International Journal of Educational Research Open 7 (2024) 100373

Contents lists available at ScienceDirect

International Journal of Educational Research Open

journal homepage: www.elsevier.com/locate/ijedro

Item- and person-level factors in test-taking disengagement: Multilevel modelling in a low-stakes context

Róbert Csányi [a,*], Gyöngyvér Molnár [b]

[a] Doctoral School of Education, University of Szeged, Petőfi sgt. 32–34, Szeged H–6722, Hungary
[b] Institute of Education, University of Szeged, MTA–SZTE Digital Learning Technologies Research Group, Szeged, Hungary

ARTICLE INFO

Keywords:
Test-taking effort
Logfile analysis
Response time
Test-taking disengagement
Multilevel modeling

ABSTRACT

The present study examines item- and person-level factors that influence test-taking disengagement. Computer-based measurement of complex problem-solving was used to eliminate the effect of factual knowledge on test performance among first-year university students in a low-stakes context. Due to the hierarchical structure of the data, multilevel modeling was used to identify item- and person-level factors that influence test-taking disengagement. Results suggested that item position and item difficulty have a significant effect on test-taking disengagement. Items presented later in test administration as well as more difficult items had a higher probability of disengaged responses. Mother's education had no significant effect on the rate of disengaged responses, while a higher proportion of disengaged responses was recorded among women. The percentage of disengaged responses was also greater among those with lower entrance scores, lower working memory capacity and lower self-reported effort (SRE). To sum up, the results suggest a relationship between the level of academic ability and test-taking disengagement, which determines how disengaged responses are treated.

1. Introduction

Students' performance on cognitive tests can be influenced by a number of affective factors, including test-taking motivation, in addition to their actual knowledge and skills (Wise et al., 2014). Several studies have shown that test performance among unmotivated students is significantly lower than that of their motivated peers (Penk et al., 2014; Silm et al., 2020; Wise et al., 2021). Akyol et al. (2021) argue that the bias in the PISA measurement due to unmotivated students' responses is significant and that only half of the bias is corrected for in the data analysis. The stakes of the tests have a significant influence on test-takers' motivation to complete the test: as the stakes increase, the effort exerted increases; however, so does the likelihood that test-takers will use unethical means or that anxiety may have a negative impact on their performance. As the stakes and the role of the test decrease, motivation to complete the test may decrease proportionally, thus potentially affecting test-takers' performance (Rios, 2021).

Research suggests that test-taking effort is influenced by a number of factors (e.g. Rios & Soland, 2022). These factors can be divided into three categories: item-related, test situation-related and test-taker-related. Research has found some contradictory results, for example, on the relation between test-taking effort and ability levels (Deribo et al., 2021; Wise & Kong, 2005). Our research was motivated by the partially inconsistent results of the research.

Data on test-taking effort have a hierarchical structure. Students' item-level answers are nested in individual level and are logically interconnected. This interdependency, the multilevel (item- and person-level) feature of the data, is often ignored in test-taking effort analyses. We fill this gap and use a multilevel framework to broaden our understanding of the phenomenon of test-taking motivation by analysing the effects of variables at different levels and how they interact (Sommet & Morselli, 2021).

1.1. Test-taking effort

A widely used model for explaining test-taking motivation is expectancy-value theory (Eccles & Wigfield, 2002; Wise & DeMars, 2005). This theory posits that a person's motivation is a function of expected performance and the value of the test. Examinees' expectations are influenced by (1) their perception of their own abilities and (2) the difficulty of the tasks. Values have four components: the attainment value, which is the importance of the test; the intrinsic value, measured

* Corresponding author.
E-mail address: csanyi.robert@edu.u-szeged.hu (R. Csányi).

https://doi.org/10.1016/j.ijedro.2024.100373
Received 27 February 2024; Received in revised form 28 June 2024; Accepted 28 June 2024
Available online 5 July 2024

by the pleasure of completing the task; the utility value, which is the relation of the task to future goals; and the cost, determined by the time spent on the task or anxiety about the tests. Test-taking motivation is manifested in the effort the examinee puts into doing the test, which is defined as the quantity of resources used to achieve the highest possible score.

