International Electronic Journal of Elementary Education
March 2025, Volume 17, Issue 2

Exploring Test-Taking Disengagement in the Context of PISA 2022: Evidence from Process Data

Başak Erdem Kara[a,*]

Received : 13 December 2024
Revised : 10 January 2025
Accepted : 2 March 2025
DOI : 10.26822/iejee.2025.380

[a,*] **Corresponding Author:** Başak Erdem Kara, Anadolu University, Education Faculty, Department of Educational Sciences, Eskisehir, Türkiye.
E-mail: basakerdem@anadolu.edu.tr
ORCID: https://orcid.org/0000-0003-3066-2892

Copyright ©
www.iejee.com
ISSN: 1307-9298

Abstract

Achievement tests are commonly used in education to evaluate students' academic performance and proficiency in specific subject areas. However, there is a major problem that threatens the validity of achievement test scores which is test-taking disengagement. Respondents provide answers that are inconsistent with their true ability level and can introduce construct irrelevant variance that threatens the validity of scores. This study examines test-taking disengagement in the context of PISA 2022 using process data to identify patterns of behavior that influence student performance. Three key indicators; response time, number of actions and self-reported effort, were used to examine engagement levels. Employing Latent Profile Analysis (LPA), distinct profiles of test-takers were identified, ranging from highly engaged to disengaged groups. Results indicate that disengagement, characterized by low self-reported effort, minimal interactions, and rapid responses, is associated with lower test performance, threatening the validity of scores. These findings highlight the significance of accounting for disengagement when interpreting the results of large-scale assessments. The implications were discussed in relation to the existing literature and recommendations for future research were provided to address identified gaps and extend the study's contributions.

Keywords:

Test-Taking Disengagement, Response Time, Number Of Action, Self-Reported Effort

Introduction

Achievement tests are a widely used tool in education to assess student performance, with the primary intention of measuring what a student knows and can do when they are fully engaged and demonstrating their maximum performance while responding to items (Cronbach, 1960; Messick 1989). Ideally, students are assumed to exert maximum effort on test items, ensuring that test scores accurately reflect the construct being measured. In practice, however, this ideal scenario is not always achieved, as some students may not put forth the effort necessary to thoroughly process an item and provide responses that are consistent with their true ability (Wise, 2017; Wise & Kingsburry, 2016, 2022).

It is recognized that a valid achievement test score requires an engaged test-taker demonstrating what they know and can do (Cronbach, 1960; Messick, 1984). However, test-takers may feel unmotivated to exert effort, particularly in low-stakes tests where they often believe their performance has no personal consequences. Consequently, when test-takers respond with inadequate effort, their test scores are likely to reflect a lower level of ability than they actually possess. This behavior, known as test-taking disengagement, introduces non-negligible, construct-irrelevant variance that poses a potential threat to score validity (Eklöf, 2010; Goldhammer et al., 2016; Kong et al., 2007; Wise, 2017). In general, test-taking disengagement is defined as providing responses that are inconsistent with those expected from engaged test-takers. It includes situations in which the individual provides a response without reference to his or her knowledge, skills, or abilities (Soland et al., 2019).

Test-Taking Disengagement and PISA

Programme for International Student Assessment (PISA) is one of the International Large Scale Assesments (ILSA) regularly administering tests and questionnaires. Its purpose is to evaluate the readiness of 15-year-old students to tackle the challenges of today's information-driven society and to draw conclusions about the effectiveness of a country's education system. The program focuses on students' ability to use their knowledge and skills to meet with real-life challenges, rather than on their mastery of a particular area of the school curriculum (OECD, 2024). In the PISA, students take a test designed to measure their skills, typically in mathematics, reading, and science. Participation is voluntary and anonymous, with minimal to no direct consequences for the students. As a result, the test is considered a low-stakes assessment at the individual respondent level (Baumert & Demmrich, 2001; Finn, 2015; Pools & Monseur, 2021).

As in other assessment situations, PISA also assumes that the scores obtained by test-takers reflect only differences in the characteristics measured, but test-takers may not give their best effort that would be desired (Buchholz et al., 2022). Thus, the validity of the inferences based on the PISA assessment needs to be controlled and demonstrated. As we discussed before, in low-stakes testing contexts, such as PISA, there are often no personal consequences for test-takers, i.e. any form of incentive, influence on academic record or feedback. Research has consistently shown that low-stakes assessments tend to produce lower levels of engagement. Disengagement is the main construct-irrelevant factor that jeopardizes the validity of low-stakes test scores, and test administrators are aware of and concerned about its potential impact (Finn, 2015; Wise, 2020; Wolf and Smith, 1995). Because PISA is also a low-stakes assessment, it is also open to the validity threat posed by disengagement.

