Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: www.tandfonline.com/journals/hame20

Measuring Survey Disengagement in International Large-Scale Assessments: One Size Does Not Fit All

Ella Anghel & Matthias von Davier

To cite this article: Ella Anghel & Matthias von Davier (27 Feb 2026): Measuring Survey Disengagement in International Large-Scale Assessments: One Size Does Not Fit All, Applied Measurement in Education, DOI: 10.1080/08957347.2026.2638581

To link to this article: https://doi.org/10.1080/08957347.2026.2638581

APPLIED MEASUREMENT IN EDUCATION
https://doi.org/10.1080/08957347.2026.2638581

Measuring Survey Disengagement in International Large-Scale Assessments: One Size Does Not Fit All

Ella Anghel [a*] and Matthias von Davier [b,c]

[a]School of Education, Ben-Gurion University in the Negev; [b]Department of Measurement, Evaluation, Statistics, and Assessment, Lynch School of Education and Human Development, Boston College; [c]TIMSS and PIRLS International Study Center, Lynch School of Education and Human Development, Boston College

ABSTRACT

Test-taking disengagement in international large-scale assessments (ILSAs) is a widely explored topic, but little is known about students' disengagement in ILSA surveys or how it compares internationally. In this study, we used data from PIRLS 2021 to compare six measures of survey disengagement across 25 countries. We also tested how the relationships between these disengagement measures, the survey scores, and reading achievement vary internationally. Countries accounted for a significant variation in the disengagement measures, especially in the percentage of slow responses ($\eta^2 = .11$). Fast responses predicted experiencing bullying and poorer reading achievement ($\beta = 0.26$, -0.37, respectively). However, there were substantial variations between countries in the relationships between the disengagement measures and students' scores. Our results suggest that there are meaningful differences between countries in survey disengagement and in the associations between disengagement and the target scores, such that uniform criteria for identifying disengaged survey-takers might not be appropriate.

1. Introduction

Identifying disengaged survey responses is key to maintaining the validity of scores (Curran, 2016). Several measures of disengagement in surveys have been suggested in the literature, including measures based on response time (RT), response patterns (e.g., straight-lining), or psychometric methods ranging from simple missing response rates to more sophisticated approaches like person fit, caution indices, and mixture models (e.g., Yamamoto, 1989). However, little is known about the international comparability of these measures, a central issue if surveys are administered globally, as is the case in international large-scale assessments (ILSAs). Using data from the Progress in International Reading Literacy Study (PIRLS) ILSA, we examined country-level differences in several measures for detecting disengagement and their relationship with survey and reading achievement scores. By doing so, we hope to inform future efforts to mitigate the effects of survey-taking disengagement on the validity of ILSAs.

CONTACT Ella Anghel anghel@bgu.ac.il School of Education, Ben-Gurion University of the Negev, David Ben Gurion Blvd 1, Be'er Sheva, Israel

*This work was performed when the author was employed at The TIMSS and PIRLS International Study Center.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/08957347.2026.2638581

2. Literature Review

Test-taking disengagement has a potentially serious impact on assessment. If test-takers do not actually try to solve the test, it does not reflect their true knowledge, threatening its validity (Wise & Smith, 2016). Disengagement is a particularly egregious issue in ILSAs because test-takers may feel like they do not have to work hard and demonstrate their knowledge given the assessment's low stakes at the student level (List et al., 2017). Therefore, researchers invest substantial efforts in identifying and perhaps controlling disengagement in ILSAs (e.g., Bulut et al., 2023).

Although most ILSAs focus on achievement testing, they also include a survey portion where test-takers report various factors relevant to the main construct targeted by the assessment, including SES, school climate, or attitudes toward the construct (e.g., math; see, for example, Reynolds et al., 2021 for TIMSS 2023). However, little attention has been dedicated to disengagement in these surveys, even though it means that the test-taker did not endorse the option that best represents their level of the construct, again threatening the assessment's validity (Silber et al., 2019). This phenomenon might be particularly severe in ILSAs because surveys are usually given after an achievement test, so they are prone to fatigue effects (Jeong et al., 2023), especially among younger test-takers who are less familiar with surveys than with tests.

Multiple measures of survey disengagement have been suggested in ILSAs and other contexts (see a review in Curran, 2016). These measures are sometimes divided into proactive and reactive (Ashley & Shaughnessy, 2023; DeSimone et al., 2015). Proactive measures are implemented a-priori by the survey designer, such as attention checks and self-reported effort. Since ILSAs try to minimize their length for logistic reasons (Rutkowski et al., 2014), proactive measures of survey disengagement are rarely available in ILSAs. One exception is PISA's self-reported effort thermometer (Kunter et al., 2002). However, this measure is limited because of social desirability and because the effort reported using the thermometer refers to parts of the achievement section, not the survey section (Pools & Monseur, 2021).

Reactive measures involve post-hoc detection measures based on response time (RT), response patterns, and psychometric analyses (e.g., person fit). Response patterns refer to skipped items or items endorsed in a pattern that is unrelated to the actual content of the items, for example, straight-lining (e.g., Reuning & Plutzer, 2020). A similar indicator is comparing items with the opposite meaning (psychometric antonyms); test takers demonstrating straight-lining or otherwise not paying attention are likely to endorse opposite items similarly (DeSimone et al., 2015). Statistical or psychometric procedures like outlier detection or person-fit statistics can detect aberrant or overly-consistent responses relative to the sample or the test-taker (Beck et al., 2019; von Davier & Molenaar, 2003).

In ILSAs, measures based on response patterns have been used to identify disengaged survey responders, including straight-lining and antonyms (Soland, Wise, et al., 2019). Others have used measures based on RTs that are too fast for meaningful engagement with the survey (e.g., Ulitzsch et al., 2024). Too slow responses are also used in ILSAs, though less frequently (e.g., Anghel et al., 2025; Lundgren & Eklöf, 2023). Missingness, which could perhaps be considered an extreme form of slow response, has also been used to detect low engagement (e.g., Zamarro et al., 2019). However, missingness may represent a different underlying disengagement process than slow responses, i.e., skipping and not interacting with the assessment at all, vs. mind wandering and low focus. Regardless, existing studies generally use a single measure for identifying disengaged test-takers, thus focusing on one type of disengagement (e.g., rapid guessing) and missing out on others (lack of attention as demonstrated by straight-lining). Using multiple measures is crucial for detecting different types of disengaged respondents and thus enhancing the responses' validity.

Another issue specific to ILSAs is international comparability; it is important to test whether the rates of disengagement are consistent across countries, both from the participating countries' and from the test designers' perspective. For the countries, knowing whether students are less likely to provide effortful responses can inform interventions that might help students take tests in a way that reflects their learning. For the test designers, it is important to know if some countries are more likely to have disengaged students, so their scores are less valid than would be expected.

We found one paper that compared multiple measures of disengagement internationally, addressing both issues. Zamarro et al. (2019) used two measures, missing and unexpected responses in the survey section of PISA (as well as one measure in the achievement section), and compared them internationally. They found that some countries have high rates of missing and inconsistent responses, while others have high rates only on one measure and not the other, supporting the idea that there are several types of disengagement that should be examined. However, their work could be expanded by including more measures of disengagement like those based on response times and response patterns, as they might capture different cognitive processes behind disengagement (e.g., random responses as opposed to skipping items).

Zamarro et al. (2019) also found that their survey-based disengagement measures predict between-country variation in achievement scores. However, as far as we know, studies have yet to examine how disengagement in surveys is associated with the scores on ILSA surveys themselves, particularly in different countries. As other studies have found that inattentive responding on surveys is associated with scores (e.g., in personality scales; Berry et al., 2019), examining this issue is key to understanding if accounting for disengagement could affect scores and, if so, to what extent this effect varies by country and scale.

