missing or non-missing response) and students' response behavior (i.e., effortless or effortful response) was examined with chi-square analysis.

Statistical Procedure

All analyses were performed using a sample of Spanish students who took the PISA 2018 Reading test Design A. Response total time and number of actions were used to measure test-taking effort on CR items. Four test-taking effort measures were compared: a) response time only (i.e., detecting students who rapidly skipped or responded to an item), b) number of actions only (i.e., identifying students who did a small number of actions while responding to an item), c) the union of both types of process data (i.e., capturing students who responded rapidly on an item, or used a small number of actions, or both), d) the intersection of response time and number of actions as a measure of effort (i.e., students who both responded rapidly and used a small number of actions on an item).

A single conservative threshold identification method (NT10) was used for response time to decrease the probability of Type I error (Wise, 2019). In the absence of any previous research to inform our decision to apply the NT10 threshold to the number of actions, we opted to adopt the identical approach as used for response time. Particularly, 10% of the mean total time and the mean number of actions were calculated on each item and the values were used as a threshold separating effortful from effortless behavior on that item. Responses that were identified as effortless on an item based on either total time (using the total time effort measure), or number of actions (using the number of actions effort measure), or both (using the union and the intersection measure) were classified as effortless. Four dichotomous variables (i.e., one for each proposed method of measuring test-taking effort) assigning student test-taking behavior on an item as effortful or effortless were created.

The effort measures were compared on the pre-specified validation criteria. When students were presented with an item, but they either did not respond to it or gave a response that does not match the item response form, this could be a sign of students' low engagement with the particular item. So, new scored variables were created for each item, where the "no response" and "invalid" responses were transformed into valid wrong responses and these new scored variables were used in the estimation of accuracy level. Accuracy levels and response informativeness (the point biserial correlation between the new scored item response and the 10 reading PVs) for effortless and effortful behaviors separately were calculated after splitting the file by the dichotomous variable dividing the responses into engaged and disengaged depending on the effort measure used.

The relationship of students' response behavior on an item (i.e., effortful or effortless) and item missingness was also used as a validation criterion for each effort measure examined. To estimate this relationship, a new dichotomous variable was created where omits and invalid responses were identified as missing responses (value = 0), while valid responses were identified as non-missing (value = 1). Then the frequency of missingness was calculated separately for effortful and effortless responses on each item for each effort measure examined. The significance of the relationship between the response missingness and effort behavior on an item was confirmed by a chi-square analysis performed on each item for each effort measure; Fisher's exact test has been applied to correct for small expected values (< 5) in the chi-square table (Freeman & Campbell, 2007).

A pilot study was conducted using the 3 CR items in the Core stage; a subsequent analysis comprised more items (i.e., 44 CR items from Stage I). Separate analyses were conducted for students who took the reading test in the first and in the second session of the assessment, to account for the item position effect on effort (Ivanova et al., 2020; Goldhammer et al., 2016). The magnitude of the effortless test-taking behavior in Stage I was also evaluated separately for the high and low testlet difficulty levels.

Results

Pilot analysis: three CR items in the Core stage

Three CR items (one computer-scored and two human-coded) were included in the Core stage of the PISA 2018 Spanish data in Reading. Four effort measures were estimated, and their validity was examined.

Number of effortless responses. Number of responses classified as effortless with the Intersection (I) measure was similar to the number of effortless responses identified with the Total Response-Time (TT) measure (Table 1). Effortless responses obtained with the Number of Action (NoA) and the Union (U) measures were also close in frequency. The I and the TT effort measures identified fewer effortless responses compared to NoA and U measures, more so for the two CR-HC than for the CR-CS item CR220Q01.

Effort measures accuracy. Effortless responses obtained with all four effort measures were less accurate than the effortful ones for all three items (Table 2). The accuracy levels of effortless responses were 0% for the I measure on all three items, and it did not exceed 1.2% on any of the items analyzed using the TT measure. As far as the NoA and the U measures were concerned, the accuracy of the effortless response did not exceed 0.5% on the CS-CR item; while the accuracy levels on both human-coded items were unexpectedly high: about 10% for the DR545Q04 item and about 45% for DR559Q08.

Response informativeness. When response informativeness was estimated for both types of behaviour, effortless responses were less informative (or undefined due to zero correct effortless responses) than solution behaviour for all three items, when I and TT effort measures were applied (Table 3). When the NoA and the U measures were used to distinguish the effortless from effortful responses, disengaged responses were either undefined or less accurate than

Table 1. Number of responses categorized as effortless and effortful across effort measures – Core stage

Session	Effort Measure	CR220Q01		DR545Q04		DR599Q08	
		Effortless (%)	Effortful	Effortless (%)	Effortful	Effortless (%)	Effortful
1	Number of actions	187 (3.7%)	4846	586 (8.8%)	6060	640 (12.7%)	4392
	Total response time	172 (3.4%)	4858	85 (1.3%)	6560	106 (2.1%)	4926
	Union	238 (4.7%)	4792	588 (8.8%)	6057	641 (12.7%)	4391
	Intersection	121 (2.4%)	4909	83 (1.2%)	6562	105 (2.1%)	4927
2	Number of actions	267 (5.3%)	4749	1074 (16.6%)	5411	855 (17.5%)	4029
	Total response time	342 (6.8%)	4674	346 (5.3%)	6139	292 (6.0%)	4593
	Union	396 (7.9%)	4620	1081 (16.7%)	5404	856 (17.5%)	4028
	Intersection	213 (4.2%)	4803	339 (5.2%)	6146	291 (6.0%)	4593

Table 2. Accuracy rates (proportion of correct responses) of response behaviours across effort measures – Core stage.

Session	Effort Measure	CR220Q01		DR545Q04		DR599Q08	
		Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
1	Number of actions	.0000	.2336	.0990	.7914	.4500	.7867
	Total response time	.0058	.2326	.0118	.7398	.0000	.7598
	Union	.0042	.2358	.1003	.7916	.4493	.7868
	Intersection	.0000	.2304	.0000	.7397	.0000	.7597
2	Number of actions	.0000	.1849	.0549	.7501	.3725	.7394
	Total response time	.0029	.1876	.0000	.6708	.0017	.7178
	Union	.0025	.1898	.0546	.7511	.3721	.7396
	Intersection	.0000	.1828	.0000	.6700	.0017	.7178

the solution behaviours on two of the three items analyzed. Surprisingly, the opposite was true for the third item (DR559Q08), where the informativeness of the effortful responses was lower than the informativeness of the effortless ones.

The point biserial correlations between PVs and item scores ranged from undefined and not significant to weak and significant for effortless responses on the CR-CS item for all four effort measures (Table 3). As far as the CR-HC items were concerned, the results were less conclusive; informativeness of the effortless responses obtained with the TT and I measures was low or undefined for both items. However, when the NoA and the U measures were used, informativeness was low to moderate for one of the CR-HC items (DR545Q04) and high for the other (DR559Q08). Informativeness of the effortful responses was moderate for all four effort measures on all three items.

Relationship of response behaviour and user-defined missingness. Generally, the proportion of user-defined missingness among the effortful responses on all three items, obtained with any of the four effort measures, was lower (ranging from 0.2% to 6.7%) than the proportion of missingness among the effortless responses (ranging from 36.4% to 100%; Table 4). For the CR-HC items, the proportion among effortless responses was higher for the I and TT measures (ranging from 91.5% to 100%) and lower for the NoA and U measures (ranging from 36.3% to 85%). Most of the disengaged responses identified with the TT and I measures were omitted or invalid; the same applied to item DR545Q04, but not to item DR599Q08 when the NoA and U measures were used. The chi-square relationship between human-defined missingness and student effort behaviour was significant for all items and all effort measures suggesting that convergent validity on the effort measures examined was established (Table 5).