Indicators of Test-Taking Disengagement

There are several measures to examine students (dis) engagement that are typically categorized as self-reported effort (SRE) data and test-takers response behavior. Response behaviors include behavioral analysis demonstrated by students while completing an assessment. In the context of ILSAs, test-based behavioral measures can be derived from either response patterns or process data collected during computer-based assessments (CBAs) (Buchholz et al., 2022). In the context of this study, log data measures and the SRE are the main focus and are discussed in detail below.

Process Data (Log Data). The use of CBAs has introduced alternative approaches leveraging log data. These assessments enable the collection of data that capture not only the answers provided by test-takers but also their observable behaviors during the test. This type of data, known as process or log data, includes metrics such as the time spent on each question, the frequency and nature of interactions, and the intervals between actions. Such data offer researcher valuable insights into both the test-takers' final responses and the cognitive processes they employed to reach those answers (Ramalingam, 2017). Recently, log file data have been utilized to identify instances of disengagement during test-taking (Gobert et al., 2015). The most widely used approach relies on the amount of time individuals spend responding to an item. These methods are based on the assumption that participants exhibiting low effort complete tasks more quickly and spend less time on them compared to those who are more motivated (Wise & Kong, 2005). Response time data is regarded as a less biased approach because it reflects actual behavior rather than self-reported evaluations and does not require any extra effort from the respondents. This approach allows for more accurate and continuous tracking of changes in engagement because response data is collected for each individual item rather than at specific points in time (Wise & Kong, 2005). In addition to response time, number of actions data from the log file could be used as complementary measure to examine disengagement. Number of actions reflects examinees' interactions with a specific item, serving as an indicator of their behavioral engagement with the task. Sahin and Colvin (2020) stated that a lower number of clicks is an indicator of lower levels of motivation and thus higher levels of disengagement.

Self-Reported Effort. One of the most widely used methods to assess engagement is to ask test-takers to directly self-report the amount of the effort they put into taking the test. For example, PISA employs an "effort thermometer" (Butler & Adams, 2007), in which test-takers rate their engagement on a scale from 1 (lowest) to 10 (highest). Despite their ease of use, self-report measures have notable limitations. First, the accuracy of the data may be questionable because self-report measures are susceptible to response bias. Second, the interpretation of self-report scores can be challenging, as these scores may not provide clear insight into the specific nature or extent of disengagement (Wise, 2020).

Test-Taking Disengagement and Test Performance

As discussed before, disengaged responding introduces a construct-irrelevant variance into the measurement process and its presence threatens the interpretation of test scores which can lead to some poor decisions (Wise & Kingsburry, 2022). Previous research has consistently highlighted a relationship between test-taking effort and achievement. In general, higher levels of engagement are associated with higher levels of test performance (Kuhfeld & Soland, 2020). Motivated students tend to perform better on tests than unmotivated ones (Wise & DeMars, 2005; Wise & Kong, 2005; Finn, 2015).

In contrast, according to Gignac et al. (2019) it is not necessary to exert maximum effort or to have a very high level of test-taking motivation to obtain valid test scores but rather reaching a sufficient level of effort. While effort generally improves performance, there are exceptions such as cases where students in low-effort clusters achieved high scores, i.e., test-taking effort had a weak negative correlation with test performance (Lundgren & Eklöf, 2020). In the context of low-stakes assessments, both motivational and cognitive factors are found to explain test performance, making the interpretation of results less straightforward. Eklöf et al. (2014) show that controlling for effort changes the ranking of countries in the TIMSS results. Zamarro et al. (2019) found that effort accounted for 32-38% of the variation in PISA 2009 scores. Similarly, Akyol et al. (2021) estimated that a country could improve its PISA ranking by up to 15 places if all students took the test seriously. These findings underscore that achievement test results are shaped by both student ability and motivation.

Present Study

Test-taking disengagement and its relationship to test performance and psychometric properties has become an important concern and significant area of interest for researchers and practitioners due to the validity challenges it poses (Wise & DeMars, 2005; Wise, 2016). Previous studies have proposed various process data-based approaches to detect unmotivated responses; however, these methods frequently produce differing outcomes when applied to the same sample (Goldhammer et al., 2016). While test-taking effort is generally positively correlated with performance, this relationship is less clear in some studies (Gignac et al., 2019; Lundgren & Eklof, 2020). Therefore, the present research aimed to examine students' test-taking effort using various indicators, specifically self-reported effort and log data, including response time and the number of actions, within the context of the PISA 2022 dataset in the Turkish sample. The Turkish sample was selected because Turkey was one of the countries that included a measure of self-reported effort and process data records in the PISA 2022 assessment, and it also ranked among the countries with the highest test effort in the PISA 2018 cycle. Turkish students had high levels of engagement based on behavioral indicators (low non-response and rapid guessing rates) and high level of self-reported effort (Buchholz, 2022). This makes Turkey a particularly relevant context for the study. In this study, Latent Profile Analysis (LPA) is used to identify the different groups that define students' test effort. This analysis will not only provide new insights into understanding student effort levels, but will also provide a deeper understanding for accurately assessing test performance. Answers were sought to the following research questions:

> *RQ1. What percentage of the sample show disengagement?*
>
> *RQ2. How does effort, as reflected in process data (response time and number of actions), self-reported effort, and test performance, relate to one another?*
>
> *RQ3. What profiles can students be classified into based on response time, number of actions, and self-reported effort data?*

In addition, some factors such as item type, demographic characteristics of the sample, item position etc. may influence the test-taking profiles and gender was taken into consideration to examine the results of LPA in depth.

