International Journal of Assessment Tools in Education
2021, Vol. 8, No. 3, 527–541
https://doi.org/10.21449/ijate.789212
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Continuity of Students' Disengaged Responding in Low-stakes Assessments: Evidence from Response Times

Hatice Cigdem Bulut [1,*]

[1]Cukurova University, Faculty of Education, Department of Educational Sciences, Adana, Turkey

ARTICLE HISTORY

Received: Sep. 01, 2020
Revised: May 20, 2021
Accepted: May 22, 2021

Keywords:
Response time,
Disengaged responding,
Insufficient effort responding,
Validity,
Low-stakes assessments.

Abstract: Several studies have been published on disengaged test respondents, and others have analyzed disengaged survey respondents separately. For many large-scale assessments, students answer questionnaire and test items in succession. This study examines the percentage of students who continuously engage in disengaged responding behaviors across sections in a low-stakes assessment. The effects on calculated scores of filtering students, based on their responding behaviors, are also analyzed. Data of this study came from the 2015 administration of PISA. For data analysis, frequencies and percentages of engaged students in the sessions were initially calculated using students' response times. To investigate the impact of filtering disengaged respondents on parameter estimation, three groups were created, namely engaged in both measures, engaged only in the test, and engaged only in the questionnaire. Next, several validity checks were performed on each group to verify the accuracy of the classifications and the impact of filtering student groups based on their responding behavior. The results indicate that students who are disengaged in tests tend to continue this behavior when responding to the questionnaire items in PISA. Moreover, the rate of continuity of disengaged responding is non-negligible as can be seen from the effect sizes. On the other hand, removing disengaged students in both measures led to higher or nearly the same performance ratings compared to the other groups. Researchers analyzing the dataset including achievement tests and survey items are recommended to review disengaged responses and filter out students who are continuously showing disengaged responding before performing further statistical analysis.

1. INTRODUCTION

Low-stakes assessments are designed to determine the achievements of students and the factors related to students' achievements. Educational stakeholders shape their strategies and make educational decisions based on the results of many low-stakes assessments, which are conducted regularly in various grade bands. Although these low-stakes assessments provide valuable information for education stakeholders, generally they are not designed to benefit students directly. Thus, students sometimes neglect to perform at their best when answering test and survey/questionnaire items in low-stakes assessments. When students do not devote their full effort, this performance is often referred to as "disengaged responding."

*CONTACT: Hatice Cigdem Bulut hcyavuz@cu.edu.tr Cukurova University, Faculty of Education, Department of Educational Sciences, Adana, Turkey

e-ISSN: 2148-7456 /

Several studies have been published on disengaged test respondents, and others have analyzed disengaged survey/questionnaire respondents; but these are usually examined separately (Maniaci & Rogge, 2014; Wise, 2017). For many large-scale assessments, such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), students answer test and questionnaire items in succession. In such cases, any low-effort responses threaten the validity of scores obtained from both the test and the questionnaire.

Rates of disengaged responses can reach up to 28% in tests (Wise *et al.*, 2019) and 50% in surveys (Buchanan & Scofield, 2018). Due to these significant numbers, researchers have used various approaches to deal with the negative effects of this threat. Notably, there has been a greater interest in detecting disengaged respondents in achievement tests than in surveys (Soland *et al.*, 2019). Regardless of which instrument of a low-stakes assessment is investigated in terms of disengaged responding, the scores obtained from measures of the students are linked and evaluated accordingly. Students who are strongly motivated in the first session of an assessment may or may not continue to be so engaged in the following sessions and vice versa. Hence, disengaged responding may be considered as a single category in low-stakes assessments. For instance, some students may devote their effort to the first session of an assessment but fail to engage when answering items in the second session of the assessment. Depending on the percentage of students who become disengaged across sessions, data quality from large-scale assessments and resulting conclusions may be considerably affected. It is critical to decide whether we should include disengaged students' responses to make conclusions when creating student profiles. Before answering such a question, it is important to know the percentage of students who are disengaged during an overall session of a low-stakes assessment. However, to the best of the author's knowledge, measures of this factor have not been presented in the literature to date. This study was designed to show the percentage of students who are disengaged during full sessions of low-stakes assessments. Also analyzed is the impact of filtering student data based on their response behaviors concerning item parameter estimation. To begin, disengaged responding behavior is defined, and an overview of prior research on surveys and psychometrics is provided.

1.1. Literature Review

Due to the utilization of technology in assessments, the issue of disengaged responding has received considerable attention. In the survey and measurement research literature, this kind of responding is variously referred to as rapid guessing, low test-taking motivation, effortless responding, disengaged responding, insufficient responding, careless responding, and inattentive responding (Huang *et al.*, 2012; Niessen *et al.*, 2016; Wise & DeMars, 2005; Wise & Kingsbury, 2016; Wise, 2017). Regardless of which term is used, the common idea is that respondents do not devote their full effort or express their real emotions/thoughts when responding on measures. "Disengaged" responding is a construct-irrelevant factor, which threatens the validity of scores (Eklöf, 2006; Wise, 2005). A great deal of research on disengaged responding has demonstrated its negative consequences for scores (Huang *et al.*, 2012; Wise & DeMars, 2005; Wise & Kingsbury, 2016). In response to these findings, new methods have been developed both in survey research and psychometric studies to handle this threat.