*Reading administered during the second test session***.** When the reading assessment was administered during the second test session, many more effortless responses (i.e., about 35% to 400% more than in the first session) were identified on all items with all effort measures (Table 1). The data from the reading assessment during the second test session yielded similar validation results for effort measures as the data from the reading test in the first test session. The accuracy level of the effortless behaviour in the second session was usually smaller or equal to the accuracy rate of the effortless behaviour in the first test session (Table 2). Accuracy levels of the solution behaviours on all items, produced by all effort measures, were lower when reading was administered in the second session than in the first one. The informativeness of both types of response behaviour (i.e., effortless and effortful; Table 3) and the proportion of the human-defined missingness on effortless responses (Table 4) were similar for both reading sessions analysed. The proportion of missingness on solution behaviour in the second session was, in most cases, higher than that of the first session. As in the first session, in session 2 the chi-square statistic between user-defined missingness and student effort was significant for all items and all effort measures evaluated (Table 5).

Main analysis of the 44 CR from Stage 1

Forty-four CR items (4 CS and 40 HC) from the Stage 1 were included in the main analysis.

Table 3. Informativeness (range of correlation coefficients) of response behaviours across effort measures- Core stage

Session	Effort Measure	CR220Q01		DR545Q04		DR599Q08	
		Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
1	Number of actions	undefined	.358**; .364**	.298**; .394**	.445**; .450**	.685**; .699**	.386**; .395**
	Total response time	.084; .188*	.358**; .364**	.056; .129	.496**; .502**	undefined	.448**; .453**
	Union	.067; .157*	.355**; .361**	.295**; .320**	.445**; .452**	.686**; .700**	.385**; .395**
	Intersection	undefined	.360**; .366**	undefined	.495**; .502**	undefined	.449**; .454**
2	Number of actions	undefined	.364**; .379**	.316**; .356**	.482**; .491**	.728**; .738**	.405**; .416**
	Total response time	.080; .135*	.360**; .376**	undefined	.541**; .549**	-.033; .010	.468**; .479**
	Union	.067; .113*	.360**; .376**	.316**; .356**	.480**; .489**	.728**; .738**	.405**; .415**
	Intersection	undefined	.364**; .380**	undefined	.543**; .551**	-.033; .010	.468**; .479**

Note: * p < .05, ** p < .01

Table 4. Proportion of user-defined missingness on response behaviour across effort measures- Core stage

Session	Effort Measure	CR220Q01		DR545Q04		DR599Q08	
		Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
1	Number of actions	96.8%	4.0%	85.0%	0.4%	36.4%	0.2%
	Total response time	86.0%	4.6%	97.6%	6.7%	91.5%	2.9%
	Union	88.7%	3.4%	84.7%	0.4%	36.3%	0.2%
	Intersection	97.5%	5.2%	100%	6.7%	92.4%	2.9%
2	Number of actions	95.1%	6.7%	88.8%	0.6%	47.4%	0.2%
	Total response time	88.3%	5.8%	97.1%	10.6%	90.8%	3.2%
	Union	87.4%	4.9%	88.4%	0.6%	47.3%	0.2%
	Intersection	98.6%	7.5%	98.5%	10.6%	91.1%	3.2%

Number of effortless responses. The I and TT effort measures yielded similar proportions of effortless responses on all items, ranging from 0% to 4.28%. Effortless responses obtained with NoA and U measures had similar relative frequencies, ranging between 0.93% and 21.85% across items. The average proportions of effortless responses for the TT and the I measures were smaller than those detected with the NoA and the U measures (see Table 6). These differences were more pronounced for most CR-HC items (on 38 out of the 40 CR-HC items) compared to all four CR-CS items. The mean proportions of effortless responses on items in the high-difficulty testlets were smaller for all four effort measures than in the low-difficulty testlets (see Table 6).

Effort measures accuracy. Effortless responses on each item (except for one[5]) were less accurate than the effortful ones for all measures investigated (see Figure 1). The effortless response accuracy rates for all CR-CS items (except for one[5]) were constant (i.e., .00) across the four effort measures. The accuracy level of the disengaged responses obtained with the TT and the I measures was lower than .03 or for all items (except for one[5]). The accuracy of the effortless responses identified with the NoA and the U measures ranged between .00 and .56 across items, with average values of .08. In terms of the average accuracy rates of disengaged responses, no meaningful differences were found across testlet difficulty levels for any effort measure examined; the average accuracy of the effortful responses was higher for low-difficulty than high-difficulty testlets.

Response informativeness. When the TT and the I measures were used, the effortless response informativeness was undefined for almost all items, as the accuracy rates of these responses were constant at .00. Informativeness of the effortful responses ranged between low and medium (see Table B1 in Appendix B), conforming to the effort measure validation criteria. When the NoA and the U measures were used,

Table 5. Chi-square relationship between missingness and student effort across effort measures- Core stage

Session	Effort Measure	CR220Q01	DR545Q04	DR599Q08
1	Number of actions	2261.145**	5271,101**	1599,120**
	Total response time	1603.790**	957,040**	1777,963**
	Union	2401,532**	5250,439**	1596,230**
	Intersection	1466,157**	983,849**	1796,411**
2	Number of actions	1960.729**	5397.979**	2020.706**
	Total response time	2152.715**	1897.251**	2710.490**
	Union	2461.119**	5382.664**	2017.818**
	Intersection	1677.043**	1921.495**	2720.547**

Note: ** p < .001, df = 1 for all tests.

Table 6. Average proportions of effortless responses across effort measures by session and difficulty level- Stage I

	Mean proportion of effortless responses (s.d.)			
	Total response time	Number of actions	Union	Intersection
Session 1- all items	1.52% (1.07%)	8.50% (5.76%)	8.52% (5.75%)	1.46% (1.03%)
High-difficulty items	0.85% (0.65%)	6.61% (5.09%)	6.63% (5.08%)	0.83% (0.65%)
Low-difficulty items	2.25% (0.96%)	10.58% (5.84%)	10.60% (5.83%)	2.16% (0.93%)
Session 2- all items	3.35% (2.47%)	11.04% (7.30%)	11.10% (7.28%)	3.17% (2.36%)
High-difficulty items	1.63% (1.11%)	8.10% (6.23%)	8.14% (6.22%)	1.59% (1.12%)
Low-difficulty items	5.22% (2.17%)	14.25% (7.14%)	14.35% (7.10%)	4.91% (2.14%)

[5] One item yielded just one effortless response when TT was used, and this response was correct resulting in accuracy rate of the effortless responses = of 1.0. This item had also 0% missing values on the effortless responses obtained with the TT measure.

Figure 1. Clustered boxplots of accuracy level for 44 items by effort measure and by response behaviour

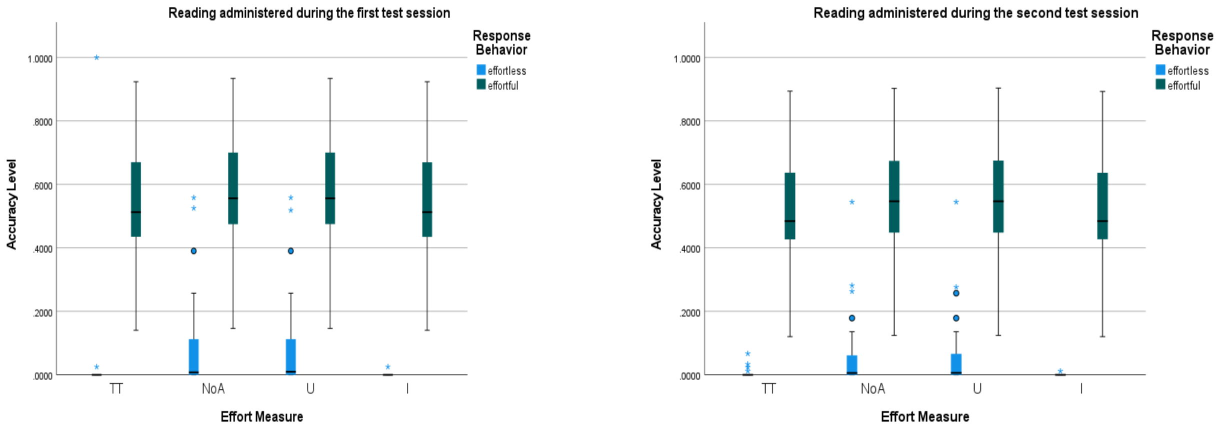
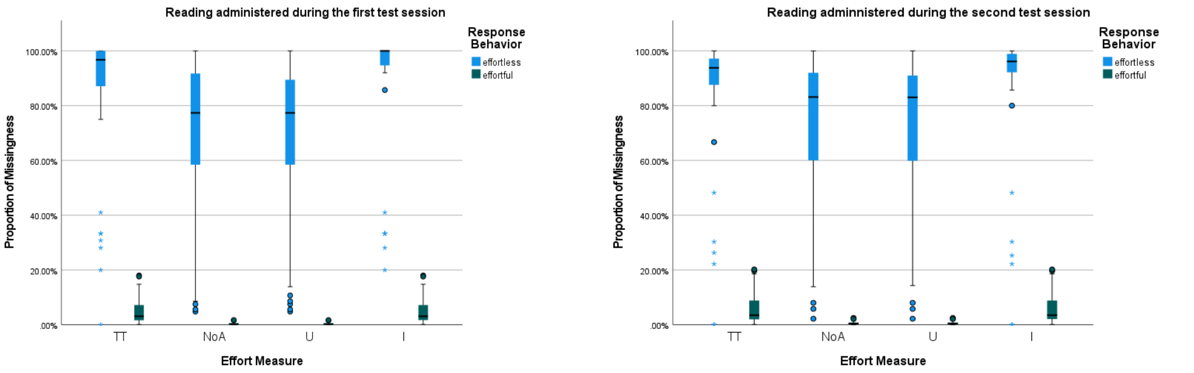