Self-Reported Effort. On the last page of the PISA assessment booklet or screen, there is a section called the PISA Effort Thermometer and students are asked to imagine a situation that they consider important and for which they would do their best and exert as much effort as possible. Students are asked to rate their self-reported effort (SRE) based on these statements using a scale of 1 to 10, with 10 being maximum effort. They are presented with the following question and asked to rate their effort (OECD, 2016).

"How much effort did you put in doing this test [PISA]?"

Here, a score of 10 indicates that students believe they put as much effort into the PISA test as they would in a real-life scenario of great importance to them (OECD, 2016).

Mathematics Performance. As mentioned above, mathematics is the main domain of the PISA 2022 assessment, so we focused on mathematical items and performance scores. The computer-based PISA 2022 assessment spanned two hours, divided into two one-hour sessions with a 5-minute break in between (OECD, 2024). Students were tasked with completing two 30-minute clusters of items in each session, amounting to four clusters in total. While two clusters were dedicated to the major domain, the remaining clusters assessed one or two of the minor domains. The PISA 2022 item pool included 99 items and a total of 234 mathematics questions (OECD, 2024).

Data Analysis

To obtain the response time (RT) and number of actions (NA) scores, we calculated the average RT and NA values for each individual. Missing values were excluded by listwise deletion and this cleaning process resulted in a sample of 6560 out of 7250 students. In addition to the raw scores of RT and NA scores, we also calculated an effort index to examine the frequency of disengaged responders on the sample. The response time effort (RTE) index was introduced by Wise and Kong (2005) and calculated as follows;

$$SB_{ij} = \begin{cases} 1, & if\ RT_{ij} \geq T_i \\ 0, & if\ RT_{ij} < T_i \end{cases} \qquad RTE = \frac{\sum SB_{ij}}{k}$$

In this formula, SBij refers to the solution behavior for the item i and person j and is calculated based on a threshold value (Ti). k refers to the number of items. In this point, RTE indicates the proportion of items in which solution behavior is shown. A higher value is assumed to be an indicator of greater test-taking effort and engagement during the test.

In our study, we examined two distinct thresholds: a 5-second threshold (Wise & Kong, 2005) and the normative threshold (NT10; Wise & Ma, 2012). The 5-sec threshold serves as a benchmark for the minimum time needed to meaningfully engage with an item. A response time below 5 seconds is interpreted as a sign of low effort or disengagement by the respondent. This threshold is useful for differentiating rapid guessing, where responses are made too quickly to demonstrate genuine effort, from intentional and effortful engagement (Wise & Kong, 2005). On the other hand, the NT10 threshold is defined as 10% of the average time test-takers spend on an item, at a maximum of ten seconds. We couldn't find an RTE-like formula used for NA in the literature. We adapted the RTE formula to NA based on the normative 10 method. Thus, we set our threshold by taking the 10% of the average NA that test takers had on an item, with the goal of following a similar logic to response time and ensuring consistency in the application of effort measures. However, we acknowledge that this is only an attempt to adapt the RTE formula. The threshold obtained may not be universally applicable, and further research is needed to refine these criteria. Readers should be aware of this and use and interpret the results with caution.