1.1.1. *Disengaged responding in achievement tests*

Before the advent of technological advancements in assessment, early studies used self-reports to measure students' engagement after a test event (Sundre & Moore, 2002; Sundre & Wise, 2003; Wise & DeMars, 2005). Self-reports are vulnerable to potential biases such as social desirability and response biases (Wise & Gao, 2017). However, these studies show that the

validity of scores improves when data cleaning is performed. The availability of recording time on computers led to the emergence of alternative methods for measuring disengaged responding. The new methods allow much more direct detection (Wise & Kong, 2005). This is because, ideally, respondents who intend to devote their full effort are supposed to spend some time on each item to understand it before offering an answer for it. If respondents quickly pass from one item to another, then their responding behavior is rated as lacking in effort (Wise, 2006, 2017). Using computerized testing, it is possible to monitor respondents' behaviors during a test event; with this information, the tester might be able to avoid the effects of disengaged responding on the validity of the scores. Experimental studies (Wise *et al.*, 2006; Wise *et al.*, 2019) show that providing a warning or notification to respondents during testing (about their low engagement with the items) is effective for increasing their engagement.

Researchers have also attempted to suppress disengaged responding behavior in achievement tests using different methods. Some have filtered out disengaged respondents' data (DeMars, 2007; Guo *et al.*, 2016; Wise, 2006, 2019; Wise & Kong, 2005; Wise & Ma, 2012). The results of those studies reveal that filtering increases the validity of the scores. The main concern about filtering data is deciding the cut-off scores while classifying respondents. Methods include fixed measures or visual inspections of items (DeMars, 2007; Wise, 2006), and normative measures to identify responding behavior (Wise & Kong, 2005; Wise & Ma, 2012; Wise, 2019). Regardless of the method, these studies yielded more valid results when they employed filtering. Other studies used response times while estimating parameters (Guo *et al.*, 2016; Meyer, 2010; van der Linden, 2009; Wang & Xu, 2015; Wise & DeMars, 2005). This method of estimating parameters with time data jointly also helped to achieve more precise item and person parameters. Several studies have assessed the consequences of cleaning the data of disengaged respondents by means of different approaches; the overall conclusion is that this is an efficient way to improve model fit and decrease biased parameter estimation in calibration and scoring (Wise & DeMars 2005; Wise & Kong 2005).

These results can help researchers to handle validity concerns for low-stakes assessments. Recent evidence suggests that the rate of disengaged responding can extend to 28% in large scale assessments; the exact rate depends on many factors, such as item positions (Wise *et al.*, 2009), time of the test event (Wise *et al.*, 2013), test structure (Setzer *et al.*, 2013), and the ethnicity and gender of the respondents (Goldhammer *et al.*, 2016). Thus, there appears to be ample evidence that disengaged responding is a validity threat for low-stakes assessments and that it affects conclusions that are based on the scores.

1.1.2. *Disengaged responding in surveys*

Disengaged responding can cause a validity threat for surveys as well. Recent interest in this threat in survey research has been sparked by advancements in online survey platforms (Huang *et al.*, 2012; Zhang, & Conrad, 2014). Disengaged responding behavior in surveys harms the accuracy of conclusions drawn from the scores. However, unlike disengaged responding in achievement tests, disengaged responding in surveys occurs in two ways: when respondents answer items in the survey randomly (Karabatsos, 2003; Meade & Craig, 2012), or when they answer in a non-random way (Johnson, 2005; Meade & Craig, 2012).

Curran (2016) discusses the most efficient methods to detect random and non-random disengaged responding in surveys: (1) *response time*, (2) *long-string analysis*, (3) *Mahalanobis distance*, (4) *odd-even consistency*, (5) *resampled individual reliability*, (6) *semantic antonyms/synonym*, (7) *psychometric antonyms/synonyms*, (8) *inter-item standard deviation*, (9) *polytomous Guttmann errors*, (10) *person total correlation*, (11) *bogus/infrequency items*, (12) *attention check item*, (13) *instructional manipulation checks*, and (14) *self-report scales*. Among these methods, response time analysis has been recently utilized by many researchers in this context (Curran, 2016; Huang *et al.*, 2012; Meade & Craig, 2012; Zhang & Conrad,

2014). Several studies show that the rate of disengaged responding varies from 10% to 50% in surveys (Huang *et al.*, 2012; Meade & Craig, 2012; Buchanan & Scofield, 2018; Soland *et al.*, 2019). As with achievement tests, many researchers are using different thresholds for response time data, such as a fixed two-second rule (Huang *et al.*, 2012), 300 milliseconds (Zhang & Conrad, 2014), and a normative method (Soland *et al.*, 2019). Mostly, these studies support the utilization of several methods, in addition to response times, to classify students' responding behavior (Buchanan & Scofield, 2018; Zhang & Conrad, 2014).