Figure 2. Clustered boxplots of missingness proportions for 44 items by effort measure and by response behaviour



the disengaged responses were less informative than the engaged ones for most of the items (on 33 items); however, 11 CR-HC items yielded more informative effortless than solution behaviors, not meeting the informativeness validation criterion when NoA and U were applied. The point biserial correlations between PVs and item scores ranged from undefined and not significant to high significant for effortless responses and from low to medium for the effortful ones (Table B1). No systematic differences were observed between low and high difficulty testlets.

Relationship of response behaviours with the user-defined missingness. Effortless responses of all items (except for one[5]) were associated with a higher proportion of human-defined missingness (ranging from 20% to 100%) than effortful ones (ranging from 0.1% to 18.0%) when the TT and the I effort measures were used. Similar results were obtained with the NoA and the U measures, but the proportions of missingness for effortless-response were lower for some items (ranging from 4.8% to 100%). The chi-square test (and the Fisher's exact test applied when small expected values were observed in the chi-square table) between human-defined missingness (missing or non-missing) and student response behaviour (effortless or effortful) was significant for all items (except for one[5]) on all effort measures examined. This implied that all four measures complied with the missingness validation criterion.

On average, the proportions of missingness on effortless responses were higher for the TT and I measures than for NoA and U measures (see Figure 2); this difference was more pronounced for low-difficulty than for high-difficulty testlets (see Table 7). Different results were observed for CR-CS items, where proportions of missingness on effortless behaviour obtained with the NoA or I measures were higher or equal to the proportion of missingness of effortless responses identified with the TT and U measures.

*Results when Reading was administered during the second test session***.** All effort measures identified a higher proportion of effortless responses on most of the Reading items (for 41 out of 44 items) administered during the second test session than when they were administered during the first one. The effortless responses in session 2 exceeded the ones in session 1 by about 2% on average for any of the effort measures applied (Table 6). These differences in the proportions of disengagement between the two sessions were higher for the low-difficulty testlets (ranging between 2.75% and 3.75% across effort measures) than for the high-difficulty ones (ranging between 0.76% and 1.51% across measures).

The accuracy rates of both response behaviours in the second test session were similar to those in the first test session (see Figure 1). Effortless response accuracy levels were lower than solution behaviour accuracy on all items for all effort measures examined. When the TT and I effort measures were used, the average accuracy rates of the effortless responses were very close to 0 (.003 for TT and .0003 for I); while the mean accuracies of the effortless responses identified with NoA and U measures were a bit higher (.052 for NoA and U). No meaningful differences were found in the average accuracy rates of the disengaged responses across testlet difficulty levels for any effort measure.

The informativeness of both types of response behaviour was similar across sessions for most of the items (Table B2 in Appendix B). When the TT measure was used, effortless behaviours were less informative than the effortful ones for most of the items (for 43 out of 44 items); similar results were obtained with the I effort measure for all items. Most of the disengaged responses identified with the NoA and the U measures were also less informative than the solution behaviours (36 items); however, 8 HC-CR items did not conform to the informativeness validation criterion.

As in the first session, the proportions of missingness on effortless responses were higher than the proportions of missingness on solution behaviour for all items; exceptions were 2 CR-HC items that did not conform to this validation criterion when the TT and I measures were applied. However, Fisher's exact test revealed a non-significant relationship between missingness and student response behaviour, and only a small number of effortless responses (i.e., 4 and 10) have been identified with TT and I measures on these two items. A comparison of the average proportions of missingness across effort measures yielded similar results in session 2 as in session 1: the proportions of missingness on the effortless responses were higher for the TT and I than for the NoA and U measures (see Figure 2), and this difference was larger for the low than for the high difficulty testlets (see Table 7). The chi-squared relationships and Fisher's exact tests showed that the relationships between user-defined missingness and student response behaviours were significant for all items (except for the two mentioned above) and for all effort measures examined.

A comparison of the two sessions showed that the average proportions of missingness on effortless responses identified with the NoA and the U measures were higher in the second test session than when Reading was administered in the first session of the assessment (Table 7); this increase in human-defined missingness was more apparent for the low than for the high difficulty testlets. On the other hand, the proportions of missingness on the effortless responses seemed to be slightly lower for the second than for the first session when the TT and I measures were applied.

Further Investigation on Why NoA Failed to Be a Valid Indicator of Effort

A couple of post hoc hypotheses were proposed to explain the failure of NoA as a valid indicator of effort. In comparison with other PISA CR items, if items where NoA failed as an indicator of effort:

a) had much higher NoA thresholds, therefore, effortful behavior can be recognized as effortless. A higher NT10 threshold means a higher mean NoA on an item; and

b) could be answered correctly with a small NoA.

Descriptive analyses of the variable NoA were performed to examine the two hypotheses. Some items for which NoA failed as an indicator had lower NoA thresholds, while others had high NoA threshold; the same results were observed for items where NoA did not fail as a measure of effort (see Graph C1 in Appendix C). Additionally, correct response could be given with just a small number of actions on some of the failing items; however, this was often the case for

Table 7. Average proportions of user-defined missingness across effort measures and response behaviors- Stage I

	Mean % of user-defined missingness (s.d.)							
	Total response time		Number of actions		Union		Intersection	
	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
Session 1- all items	84.83 (26.71)	4.85 (4.68)	67.97 (29.40)	0.40 (0.46)	67.57 (29.08)	0.39 (0.46)	88.73 (23.63)	4.86 (4.69)
High-difficulty items	78.52 (31.16)	4.36 (4.69)	67.59 (34.11)	0.47 (0.51)	66.87 (33.60)	0.47 (0.51)	85.42 (27.11)	4.36 (4.69)
Low-difficulty items	91.75 (19.24)	5.39 (4.72)	68.39 (24.04)	0.31 (0.38)	68.33 (23.99)	0.31 (0.38)	92.20 (19.40)	5.41 (4.74)
Session 2- all items	83.58 (26.54)	5.89 (5.58)	72.60 (28.64)	0.57 (0.55)	71.95 (28.16)	0.54 (0.55)	85.73 (27.09)	5.93 (5.61)
High-difficulty items	77.14 (31.33)	5.05 (5.54)	70.16 (33.64)	0.63 (0.66)	69.15 (32.89)	0.62 (0.65)	80.39 (32.45)	5.06 (5.54)
Low-difficulty items	90.62 (18.29)	6.80 (5.62)	75.28 (22.47)	0.51 (0.40)	75.01 (22.27)	0.46 (0.42)	91.58 (18.75)	6.88 (5.65)

some of the non-failing items as well (see Graph C2 in Appendix C). So, neither of the two post hoc hypothesis was supported.