Various methods can be used to measure test-taking effort. *Self-report questionnaires* were initially used, generally measuring the test-taking effort components on a Likert scale. Students are usually asked to rate their effort in doing the test after finishing it. This approach assumes that students' test-taking effort represents a constant value, while a number of studies have found that test-taking effort tends to decrease during the test (Attali, 2016; Goldhammer et al., 2016; Penk & Richter, 2017; Wise et al., 2009). Changes in test-taking effort can also be tracked by asking the same questions more than once during the test. However, it is not possible to measure test-taking effort after each task because asking students to answer related questions too many times will in itself reduce test-taking effort. An important advantage of self-report questionnaires is that they are easy to use in traditional paper-and-pencil testing and easy to evaluate. Among their many limitations, they are subjective and there is no way of knowing how honest the test-takers' responses were, as they can be influenced by many factors (Wise & Kong, 2005).

The expansion of computer-based assessments has made it the basis for the development of *response time-based methods*. Response time is the time the test-taker spends on a given task from the time the task is presented until the "next" button is clicked. Response time-based methods assume that disengaged participants spend less time on tasks and therefore respond faster than their engaged counterparts (Wise & Kong, 2005). One of the main advantages of response time-based methods is that the actual behavior of the examinees is measured, not their perceptions. An additional advantage is that it does not require extra work for the examinee and changes in motivation can be tracked from item to item (Wise & Ma, 2012). The first step in applying response time-based methods is to define a threshold in a certain way. As a second step, if the response time is shorter than the threshold, the response is identified as disengaged; if it is longer, it is identified as engaged (Wise & Kong, 2005). The simplest way to determine the threshold is to use a predefined threshold (e.g. 3 or 5 s) for each item, called a constant threshold. However, this method can be biased, as the minimum time required to complete certain tasks is different from item to item. Therefore, item-specific thresholds have been introduced, which means that the threshold differs from item to item (Goldhammer et al., 2016). In this study, both self-report questionnaires and a response time-based method were used to measure test-taking effort.

1.2. Factors influencing test-taking effort

Research has identified several factors that influence test-taking effort. These factors are related to the items, the test situation and the test-takers. By modifying these factors, test-taking effort can be significantly influenced.

1.2.1. Item-related factors

Item position. Various research results indicate that test-taking effort tends to decrease during the test (Attali, 2016; Nuutila et al., 2021; Penk & Richter, 2017; Wise et al., 2009). Multistage testing design is used in the most important large-scale assessments to eliminate the item position effect (Buchholz et al., 2022; Goldhammer et al., 2016).

Item difficulty. Research has found that test-taking effort generally decreases as item difficulty increases (Lindner et al., 2017; Pools & Monseur, 2021). Another approach is that test-takers put more effort into completing tasks that match their ability levels, i.e. tasks that are neither too difficult nor too easy (Asseburg & Frey, 2013). The optimal challenge provided by adaptive testing is based on ability-matched items, thus providing a flow experience for test-takers (Molnár, 2021).

Item type. Students demonstrate greater test-taking effort on selected-response items than on constructed-response items (DeMars, 2000; Guo et al., 2022; Michaelides & Ivanova, 2022) because the latter are more cognitively demanding (Lindner et al., 2020).

Item length. In the case of longer item stems, students show less test-taking effort (Setzer et al., 2013; Wise et al., 2009), which can also be explained by cognitive load (Wise, 2006).

Illustrations. The use of representational pictures or illustrations increases students' test-taking effort (Lindner et al., 2017; Lindner, 2020). These schematic pictures represent the task and illustrate the important information provided in the text but do not offer any additional information beyond what is supplied in the text (Lindner et al., 2017). In contrast to representational pictures, seductive details are entertaining and interesting but not relevant to the task (Eitel et al., 2020). Therefore, the use of representational pictures and the reduction of seductive details improves test-taking effort.