Similar to research on disengaged responding behavior in achievement tests, there is a large volume of published studies that discuss removing invalid data in surveys. These studies report that removing that data helps to reduce measurement errors, so that more valid results regarding means, variance, and the reliability of scales may be obtained (Huang *et al.*, 2015; Maniaci & Rogge, 2014; Woods, 2006).

1.1.3. *Study objectives*

The current state of research indicates that disengaged responding negatively affects both achievement tests and questionnaires in low-stakes assessments. Moreover, consideration of continuity in disengaged responding by students across sections of assessments is lacking in all the aforementioned studies. This indicates a need for investigation across all large-scale assessment events because some researchers use students' responses for all measures. To address this need, this study examines the percentage of students who continuously engage in disengaged responding behaviors across sections in a low-stakes assessment. The effects on calculated scores of filtering students, based on their responding behaviors, are also analyzed. The goal is to assess whether the degree of disengaged responding continuity is significant or negligible and to document the effects on scores obtained from achievement tests and questionnaires.

2. METHOD

The data of this study came from the 2015 administration of PISA (Organisation for Economic Cooperation and Development [OECD], 2017). PISA is a large-scale, international assessment that measures 15-year-old students' achievement in reading, mathematics, and science literacy. After completing these cognitive assessments, students also take a questionnaire that focuses on students' attitudes toward their homes, schools, and learning experiences. Although around 500,000 students took PISA 2015, only 69,426 of the students were included in the analysis based on some selection criteria. These criteria will be explained in the method section. Table 1 shows the selected students' frequency and percentage across countries.

Table 1. *Students' frequency and percent across countries.*

Country	N	%		N	%
United Arab Emirates	2215	3.2	Lithuania	998	1.4
Australia	2261	3.3	Luxembourg	819	1.2
Austria	1109	1.6	Latvia	770	1.1
Belgium	1471	2.1	Macao	681	1.0
Bulgaria	942	1.4	Mexico	1158	1.7
Brazil	3582	5.2	Montenegro	884	1.3
Canada	3100	4.5	Malaysia	1399	2.0
Switzerland	682	1.0	Netherlands	796	1.1
Chile	1094	1.6	Norway	836	1.2
COL	1830	2.6	New Zealand	742	1.1
Colombia	956	1.4	Peru	1079	1.6
Czech Republic	1039	1.5	Poland	537	.8
Germany	1008	1.5	Portugal	1120	1.6

Table 1. *Continues.*

Denmark	1093	1.6	Qatar	1377	2.0
Dominican Republic	558	.8	B-S-J-G (China)	1524	2.2
Spain	1049	1.5	Spain (Regions)	5046	7.3
Estonia	863	1.2	Massachusetts	256	.4
Finland	910	1.3	North Carolina	270	.4
France	930	1.3	Russian Federation	936	1.3
United Kingdom	2223	3.2	Singapore	944	1.4
Greece	859	1.2	Slovak Republic	956	1.4
Hong Kong	842	1.2	Slovenia	967	1.4
Croatia	924	1.3	Sweden	857	1.2
Hungary	894	1.3	Chinese Taipei	1188	1.7
Ireland	681	1.0	Thailand	1280	1.8
Iceland	498	.7	Tunisia	845	1.2
Israel	1016	1.5	Turkey	928	1.3
Italy	1847	2.7	Uruguay	965	1.4
Japan	1043	1.5	United States	872	1.3
Korea	877	1.3	Total	69426	100.0

2.1. Measures

Science literacy tests: In PISA 2015, the major domain was science literacy, in which there were 67 forms (i.e., booklets), each containing seven science clusters and items related to other domains. Only 21 of these forms prioritized science clusters over the other domains. In this study, five forms (33, 44, 45, 91, 93) were randomly selected from the 21 forms to avoid position and other types of contextual effects among the forms. Data analysis was undertaken in each of the clusters in every five forms.

Student questionnaire: Students' responses in the cognitive assessments were combined with their questionnaire responses. From the questionnaire, a science-related module including eight scales (see Table 2) with 51 items was selected.

Table 2. *The science-related module in PISA 2015.*

Scales	Number of Items	Description
ENVAWARE	7	Environmental awareness
ENVOPT	7	Environmental optimism
ENVOPT	5	Enjoyment of science
INTBRSCI	5	Interest in broad science topics
INSTSCIE	4	Instrumental motivation
SCIEEFF	8	Science self-efficacy
EPIST	6	Epistemological beliefs
SCIEACT	9	Science activities