Discussion

The purpose of the current study was to explore behavioural indicators of test-taking effort on CR items based on item response time and/or number of actions and examine their validity. The rationale behind this study was that, while response time effort has been demonstrated as a valid indicator of test-taking effort on multiple-choice items (Wise & Kong, 2005), its effectiveness on constructed response items may be diminished (Wise & Gao, 2017). Beyond response time, NoA is an alternative indicator that could hold promise in capturing test engagement, because the frequency of active interactions with an item was supposed to reflect a goal-directed behavior on that item (Greiff et al., 2016). Measures of effort on CR items have been understudied, even though such items have been related to lower levels of effort compared to selected-response items (DeMars, 2000; Eklöf & Knekta, 2017). Change in student response behaviour throughout the test was taken into consideration by analysing results separately for the two sessions of the PISA reading test. Another innovation of the present study was the comparison of examinee behaviour across testlet difficulty levels. Such comparison may provide a better understanding of student test-taking behaviour across examinee ability groups responding to testlets of different difficulty levels.

A pilot study focused on the three core stage CR items included in the PISA 2018 Reading domain to perform a detailed description and analysis of the validity of the effort measures, since it consisted of fewer items and was administered to a larger number of students. The pilot study results were confirmed and enhanced by the main analysis, using data from 44 additional CR items. Pilot and main analysis results showed that the NoA yielded a similar number of disengaged responses as the U measure. The number of effortless responses obtained with the TT and I effort indicators were also similar and lower than those identified by the NoA and U measures. Most of the responses classified as effortless by the TT measure seemed to fall also into the same category of disengaged behaviour when the NoA data were used; however, NoA detected additional effortless responses not identified by the TT measure. The difference between the proportions of disengaged responses obtained with NoA and effortless behaviour identified by TT was larger for the HC-CR than for the CS-CR items.

Similarities were observed also in the validity results, between NoA and U measures and between the TT and I measures. All four effort measures conformed to all of the validity requirements for most of the items analysed. Effortless responses had a lower accuracy rate and were associated with significantly more missing responses than the effortful ones when there were enough disengaged responses to ensure robust validity estimates. However, about one-third of the CR items (all of them were CR-HC) yielded an unexpectedly high accuracy level (higher than 5%) on the disengaged behaviour and higher informativeness on the effortless than on the effortful responses when using the NoA and U indicators. These results suggested that the NoA measure, when used as a binary indicator of engagement with NT10 threshold, may lead to misclassification of effortful behaviour as effortless (Type II error) for multiple CR-HC items, while the TT did not seem to be associated with such a drawback. So, responses flagged as disengaged with NoA due to a small number of actions may include students who performed enough actions to get credit on an item. The NT10 was known to be a very conservative threshold identification method when used with response time data (Wise, 2019); but when applied to NoA, it might be more liberal in identifying disengagement, especially on CR-HC items. Additionally, lower average proportions of user-defined missingness were observed when the NoA and the U measures were used. Note that the unexpected validity results for some items obtained with the NoA and U measures could not be explained by: the difficulty level of the testlet in which the items were included, or, as post-hoc analyses have shown, a high number of actions threshold, or a high probability to respond correctly to an item with a small number of actions.

Items included in the high-difficulty testlets had, on average, lower proportions of disengagement than items in easier testlets; this was true for all the effort indicators used in the study. In the multi-stage testing approach, most of the students in the low-difficulty testlets in Stage I had low or moderate scores in the Core stage, while examinees taking the high-difficulty testlets had mainly moderate or high results in the Core stage (OECD, 2020). This implies that, in general, students presented with an easy testlet at Stage I had lower ability than students getting a difficult testlet. Therefore, higher ability students engaged in less effortless behaviour than lower-ability students, despite responding to more demanding material. This was in line with previous studies arguing that lower-ability college students exhibit more rapid-guessing behavior than higher-ability students (Wise et al., 2009).

As expected, effort measures showed overall similar validity results, leading to the same conclusions about the validity of the indicators across testlet difficulty levels. However, the disengaged responses in low-difficulty testlets had, on average, a higher proportion of missingness than the effortless responses in high-difficulty testlets, especially when the NoA and the U were used as indicators of effort. This suggests that low-ability students who invest less effort on an item solution tend to skip the items more frequently without giving a response than the higher ability rapid guessers. Higher-ability students are more likely to make guesses and give "perfunctory responses" (Wise & Gao, 2017) when they are not fully engaged with the item solution.

As hypothesized, the validity results for effort measures were similar across sessions, even though items administered later in the test yielded more effortless responses than the same items administered in the first half of the assessment, especially in low-difficulty testlets. The difference in the level of disengagement across test sessions was in line with previous literature confirming the item position effect on engagement (e.g., Ivanova et al., 2020; Debeer et al., 2014; Goldhammer et al., 2016; Wise & Kingsbury, 2016).

The results implied that high-ability examinees may start with a higher level of engagement in item solution and were less likely to give up throughout the assessment than the low-ability examinees. The proportion of missingness on effortless responses was higher in the second than in the first session when the NoA and the U measures were used, especially for items belonging to the low-difficulty testlets, suggesting that low-ability students, who tend to give brief disengaged responses, may persist less and more often skip items when they get tired or bored throughout the test than the higher-ability disengaged examinees.

Limitations and Future Directions

To examine the validity of different effort measures, the present study focused on data from a single country (Spain) and subject (Reading), selected because of its large sample size and potential to provide more stable estimates. However, results may be influenced by the particular nation, language, subject, test administration, or student age. Future research may replicate the study using different samples, tests, and contexts.

The zero accuracy rate validation criteria of disengaged responses on CR items and the 10% normative threshold identification method were selected arbitrarily. The conservative nature of the NT10 threshold, when applied on response time data, has been well documented (Kroehne et al., 2020; Lindner et al., 2017; Wise, 2019), however it does not exclude the possibility of misclassification of effortful behaviour as effortless: some examinees may spend a long time on an item without truly investing effort in solving it. Additionally, NT10 had been previously applied mostly to response time data and it was used primarily with multiple-choice items. Its applicability to other types of process data has not been studied. Additional studies may focus on examining different threshold identification methods for TT and NoA across item types.

The simple binary splitting of item-level test-taking behaviour into rapid-guessing and solution behaviour has led to progress in the literature examining test-taking effort; however, student effort may be also comprehended as a continuous phenomenon where examinees can invest partial effort on an item (Goldhammer, Martens, & Lüdtke, 2017; Wise & Kuhfeld, 2021). Future studies can explore continuous indicators of item-level test-taking effort. Two types of student behaviour were selected to indicate effortless responses (TT and NoA); other process data (e.g., time elapsed to the first action, item revisits, text reread, etc.) may also provide valuable information about examinee test-taking behaviour and future research may focus on investigating other behavioural indicators of student engagement.

The PISA items were not publicly available, so the information available about the item format was limited. Different dimensions, such as item types and testlet difficulty levels, had been taken into consideration in an attempt to understand the unexpected NoA validity results, but no plausible explanation had been revealed. Additional item characteristics, such as the length of the item or the required response, and the inclusion of graphs, images, or ancillary reading materials, may also influence examinee test-taking behaviour (Setzer et al., 2013) and explain the large level of misclassification on some items when NoA was used. For example, a brief response may be sufficient for some CR items, but extended answers with multiple NoA may be preferred by many students and accepted by the raters as correct. Future research examining the effect of different item characteristics on the student efficiency will assist in a better understanding of examinees' test-taking behaviour on CR items.