Our first purpose in the current study was to examine how different disengagement measures vary across countries, expanding previous works by including measures based on fast and slow RT, response patterns, and psychometric approaches. Our second purpose was to test the relationship between disengagement measures and survey as well as achievement scores, to see if our expanded set of disengagement measures predicted them. Doing so could inform test developers and policy makers on the rates of disengaged survey respondents in ILSAs. This will result in flagging countries with high rates of disengagement of different types, informing future interventions, as well as in a better understanding of if and how disengagement is related to the survey scores, thus impacting their validity.

In summary, our research questions are:

RQ1: Do the levels of survey engagement measures vary by country?

RQ2: Do survey engagement measures predict survey responses and reading achievement, and does this vary by country?

3. Methods

3.1. Data

The data used in this study were obtained using the PIRLS 2021 context questionnaire (Yin & Reynolds, 2023). PIRLS is a low-stakes assessment designed to inform countries and educational systems about students' overall performance rather than track individual students' achievement. The assessment is administered at the sampled schools, proctored by a supervisor trained to follow standard test administration procedures to ensure comparability of the conditions across schools and countries. More information on test administration procedures is available in Johansone (2023).

In this study, we included all participating countries that had data on all of the PIRLS' survey scales (see below) and allowed for their data to be used for research purposes (with non-country entities like Quebec included in the appropriate country's sample, in this case, Canada). We also only included test-takers participating in the computer-based version of PIRLS (digital PIRLS), where response time data were collected. Table 1 presents the final sample sizes by country.

Table 1. Sample sizes and gender by country.

Country	Sample size	% Girl	% Boy	% Other
Belgium – Flemish	5,049	48	50	2
Canada	15,403	45	46	10
Croatia	3,996	49	51	0
Czechia	7,386	50	50	0
Denmark	4,769	51	48	1
Finland	6,992	49	49	2
Germany	4,409	44	46	9
Hungary	5,293	50	49	0
Italy	5,435	50	50	0
Kazakhstan	7,015	50	50	0
Lithuania	4,612	50	50	0
Malta	3,000	46	53	2
New Zealand	5,485	44	48	8
Norway	5,263	49	51	0
Portugal	6,098	48	52	0
Qatar	5,258	52	47	0
Russian Federation	5,213	49	51	0
Saudi Arabia	4,707	50	49	0
Singapore	6,710	49	51	0
Slovakia	4,823	51	49	0
Slovenia	5,097	49	51	0
Spain	8,490	48	52	0
Sweden	5,215	49	49	1
Taiwan	5,545	48	51	1
United Arab Emirates	9,314	52	48	0

Note: Percentages may not add up to 100 due to rounding and missing data on gender.

Before taking the questionnaire, each participating PIRLS student completed a booklet consisting of two blocks, each containing a reading passage and related items. Test-takers are given 40 minutes to complete each of the blocks. Then, the students are asked to complete a survey about their attitudes toward school and reading (see below for specific scales), dedicating at most 30 minutes to this survey. The survey scales are organized by screens such that each screen generally includes items from a single scale, with the exception of single items (e.g., gender) or very short scales that were not included in our analyses. Scales were spread across one or two screens each. Overall, the survey included 80 items across 23 screens. Tables 1 and 2 present information about the participants' gender and home language, respectively.

3.2. Instruments

3.2.1. Engagement Measures

We extracted several engagement indicators from our data, which are described below. We chose to focus on these measures due to their ubiquity in the literature and evidence of validity (see details below), as well as data availability. The measures were based on all of the items unless otherwise specified. While most measures were initially based on individual screens or scales, we used them to calculate continuous engagement metrics at the participant level, for example, the percentage of screens that were determined to be fast, slow, and so on. Table 3 presents descriptive statistics for those measures by country.

3.2.1.1. Fast and Slow Response Times. Both fast and slow RTs are measures often used in the study of disengagement in ILSAs' achievement and survey portions (e.g., Anghel et al., 2025; Lundgren & Eklöf, 2023) and in other surveys (Kumar et al., 2022). There is some theoretical and empirical evidence suggesting that fast and slow RTs represent rapid and distracted response patterns, respectively. In this study, RTs were recorded at the screen level. We calculated the total time spent on each screen across all visits. We then identified screens with an RT lower than two seconds per item (i.e.,

Table 2. Sample sizes and percentage of students by how often they speak the test's language at home by country.

Country	Sample size	% Always	% Almost always	% Sometimes	% Never
Belgium – Flemish	5,049	52	14	27	6
Canada	15,403	60	16	21	3
Croatia	3,996	71	19	9	1
Czechia	7,386	69	19	10	1
Denmark	4,769	63	20	15	1
Finland	6,992	65	18	14	2
Germany	4,409	56	15	18	2
Hungary	5,293	71	25	4	1
Italy	5,435	69	14	14	2
Kazakhstan	7,015	68	12	17	1
Lithuania	4,612	55	28	15	1
Malta	3,000	27	18	45	9
New Zealand	5,485	–	–	–	–
Norway	5,263	58	22	17	3
Portugal	6,098	76	11	10	1
Qatar	5,258	34	14	41	9
Russian Federation	5,213	76	12	9	2
Saudi Arabia	4,707	54	14	24	5
Singapore	6,710	32	20	44	3
Slovakia	4,823	60	20	14	5
Slovenia	5,097	66	16	12	4
Spain	8,490	60	13	18	8
Sweden	5,215	51	21	24	3
Taiwan	5,545	52	18	29	1
United Arab Emirates	9,314	42	16	33	7

Note: No data were available on this question in New Zealand. Percentages may not add up to 100 due to rounding and missing data.

Table 3. Means (SD) of the engagement measures by country.