To examine the consistency of different measures and their relationship with achievement, Pearson correlations were examined. In addition, the presence of different subgroups of disengaged responders were investigated with LPA using the following indices: response time, number of actions, self-reported effort. Latent profile analysis (LPA) is a statistical technique used to uncover and characterize hidden groups of individuals (referred to as profiles in LPA) who exhibit similar patterns across one or more indicator variables. These groups, often referred to as unobserved latent mixture components, can be conceptualized as distinct classes or profiles of individuals. LPA falls under the broader category of Mixture Models (Ferguson et al., 2020, Hofverberg et al., 2022). Because LPA, unlike many traditional statistical methods, emphasizes the grouping of individuals rather than variables, it is often referred to as a person-centered approach to statistical analysis, as opposed to a variable-centered approach. Prior to conducting the analysis, multivariate normality was assessed using the Mardia test via the psych package in R (Revelle, 2022) in order to account for potential violations. Due to significant departures from normality, with both skewness and kurtosis showing p-values less than 0.01, the MLR estimator was chosen for its robustness to normality violations and its ability to produce more stable results (Li, 2015; Vermunt & Magidson, 2002). When using the MLR estimator, the inclusion of various fit indices contributes to a clearer interpretation and more robust model evaluation. While aBIC is particularly relevant due to its sample size adjustment, it is also important to consider other indices such as BIC, AIC and entropy when evaluating model fit and classification accuracy. Lower aBIC, BIC and AIC values indicate a better fitting model, while entropy values closer to 1 indicate a more accurate classification. In addition, likelihood ratio tests (e.g., LMR-LRT, BLRT) are useful for comparing models with different numbers of latent profiles to assess whether additional profiles significantly improve model fit (Morgan, 2015; Nylund et al., 2007; Spurk et al., 2020). Briefly, the number of groups was determined based on AIC, BIC, aBIC, entropy value, Lo-Mendell-Rubin likelihood ratio test (LMR), interpretability of the resulting groups, and the parsimony principle. Both the descriptive analysis and the LPA (using the MplusAutomation package (Hallquist & Wiley, 2018) with Mplus7 (Muthén & Muthén, 2014)) were performed in R statistical software (v2024.09.1+394; R Core Team, 2024).

Results

In this section, we first present descriptive statistics and correlations between different measures. Next, we interpret the results of the latent profile analysis, including how we classified students into profiles, how we determined the optimal model, and how we described the resulting profiles. Finally, we examine the relationship between the profiles and students' mathematics achievement and effort.

What percentage of the sample show disengagement?

Table 1 presents the distribution of students' engagement across three metrics: RTE_5sec, RTE_10p, and NA_10p. Engagement is categorized as Fully Engaged (=1), Highly Engaged (>.90), Moderately Engaged (.90 - .80), and Low Engaged (<.80).

Table 1.
Number of engaged and disengaged students under three different threshold system

	RTE_5sec	RTE_10p	NA_10p
Fully engaged (1.00)	5735 (82.37%)	4526 (65.00%)	382 (5.49%)
Highly engaged (>.90)	977 (14.03%)	1782 (25.59%)	1041 (14.95%)
Moderately engaged (.80 - .90)	173 (2.48%)	406 (5.83%)	1919 (27.56%)
Low engaged (<.80)	78 (1.12%)	249 (3.58%)	3621 (52.00%)

The data in Table 1 shows that under 5-sec threshold, the number of low engaged respondents was 78 (1.12%) and the number of medium engaged respondents was 173 (2.48%). The RTE_10 percent method provided more conservative results than the common threshold method. The number of fully engaged students were fewer on this normative method. On the other hand, the number of actions methods classified most of the examinees (52.00%) as low engaged test takers. The NA_10p metric, likely reflecting a call for further investigation and try with another threshold method due to its much lower engagement distribution.

How does effort, as reflected in process data (response time and number of actions), self-reported effort, and test performance, relate to one another?

Descriptive statistics and Pearson correlations for each pair of measures, that have the potential to serve as indicators of disengaged responding: response time (RT), number of actions (NA), self-reported effort (SRE) and mathematics achievement (Ach), are provided in Table 2.

Table 2.
Correlations and descriptive statistics between variables

	RT	NA	SRE	Ach	Mean	*SD*
Response Time (RT)	1.00				93.66	23.17
Number of Actions (NA)	.37	1.00			20.4	11.47
Self-Reported Effort (SRE)	.09	.03	1.00		8.14	2.12
Math Achievement (Ach)	.40	.43	.02	1.00	452.24	89.29

The mean response time for the Turkey sample is 93.66 seconds (*SD* = 23.17) and the mean number of actions is 20.4 (*SD* = 11.47) for an item. The self-reported effort (SRE) item has an average of 8.14 out of 10 which is indicating a high level of self-effort. Lastly, the average mathematics achievement mean score is 452.24.

Notable relationships are observed between RT, NA, SRE, and mathematics achievement. To illustrate, the strongest correlation with achievement is observed for the NA (*r* = .43). The correlation between RT and achievement is relatively low (*r* = .40). Notably, SRE has the lowest correlation with performance, with correlation coefficients of .02. Similarly, the correlations between the SRE and RT (*r* = .09) and number of actions (*r* = .03) are weak, suggesting that these items may have a limited relationship with process data based methods for identifying disengaged responses. Conversely, the positive correlation between response time (RT) and the number of actions (NA) (*r* = .37) suggests that longer RT are associated with a higher NA, which may indicate a higher level of engagement in the test taking process. These findings highlight the importance of considering RT, NA, SRE, and performance-related variables in understanding disengaged responding.

What profiles can students be classified into based on response time, number of actions and self-reported effort data?

In the context of this study, LPA was used to classify students into subgroups based on different measures of disengagement. As stated in the methods section, the Mardia test results revealed significant deviations from multivariate normality, with both skewness and kurtosis showing p-values less than .01. Consequently, the MLR estimator was preferred for LPA and the results of the analysis are presented in Table 3.