Conclusion

Overall, the validity results were in favour of relying on response time over NoA as a dichotomous effort indicator on CR items in PISA when the NT10 threshold was applied, since NoA and the union of NoA and TT did not conform to some of the pre-defined validation criteria used in the study. This finding establishes an empirical foundation for researchers to employ response time effort indicators in tests that encompass both multiple-choice and constructed response items. However, further research may focus on examining the validity of NoA as a potentially useful indicator of engagement when used as a continuous variable (Ivanova et al., 2020) or with a different, less liberal threshold identification method. Obtaining valid indicators of test-taking effort for CR items can assist professionals in investigating and comparing examinee behavior on different kinds of items, thereby improving test properties and item specifications. A measure of effort on different types of items is also essential when examining changes in effort throughout the test and across items.

References

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*(3), 279-304.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502–523.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77.

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10*(3), 207-229.

Eklöf, H. (2010a). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345-356.

Eklöf, H. (2010b). *Student motivation and effort in the Swedish TIMSS advanced field study* [Paper presentation]. 4th IEA International Research Conference, Gothenburg, Sweden. https://www.iea.nl/sites/default/files/2019-04/IRC2010_Eklof.pdf

Eklöf, H., & Knekta, E. (2017). Using large-scale educational data to test motivation theories: A synthesis of findings from Swedish studies on test-taking motivation. *International Journal of Quantitative Research in Education, 4*(1-2), 52-71.

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in Education*, *27*(1), 31-45.

Freeman, J.V., & Campbell, M.J. (2007). The analysis of categorical data: Fisher's exact test. *Scope, 33*(5), 11-12.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist, 50*(1), 43-57.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers, No. 133). Paris: OECD Publishing.

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

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating Product Data to Process Data from Computer-Based Competency Assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence Assessment in Education: Research, Models and Instruments*. Switzerland: Springer.

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36-46.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*, 173–183. doi:10.1080/08957347.2016.1171766

Ivanova, M. G., Michaelides, M., & Eklöf, H. (2020). How Does the Number of Actions on Constructed-Response Items Relate to Test-Taking Effort and Performance? *Journal of Educational Research and Evaluation, 26*(5-6), 252-274. DOI: 10.1080/13803611.2021.1963939

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103-129.

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.

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling, 62*(2), 147-177.

Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology, 51*, 482-492.

Liu, Y., & Hau, K. T. (2020). Measuring Motivation to Take Low-Stakes Large-Scale Test: New Model Based on Analyses of "Participant-Own-Defined" Missingness. *Educational and Psychological Measurement*, 1-30.

Michaelides, M. & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: country and item-type differences. *Psychological Test and Assessment Modeling, 64*(3), 304-338.

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items. *International Journal of Testing*, 1-19.

Organization for Economic Co-operation and Development (2010). *PISA 2009 Results: What Students Know and Can Do Student Performance in Reading, Mathematics and Science (Volume I)*. Paris: OECD Publishing.

Organization for Economic Co-operation and Development (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*. Paris: OECD Publishing.

Organization for Economic Co-operation and Development (2017). *PISA Technical Report*. Paris: OECD Publishing.

Organization for Economic Co-operation and Development (2019a). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing. https://doi.org/10.1787/b25efab8-en

Organization for Economic Co-operation and Development (2019b). *PISA 2018 Results: What Students Know and Can Do* (Vol. I). Paris: OECD Publishing. https://doi.org/10.1787/5f07c754-en

Organization for Economic Co-operation and Development (2020). *PISA Technical Report*. Paris: OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Provasnik, S. (2021). Process data, the new frontier for assessment development: rich new soil or a quixotic quest?. *Large-scale Assessments in Education, 9*(1), 1-17.

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

Schleicher, A. (2019). *PISA 2018: Insights and Interpretations.* Paris: OECD Publishing.

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

Silm, G., Pedaste, M., & Täht, K. (2020) The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review *Educational Research Review.* doi: https://doi.org/10.1016/j.edurev.2020.100335.

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 presentation]. Annual meeting of the National Council on Measurement in Education, IL, Chicago.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*(6), 2568-2572.

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. (2019). An Information-Based Approach to Identifying Rapid-Guessing Thresholds. *Applied Measurement in Education, 32*(4), 325-336.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment*, *10*(1), 1-17.

Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's "Reliability of speeded number-right multiple-choice tests". *Applied Psychological Measurement, 33*(6), 488-490.

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. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S., & Kuhfeld, M. (2021). *A method for identifying partial test-taking engagement. Applied Measurement in Education, 34*(2), 150-161.

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

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). *An Investigation of the Relationship between Time of Testing and Test-Taking Effort* [Paper presented]. Annual meeting of the National Council on Measurement in Education, CO, Denver.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185-205.

Yavuz, H. C. (2019). The effects of log data on students' performance. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 10*(4), 378-390.

Citation:
Ivanova, M. G., & Michaelides, M. P. (2023). Measuring test-taking effort on constructed-response items with item response time and number of actions. *Practical Assessment, Research, & Evaluation*, 28(15). Available online: https://doi.org/10.7275/pare.1921

Corresponding Author:
Militsa G. Ivanova
University of Cyprus
Email: militsagi [at] yahoo.com

Appendix A

Table A1. Descriptive statistics of time on last visit, total time and number of action variables in PISA 2018.

Item Name	Type of Item	Mean Time on Last Visit		Mean Total Time		Mean Number of Actions	
		One Visit	Revisits	One Visit	Revisits	One Visit	Revisits
CR220Q01	OR-CS	198 279.08	82 728.42	197 873.09	238 678.67	16.47	22.52
DR545Q04	OR-HC	120 968.11	54 710.98	120 469.05	162 462.59	129.11	153.80
DR559Q08	OR-HC	170 438.05	87 777.20	146 184.08	153 557.63	223.42	258.14
CR424Q02	CMC	89 826.32	54 498.94	89 528.67	122 793.85	5.32	8.17
CR545Q03	CMC	84 605.34	36 843.95	84 300.03	105 845.06	5.36	6.27
CS424Q03	SMC	44 216.67	27 829.25	43 905.85	59 031.54	3.35	5.39
CR424Q07	SMC	42 775.88	30 762.49	42 444.98	52 123.47	2.78	3.16
CR220Q02	SMC	66 884.46	36 841.92	66 593.94	85 714.12	4.66	6.03
CR220Q04	SMC	48 973.71	31 339.58	48 664.59	57 987.00	3.21	4.01
CR220Q05	SMC	24 550.54	13 674.05	24 246.84	34 605.48	1.73	2.16
CR220Q06	SMC	36 558.46	21 399.79	36 237.96	54 096.23	2.26	3.26
CR560Q10	SMC	86 807.20	22 327.29	86 127.52	104 752.90	6.27	12.29
CR560Q03	SMC	149 676.85	77 869.38	149 344.20	168 559.74	3.25	4.02
CR560Q06	SMC	49 846.90	25 089.65	49 535.96	67 883.15	2.33	3.26
CR560Q08	SMC	60 599.00	45 978.69	60 283.40	83 949.69	2.53	3.06
CR545Q02	SMC	197 152.56	46 040.74	196 540.83	234 248.16	4.24	4.73
CR545Q06	SMC	37 274.45	17 309.82	197 836.20	230 998.33	2.41	2.76
CR545Q07	SMC	29 229.75	16 234.95	28 918.60	44 065.40	2.23	2.66
CR559Q01	SMC	145 761.79	35 391.50	145 414.00	156 970.93	2.77	4.03
CR559Q04	SMC	42 528.93	21 781.90	42 233.12	53 953.82	2.16	2.52
CR559Q03	SMC	45 119.80	17 863.06	44 821.11	60 162.35	2.17	2.91
CR559Q06	SMC	62 098.71	30 846.78	61 800.18	80 036.47	2.59	3.26

Note: OR-CS- open response computer scored items, OR-HC- open response human coded item, CMC- complex multiple-choice, SMC- simple multiple-choice items.

Appendix B

Table B1. Informativeness (range of correlation coefficients between item scores with 10 plausible values) of response behaviours across effort measures-Stage I, first session.