Country	% Fast screens	% Slow screens	% Missing screens	% Straight-lining screens	% Misfit Zh scales	% Overfit Zh scales
Belgium – Flemish	0.83 (3.30)	9.99 (9.35)	0.86 (2.98)	19.47 (16.37)	2.81 (6.81)	0.02 (0.53)
Canada	1.16 (5.21)	5.59 (7.20)	1.07 (5.03)	23.58 (17.75)	2.60 (6.70)	0.01 (0.43)
Croatia	1.30 (6.02)	4.92 (6.12)	0.78 (4.92)	25.21 (17.11)	3.88 (8.86)	0.00 (0.00)
Czechia	0.88 (4.24)	8.47 (11.41)	0.64 (3.23)	23.17 (17.01)	3.37 (7.94)	0.05 (0.83)
Denmark	0.76 (5.07)	15.92 (14.63)	0.86 (5.15)	19.64 (15.05)	3.30 (7.73)	0.00 (0.21)
Finland	0.77 (4.31)	10.34 (11.43)	0.68 (3.55)	25.98 (17.17)	3.82 (8.71)	0.00 (0.00)
Germany	10.30 (29.18)	9.21 (8.89)	10.79 (29.41)	19.99 (17.09)	3.13 (7.64)	0.02 (0.48)
Hungary	1.08 (5.91)	5.03 (6.30)	0.61 (4.91)	23.36 (17.55)	3.35 (7.83)	0.04 (0.73)
Italy	0.58 (3.64)	5.60 (6.53)	0.79 (3.75)	20.46 (16.25)	2.60 (6.94)	0.05 (0.84)
Kazakhstan	1.30 (5.41)	6.30 (7.88)	0.57 (2.37)	29.28 (22.07)	2.06 (6.00)	0.05 (0.85)
Lithuania	2.35 (7.14)	4.06 (5.41)	0.81 (4.31)	22.15 (17.92)	3.63 (8.44)	0.01 (0.42)
Malta	1.68 (6.93)	14.16 (13.75)	0.99 (5.40)	25.81 (18.60)	1.83 (5.72)	0.06 (0.94)
New Zealand	2.24 (7.02)	7.66 (6.71)	1.35 (5.92)	21.86 (17.85)	2.92 (7.30)	0.02 (0.51)
Norway	0.95 (5.45)	11.46 (12.68)	0.93 (5.35)	22.03 (16.11)	3.36 (7.84)	0.03 (0.68)
Portugal	0.72 (4.36)	6.64 (8.13)	0.69 (3.82)	25.14 (17.37)	2.02 (5.63)	0.05 (0.84)
Qatar	4.46 (14.13)	10.17 (11.06)	2.28 (11.74)	27.06 (21.32)	1.52 (5.10)	0.15 (1.45)
Russian Federation	0.86 (4.47)	4.35 (6.60)	0.54 (3.37)	19.32 (16.44)	2.79 (7.06)	0.08 (1.06)
Saudi Arabia	10.06 (17.58)	9.55 (8.29)	3.95 (11.61)	29.56 (24.81)	1.86 (5.44)	0.19 (1.66)
Singapore	1.48 (5.07)	2.06 (4.30)	0.34 (2.96)	22.76 (17.45)	3.07 (7.54)	0.01 (0.43)
Slovakia	1.31 (6.90)	9.78 (11.90)	0.82 (5.23)	23.99 (17.74)	2.86 (7.01)	0.06 (0.92)
Slovenia	1.10 (4.84)	6.04 (7.27)	0.83 (3.84)	25.74 (18.43)	3.39 (8.02)	0.04 (0.80)
Spain	0.85 (4.10)	9.17 (8.97)	0.97 (4.18)	26.88 (18.70)	2.00 (5.89)	0.06 (0.90)
Sweden	1.02 (5.12)	15.20 (14.46)	0.98 (5.12)	22.35 (16.19)	3.90 (8.80)	0.07 (0.97)
Taiwan	1.28 (3.67)	5.23 (7.83)	0.30 (1.22)	27.67 (18.61)	3.58 (7.86)	0.04 (0.81)
UAE	5.45 (13.08)	9.99 (10.18)	1.44 (7.18)	35.87 (25.35)	1.59 (5.14)	0.20 (1.68)

Note: UAE – United Arab Emirates.

fast) or higher than thirty seconds per item (i.e., slow). We preferred a consistent cutoff rather than a relative cutoff (e.g., 5% of responses) to facilitate interpretation and comparison across countries. The two-second cutoff was selected because of its support in the literature. For example, Bowling et al. (2023) found evidence of its construct validity, including a correlation with other measures of low effort in surveys. However, we could not find a validated cutoff for slow RTs. To select the cutoff, we visually inspected the item duration distributions (Supplemental material A), an established practice for short RTs (e.g., Kong et al., 2007). As seen in Supplemental material A, the item durations had a similar unimodal distribution in all countries, with a long right tail representing outlying slow responses. The visual inspection revealed a natural inflection point around 25 seconds, and very few test-takers had RTs longer than 30 seconds per item (2.25%). Therefore, we selected 30 seconds as the cutoff for slow RTs that likely represent disengagement.

3.2.1.2. Missing Screens. The rate of missing items is a common measure of disengagement in surveys (e.g., Steyn, 2017) and in ILSAs (e.g., Pools, 2022). Evidence from surveys shows that non-response rates predict non-cognitive constructs like engagement and motivation (Hitt et al., 2016), making them an appropriate measure in our case. We identified screens that were skipped entirely (had no responses) and calculated their percentage out of the total number of screens. We focused on screens rather than individual items because the other disengagement measures were at the screen level.

3.2.1.3. Straight-Lining. An overly consistent pattern of responses, especially repeating the same response several times in a row, is often indicative of carelessness (Soland, Zamarro, et al., 2019). In this work, we identified screens where the same response was used for all items (before reverse coding), similar to the approach proposed by Soland, Jensen, et al. (2019). Then, we calculated the percentage of those screens out of the total. We limited this to screens with at least three items and at least three response options, as with screens with fewer items or options, endorsing the same option is likely reflective of a consistent level of the construct rather than low effort. Therefore, the maximum possible number of straight-line screens was 12. Empty screens (all null responses) were not considered straight-lining. In the subset of screens we examined, there was only one scale per screen.

As a result of how the variables are designed, it is more likely to have a higher percentage of straight-lining than the other variables (even one screen with straight-lining would result in a value of 8% on this indicator). Since our analyses combine multiple indicators and are relative to other participants rather than relying on a consistent cutoff to identify disengaged test-takers, this issue is unlikely to impact our main research questions.

3.2.1.4. Person Fit Statistics. Both overly-consistent responses (e.g., straight-lining) and inconsistent responses (e.g., endorsing items that are psychometric antonyms) represent careless or disengaged response patterns (Soland, Zamarro, et al., 2019). One of the ways of capturing the responses' consistency is using person fit statistics, the extent to which the responses fit the statistical model used to describe the scores. Indeed, person fit statistics can identify over-consistency and inconsistency in simulated data, and they are associated with other measures described here like straight-lining and RT (Jones et al., 2023), supporting their validity as measures of disengagement.

In this study, we fit a one-factor rating scale model (Andrich, 1978) for each of the scales used in this study (see below) using the MIRT R package (Chalmers, 2012). Then, MIRT reports Zh person fit statistics (Drasgow et al., 1985). Zh is a person's standardized log likelihood, measuring their deviation from the response pattern expected based on the model (in our case, a one-factor rating scale model). Negative values of this statistic can be used for flagging misfitting respondents (overly inconsistent), while positive values represent overfitting, overly consistent respondents (e.g., tend to always give the same response). Following Felt et al. (2017), we used values below −2 to represent misfit and values above 2 to represent overfit. We calculated the percentage of scales with extreme negative (misfitting) and positive (overfitting) Zh values for each student out of the possible seven scales.

3.2.2. Survey Scales

Below, we describe the scales we used to see whether the aforementioned engagement measures predict the survey's results; since we were interested in the association between disengagement and survey scores in general, we did not focus on a specific variable, and included all available survey scales that represented a latent construct (i.e., not questions about gender and language, etc.). In most scales, the original survey is structured such that low scores represent agreement with the items. Where relevant, we reverse-scored the scales to facilitate interpretation; in this study, high scores always represent high levels of the construct. Yin and Reynolds (2023) provide more information on those scales. Sample sizes, mean scale scores, and reliabilities by country in this sample are presented in Table 4. In Table 4 and throughout all analyses, students who skipped an entire scale were excluded from the calculations with respect to that particular scale but not other scales. In cases of some missing items without skipping the whole scale, scale scores were calculated by averaging the remaining items.

3.2.2.1. Digital Self-Efficacy. This scale describes students' sense that they successfully use computers or smart devices to complete tasks. The scale includes eight items, for instance, "It is easy for me to find information on the internet," rated from 1—*Agree a lot* to 4—*Disagree a lot*.

3.2.2.2. Students' Sense of School Belonging. The students were asked to explain their thoughts about their school in five items, e.g., "I like being in school" (1 – *Agree a lot* to 4 – *Disagree a lot*).

3.2.2.3. Student Bullying. In this ten-item scale, the students described the frequency of experiencing bullying from 1 – *Never* to 4 – *At least once a week*. Such behaviors include cases when someone "Made fun of me or called me names" or "Hit or hurt me (e.g., shoving, hitting, kicking)."

3.2.2.4. Students Engaged in Reading Lessons. The scale represents different attitudes toward their reading classes, including "I like what I read about in school" and "My teacher is easy to understand." It has nine items and is scored on a 1 – *Agree a lot* to 4 – *Disagree a lot* scale.