Table 3 shows the fit indices of the LPA models for the different profile solutions. When deciding on the optimal solution, the lower AIC, BIC and aBIC values indicate a better fit and higher entropy values indicate a higher classification confidence. The p-value of the LMR test is also taken into account. Considering all these indicators, the three-profile model was considered as the optimal solution. The model fit statistics presented in Table 3 indicate that the three-profile solution provides the optimal balance between statistical fit and interpretability. The three-profile solution shows a significant improvement in model fit as evidenced by a significant reduction in AIC (51750.54), BIC (51845.58) and adjusted BIC (51801.09) compared to the two-profile model. Besides, the entropy value of the three-profile solution (0.883) is also high, indicating high classification accuracy. The Lo-Mendell-Rubin (LMR) test also yielded a significant result for the three-profile solution ($p < .05$), further supporting the addition of a third profile. Although the four and five-profile solutions have lower AIC, BIC, and ABIC values, the entropy value (0.883) drops significantly, indicating that the classification is less accurate. In addition, the LMR test results indicated that there was no further support for the addition of the fourth profile ($p > .05$). The 3-profile solution provides a balanced and meaningful structure and was selected as the most appropriate model for further analysis. After the 3-profile model was selected as the optimal solution, a closer look at this model was taken.

The data presented in Table 4 highlight the means for each profile across response time, number of actions, and self-report items. Figure 1 also shows the average standardized scores for three variables across different profiles.

The ANOVA results indicated that there were statistically significant differences between the profiles for all three variables ($p < .05$). In post-hoc analyses, the Tukey test was performed to examine the differences between profiles. Tukey test results indicated that all profiles were significantly different on all variables (RT, NA and SRE, $p < .05$).

The first profile (Profile 1) consists of 5431 students representing 82.79% of the sample and is characterized by a low number of actions within a short time period, i.e. they didn't put a high amount of effort, but they have the highest level of SRE among the three profiles ($p < .05$). They have lower RT and NA scores than Profile 2, but they are higher than Profile 3. Profile 2 consists of 472 students (7.20%) who have the highest mean response time ($p < .05$) and number of actions ($p < .05$), indicating that the test-takers exerted a high level of effort and demonstrated a low level of disengagement. Although they rated their effort lower than in the first profile ($p < .05$), it is at a moderate level and much higher than in Profile 3. Profile 3 ($n = 657$; 10.01%) had the lowest RT, NA, and SRE scores, all of which were statistically significantly different from the other profiles. This profile had the characteristics of disengaged responders and was labeled "Disengaged". Although Profile 2 had a slightly lower SRE than Profile 1, it has the highest RT and NA scores, and this pattern indicates the characteristics of "highly engaged" responders. Profile 1, with the largest number of students, had scores very close to the mean. It shows signs of engagement, but the level of engagement is lower than Profile 2, which results in the label of "Moderately-Engaged". Finally, the three-profile solution clearly distinguishes between engaged and disengaged individuals. It proved effective in differentiating between engaged and disengaged individuals. P3 is the group with the highest level of disengagement, while P2 has the highest level of engagement and P1 has moderately engaged individuals.

Table 3.
LPA models fit indices with different latent profiles

	Two-Profile	Three-Profile	Four-Profile	Five-Profile
Fit Statistics				
AIC	52853.5	51750.54	51205.58	50719.29
BIC	52921.39	51845.58	51327.77	50868.65
ABIC	52889.61	51801.09	51270.57	50798.74
Entropy	0.929	0.883	0.797	0.806
LMR (p)	2171.753 (.00)	1080.232 (.016)	537.668 (.379)	480.611 (.035)
Profile size (%)				
P1	688 (10.49%)	5431 (82.79%)	4424 (67.44%)	4453 (67.88%)
P2	5872 (89.51%)	472 (7.20%)	107(1.63%)	558 (8.51%)
P3		657 (10.01%)	640 (9.76 %)	1145 (17.45%)
P4			1389 (21.17%)	364 (5.55%)
P5				40 (0.61%)

Table 4.
Average Standardized Scores for Three Profiles

	P1	P2	P3
Response Time Mean	-.037	.813	-.384
Number of Actions Mean	-.164	1.996	-.357
Self-Reported Effort Item Mean	.267	.001	-2.308

Figure 1.