Item in MS Analysis	Total Time Threshold		Number of Actions Threshold		Union Threshold		Intersection threshold	
	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
DR067Q04C	undefined	[.381**; .416**]	undefined	[.332**; .371**]	undefined	[.332**; .371**]	undefined	[.381**; .416**]
DR067Q05C	undefined	[.431**; .448**]	undefined	[.366**; .386**]	undefined	[.366**; .386**]	undefined	[.431**; .448**]
DR456Q02C	undefined	[.322**; .344**]	undefined	[.273**; .298**]	undefined	[.273**; .298**]	undefined	[.322**; .344**]
DR456Q06C	undefined	[.413**; .442**]	[.142; .272**]	[.368**; .395**]	[.142; .272**]	[.368**; .395**]	undefined	[.413**; .442**]
DR547Q09C	undefined	[.506**; .534**]	[.365**; .463**]	[.494**; .524**]	[.365**; .463**]	[.494**; .524**]	undefined	[.506**; .534**]
DR540Q04C	undefined	[.346**; .365**]	[.353**; .437**]	[.296**;.320**]	[.354**; .437**]	[.294**; .318**]	undefined	[.347**; .365**]
DR542Q02C	undefined	[.221**; .247**]	[.178; .326]	[.206**; .232**]	[.178; .326]	[.206**; .232**]	undefined	[.221**; .247**]
DR420Q02C	undefined	[.438**;.465**]	[.652**; .702**]	[.410**; .442**]	[.652**; .702**]	[.410**; .442**]	undefined	[.438**; .465**]
DR420Q10C	undefined	[.563**; .587**]	undefined	[.490**; .523**]	undefined	[.490**; .523**]	undefined	[.563**; .587**]
DR420Q06C	undefined	[.201**; .229**]	[.020; .093]	[.113**; .138**]	[.020; .093]	[.113**; .136**]	undefined	[.201**; .229**]
DR420Q09C	undefined	[.362**; .391**]	undefined	[.344*; .373**]	undefined	[.341**; .370**]	undefined	[.364**; .394**]
DR455Q02C	undefined	[.284**; .302**]	undefined	[.267**; .285**]	undefined	[.267**; .285**]	undefined	[.284**; .302**]
DR455Q03C	undefined	[.179**; .203**]	[.442**; .543**]	[.151**; .172**]	[.442**; .543**]	[.151**; .172**]	undefined	[.179**; .203**]
DR550Q09C	undefined	[.324**; .341**]	[.339*; .496**]	[.301**; 317**]	[.339*; .496**]	[.301**; 317**]	undefined	[.324**; .341**]
DR550Q10C	undefined	[.262**; .286**]	[.287**; .406**]	[.228**; .252**]	[.287**; .406**]	[.228**; .252**]	undefined	[.262**; .286**]
DR550Q07C	undefined	[.309**; .340**]	[.309**; .341**]	[.272**; .305**]	[.309**; .341**]	[.272**; .305**]	undefined	[.309**; .340**]
DR055Q02C	undefined	[.367**; .388**]	[.404**; .478**]	[.323**; .346**]	[.404**; .478**]	[.323**; .346**]	undefined	[.371**; .391**]
DR055Q03C	undefined	[.450**; .470**]	[.578**; .639**]	[.443**; .462**]	[.578**; .639**]	[.443**; .462**]	undefined	[.450**; .470**]
DR055Q05C	undefined	[.500**; .530**]	[.244*; .333**]	[.416**; .442**]	[.244*; .333**]	[.416**; .442**]	undefined	[.500**; .530**]
DR111Q02BC	undefined	[.340**; .362**]	[.124; .206**]	[.301**; .320**]	[.124; .206**]	[.301**; .320**]	undefined	[.340**; .362**]
DR111Q06C	undefined	[.430**; .445**]	[.062; .143]	[.391**; .407**]	[.062; .143]	[.390**; .407**]	undefined	[.430**; .446**]
DR446Q06C	undefined	[.402**; .422**]	[.088; .144]	[.356**; .381**]	[.088; .144]	[.356**; .381**]	undefined	[.402**; .422**]
DR546Q03C	undefined	[.402**; .422**]	[.576**; .645**]	[.387**; .410**]	[.567**; .639**]	[.388**; .410**]	undefined	[.402**; .421**]
DR549Q05C	undefined	[.413**; .437**]	[.545**; .645**]	[.361**; .389**]	[.545**; .645**]	[.361**; .389**]	undefined	[.413**; .437**]
DR558Q04C	undefined	[.413**; .439**]	[.239; .386**]	[.380**; .408**]	[.239; .386**]	[.380**; .408**]	undefined	[.413**; .439**]
DR558Q12C	undefined	[.385**; .401**]	undefined	[.362**; .379**]	undefined	[.362**; .379**]	undefined	[.385**; .401**]
DR437Q07C	undefined	[.220**; .245**]	[-.013; .041]	[.208**; .230**]	[-.013; .041]	[.208**; .230**]	undefined	[.220**; .245**]
DR561Q07C	undefined	[.169**; .201**]	undefined	[.137**; .172**]	undefined	[.137**; .172**]	undefined	[.169**; .201**]

Item in MS Analysis	Total Time Threshold		Number of Actions Threshold		Union Threshold		Intersection threshold	
	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
DR562Q03C	undefined	[.279**; .312**]	undefined	[.276**; .310**]	undefined	[.276**; .310**]	undefined	[.279**; .312**]
DR562Q06C	undefined	[.303**; .332**]	[.083; .195*]	[.277**; .302**]	[.083; .195*]	[.277**; .302**]	undefined	[.303**; .332**]
DR564Q05C	undefined	[.511**; .531**]	[.513**; .554**]	[.461**; .482**]	[.513**; .554**]	[.461**; .482**]	undefined	[.511**; .531**]
DR565Q02C	undefined	[.349**; .369**]	undefined	[.319**; .343**]	undefined	[.318**; .342**]	undefined	[.350**; .369**]
DR565Q05C	undefined	[.349**; .380**]	undefined	[.317**; .348**]	undefined	[.317**; .348**]	undefined	[.349**; .380**]
DR404Q10C	undefined	[.549**; .562**]	undefined	[.505**; .521**]	undefined	[.505**; .521**]	undefined	[.549**; .562**]
DR453Q04C	undefined	[.383**; .402**]	[-.100; -.026]	[.347**; .368**]	[-.100; -.026]	[.347**; .368**]	undefined	[.383**; .402**]
DR453Q06C	undefined	[.396**; .413**]	undefined	[.349**; .368**]	undefined	[.347**; .366**]	undefined	[.398**; .414**]
DR553Q04C	[.046; .394]	[.510**; .533**]	[.236; .373**]	[.485**; .510**]	[.236; .373**]	[.485**; .510**]	[.046; .394]	[.510**; .533**]
CR104Q01S	undefined	[.409**; .427**]	undefined	[.402**; .422**]	undefined	[.402**; .422**]	undefined	[.409**; .427**]
CR104Q02S	undefined	[.183**; .199**]	undefined	[.181**; .198**]	undefined	[.180**; .197**]	undefined	[.184**; .200**]
CR104Q05S	undefined	[.340**; .374**]	undefined	[.336**; .370**]	undefined	[.335**; .369**]	undefined	[.341**; .375**]
DR569Q06C	undefined	[.445**; .474**]	[.200; .277*]	[.406**; 438**]	[.200; .277*]	[.406**; 438**]	undefined	[.445**; .474**]
DR466Q02C	undefined	[.425**; .437**]	[.351**; .426**]	[.412**; .428**]	[.351**; .426**]	[.412**; .428**]	undefined	[.425**; .437**]
CR466Q06S	undefined	[.322**; .336**]	undefined	[.288**; .307**]	[-.128; .027]	[.288**; .308**]	undefined	[.321**; .335**]
DR412Q08C	undefined	[.374**; .389**]	undefined	[.352**; .375**]	undefined	[.351**; .375**]	undefined	[.375**; .389**]

Note: * p < .05, ** p < .01. Green boxes represented items belonging to low-difficulty testlets, while brown boxes stand for items belonging to high-difficulty testlets. CR-CS items were highlighted in yellow. Unexpected informativeness results, where informativeness was higher for effortless than for effortful responses were presented in red.