2.2. Procedure

To classify the students as either disengaged or engaged respondents, their response times from the PISA 2015 database were used. Disengaged students were determined based on the normative threshold (NT10) method (Wise & Ma, 2012). NT10 method is one of the most effective methods for determining disengaged respondents in achievement tests (Wise, 2020). For each item, the time threshold is calculated "as a percentage of the elapsed time between when the item is displayed and the mean of the response time distribution for the item, up to a maximum threshold value of 10 seconds" (Wise & Ma, 2012; p. 9). Setzer *et al.* (2013) suggested that spending longer than 10 seconds on an item should not be defined as disengaged responding. By utilizing NT10 method, we classified the students' engagement for each item

(i.e., 1 = engaged in answering the item; 0 = disengaged in answering the item). Then, we calculated students' total engagement scores (ESs) by summing all the binary classifications generated from the items. Finally, students were classified as "disengaged in the test" if they showed disengaged responding on more than 90% of the items (i.e., .90 threshold) in the test (Wise & Kong, 2005). For example, assume a respondent answered 20 items and this respondent showed disengaged behavior in 10 items based on the NT10 method. Then, this respondent's engagement score would be 10, meaning that the respondent would be classified as "engaged in the test" based on .90 threshold as the score was less than 18.

Disengaged students in the questionnaire were determined using the two-second method proposed by Huang *et al.* (2012). In PISA 2015, there were eight scales in the science-related module presented on a single page. As a result, students' response times included the time spent per scale, not the time per item. Hence, we followed Soland *et al.*'s (2019) approach by calculating the response time for each item as the time spent on the scale divided by the number of items in the scale. Then, we classified the students' engagement in the items separately by using the two-second threshold. Then, we calculated students' total engagement scores (ESs) for the questionnaire and used the .90 threshold again. In this way, students were classified as disengaged in the questionnaire if they showed disengaged responding behavior to more than 90% of the items in the questionnaire. To investigate the impact of filtering disengaged respondents on parameter estimation, three groups were created, namely engaged in both measures (the test and the questionnaire), engaged only in the test, and engaged only in the questionnaire. The group of engaged in both measures will be mentioned as 2, engaged in the test as 3, engaged in the questionnaire as 4, and full sample as 1 in the remainder of the paper.

For data analysis, frequencies, and percentages of engaged students in the sessions were initially calculated. Next, several validity checks were performed on each group to verify the accuracy of the classifications. The idea behind this step was to learn whether or not removing disengaged students made a difference in the parameter estimation, and which group had the highest quality data across the three engagement classification groups. First, all parameter estimations were conducted separately on all groups using the same item response theory modeling approach as PISA utilized, namely the two-parameter-logistic model (2PLM; Birnbaum, 1968) for dichotomously scored responses and Generalized Partial Credit Model (Muraki, 1992) for polytomously scored responses. Besides, classical test theory analysis was carried out with test items. Second, reliability coefficients, effect sizes, and correlations between scores were calculated. Third, fit indices related to factor structures of scales were compared. Note that, the second group was used for all comparisons while reporting results, however, the only third group was taken into consideration when comparisons were done for tests and the fourth group when comparisons were done for questionnaires. All analyses were conducted in R (R Core Team, 2019) using ShinyItemAnalysis (Martinkova *et al.*, 2017), ltm (Rizopoulos, 2006), and lavaan (Rosseel, 2011) packages.

3. FINDINGS

Only the results from the first cluster of Form 33 were reported as similar results were found for the other clusters. The results obtained from other clusters are available from the author upon request. The results showed that although the proportion of disengaged respondents changed across the clusters, a great number of disengaged students in the test also continued their disengaged responding behavior in the questionnaire session. Among the disengaged students who took the test, approximately 38-43% followed the same type of disengagement when responding to the questionnaire items. Specifically, when we look at the proportion in the first cluster (see Table 3), only 49% of students appear to be engaged in both measures. Most students (80%) were engaged in the test session while only 60% of students were engaged in the questionnaire session. This finding suggests that some disengaged students in the test

became engaged respondents in the questionnaire session. Furthermore, Appendix 1 shows student percentages based on responding behaviors across countries in PISA 2015. When we look at the countries in Appendix 1, especially the most successful East Asian countries, they have relatively smaller percentages of disengaged students in the test, but mostly higher percentages of disengaged students in the questionnaire.

Table 3. *Descriptive Statistics of Ability Estimates and Engagement Scores (ESs).*

Group	N	Ability estimates		ESs based on the test		ESs based on the questionnaire	
		$\overline{X}$	*SD*	$\overline{X}$	*SD*	$\overline{X}$	*SD*
1	2182	-.01	.96	.93	.11	.85	.25
2	1075	.27	.85	.98	.03	1	0
3	1745	.11	.93	.97	.03	.88	.22
4	1278	.19	.88	.94	.10	1	0

Note: 1 = Full sample; 2 = Engaged in both measures; 3= Engaged in the test; 4= Engaged in the questionnaire.