3.2.2.5. Disorderly Behavior During Reading Lessons. In five items (1 – *Never* to 4 – *Every lesson*), students reported the frequency of disruptions in reading lessons. Examples include too much noise or students' interruptions.

3.2.2.6. Students Like Reading. The scale represents different attitudes toward reading in general, including "I enjoy reading" and "I learn a lot from reading." It has eight items and is scored on a 1 – *Agree a lot* to 4 – *Disagree a lot* scale. One of the items is reverse scored ("I think reading is boring").

3.2.2.7. Students Confident in Reading. The students describe how well they can read in six items (1 – *Agree a lot* to 4 – *Disagree a lot*). Items include, for example, "Reading is easy for me." Four of the items in this scale are reverse-coded (e.g., "I am just not good at reading").

3.2.3. Reading Achievement

In PIRLS, the raw performance score does not represent the student's achievement level as different countries took different booklets with varying difficulties (for more detail, see Martin et al., 2019). Instead, PIRLS provides a set of five plausible values representing reading achievement that account for these differences (National Center for Education Statistics [NCES], 2025). The plausible values are randomly drawn from an empirical distribution of scores derived from the students' responses, their background characteristics, and information on the items they completed; this is done instead of a direct ability estimation because, to keep the assessment brief, each student answers only a subset of items and thus does not have enough data to determine their individual ability. Using plausible values, therefore, allowed us to directly compare the reading performance across countries. We used all five plausible values in our models predicting reading achievement (see below).

Table 4. Sample sizes, means (SD), and reliabilities of the scales by country.

Country		Digital self-efficacy	Sense of school belonging	Student bullying	Engagement in reading lessons	Disorderly behavior in reading lessons	Liking reading	Reading confidence
Belgium – Flemish	*n*	4,622	4,802	4,506	4,629	4,729	4,570	4,514
	Mean (*SD*)	3.26 (0.50)	3.48 (0.56)	1.67 (0.65)	3.43 (0.48)	2.59 (0.75)	2.94 (0.75)	3.16 (0.70)
	α	.73	.79	.88	.82	.81	.87	.83
Canada	*n*	14,134	14,382	13,579	14,048	14,365	13,992	14,019
	Mean (*SD*)	3.30 (0.53)	3.57 (0.51)	1.60 (0.64)	3.55 (0.47)	2.53 (0.77)	3.17 (0.65)	3.24 (0.66)
	α	.76	.79	.88	.85	.84	.85	.82
Croatia	*n*	3,716	3,761	3,583	3,698	3,728	3,594	3,575
	Mean (*SD*)	3.35 (0.46)	3.34 (0.54)	1.44 (0.58)	3.33 (0.49)	2.58 (0.83)	2.95 (0.67)	3.26 (0.61)
	α	.77	.80	.89	.85	.86	.87	.81
Czechia	*n*	6,735	6,835	6,552	6,670	6,743	6,668	6,675
	Mean (*SD*)	3.21 (0.53)	3.49 (0.50)	1.50 (0.61)	3.41 (0.52)	2.60 (0.83)	2.98 (0.67)	3.15 (0.63)
	α	.77	.76	.87	.86	.83	.85	.79
Denmark	*n*	4,532	4,589	4,398	4,443	4,514	4,426	4,414
	Mean (*SD*)	3.32 (0.46)	3.54 (0.52)	1.47 (0.52)	3.27 (0.53)	2.41 (0.67)	2.79 (0.69)	3.22 (0.66)
	α	.76	.81	.87	.86	.81	.86	.85
Finland	*n*	6,597	6,734	6,463	6,522	6,625	6,469	6,505
	Mean (*SD*)	3.43 (0.45)	3.60 (0.47)	1.37 (0.49)	3.35 (0.52)	2.31 (0.71)	2.88 (0.72)	3.36 (0.60)
	α	.79	.81	.88	.87	.86	.89	.83
Germany	*n*	3,630	3,728	3,453	3,532	2,561	3,402	3,427
	Mean (*SD*)	3.18 (0.56)	3.57 (0.49)	1.54 (0.62)	3.48 (0.51)	2.54 (0.78)	3.03 (0.72)	3.32 (0.63)
	α	.79	.77	.89	.86	.83	.87	.81
Hungary	*n*	4,849	4,980	4,779	4,857	4,951	4,801	4,773
	Mean (*SD*)	3.31 (0.53)	3.45 (0.54)	1.56 (0.60)	3.40 (0.55)	2.55 (0.81)	2.93 (0.74)	3.20 (0.65)
	α	.76	.77	.87	.87	.85	.87	.80
Italy	*n*	4,851	5,072	4,721	4,900	5,047	4,841	4,843
	Mean (*SD*)	3.15 (0.57)	3.56 (0.49)	1.56 (0.61)	3.48 (0.48)	2.83 (0.74)	3.24 (0.65)	3.28 (0.61)
	α	.74	.74	.86	.81	.79	.86	.77
Kazakhstan	*n*	6,203	6,411	5,916	6,187	6,364	6,265	6,242
	Mean (*SD*)	3.25 (0.53)	3.62 (0.49)	1.58 (0.67)	3.60 (0.46)	2.35 (0.95)	3.38 (0.57)	3.12 (0.66)
	α	.76	.77	.87	.83	.86	.82	.74
Lithuania	*n*	4,290	4,338	4,095	4,225	4,299	4,216	4,214
	Mean (*SD*)	3.30 (0.48)	3.42 (0.52)	1.58 (0.65)	3.32 (0.54)	2.54 (0.83)	2.87 (0.69)	3.12 (0.63)
	α	.76	.77	.89	.87	.87	.86	.79
Malta	*n*	2,783	2,837	2,737	2,745	2,771	2,746	2,747
	Mean (*SD*)	3.26 (0.53)	3.64 (0.48)	1.52 (0.62)	3.63 (0.45)	2.65 (0.81)	3.27 (0.65)	3.12 (0.68)
	α	.71	.75	.87	.82	.79	.84	.78
New Zealand	*n*	4,895	5,081	4,787	4,891	5,019	4,917	4,909
	Mean (*SD*)	3.25 (0.55)	3.52 (0.54)	1.76 (0.70)	3.42 (0.52)	2.81 (0.74)	3.16 (0.66)	3.04 (0.67)
	α	.77	.79	.89	.85	.83	.85	.79
Norway	*n*	4,946	5,050	4,800	4,822	4,972	4,836	4,860
	Mean (*SD*)	3.44 (0.42)	3.59 (0.47)	1.47 (0.53)	3.35 (0.53)	2.32 (0.68)	2.75 (0.70)	3.21 (0.66)
	α	.73	.80	.88	.87	.84	.86	.82
Portugal	*n*	5,623	4,299	4,415	2,593	5,665	5,549	4,983
	Mean (*SD*)	3.37 (0.49)	3.79 (0.36)	1.61 (0.62)	3.71 (0.36)	2.56 (0.80)	3.48 (0.51)	3.19 (0.64)
	α	.75	.73	.86	.79	.83	.82	.79

(*Continued*)

Table 4. (Continued).