1.2.2. Factors related to the test situation

Stakes of the test. The stakes of a test indicate the consequences for the test-taker of their test performance (Wise, 2006). Low-stakes tests have no significant consequences for a person's academic performance, while high-stakes tests have significant consequences (Lindner et al., 2019). Low-stakes tests are often correlated with lower test-taking motivation (Wise et al., 2014).

Time of testing. Wise et al. (2010) investigated the effects of testing time related to test-taking effort. They found that test-taking effort decreased within a given day; that is, it was higher in the morning than in the afternoon. However, there was no difference in test-taking effort depending on when testing took place within a year. Test-taking effort also did not vary depending on which day of the week the testing took place.

Motivational instructions. Low-stakes tests have no significant consequences for students but may have significant consequences at the institutional or national level. Test-taking effort increased when invigilators made students aware that test scores have significant institutional relevance (Liu et al., 2012, 2015).

Monetary incentives. Various studies have shown that the use of monetary incentives increases test-taking effort as well as test performance (Braun et al., 2011; Wise & DeMars, 2005). Their use also appears in international large-scale assessments, for example, in the Programme for the International Assessment of Adult Competencies (PIAAC), where participating member countries can decide to use them (Martin et al., 2014). Rios (2021) conducted a meta-analysis of data from 53 studies to investigate the methods used to increase test-taking effort. He concluded that the use of financial incentives has the greatest impact. The disadvantages of using monetary incentives are that they are costly and unlikely to have the same motivational effect on examinees from different financial backgrounds (Lau et al., 2009).

1.2.3. Person-related factors

Ability level. Several studies have investigated the relation between ability levels and test-taking disengagement. Most studies concluded that test-taking disengagement was unrelated to ability levels (Kong et al., 2007; Rios et al., 2014; Wise & DeMars, 2005; Wise & Kong, 2005), but some suggest there is a relation (Deribo et al., 2021; Rios et al., 2017b).

Working memory capacity. There is hardly any literature on the relation between test-taking disengagement and working memory capacity. Lindner et al. (2019) investigated the factors influencing test-taking effort among fifth- and sixth-grade German students in a low-stakes science test context. It was observed that participants with a higher working memory capacity had a higher test-taking effort.

Educational attainment. In PIAAC, lower educational attainment was associated with lower test-taking effort (Goldhammer et al., 2016, 2017; Wang et al., 2023).

Gender. Various research results indicate that women are characterised by higher test-taking effort than men (Goldhammer et al., 2016;

Wise & DeMars, 2010). According to DeMars et al. (2013), the gender gap is not evenly distributed. More men are at the low end of the effort scale than at the higher end; that is, more men with extremely low effort were found among test-takers, while there was not such a difference in effort levels among women. However, not all studies demonstrated a significant relationship between gender and test-taking disengagement (Lindner et al., 2019; Wise et al., 2009).

Age. Test-taking effort tends to decrease with age. This trend can be observed for a number of age groups. Rosenzweig et al. (2019) showed a decrease in motivation among K–12 students. Juniors and seniors have lower test-taking effort than freshers and sophomores (Rios & Guo, 2020). In the measurement of adult competencies (PIAAC), older age groups were characterised by lower test-taking effort (Goldhammer et al., 2016).

Ethnicity. Ethnic minorities tend to show lower test-taking effort than the majority (Soland, 2018; Wise et al., 2021). This effect was demonstrated by Wise et al. (2021) on tests taken by eighth-grade students in maths, English and science and by Soland (2018) on MAP Growth tests taken by fifth- to ninth-grade students.

Native language. Test-taking effort is lower for test-takers whose native language is different from the test language (Deribo et al., 2021; Goldhammer et al., 2017; Rios & Soland, 2022).