RT, NA and SRE averages by profile

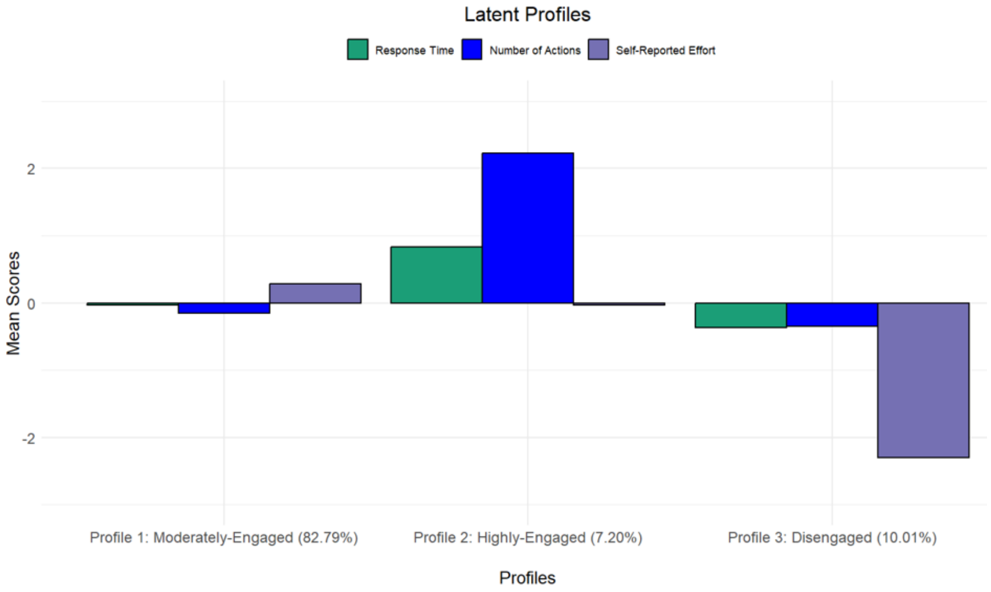

Figure 2.

Profile - specific mathematics achievement means

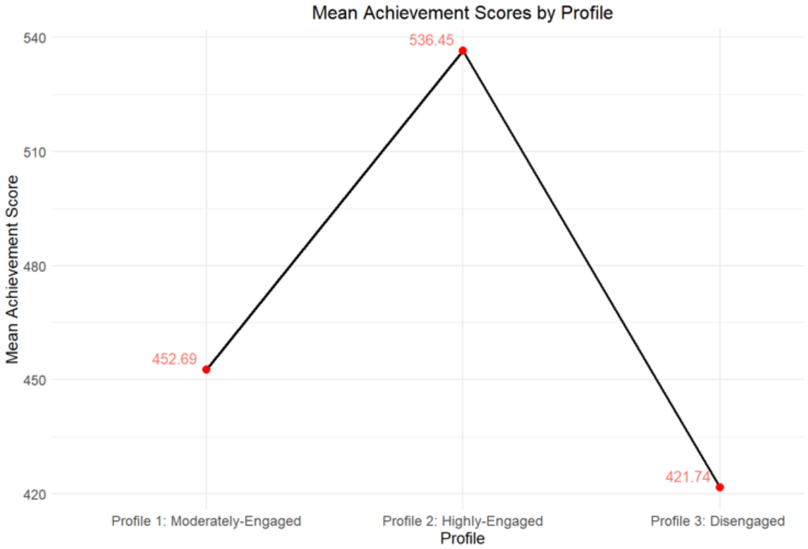

Table 5.

Gender distribution at student profiles

Profile	n	% of Total	# of Females
Moderately-Engaged	5431	82.79%	2786 (51.30%)
Highly-Engaged	472	7.20%	216 (45.76%)
Disengaged	657	10.01%	212 (32.27%)

N: Sample Size

While examining the mathematics achievement scores of three different profiles, it is indicated that highly engaged group (Profile 2) has the highest achievement score of 536.453. While the disengaged group has the lowest achievement score (421.74), Profile 1 has an average achievement score of 452.689. The differences were at a significant level for each group. Figure 2 shows the mean achievement scores across the different profiles.

After these interpretations, the distribution of gender was also studied in three profiles. Table 5 shows the corresponding information.

While the number of men and women in the moderately-engaged group is close, the proportion of men in the disengaged groups is almost double that of women. In the highly engaged group, the numbers of men and women are close, but tend to be dominated by men.

Discussion

The purpose of the present study is to examine the test-taking disengagement behaviors of responders based on response time, number of actions, and self-reported effort data from PISA 2022 data. The results of this study provide valuable insights into student engagement and its relationship with test performance and demographic factors. Through a combination of descriptive statistics, correlation analysis, and latent profile analysis (LPA), several important conclusions emerge regarding the nature of student disengagement and its implications for educational assessment.