Country		Digital self-efficacy	Sense of school belonging	Student bullying	Engagement in reading lessons	Disorderly behavior in reading lessons	Liking reading	Reading confidence
Qatar	*n*	4,376	4,557	4,232	4,435	4,464	4,468	4,443
	Mean (*SD*)	3.19 (0.55)	3.56 (0.57)	1.57 (0.72)	3.59 (0.49)	2.35 (0.87)	3.34 (0.62)	3.21 (0.67)
	α	.69	.78	.90	.83	.80	.83	.74
Russian Federation	*n*	4,741	3,408	4,541	4,707	4,839	4,680	4,736
	Mean (*SD*)	3.29 (0.52)	3.38 (0.54)	1.82 (0.69)	3.47 (0.50)	2.53 (0.88)	3.11 (0.63)	3.14 (0.61)
	α	.75	.76	.85	.85	.85	.85	.77
Saudi Arabia	*n*	3,118	3,391	3,089	3,180	3,251	3,213	3,177
	Mean (*SD*)	3.02 (0.63)	3.67 (0.52)	1.60 (0.83)	3.61 (0.52)	2.33 (0.94)	3.47 (0.54)	3.07 (0.70)
	α	.74	.80	.93	.84	.82	.81	.71
Singapore	*n*	6,340	6,481	6,258	6,333	6,447	6,357	6,344
	Mean (*SD*)	3.16 (0.57)	3.43 (0.58)	1.60 (0.65)	3.38 (0.53)	2.75 (0.78)	3.11 (0.68)	3.27 (0.65)
	α	.77	.80	.88	.86	.84	.86	.81
Slovakia	*n*	4,419	4,516	4,268	4,378	4,422	4,323	4,300
	Mean (*SD*)	3.19 (0.59)	3.53 (0.53)	1.52 (0.62)	3.46 (0.53)	2.61 (0.85)	3.01 (0.71)	3.15 (0.69)
	α	.80	.79	.88	.87	.84	.86	.82
Slovenia	*n*	4,682	4,775	4,558	4,613	4,694	4,578	4,618
	Mean (*SD*)	3.28 (0.54)	3.46 (0.57)	1.56 (0.65)	3.43 (0.52)	2.68 (0.82)	2.98 (0.68)	3.26 (0.65)
	α	.79	.82	.89	.86	.83	.85	.82
Spain	*n*	7,568	7,839	7,235	7,542	7,749	7,569	7,534
	Mean (*SD*)	3.33 (0.51)	3.68 (0.46)	1.57 (0.64)	3.62 (0.45)	2.76 (0.77)	3.43 (0.58)	3.15 (0.60)
	α	.72	.75	.85	.82	.79	.84	.69
Sweden	*n*	4,830	4,952	4,702	4,779	4,871	4,721	4,672
	Mean (*SD*)	3.37 (0.44)	3.50 (0.51)	1.49 (0.57)	3.40 (0.51)	2.52 (0.75)	2.76 (0.72)	3.38 (0.60)
	α	.75	.80	.88	.87	.85	.87	.83
Taiwan	*n*	5,168	5,315	5,183	5,193	5,295	5,215	5,192
	Mean (*SD*)	3.01 (0.62)	3.38 (0.59)	1.45 (0.57)	3.35 (0.62)	2.45 (0.85)	3.21 (0.69)	3.01 (0.69)
	α	.79	.77	.86	.89	.86	.88	.77
UAE	*n*	7,871	8,202	7,668	7,960	8,103	8,004	7,867
	Mean (*SD*)	3.32 (0.53)	3.67 (0.50)	1.64 (0.84)	3.66 (0.48)	2.47 (1.01)	3.47 (0.53)	3.13 (0.73)
	α	.72	.78	.93	.85	.88	.81	.74

Note: UAE – United Arab Emirates.

3.3. Analysis

To answer RQ1, we used a set of ANOVAs to compare the disengagement measures by country. As the disengagement measures are expected to be correlated, we applied a Bonferroni-adjusted $\alpha = .05/6 = .008$ to account for the multiple comparisons (6 ANOVAs). Then, for each ANOVA, we used Tukey HSD tests (Tukey, 1949) to identify the source of the differences among the countries.

To answer RQ2, we performed multivariate regression analyses predicting the scores of the scales we described above, first from the disengagement indicators and then adding the countries and their interactions. In the country variable, the reference group was arbitrarily selected as the United Arab Emirates because it was the last country alphabetically. A similar analysis was performed for reading achievement. However, as there were multiple plausible values representing achievement, we combined separate models predicting each plausible value using the *mitools* R package (Lumley, 2019) and using the student-level weights provided by PIRLS to ensure the representativeness of the sample. The package performs a separate regression analysis for each set of plausible values and then combines the resulting estimates using Rubin's (2018) rules. Since p-values are not automatically provided by the package, they were calculated by hand from the estimates and standard errors. The models' coefficients and adjusted R^2 were used to evaluate and interpret the models. Since the adjusted R^2 was not reported for the combined model, we report the mean adjusted R^2 across all models.

4. Results

RQ1: Do the levels of survey engagement measures vary by country?

The ANOVAs suggest that all five engagement measures vary significantly by country (all $p < .001$), with η^2 effect sizes of .07 for percent fast screens, .11 for percent slow screens, .06 for percent missing screens, .04 for percent straight-lining, .01 for percent misfitting, and <.01 for percent overfitting. We then performed a series of Tukey HSD tests to identify specific country differences. The results of the Tukey HSD tests are available in Supplemental material B. Table 5 presents the order of the countries on each of the measures (the values on each measure by country are given in Table 3).

We will briefly discuss here some cross-national patterns that emerged from the results. Most notably, the three Gulf states included in our sample (Saudi Arabia, Qatar, and the United Arab Emirates) showed a unique response pattern that distinguishes them from the other countries. These countries had elevated rates of fast, missing, straight-lining, and overfitting scales, as well as relatively low misfit rates. Interestingly, Germany was higher (ranked 1st) in both fast and missing screens, but

Table 5. The countries' order in the disengagement measures.

	% Fast screens	% Slow screens	% Missing screens	% Straight-lining screens	% Misfit Zh scales	% Overfit Zh scales
1	Germany	Denmark	Germany	UAE	Sweden	UAE
2	Saudi Arabia	Sweden	Saudi Arabia	Saudi Arabia	Croatia	Saudi Arabia
3	UAE	Malta	Qatar	Kazakhstan	Finland	Qatar
4	Qatar	Norway	UAE	Taiwan	Lithuania	Russian Federation
5	Lithuania	Finland	New Zealand	Qatar	Taiwan	Sweden
6	New Zealand	Qatar	Canada	Spain	Slovenia	Slovakia
7	Malta	UAE	Malta	Finland	Czechia	Spain
8	Singapore	Belgium – Flemish	Sweden	Malta	Norway	Malta
9	Slovakia	Slovakia	Spain	Slovenia	Hungary	Czechia
10	Kazakhstan	Saudi Arabia	Norway	Croatia	Denmark	Italy
11	Croatia	Germany	Denmark	Portugal	Germany	Kazakhstan
12	Taiwan	Spain	Belgium – Flemish	Slovakia	Singapore	Portugal
13	Canada	Czechia	Slovenia	Canada	New Zealand	Taiwan
14	Slovenia	New Zealand	Slovakia	Hungary	Slovakia	Slovenia
15	Hungary	Portugal	Lithuania	Czechia	Belgium – Flemish	Hungary
16	Sweden	Kazakhstan	Italy	Singapore	Russian Federation	Norway
17	Norway	Slovenia	Croatia	Sweden	Canada	Germany
18	Czechia	Italy	Portugal	Lithuania	Italy	New Zealand
19	Russian Federation	Canada	Finland	Norway	Kazakhstan	Belgium – Flemish
20	Spain	Taiwan	Czechia	New Zealand	Portugal	Lithuania
21	Belgium – Flemish	Hungary	Hungary	Italy	Spain	Singapore
22	Finland	Croatia	Kazakhstan	Germany	Saudi Arabia	Canada
23	Denmark	Russian Federation	Russian Federation	Denmark	Malta	Croatia
24	Portugal	Lithuania	Singapore	Belgium – Flemish	UAE	Finland
25	Italy	Singapore	Taiwan	Russian Federation	Qatar	Denmark

Note: UAE – United Arab Emirates.

unlike these countries, it had relatively low rates of straight-lining (ranked 22nd out of 25), and was not particularly high or low in the rates of overfitting or misfitting scales.