Table B2. Informativeness (range of correlation coefficients between item scores with 10 plausible values) of response behaviours across effort measures-Stage I, second session

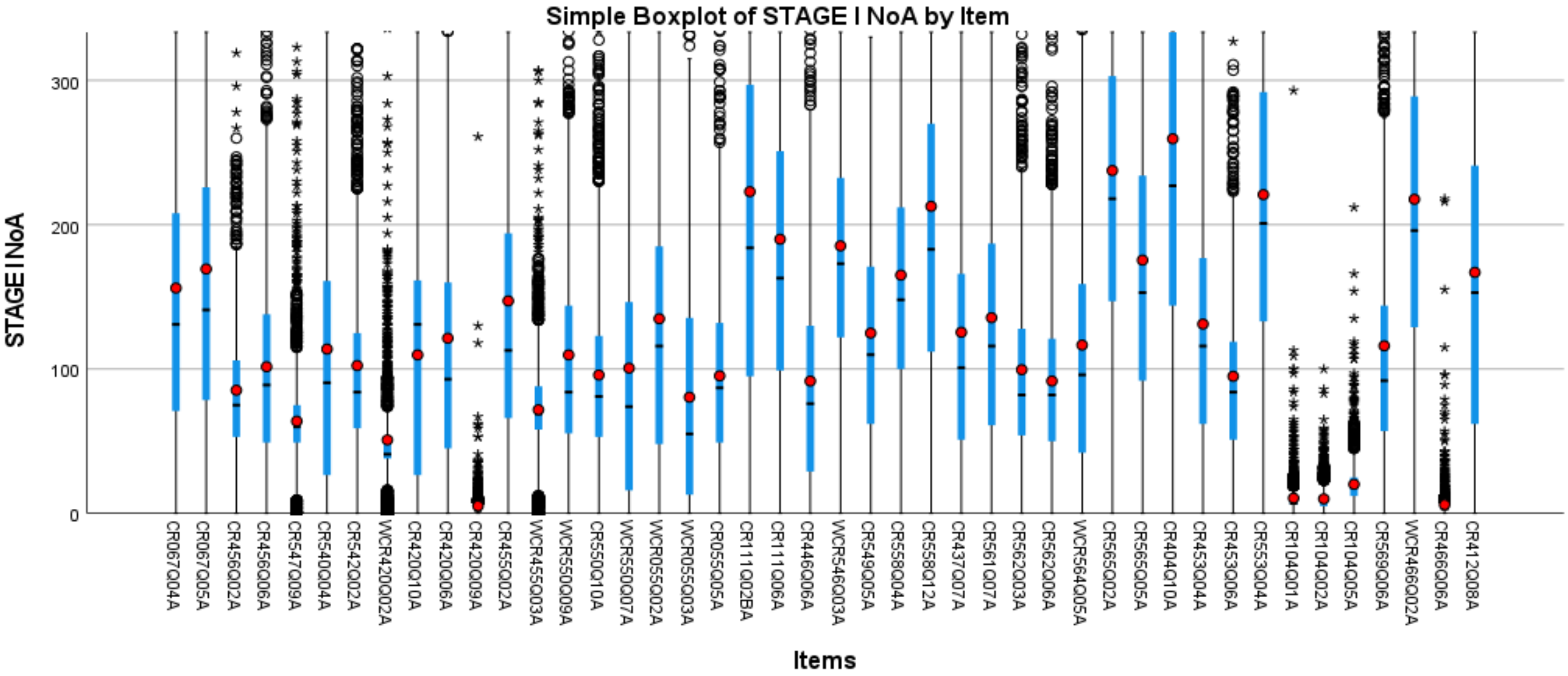
Item in MS Analysis	Total Time Threshold		Number of Actions Threshold		Union Threshold		Intersection threshold	
	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
DR067Q04C	undefined	[.421**; .443**]	[-.064; .005]	[.374**; .397**]	[-.064; .005]	[.374**; .397**]	undefined	[.421**; .443**]
DR067Q05C	undefined	[.450**; .465**]	undefined	[.390**; .406**]	undefined	[.388**; .404**]	undefined	[.452**; .466**]
DR456Q02C	undefined	[.370**; .400**]	[.210*; .329**]	[.330**; .360**]	[.210*; .329**]	[.328**; .358**]	undefined	[.372**; .402**]
DR456Q06C	undefined	[.449**; .464**]	[.214*; .297**]	[.412**; .432**]	[.214*; .297**]	[.411**; .430**]	undefined	[.450**; .465**]
DR547Q09C	[.360; .551*]	[.477**; .492**]	[.511**; .571**]	[.466**; .480**]	[.509**; .592**]	[.466**; .480**]	undefined	[.477**; .492**]
DR540Q04C	undefined	[.383**; .409**]	[.360**; .394**]	[.359**; .390**]	[.360**; .394**]	[.358**; .390**]	undefined	[.384**; .410**]
DR542Q02C	undefined	[.253**; .270**]	undefined	[.214**; . 233**]	undefined	[.214**; . 232**]	undefined	[.254**; .270**]
DR420Q02C	undefined	[.410**; .430**]	[.561**; .614**]	[.370**; .392**]	[.563**; .616**]	[.367**; .388**]	undefined	[.413**; .433**]
DR420Q10C	[-.074; .030]	[.574**; .585**]	[-.077; -.026]	[.511**; .527**]	[-.077; -.026]	[.511**; .527**]	[-.074; .030]	[.574**; .585**]
DR420Q06C	undefined	[.180**; .207**]	undefined	[.103**; .132**]	undefined	[.103**; .132**]	undefined	[.180**; .207**]
DR420Q09C	undefined	[.323**; .346**]	undefined	[.308**; .330**]	undefined	[.301**; .324**]	undefined	[.329**; .351**]
DR455Q02C	undefined	[.302**; .316**]	undefined	[.285**; .298**]	undefined	[.285**; .298**]	undefined	[.302**; .316**]
DR455Q03C	undefined	[.250**; .272**]	[.550**; .608**]	[.220**; .246**]	[.550**; .608**]	[.220**; .246**]	undefined	[.250**; .272**]
DR550Q09C	undefined	[.329**; .353**]	[.188; .322*]	[.307**; .332**]	[.188; .322*]	[.307**; .332**]	undefined	[.329**; .353**]
DR550Q10C	undefined	[.294**; .324**]	[.171; .261**]	[.269**; .296**]	[.171; .261**]	[.269**; .296**]	undefined	[.294**; .324**]
DR550Q07C	undefined	[.365**; .397**]	[.269**; .319**]	[.350**; .393**]	[.269**; .319**]	[.350**; .393**]	undefined	[.365**; .397**]
DR055Q02C	undefined	[.428**; .444**]	[.396**; .434**]	[.390**; .410**]	[.396**; .433**]	[.389**; .409**]	undefined	[.430**; .446**]
DR055Q03C	undefined	[.488**; .509**]	[.552**; .594**]	[.471**; .497**]	[.551**; .592**]	[.470**; .496**]	undefined	[.489**; .510**]
DR055Q05C	undefined	[.514**; .528**]	[.197**; .270**]	[.433**; .452**]	[.197**; .270**]	[.429**; .449**]	undefined	[.515**; .529**]
DR111Q02BC	undefined	[.382**; .404**]	[.059; .124]	[.339**; .363**]	[.059; .124]	[.339**; .363**]	undefined	[.382**; .404**]
DR111Q06C	undefined	[.420**; .436**]	[.042; .126]	[.368**; .388**]	[.042; .126]	[.368**; .388**]	undefined	[.420**; .436**]
DR446Q06C	undefined	[.384**; .406**]	undefined	[.352**; .370**]	undefined	[.350**; .368**]	undefined	[.386**; .407**]
DR546Q03C	undefined	[.439**; .458**]	[.548**; .590**]	[.426**; .444**]	[.552**; .592**]	[.423**; .442**]	undefined	[.441**; .460**]
DR549Q05C	undefined	[.462**; .484**]	[.365**; .418**]	[.411**; .435**]	[.362**; .415**]	[.409**; .433**]	undefined	[.463**; .485**]
DR558Q04C	undefined	[.481**; .492**]	[-.013; .119]	[.456**; .468**]	[-.013; .119]	[.456**; .468**]	undefined	[.481**; .492**]
DR558Q12C	undefined	[.394**; .414**]	undefined	[.373**; .393**]	undefined	[.373**; .393**]	undefined	[.394**; .414**]
DR437Q07C	undefined	[.237**; .256**]	undefined	[.219**; .241**]	undefined	[.219**; .241**]	undefined	[.237**;. 256**]
DR561Q07C	undefined	[.178**; .199**]	undefined	[.157**; .178**]	undefined	[.156**; .178**]	undefined	[.178**; 199**]
DR562Q03C	undefined	[.281**; .299**]	undefined	[.278**; .297**]	undefined	[.278**; .297**]	undefined	[.281**; .299**]
DR562Q06C	undefined	[.351**; .373**]	[.099; .169]	[.314**; .340**]	[.099; .169]	[.314**; .340**]	undefined	[.351**; .373**]
DR564Q05C	undefined	[.470**; .490**]	[.375**; .406**]	[.413**; .440**]	[.375**; .406**]	[.413**; .440**]	undefined	[.470**; .490**]
DR565Q02C	undefined	[.338**; .354**]	undefined	[.315**; .334**]	undefined	[.315**; .334**]	undefined	[.338**; .354**]
DR565Q05C	undefined	[.391**; .417**]	undefined	[.356**; .381**]	undefined	[.356**; .381**]	undefined	[.391**; .417**]
DR404Q10C	[.320; .474**]	[.577**; .597**]	undefined	[.535**; .559**]	[.077; .153*]	[.536**; .559**]	undefined	[.577**; .597**]
DR453Q04C	undefined	[.362**; .386**]	[-.076; -.011]	[.324**; .348**]	[-.074; -.010]	[.324**; .347**]	undefined	[.363**; .387**]
DR453Q06C	undefined	[.417**; .440**]	undefined	[.338**; .364**]	undefined	[.337**; .362**]	undefined	[.417**; .440**]
DR553Q04C	[.225; .323**]	[.538**; .561**]	[.435**; .521**]	[.515**; .540**]	[.443**; .522**]	[.516**; .539**]	undefined	[.538**; .562**]
CR104Q01S	undefined	[.443**; .466**]	undefined	[.435**; .458**]	undefined	[.433**; .456**]	undefined	[.445**; .468**]