First, we observed that the proportion of engaged behaviors in the dataset differed significantly depending on the metrics used (RTE_5sec, RTE_10p, and NA_10p). Between response time methods, the RTE_10p method produced more conservative results compared to RTE_5sec and fewer individuals were classified as fully engaged. In the literature, item-specific threshold methods (such as normative methods), are recommended as a useful criterion to find the invalid results due to low effort (Goldhammer et. al, 2017; Wise & Ma, 2012) because they use the item characteristics too. However, it should be noted that the thresholds coinciding with 10 percent were too high and the 10 second threshold which was set as the maximum was used for all of the questions. Thus, the normative method became a common method using the 10-second threshold. On the other hand, the NA_10p method classified most students (52%) as low-engaged, suggesting that it captures a broader, potentially inflated range of disengaged behaviors. Unlike response time, where minimal time clearly signals disengagement, the number of actions (NA) may not have a straightforward relationship with cognitive effort. Certain items in the assessment may naturally require fewer actions to complete, regardless of the level of engagement or cognitive effort. On the other hand, the observed discrepancy may also stem from the threshold setting process since we have just adapted the RTE formula into the number of actions. Therefore, the method has some limitations, as the threshold used may not be universally applicable and reliable. Further research is necessary to refine these criteria. Readers should be mindful of these limitations and interpret the results with caution. These factors highlight the complexity of using the number of actions as a sole indicator of engagement and the need for careful selection of thresholds and the potential benefits of combining multiple metrics for a more comprehensive understanding of student engagement.

The correlations between response time (RT), number of actions (NA), self-reported effort (SRE), and mathematics achievement reveal some important patterns in test-taking disengagement. In particular, number of actions had the strongest correlation with performance (r = .43), suggesting that higher levels of interaction with the test are positively associated with performance. The relationship between response time and achievement was also positive and at a moderate level (r = .40), as observed in recent literature (Eichmann et al., 2020; Kuhfeld & Soland, 2020; Wise&Kong, 2005). However, the correlation was relatively weaker compared to the NA, in line with the findings of Csányi & Molnár (2023). Conversely, self-reported effort has the weakest correlation with performance (r = .02), highlighting a potential gap between perceived and actual effort. The moderate correlation between RT and NA (r = .37) suggests that students who spend more time on tasks also tend to perform more actions, which is consistent with higher engagement. These results indicate that log data based measures such as RT and NA are more reliable indicators of engagement and effort than self-reported measures. Previous studies have consistently shown that test-taking effort, especially when assessed using response time effort, has a stronger correlation with performance than self-reported effort (Rios et al., 2014; Silm et al., 2020 Wise & Kong, 2005).

The latent profile analysis identified three distinct engagement profiles: Moderately Engaged (Profile 1), Highly Engaged (Profile 2), and Disengaged (Profile 3). The Moderately-Engaged group, which comprised the majority (82.79% of the sample), was characterized by average RT and NA scores but the highest self-reported effort. The Highly Engaged group (7.20%) has the highest RT and NA scores, indicating sustained effort on the task, despite slightly lower self-reported effort than the Moderately Engaged group. The Disengaged group (10.01%) has the lowest RT, NA, and SRE scores, highlighting their lack of effort and interaction during the test. Math achievement scores varied significantly across the engagement profiles, further validating the LPA results. These performance differences underscore the critical role of engagement in academic success and suggest that targeted interventions to increase engagement could significantly improve achievement.

An analysis of the gender distribution also reveals interesting trends. While the ‘Moderately Engaged’ group includes almost equal numbers of men and women, the ‘Disengaged’ group is more prevalent among men, with almost twice as many men as women in the Disengaged profile. Conversely, the Highly Engaged group shows a slight male predominance, although the difference is not as great. These patterns suggest potential gender differences in engagement behaviors, in line with the findings in the literature (Buchholz et al., 2022; DeMars et al., 2013; Wise et al., 2010) which warrant further investigation to understand the underlying causes and address inequalities.

In conclusion, this study highlights the multifaceted nature of student engagement and its critical influence on academic outcomes. This study provides an important step towards a better understanding of students' behavior and effort during the exam process. The findings obtained with the LPA method suggest that test-taking effort can be modeled in different profiles and that these profiles should be taken into account in exam design and assessments. Rather than focusing solely on exam outcomes, educational systems should devise more equitable and efficient assessment approaches by considering students' effort and motivation throughout the examination process. Policymakers and educators should consider using multiple engagement metrics, such as response time and number of actions, alongside measures of motivation, to create a more holistic picture of student performance. By addressing both effort and motivation across diverse contexts, education systems can better support student learning and equity worldwide.

Future research should explore alternative threshold settings for the number of action and focus on refining response time and action-based metrics to better identify disengagement, particularly through the use of item-specific thresholds for both number of action and response time which could provide more accurate and context-sensitive measures of engagement. This would help refine our understanding of how cognitive engagement is reflected across different types of test items and lead to more valid and reliable classifications of engagement. A crucial dimension to explore further is the role of motivational factors in engagement behaviors. Investigating these factors across different demographic groups, including gender, socio-economic status, and cultural contexts, can provide insights into disengagement and help develop targeted interventions. Another area of interest is cross-national comparisons of engagement behavior. Our study was limited to the Turkish sample, but examining how students' engagement and motivational factors differ across countries could provide a broader perspective on how educational systems, cultural values and socio-economic conditions shape test-taking behavior. By identifying best practices in countries with higher levels of participation, such analyses can provide actionable strategies for improvement in other regions.