The Nordic countries in our sample (Denmark, Finland, Norway, and Sweden) exhibited different, slower engagement patterns; they were four out of the five countries with the highest rates of slow screens. They also tended to have high misfit levels. Several individual countries also had unique engagement profiles. For example, the Russian Federation and Singapore demonstrated low levels of both slow and missing screens, potentially revealing rapid but complete survey responses. However, most countries exhibited heterogeneous patterns across measures without consistent patterns.

RQ2: Do survey engagement measures predict survey responses and reading achievement, and does this vary by country?

We then tested whether the disengagement measures predict the scales' scores and reading achievement (see Table 6). First, let us examine the results with respect to the scale scores. Without accounting for variations by country, the models' explained variances varied, with the highest $R^2 = .21$ for engagement in reading lessons. Some individual predictors had quite meaningful effect sizes. In particular, straight-lining had a relatively strong effect on several of the surveys' scores, including liking reading (β = .29) and reading lessons engagement (β = .26). Interestingly, the percentage of fast screens significantly predicted experiencing bullying (β = .26). The explained variance in reading achievement was $R^2 = .13$. To illustrate the practical significance of these effects, we compared the predicted achievement of students with low levels of disengagement ($Z = -1$ on each indicator) with students with average levels of disengagement ($Z = 0$ on each indicator). We found that the reading scores of the former students were predicted to be 0.75 higher, a substantial difference. Specifically, the association between reading achievement and the percentage of fast (β = −.37), slow (β = −.18), and missing screens (β = .16) was stronger than the association between these measures and the survey scores themselves.

Adding the countries and the interactions as predictors (see Supplemental material C) did not improve the explained variance by much, with the highest $\Delta R^2 = .11$ for reading achievement and $\Delta R^2 = .07$ for liking reading. There were some significant differences between the countries at baseline (i.e., among test-takers with zeros in all disengagement predictors, as represented by the coefficients of the countries), suggesting that there were true differences in the scale scores and in achievement even among fully engaged students. However, all |βs| for the survey scores were smaller than .20 except for β = −0.26 in Sweden in the liking reading scale. The effect sizes of the countries on achievement were much stronger, and ranged from −0.08 (Saudi Arabia) to 1.16 (The Russian Federation).

Table 6. Standardized coefficients (β) in a multivariate regression model predicting scale scores from the disengagement measures.

	Digital self-efficacy	School belonging	Bullying	Reading lessons engagement
% Fast screens	−0.02	−0.07	0.26	−0.06
% Slow screens	−0.07	0.02	−0.03	−0.01
% Missing screens	−0.03	ns	ns	ns
% Straight-lining	0.17	0.21	−0.22	0.26
% Misfitting Zh	−0.05	−0.16	0.12	−0.14
% Overfitting Zh	−0.01	−0.02	0.05	−0.01
Adjusted R^2	.08	.16	.12	.21
	Disorderly behavior	Like reading	Confident in reading	Reading achievement
% Fast screens	0.15	−0.04	−0.19	−0.37
% Slow screens	−0.07	−0.05	−0.11	−0.18
% Missing screens	ns	−0.01	−0.03	0.16
% Straight-lining	−0.22	0.29	0.08	−0.03
% Misfitting Zh	0.12	−0.16	−0.09	−0.16
% Overfitting Zh	0.02	−0.01	−0.09	−0.02
Adjusted R^2	.06	.14	.05	.13

Note: Only significant coefficients ($p < .001$) are presented; other results are denoted not significant (ns).

Table 7. The ranges of the countries' aggregated βs when predicting the scale scores.

	Digital self-efficacy	School belonging	Bullying	Reading lessons engagement	Disorderly behavior	Like reading	Confident in reading
% Fast screens	Min: −0.07 (Taiwan), Max: 0.06 (Finland, Sweden)	Min: −0.13 (Russia), Max: −0.01 (Portugal)	Min: 0.15 (Finland, Taiwan), Max: 0.28 (Russia)	Min: −0.13 (Finland, Taiwan), Max: 0 (Russia)	Min: 0.01 (Finland), Max: 0.19 (Spain)	Min: −0.32 (Sweden), Max: −0.06 (Spain, UAE)	Min: −0.29 (Kazakhstan), Max: −0.12 (Sweden)
% Slow screens	Min: −0.13 (Taiwan), Max: 0.02 (Finland, Norway, Sweden)	Min: −0.07 (Russia), Max: 0.04 (Portugal)	Min: −0.11 (Finland), Max: 0.03 (Russia)	Min: −0.11 (Finland), Max: 0.03 (Russia)	Min: −0.29 (Canada), Max: −0.09 (Denmark)	Min: −0.26 (Sweden), Max: 0 (Portugal, Saudi Arabia, Spain, UAE)	Min: −0.25 (Kazakhstan), Max: −0.04 (Sweden)
% Missing screens	Min: −0.12 (Taiwan), Max: 0.03 (Finland, Sweden)	Min: −0.08 (Russia), Max: 0.04 (Portugal)	Min: −0.03 (Finland), Max: 0.11 (Russia)	Min: −0.10 (Finland), Max: 0.04 (Russia)	Min: −0.17 (Canada, Finland), Max: 0 (Italy, Spain, UAE)	Min: −0.26 (Sweden), Max: 0 (Portugal, Saudi Arabia, Spain, UAE)	Min: −0.20 (Kazakhstan), Max: 0.02 (Germany)
% Straight-lining	Min: 0.10 (Taiwan), Max: 0.25 (Finland, Sweden)	Min: 0.13 (Croatia, Taiwan), Max: 0.23 (Portugal)	Min: −0.31 (Finland), Max: −0.17 (Russia)	Min: 0.10 (Finland), Max: 0.24 (Russia)	Min: −0.45 (Finland), Max: 0.11 (New Zealand)	Min: −0.03 (Sweden), Max: 0.24 (Spain)	Min: −0.04 (Taiwan), Max: 0.09 (Sweden)
% Misfitting Zh	Min: −0.22 (Taiwan), Max: −0.03 (Finland)	Min: −0.29 (Russia), Max: −0.17 (Portugal)	Min: 0.03 (Finland), Max: 0.19 (Russia)	Min: −0.31 (Taiwan), Max: −0.21 (Spain)	Min: −0.15 (Norway), Max: 0 (Italy, Spain, UAE)	Min: −0.46 (Sweden), Max: −0.24 (Spain)	Min: −0.31 (Kazakhstan), Max: −0.12 (Sweden)
% Overfitting Zh	Min: −0.08 (Taiwan), Max: 0.07 (Finland, Sweden)	Min: −0.08 (Russia), Max: 0.04 (Portugal)	Min: −0.02 (Finland), Max: 0.12 (Russia)	Min: −0.11 (Finland), Max: 0.02 (Spain)	Min: −0.17 (Finland), Max: 0 (Italy, Spain, UAE)	Min: −0.26 (Sweden), Max: 0 (Portugal, Saudi Arabia, UAE)	Min: −0.16 (Kazakhstan), Max: 0.03 (Sweden)

Note: UAE – United Arab Emirates.

To understand the net effect of each measure in each country, we computed country-specific total effects by summing the coefficients for the main effects (of the disengagement measures) with the country effect and the relevant interaction term. Table 7 presents the ranges of these aggregated coefficients across countries. Examining these interactions reveals some differences in how the disengagement measures predict the scale scores. The largest gap between countries was in the effect of straight-lining on disorderly behavior in reading lessons; this relationship was strong and negative in Finland ($\beta = -0.45$) but positive in New Zealand ($\beta = 0.11$). In another example, while in all countries the percentage of misfitting responses negatively predicted liking reading, this effect was much stronger in Sweden ($\beta = -0.46$) than in Spain ($\beta = -0.24$).