Table 3 shows that the difference in the mean ability estimates between the groups were not negligible, especially between the full sample (group 1) and the group of students engaged in both measures (group 2); Cohen's d ranged from 0.12 to 0.29 [d1-2=−0.29, d1-3=−0.12, d2-3=.18]. The ability estimates were lowest in the group of students engaged in both measures. This is because easy items tended to get even easier after filtering out students based on their response behaviors. Furthermore, as seen in Figure 1, the test information appears to be much greater for the group of students engaged in both measures in the lower ability range. This is to be expected, given the removal of low-accuracy responses by disengaged students. The same results apply to the scores obtained from the questionnaires. For example, Figure 2 shows the test information functions of EPIST. Since the thresholds tended to be lower after filtering out the related students, the information appears to be generally less in the lower theta range. However, Figures 1 and 2 show that the test information appears to be much greater for the full sample between the -2 and 2 theta range. Therefore, it is possible that item parameters might be overestimated, and the measurement model inflated test information in this range due to the presence of disengaged responses.

Figure 1. *Test information functions of the first cluster of the 33rd form*

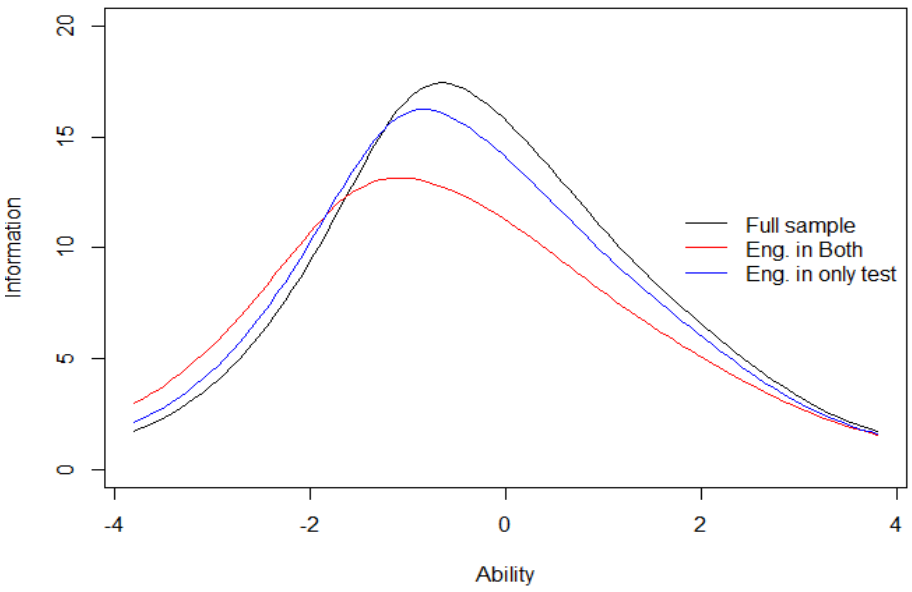
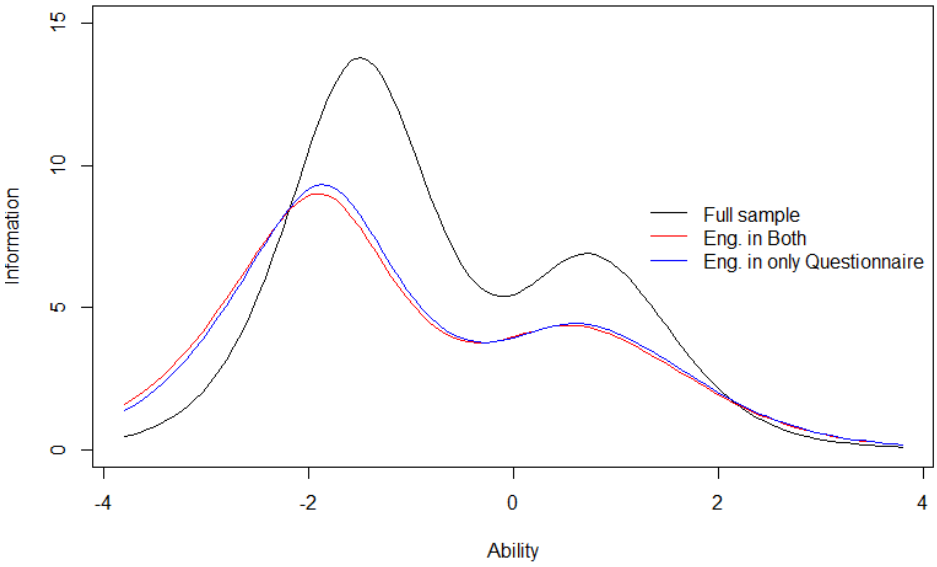


Figure 2. *Test information functions of EPIST.*



The alpha reliability coefficients for the test (.88, .86, .88 for groups 1, 2, and 3 respectively) and questionnaires slightly changed for the groups (see Table 4), but there were generally no big differences except that the reliability coefficient calculated from the group of students engaged in both measures was significantly lower in value than the others.