Item in MS Analysis	Total Time Threshold		Number of Actions Threshold		Union Threshold		Intersection threshold	
	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful	Effortless	Effortful
CR104Q02S	undefined	[.163**; .195**]	undefined	[.155**; .187**]	undefined	[.155**; .187**]	undefined	[.164**; .195**]
CR104Q05S	undefined	[.402**; .423**]	undefined	[.398**; .419**]	undefined	[.397**; .418**]	undefined	[.403**; .424**]
DR569Q06C	undefined	[.437**; .456**]	[.129; .189*]	[.369**; .388**]	[.129; .189*]	[.369**; .388**]	undefined	[.437**; .456**]
DR466Q02C	undefined	[.442**; .461**]	[.320**; .362**]	[.429**; .451**]	[.320**; .362**]	[.429**; .451**]	undefined	[.442**; .461**]
CR466Q06S	undefined	[.367**; .389**]	undefined	[.362**; .382**]	undefined	[.358**; .377**]	undefined	[.372**; .394**]
DR412Q08C	undefined	[.359**; .374**]	undefined	[.331**; .349**]	undefined	[.329**; .347**]	undefined	[.360**; .374**]

Note: * p < .05, ** p < .01. Green boxes represented items belonging to low-difficulty testlets, while brown boxes stand for items belonging to high-difficulty testlets. CR-CS items were highlighted in yellow. Unexpected informativeness results, where informativeness was higher for effortless than for effortful responses were presented in red.

Appendix C

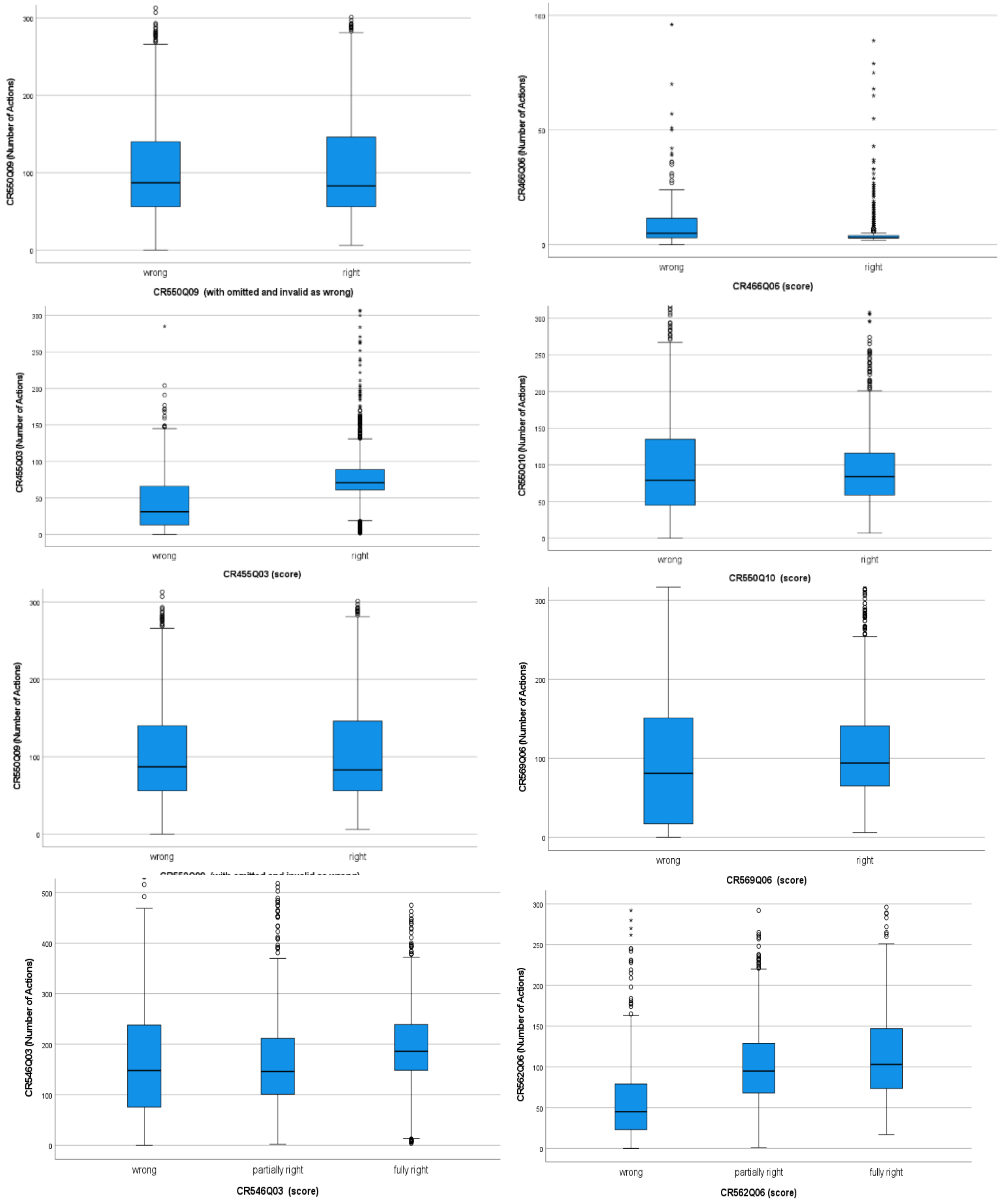
Graph C1. Boxplots of Number of Actions on Items from Stage I



Notes: NoA = number of actions. Items where NoA failed as an indicator of effort start with the letter "W". The red dot stands for the mean NoA on an item.

Graph C2. Examples of boxplots of Number of Actions by item score in Stage I

Items where NoA failed	**Items where NoA did not fail**



Note: NoA- Number of Actions