In general, the same countries tended to have minimal or maximal effects within the same scale. For example, considering all measures, Finland had the minimal and the Russian Federation had the maximal effects on the bullying scale. However, these trends were not always consistent across scales; the effects of almost all disengagement measures were the lowest (most negative) in Sweden when predicting liking reading, but highest when predicting confidence in reading and digital self-efficacy.

There were meaningful disengagement-country interactions when predicting reading achievement, as well, such that the association between disengagement and reading achievement varied substantially between countries (see Table 8). Specifically, higher rates of disengagement, especially fast and slow responses, negatively or very weakly predicted achievement in Saudi Arabia (with similar trends in Qatar and Kazakhstan, not presented in the table). However, they were positively associated with achievement in Russia and Singapore, reaching some substantial effect sizes ($\beta > 1$ for the percent of slow, missing, and straight-lined screens).

Table 8. The ranges of the countries' aggregated βs when predicting reading achievement.

	Ranges
% Fast screens	Min: −0.28 (Saudi Arabia) Max: 0.86 (Singapore)
% Slow screens	Min: −0.27 (Saudi Arabia) Max: 1.05 (Russia)
% Missing screens	Min: 0 (Saudi Arabia) Max: 1.24 (Russia)
% Straight-lining	Min: 0.07 (Saudi Arabia) Max: 1.24 (Russia)
% Misfitting Zh	Min: −0.23 (Saudi Arabia) Max: 0.99 (Russia)
% Overfitting Zh	Min: −0.10 (Saudi Arabia) Max: 1.14 (Russia)

5. Discussion

Disengagement in ILSA surveys presents several threats to the scores' validity. In this work, we examined country-based differences in rates of fast, slow, missing, and straight-lined screens, as well as in rates of scale-level person misfit and overfit. We also compared these measures' relationships with survey and achievement scores. Doing so may expand our understanding of disengagement as a threat to score validity and comparability across countries.

We found significant differences among countries in all of the disengagement measures we examined, and they were the most substantial in the percentage of slow screens. These differences do not necessarily mean that these measures are inappropriate for detecting disengagement, as some countries may have more disengaged students than others. Multiple studies on achievement tests in ILSAs reported this (e.g., Goldhammer et al., 2016; Rios & Guo, 2020; Rios & Soland, 2022). This may be due to a true difference in disengagement, for instance, due to variations in the test's perceived value or in difficulties reading and understanding the survey questions, leading to carelessness. Alternatively, the universal cutoffs for disengagement used in this study might miscategorize engaged students if some countries encourage fast, slow, or consistent responses, so they are not necessarily indicative of disengagement.

While we did not directly examine country-level disengagement profiles, some countries' ranks in the disengagement metrics revealed distinct and consistent "types" of disengagement. Some common types were fast and consistent, and slow and inconsistent. These groups somewhat support findings among individuals that demonstrate at least two types of disengagement: rapid guessing, characterized by completing the assessment as fast as possible (e.g., by endorsing the same option on all items), and slow disengagement, characterized by distractedness and randomness (Anghel et al., 2025; Read et al., 2021). However, as most countries did not demonstrate such clear profiles, these groups should not be overinterpreted without further research.

We also found several significant associations between the disengagement metrics and the survey and performance scores. The strongest associations for the survey scores were between straight-lining and liking reading. Further, straight-lining had relatively strong, positive associations with the positive scales and negative associations with the negative scales (bullying and disorderly behaviors in reading lessons). This means that straight-lining participants did not necessarily always agree with the items, but rather endorsed the more socially desirable response option.

A surprising finding was the relatively strong association between fast RTs and experiencing bullying at school; this association emerged in all of the countries in the sample, suggesting that this is likely not a spurious correlation. Under the strong assumption that the scores of fast and inattentive responses reflect true bullying experiences, this link might be explained by lower school engagement, and thus low test-taking engagement, of students who experience bullying (Soland, Jensen, et al., 2019; Yang et al., 2018).

The survey-based disengagement measures, especially the percentage of fast responses, also predicted the achievement scores. Our effect size is quite substantial, and similar to that reported by Zamarro et al. (2019), who used missing and inconsistent survey items to predict reading achievement. These effect sizes are also similar to those reported in studies exploring the link between achievement and disengagement in achievement tests as measured using proactive measures or known groups (i.e., not survey-based measures), but lower than the correlation when disengagement is measured using RT on the assessment itself (see meta-analyses in Silm et al., 2020; Wise & DeMars, 2005). Practically, this means that the measures proposed here can be used in lieu of proactive measures of disengagement that are rarely available in ILSAs, but not instead of measures of effort based on the achievement test itself. However, conceptually, the direction of this relationship remains unclear; it is possible that disengagement is global, such that those who were not engaged with the survey were also not engaged in the achievement test, affecting their performance, that students with low literacy skills are more likely to get frustrated or careless in the survey, or that there is an alternative variable accounting for both (e.g., low school engagement; Soland, Jensen, et al., 2019). Future research could further examine these options, thus informing interventions designed to address test-taking disengagement so that ability estimates are more accurate.

Beyond these general trends in the relationship between disengagement and scores, we found some important cross-country variations. Most of them involved the strength of the relationship, but in some cases, its direction also varied between countries. This was especially notable when predicting achievement, where in some cases, disengagement was positively associated with achievement. Generally speaking, this means that test-taking disengagement is a meaningful factor for international comparability, and accounting for it could affect the conclusions made based on the scores. However, our findings do not reveal the reason behind these differences; Again, they may be attributed to real international differences (for example, in social desirability bias; Keillor et al., 2001) or to issues with the cutoff scores we selected to define disengaged participants. Future studies could address this issue, as well.

It is important to note that our results reflect the behaviors of PIRLS test-takers, namely, fourth graders. At this stage, children are transitioning from learning to read to "reading to learn" (Chall, 1983), and as such they might not yet be proficient readers, potentially affecting their disengagement patterns. For example, they might take longer to respond to items than other readers or might get frustrated more easily, contributing to their disengagement. Although the survey items were designed specifically for this age group and their reading level, the relationships among different engagement indicators might look different among more advanced readers.

Beyond this point, our study has several limitations. First, we were unable to examine proactive measures of disengagement. While they are rarely used in ILSAs, it is still interesting to test their properties in an international sample to better understand test-taking engagement. We also used universal and absolute criteria for identifying fast and slow RT. Some suggest that item-specific or relative criteria (Soland, Jensen, et al., 2019) might be more appropriate, especially in this diverse sample. Finally, we looked at the number of disengagement issues per student (i.e., number of missing screens, straight-lining, etc.) rather than directly categorizing students as engaged or disengaged. In ILSA achievement tests, disengagement is usually identified using dichotomous thresholds, for example, students being disengaged on a certain percentage of items are considered disengaged test-takers (e.g., Pastor et al., 2019; Wise, 2014). Such criteria do not currently exist for survey research (Ulitzsch et al., 2024) but should be developed in the future.

In conclusion, this study shows that disengagement in the survey portion of an ILSA is associated not only with the survey scores but also with the test-takers' reading achievement. We also found that these trends vary across countries, and in some cases, accounting for disengagement could have a substantial impact on scores. Therefore, we recommend that test designers consider and ideally test empirically how specific scales are likely to be impacted by disengagement, and which disengagement measures and cutoffs are the most relevant to specific countries.

Acknowledgments

We want to thank Bethany Fishbein for her help in preparing and interpreting the data and Katherine Reynolds for her helpful comments on an earlier draft of this paper.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Our work was funded by TIMSS 2019 and PIRLS 2021.