Table 4. *The reliability coefficients of the first cluster of 33rd form.*

	Reliabilities			
	1	2	4	Significance
ENVAWARE	.86	.84	.85	1-2, 1-4
ENVOPT	.87	.84	.85	1-2, 1-4
JOYSCIE	.94	.93	.93	1-2, 1-4
INTBRSCI	.81	.73	.74	1-2, 1-4
INSTSCIE	.92	.92	.92	-
SCIEEFF	.89	.86	.87	1-2, 1-4
EPIST	.88	.83	.83	1-2, 1-4
SCIEACT	.93	.90	.90	1-2, 1-4

Note: 1 = Full sample; 2 = Engaged in both measures; 4= Engaged in the questionnaire

Table 5 shows the correlations of domains and ES which can be interpreted regarding the validity evidence for calculated ESs. The correlations of ability estimates and ES are .24 (p<.01), .01 (p>.05), .08 (p<.05) in groups 1, 2, and 3 respectively. There was a significant low correlation between the ESs obtained from the tests and scales (.23, p <.01). In the full sample, calculated ES in the questionnaires was significantly correlated with the students' thetas estimated from the questionnaires. However, these correlations were not significant within each group. As expected, those correlations were significantly lower in the opposite direction in the group of students engaged in both measures. This suggests that both ESs were effective in removing disengaged students who caused a negative significant correlation between the overall ability estimates and thetas.

Table 5. *The correlation coefficients*.

Domains	ES in the questionnaire	Correlations with ability estimates			Cohen's q (1-2), (1-4), (2-4)
		1	2	4	
ENVAWARE	-.17**	-.05*	.01	.03	-.06, -.08, -.02
ENVOPT	-.25**	-.09*	.01	.04	-.09, -.13, -.04
JOYSCIE	-.22**	-.11**	-.02	-.05	-.09, -.06, -.02
INTBRSCI	-.23**	-.10*	.03	-.03	-.14, -.07, .07
INSTSCIE	-.14**	-.03	.05	.07*	-.07, -.10, -.03
SCIEEFF	-.33**	-.13**	.01	.06*	-.14, -.07, .07
EPIST	-.25**	-.14*	.06*	.01	-.08, -.15, -.06
SCIEACT	-.31**	-.16*	.02	.02	-.19, -.18, -.01

Note: 1 = Full sample; 2 = Engaged in both measures; 4= Engaged in the questionnaire

Table 6 shows the model fit indices obtained from the confirmatory factor analysis conducted separately for all the domains in the questionnaire. The confirmatory factor analysis of the fit indices of all the domains shows that a 1-factor model fits the data well. Although there were no big differences between the indices, the indices obtained from the fourth group are slightly better, suggesting that the method based on calculating ES provides good performance for the underlying construct.

Table 6. *The model fit indices of the first cluster of 33rd form.*

	1			2			4		
Domains	RMSEA	CFI	TLI	RMSEA	CFI	TLI	RMSEA	CFI	TLI
ENVAWARE	.11	.94	.91	.09	.96	.94	.10	.94	.91
ENVOPT	.10	.95	.93	.11	.93	.90	.10	.94	.91
JOYSCIE	.05	.99	.99	.03	.1	.1	.04	.1	.1
INTBRSCI	.17	.91	.83	.17	.88	.75	.17	.88	.77
INSTSCIE	.16	.98	.95	.14	.99	.96	.13	.99	.96
SCIEEFF	.07	.97	.96	.05	.98	.97	.05	.98	.97
EPIST	.16	.92	.87	.17	.86	.76	.17	.86	.77
SCIEACT	.17	.87	.83	.18	.82	.76	.18	.83	.77

Note: 1 = Full sample; 2 = Engaged in both measures; 4= Engaged in the questionnaire

Overall, the second group performed better or nearly the same as the third and fourth groups in terms of obtained results. This suggests that even if conservative methods are selected for identifying disengaged respondents, as in this study, some students still may not be assigned to the correct group. That is why the third and fourth groups did not appear to perform much better than the second group. The decision not to filter disengaged students may significantly affect the estimation of the scores in both measures.

4. DISCUSSION and CONCLUSION

Because disengaged responding behavior in tests and questionnaires causes a validity threat, this study was designed to examine the percentage of students who continuously demonstrate disengaged responding behaviors across the sessions of a low-stakes assessment. This paper contributes to research in the field of both questionnaires and tests and applies to disengaged responding generally in large-scale assessments. Another question asked is whether the effects of continuously disengaged behavior are significant or negligible in scores obtained from achievement tests and scales.

The results indicate that students who are disengaged in tests tend to continue this behavior when responding to the questionnaire items in PISA. Moreover, the rate of continuity of disengaged responding is non-negligible as can be seen from the effect sizes. This makes it critical to use large-scale assessments' data for educational decisions and policies without first screening for disengaged responding. Recent studies that focused on data from achievement tests reveal that disengaged responding behaviors affect the country rank orderings of international assessments (Eklöf *et al.*, 2014; Zamarro *et al.*, 2019). Hence, when we consider both cognitive and non-cognitive data sets together, disengaged responding may cause validity issues.