1.3. Research purpose, questions and hypotheses

Research results suggest that test-taking effort depends on many factors. Some factors are under-researched, such as working memory capacity, while studies have found contradictory results for other factors, such as item difficulty, ability level and gender. Furthermore, many studies have not taken into account the multilevel nature of data, as they have focused on either the item or the person being studied.

To address these limitations, the objective of this study was to model disengaged responses observed in the evaluation using hierarchical linear models as a function of characteristics at the level of items and individuals. We investigated students' test-taking effort with self-report and log data-based methods using interactive tasks and situations in which already existing factual knowledge could not be used during the problem-solving process. Students' test-taking effort was measured with the $P+>0$ % time-on-task method and by asking students to rate their test-taking effort. These objectives were addressed via the following research questions, with the following hypotheses being formulated:

RQ1: How much of the variation in disengaged responses can be detected at the item and person levels?
H1: Research suggests that disengaged responses are associated partly with items and partly with test-takers (Rios & Soland, 2022). We thus hypothesised that multilevel modeling would be warranted.
RQ2a: Can we define item-level factors which result in disengaged responses?
H2a: Research has identified various item-level factors that influence test-taking disengagement, such as item position, item difficulty, item type, item length and illustrations (Attali, 2016; Guo et al., 2022; Lindner et al., 2020; Pools & Monseur, 2021; Wise, 2006). We thus hypothesised that there would be item-level factors that influence test-taking disengagement.
RQ2b: Which item-level factors are predictive of disengaged responses?
H2b: In our research, we examined two item-level factors: item position and item difficulty. According to several studies, test-takers exhibited higher test-taking disengagement for later tasks and for more difficult tasks (e.g. Penk & Richter, 2017; Pools & Monseur, 2021). Based on these findings, we hypothesised that both factors would be predictive of disengaged responses.
RQ3a: Can we determine person-level factors which result in disengaged responses?
H3a: Based on the research, there are several person-level factors that influence test-taking disengagement, such as ability level, working memory capacity, educational attainment, gender, age, ethnicity and native language (Goldhammer et al., 2016; Lindner et al., 2019; Rios et al., 2014; Soland, 2018). We thus hypothesised that there would be person-level factors that influence test-taking disengagement.
RQ3b: Which person-level factors are predictive of disengaged responses?
H3b: Our research investigated five person-level factors. A majority of studies have found that men show higher test-taking disengagement (e.g. Wise & DeMars, 2010); hence, this is what we hypothesised. Mother's education level was an indicator of family background. We have found no research on the effect of mother's education on test-taking disengagement, but several studies suggest that it has an effect on academic performance (e.g. Csapó & Molnár, 2017). We hypothesised that test-takers with disadvantaged family backgrounds would demonstrate higher test-taking disengagement. We used the entrance score as a proxy for academic ability. Several studies have investigated the relationship between academic ability and test-taking disengagement. The results are contradictory, but more recent research suggests that test-taking disengagement is related to academic ability (e.g. Deribo et al., 2021). Therefore, we hypothesised that people with lower ability levels would exhibit higher test-taking disengagement. Based on Lindner et al. (2019) research, we hypothesised that people with lower working memory capacity would have higher test-taking disengagement. Research (e.g. Silm et al., 2020) has indicated that self-reported effort (SRE) correlates with response time-based effort, so we hypothesised that students who rate their effort higher would have lower test-taking disengagement.

2. Materials and methods

2.1. Participants

The sample consisted of first-year undergraduate students who were commencing their studies at one of the largest Hungarian universities. The assessment took place just after the start of their studies. The university has twelve faculties (e.g. faculties of humanities and social sciences, natural sciences, law and medicine), all of which were included in the assessment. All full-time, first-year students were informed of the details before the assessment via the university's learning management system. Participation was voluntary, but students who successfully completed the test received one credit as an incentive. Students who participated in the assessment were assigned to a specific course, Career Development. This was due to the administrative requirements of the university. A total of 1751 students (46.2 % of the target population) participated in the study (mean age = 19.80, SD = 1.92), 53.0 % of them being female.