References

Akyol, P., Krishna, K. & Wang, J. (2021) Taking PISA seriously: How accurate are low-stakes exams?. *Journal of Labor Research*, 42, 184–243 (2021). https://doi.org/10.1007/s12122-021-09317-8

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16, 441-462. https://doi.org/10.1007/BF03173192

Buchholz, J. (2022), "Are students trying hard to succeed in PISA?", *PISA in Focus*, No. 119, OECD Publishing, Paris, https://doi.org/10.1787/16c159b2-en

Buchholz, J. , Cignetti, M. & Piacentini, M. (2022). Developing measures of engagement in PISA. *OECD Education Working Papers* (279). https://dx.doi.org/10.1787/2d9a73ca-en

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper & Row.

Csányi, R. & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences*, *106*(2023). https://doi.org/10.1016/j.lindif.2023.102340

DeMars, C. E., Bashkov, B. M. & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, 8, 69-82.

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning*, *36*(6), 933–956. https://doi.org/10.1111/jcal.12451

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

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

Ferguson, S. L., G. Moore, E. W., & Hull, D. M. (2020). Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *International Journal of Behavioral Development*, *44*(5), 458-468. https://doi.org/10.1177/0165025419881721

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1-17. https://doi.org/10.1002/ets2.12067

Gignac, G. E., Bartulovich, A., & Salleo, E. (2019). Maximum effort may not be required for valid intelligence test score interpretations. *Intelligence*, 75, 73–84. https://doi.org/10.1016/j.intell.2019.04.007

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. *OECD Education Working Papers*, (No. 133). https://dx.doi.org/10.1787/5jlzfl6fhxs2-en

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 Assessment in Education*, *5*(18). https://doi.org/10.1186/s40536-017-0051-9

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hofverberg, A., Eklöf, H., & Lindfors, M. (2022). Who makes an effort? A person-centered examination of motivation and beliefs as predictors of students' effort and performance on the PISA 2015 science assessment. *Frontiers in Education*, *6* (2021). https://doi.org/10.3389/feduc.2021.791599

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*, 606–619. doi: 10.1177/0013164406294779

Kuhfeld, M. & J. Soland (2020). Using Assessment Metadata to Quantify the Impact of Test Disengagement on Estimates of Educational Effectiveness, *Journal of Research on Educational Effectiveness*, *13*(1), 147-175, https://doi.org/10.1080/19345747.2019.1636437

Li, C. H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Lundgren, E. & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation*, *26*(5–6), 275–301. https://doi.org/10.1080/13803611.2021.1963940

Messick, S. (1984). The nature of cognitive styles: problems and promise in educational practice. *Educational Psychologist*, *19*, 59-74. https://doi.org/10.1080/00461528409529283

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. doi:10.2307/1175249

Morgan, G. B. (2015). Mixed mode latent class analysis: An examination of fit index performance for classification. *Structural Equation Modeling*, *22*(1), 76–86. https://doi.org/10.1080/10705511.2014.935751

Muthén, L. K., & Muthén, B. O. (2014). *Mplus: Statistical Analysis with Latent Variables: User's Guide* (Version 7).

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569. https://doi.org/10.1080/10705510701575396

OECD (2016). *Low-performing students: Why they fall behind and how to help them succeed*. PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264250246-en

OECD (2024). *PISA 2022 technical report*. PISA, OECD Publishing. https://doi.org/10.1787/01820d6d-en

Pools, E. & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-scale Assessment in Education*, *9* (10),. https://doi.org/10.1186/s40536-021-00104-6

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rios J. A., Liu, O. L. & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014* (161). https://doi.org/10.1002/ir.20068

Sahin, F., & Colvin, K.F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessment in Education*, *8*(5). https://doi.org/10.1186/s40536-020-00082-1

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, *32*(2), 151-165. https://doi.org/10.1080/08957347.2019.1577244

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior, 120,* 103445. https://doi.org/10.1016/j.jvb.2020.103445

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press.

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

Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, *33*(2), 83–94. https://doi.org/10.1080/08957347.2020.1732386

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

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. https://doi.org/10.1111/jedm.12102

Wise, S. L., & Kingsbury, G. G. (2022). Performance decline as an indicator of generalized test-taking disengagement. *Applied Measurement in Education*, *35*(4), 272–286. https://doi.org/10.1080/08957347.2022.2155651

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. https://doi.org/10.1207/s15324818ame1802_2

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 at the annual meeting of the National Council on Measurement in Education, Denver.

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 (pp. 163-183).

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