The percentage of students who were engaged in the cognitive part of the assessment in the current study was higher than the percentage of students who were engaged in the non-cognitive part of the assessment. This can be explained using the expectancy-value theory (see Wigfield & Eccles, 2000). According to expectancy-value theory, students' engagement in measures depends on their perceived value for the measure or expectancy for the test. For example, some students might assign more importance or value to the cognitive session (e.g., achievement tests) in the large-scale assessment. Ultimately, this influences their engagement across the sessions. Wise *et al.* (2019) reported a similar situation concerning the initial and final parts of a test. Inconsistencies in engaged responding across sessions of PISA 2015 are more obvious for some countries. Furthermore, several studies support that respondents' cultural backgrounds affect the occurrence of disengaged responses in questionnaires (e.g., Palaniappan & Kum, 2019). Respondents coming from collectivistic cultures tend to show more disengaged responding in questionnaires.

The results of this study also show that the information obtained from both measures appeared to be generally less in the lower theta range within the full sample. Removing disengaged students in both measures led to higher or nearly the same performance ratings compared to the other groups. These results are similar to those of several studies in the literature (Maniaci & Rogge, 2014; Meade & Craig, 2012; Wise & DeMars, 2006; Wise & Kingsbury, 2016), which all suggest that the removal of disengaged respondents' data provides more valid results. Alternatively, methods such as sending warning notifications (see Wise *et al.*, 2006; Wise *et al.*, 2019) to disengaged respondents before upcoming sessions of the assessment can be adopted to promote engagement in those upcoming measures. Further results of this study suggest that removing disengaged students can change the negative significant correlation to a non-significant correlation between the overall ability estimates and thetas. These results highlight an important area of further research.

Although the current study has yielded important results, the examination was constrained by several limitations. The main limitation in this study involves the use of a limited number of (randomly selected) science achievement tests and only the science module in the student questionnaire in PISA 2015. Another limitation involves the methods used to classify the students into engagement groups. As reported by Curran (2016), incurring a Type I error when using conservative methods is inevitable. A further limitation of the study relates to the use of response times for each scale, rather than for each item, during the process of classifying the students in the questionnaire session. This limitation can cause several problems, as Soland *et al.* (2019) indicated, and might ultimately limit the generalizability of the results. Therefore, more research should be conducted, using different low-stakes assessment data that include response times for each item in the questionnaire, and different methods and measures for classifying the students.

In conclusion, the present study unveils that disengaged respondents become a validity threat not only for the inferences of achievement scores but also for the information gathered from student questionnaires. Therefore, researchers analyzing the PISA dataset are recommended to

review disengaged responding behaviors. More importantly, researchers intended to use students' both cognitive and non-cognitive data sets are strongly recommended to filter out students who are continuously showing disengaged responding before performing further statistical analysis.

Declaration of Conflicting Interests and Ethics

The author declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

ORCID

Hatice Cigdem Bulut https://orcid.org/0000-0003-2585-3686

5. REFERENCES

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord and M.R. Novick (eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low-quality data and its implication for psychological research. *Behavior Research Methods, 2018*, (50), 2586–2596. https://doi.org/10.3758/s13428-018-1035-6

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

DeMars, C. E. (2007). Changes in rapid-guessing behavior series of assessments. *Educational Assessment, 12*(1), 23–45. https://doi.org/10.1080/10627190709336946

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual review of psychology*, *53*(1), 109-132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement, 66,* 643–656. https://doi.org/10.1177/0013164405278574

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. https://doi.org/10.1080/08957347.2013.853070

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

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. https://doi.org/10.1080/08957347.2016.1171766

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J.L., Bowling, N.A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*, 299–311. https://doi.org/10.1007/s10869-014-9357-6

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99–114. https://doi.org/10.1007/s10869-011-9231-8

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*, 103-129. https://doi.org/10.1016/j.jrp.2004.09.009

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-
fit statistics. *Applied Measurement in Education, 16*, 277-298. https://doi.org/10.1207/S15324818AME1604

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

Martinkova, P., Drabinova, A., Leder, O., & Houdek, J. (2017). ShinyItemAnalysis: Test ´and item analysis via shiny [Computer software manual]. *https://CRAN.R-project.org/package=ShinyItemAnalysis*.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455. https://doi.org/10.1037/a0028085

Meyer, P. J. (2010). A mixture Rasch model with response time components. *Applied Psychological Measurement, 34,* 521–538. https://doi.org/10.1177/0146621609355451

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use?. *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving.* OECD Publishing. https://doi.org/10.1787/9789264281820-en

Palaniappan, K., & Kum, I. Y. S. (2019). Underlying Causes behind Research Study Participants' Careless and Biased Responses in the Field of Sciences. *Current Psychology*, *38*(6), 1737–1747. https://doi.org/10.1007/s12144-017-9733-2

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL, *https://www.R-project.org/.*

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25.

Rosseel, Y. (2011). *lavaan: An R package for structural equation modeling and more* (Version 0.4-10 beta).

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

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, *14*(1), 8–9.