ORCID

Ella Anghel http://orcid.org/0000-0001-6332-7826

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Anghel, E., Wry, E., & von Davier, M. (2025). Affective, behavioral, and cognitive engagement in ePIRLS: A latent profile analysis. *Educational Technology Research & Development*, *73*(3), 1–27. https://doi.org/10.1007/s11423-025-10468-z

Ashley, M., & Shaughnessy, K. (2023). Predicting insufficient effort responding: The relation between negative thoughts, emotions, and online survey responses. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement*, *55*(3), 198–209. https://doi.org/10.1037/cbs0000308

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-fit as an index of inattentive responding: A comparison of methods using polytomous survey data. *Applied Psychological Measurement*, *43*(5), 374–387. https://doi.org/10.1177/0146621618798666

Berry, K., Rana, R., Lockwood, A., Fletcher, L., & Pratt, D. (2019). Factors associated with inattentive responding in online survey research. *Personality & Individual Differences*, *149*, 157–159. https://doi.org/10.1016/j.paid.2019.05.043

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2023). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, *26*(2), 323–352. https://doi.org/10.1177/10944281211056520

Bulut, O., Gorgun, G., & Karamese, H. (2023). Incorporating test-taking engagement into multistage adaptive testing design for large-scale assessments. *Journal of Educational Measurement*, *62*(1), 57–80. https://doi.org/10.1111/jedm.12380

Chall, J. S. (1983). *Stages of reading development*. McGraw-Hill.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

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

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. https://doi.org/10.1002/job.1962

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *The British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, *8*, 863. https://doi.org/10.3389/fpsyg.2017.00863

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. *Oecd education working papers*, no. 133. OECD Publishing. https://doi.org/10.1787/5jlzfl6fhxs2-en

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, *52*, 105–119. https://doi.org/10.1016/j.econedurev.2016.02.001

Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., & Park, D. S. (2023). Exhaustive or exhausting? Evidence on respondent fatigue in long surveys. *Journal of Development Economics*, *161*, 102992. https://doi.org/10.1016/j.jdeveco.2022.102992

Johansone, I. (2023). Survey operations procedures in PIRLS 2021. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and procedures: PIRLS, 2021 technical report* (pp.4.1–4.26). TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb8234

Jones, E. A., Wind, S. A., Tsai, C.-L., & Ge, Y. (2023). Comparing person-fit and traditional indices across careless response patterns in surveys. *Applied Psychological Measurement*, *47*(5–6), 365–385. https://doi.org/10.1177/01466216231194358

Keillor, B. D., Owens, D., & Pettijohn, C. (2001). A cross-cultural/cross national study of influencing factors and socially desirable response biases. *International Journal of Market Research*, *43*(1), 1–19. https://doi.org/10.1177/147078530104300101

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kumar, M., Valacich, J., Jenkins, J., & Kim, D. (2022). Too fast? Too slow? A novel approach for identifying extreme response behavior in online surveys. *SIGHCI, 2022 Proceedings*, *4*. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1003&context=sighci2022

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., & Neubrand, M. (2002). *Ger-man scale handbook for PISA 2000*. Max-Planck-Institut für Bildungs-forschung.

List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., & Nagy, G. (2017). Performance decline in low-stakes educational assessments: Different mixture modeling approaches. *Large-Scale Assessments in Education*, *5*(1), 1–25. https://doi.org/10.1186/s40536-017-0049-3

Lumley, T. (2019). *Mitools: Tools for multiple imputation of missing data*. R package version 2.4. https://CRAN.R-project.org/package=mitools

Lundgren, E., & Eklöf, H. (2023). Questionnaire-taking motivation: Using response times to assess motivation to optimize on the PISA, 2018 student questionnaire. *International Journal of Testing*, *23*(4), 231–256. https://doi.org/10.1080/15305058.2023.2214647

Martin, M. O., von Davier, M., Foy, P., & Mullis, I. V. S. (2019). PIRLS, 2021 assessment design. In I. V. S. Mullis, M. O. Martin, & M. O (Eds.), *PIRLS 2021 assessment frameworks* (pp. 57-74). Boston College, TIMSS & PIRLS International Study Center website. https://timssandpirls.bc.edu/pirls2021/frameworks/

National Center for Education Statistics. (2025). Progress in International Reading Literacy Study (PIRLS) methodology and technical notes. Retrieved May 14, 2025, from https://nces.ed.gov/surveys/pirls/pirls2021/technotes.asp#top

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189–212. https://doi.org/10.1080/10627197.2019.1615373

Pools, E. (2022). Not-reached items: An issue of time and of test-taking disengagement? The case of PISA, 2015 reading data. *Applied Measurement in Education*, *35*(3), 197–221. https://doi.org/10.1080/08957347.2022.2103136

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*(1), 10. https://doi.org/10.1186/s40536-021-00104-6

Read, B., Wolters, L., & Berinsky, A. J. (2021). Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys. *Political Analysis*, *30*(4), 550–569. https://doi.org/10.1017/pan.2021.32

Reuning, K., & Plutzer, E. (2020). Valid vs. invalid straightlining: The complex relationship between straightlining and data quality. *Survey Research Methods*, *14*(5), 439–459. https://doi.org/10.18148/srm/2020.v14i5.7641

Reynolds, K. A., Mullis, I. V. S., & Martin, M. O. In Mullis, I. V. S., Martin, M. O., and von Davier, M. (2021). Timss, 2023 context questionnaire framework. Timss, 2023 assessment frameworks. https://timssandpirls.bc.edu/timss2023

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, *33*(4), 263–279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, *22*(2), 154–184. https://doi.org/10.1080/15305058.2022.2036161

Rubin, D. B. (2018). Multiple imputation. In Stef van Buuren (Ed.), *Flexible imputation of missing data, second edition* (pp. 29–62). Chapman and Hall/CRC.

Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 75–95). Chapman and Hall/CRC.

Silber, H., Danner, D., & Rammstedt, B. (2019). The impact of respondent attentiveness on reliability and validity. *International Journal of Social Research Methodology*, *22*(2), 153–164. https://doi.org/10.1080/13645579.2018.1507378

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335

Soland, J., Jensen, N., Keys, T. D., Bi, S. Z., & Wolk, E. (2019). Are test and academic disengagement related? Implications for measurement and practice. *Educational Assessment*, *24*(2), 119–134. https://doi.org/10.1080/10627197.2019.1575723

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

Soland, J., Zamarro, G., Cheng, A., & Hitt, C. (2019). Identifying naturally occurring direct assessments of social-emotional competencies: The promise and limitations of survey and assessment disengagement metadata. *Educational Researcher*, *48*(7), 466–478. https://doi.org/10.3102/0013189X19861356

Steyn, R. (2017). How many items are too many? An analysis of respondent disengagement when completing questionnaires. *African Journal of Hospitality, Tourism & Leisure*, *6*(2), 1–11.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99–114. https://doi.org/10.2307/3001913

Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2024). Accounting for careless and insufficient effort responding in large-scale survey data-development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*, *56*(2), 804–825. https://doi.org/10.3758/s13428-022-02053-6

von Davier, M., & Molenaar, I. W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika*, *68*(2), 213–228. https://doi.org/10.1007/BF02294798

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*, 1–17. https://doi.org/10.7333/1401-02010001

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., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. K. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204–220). Routledge.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (In ETS Research Report RR89-41). Educational Testing Service.

Yang, C., Sharkey, J. D., Reed, L. A., Chen, C., & Dowdy, E. (2018). Bullying victimization and student engagement in elementary, middle, and high schools: Moderating role of school climate. *School Psychology Quarterly*, *33*(1), 54–64. https://doi.org/10.1037/spq0000250

Yin, L., & Reynolds, K. A. (2023). Creating and interpreting the PIRLS, 2021 context questionnaire scales. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and procedures: PIRLS, 2021 technical report* (pp..15.1–.15.161). TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb6994

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