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.*

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*(3), 247-272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

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

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretations, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L. (2019). An Information-Based Approach to Identifying Rapid-Guessing Thresholds. *Applied Measurement in Education*, *32*(4), 325-336, https://doi.org/10.1080/08957347.2019.1660350

Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-18. https://doi.org/10.1207/s15326977ea1001_1

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

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343-354. https://doi.org/10.1080/08957347.2017.1353992

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*, 86–105. https://doi.org/10.1111/jedm.2016.53.issue-1.

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

Wise, S. L., Soland, J., & Bo, Y. (2019). The (Non) Impact of Differential Test Taker Engagement on Aggregated Scores. *International Journal of Testing*, 1–21. https://doi.org/10.1080/15305058.2019.1605999

Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 189–194. https://doi.org/10.1007/s10862-005-9004-7

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining ınternational differences in achievement and student effort. *Journal of Human Capital*, *13*(4), 519–552. https://doi.org/10.1086/705799

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *In Survey Research Methods, 8*, 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

6. APPENDIX

Appendix 1. *Student percentages based on responding behaviors across countries in PISA 2015.*

Country	Disengaged respondents' percent in test	Disengaged respondents' percent in the questionnaire	Disengaged respondents' percent in both measures	Performance means estimated in this study	Performance means estimated in PISA 2015
Singapore	14.09	56.91	8.47	0.60	556
Estonia	16.51	76.52	13.02	0.51	523
Hong Kong	9.97	29.91	2.60	0.51	534
Chinese Taipei	17.59	53.92	10.61	0.50	532
Japan	16.30	43.37	6.92	0.49	538
Massachusettes	11.33	41.11	5.53	0.45	-
B-S-J-G (China)	24.28	53.75	10.84	0.44	518
Finland	9.56	32.69	3.97	0.41	531
Macao	29.66	51.03	13.82	0.40	529
Germany	16.17	45.96	7.52	0.33	513
New Zealand	13.10	36.13	5.14	0.33	509
Canada	11.48	51.47	6.10	0.31	528
Netherlands	9.80	33.29	6.56	0.31	509
Belgium	13.60	31.63	4.04	0.28	502
Korea	13.11	80.83	12.06	0.28	516
Ireland	12.19	21.94	3.13	0.26	503
United Kingdom	9.36	38.59	4.92	0.24	509
Spain (Regions)	18.23	23.48	3.66	0.22	-
Switzerland	16.57	45.12	10.80	0.21	493
Spain	14.90	24.25	4.29	0.21	501
Norway	15.07	39.53	6.77	0.21	498
Poland	20.02	20.95	3.49	0.21	493
Austria	16.13	36.40	7.33	0.20	495
France	8.93	34.40	3.67	0.20	495
Czech Republic	10.61	31.66	2.96	0.18	481
Italy	10.49	30.71	3.50	0.18	493
Australia	11.63	44.65	4.92	0.17	510
Slovenia	11.07	33.86	4.81	0.16	513
Sweden	22.87	38.86	8.53	0.14	493
Russian Federation	20.51	32.56	6.64	0.12	487
Portugal	18.84	38.10	5.43	0.10	501
Denmark	16.74	41.38	6.64	0.09	496
United States	15.37	30.22	6.16	0.09	502
Hungary	9.62	35.21	2.94	0.08	477
Luxembourg	24.54	34.89	9.07	0.06	483
Iceland	13.33	49.07	4.83	0.05	-
Latvia	19.08	42.74	8.38	0.05	473
North Carolina	10.52	29.27	3.31	0.05	490
Israel	34.55	45.58	15.16	0.00	467
Croatia	10.28	34.53	2.35	-0.01	475
Lithuania	9.02	37.92	3.80	-0.09	475
Slovak Republic	13.08	43.19	5.57	-0.10	461
Greece	22.12	29.44	7.83	-0.11	455
Chile	23.40	28.71	7.03	-0.14	447
Malaysia	17.87	23.53	4.07	-0.17	-
Bulgaria	24.52	47.21	11.72	-0.27	446
Uruguay	28.91	40.86	10.61	-0.36	435
United Arab Emirates	21.81	45.35	12.60	-0.39	437

Appendix 1. *Continues.*

Country	Disengaged respondents' percent in test	Disengaged respondents' percent in the questionnaire	Disengaged respondents' percent in both measures	Performance means estimated in this study	Performance means estimated in PISA 2015
Thailand	17.66	53.21	7.46	-0.40	421
Turkey	14.66	53.10	7.94	-0.48	425
Montenegro	22.17	56.53	13.64	-0.52	411
COL	27.10	22.39	5.44	-0.53	416
Colombia	34.72	21.99	5.38	-0.53	416
Mexico	26.99	20.87	6.28	-0.53	420
Qatar	38.42	57.61	27.75	-0.64	418
Brazil	58.82	58.16	35.97	-0.68	401
Peru	51.99	13.72	6.24	-0.72	397
Tunisia	34.91	43.22	14.82	-0.85	386
Dominican Republic	54.12	32.31	18.00	-1.11	332