2.2. Data collection procedure

The assessment was administered via the eDia system (Csapó & Molnár, 2019) and conducted in the main computer room of the university learning and information center. Test administration was supervised by invigilators. Students were allowed to choose their own schedule, so the number of participants varied between 10 and 150 at each session. Students who registered for the assessment were required to attend two-hour sessions in which they completed a complex problem-solving test and other cognitive tests related to learning. At the beginning of the test, participants were introduced to the user interface and given a warm-up exercise. After signing into eDia, students were given 60 min to complete all the tasks and the questionnaire. If they used up the full 45 min on the problem-solving activities, they still had 15 min left for the questionnaire. Students received immediate feedback on

their average performance after completing the test as well as detailed feedback a week later.

The study rigorously conformed to the regular standards of approved research ethics. The research was approved by the University of Szeged Doctoral School IRB (No. 11/2023). However, (1) the data collection was an integral part of the educational processes at the university, (2) participation was voluntary, (3) all of the students in the assessment had turned 18, and (4) all of the participants confirmed with their signature that they understood that their data would be used for educational and research purposes at both the faculty and university levels.

2.3. The problem-solving tasks

We used a complex problem-solving test based on the MicroDYN approach. These tasks center on fictional situations and are thus independent of the impact of previous school learning (Funke, 2014; Greiff et al., 2013). MicroDYN has proved to be a reliable and effective method for evaluating complex problem-solving (Greiff et al., 2013, 2018; Molnár & Csapó, 2018).

The tasks are divided into two phases: the knowledge acquisition phase and the knowledge application phase (Greiff et al., 2013). In the first phase, students were asked to work out the relationships between the variables. They were expected to change the values of the input variables (e.g. two different kinds of paint) and then observe the effect of the changes on the values of the output variables (the color of the paint). It was possible to carry out this process several times because the number of clicks was unlimited in this phase, but the time available was a maximum of 180 s. Based on the information collected and interpreted, the relationships between input and output variables were drawn on the concept map displayed on the screen (Molnár & Csapó, 2018). In the second phase, based on the information obtained, students were asked to reach the predefined values of the output variables by changing the values of the input variables. In the second phase of the test, they were given a time limit of 90 s, with a maximum of four trials, i.e. four ways to configure the input variables. The test consisted of ten increasingly complex tasks, i.e. more and more input and output variables and an increasing number of relations. The reliability of the tasks was good ($\alpha = 0.88$).

In this study, we focused on data collected during the first phase of the problem-solving process, as we were less limited by the maximum time, an important indicator of test-taking effort. Consequently, the time data differed between students to a greater extent than the log data collected in the second phase of the problem-solving process.

2.4. Data collected

Two distinct methodologies were integrated to quantify test-taking effort: the questionnaire-based self-report design and the time-on-task-based approach. Students were requested to evaluate their test-taking effort (*self-reported effort; SRE*) based on a statement ("I put a lot of effort into the tasks") using a five-point Likert scale ranging from 1 (not true at all) to 5 (completely true). Previous research has shown that test-taking effort decreases during the test (Penk & Richter, 2017; Wise et al., 2009); therefore, we administered the self-report questionnaire six times during the cognitive examination to obtain a more accurate value. The initial assessment was conducted after the warm-up task, followed by four subsequent evaluations after every other problem scenario and finally after the last problem.

In response to time-based methods, the metric measured refers to the amount of time spent by the respondent on a given task, usually referred to as time-on-task. If the response time for an item is less than the threshold, it is considered a non-effortful response. If greater than or equal to the threshold, it is considered an effortful response. Based on the work of Wise and Kong (2005), the following relationship is used to measure the *disengaged response* associated with item i and examinee j:

$$\textit{disengaged response}_{ij} = \begin{cases} 1, & \textit{if } RT_{ij} < T_i \\ 0, & \textit{if } RT_{ij} \geq T_i \end{cases}, \quad (1)$$

where T_i = threshold value for item i and RT_{ij} = response time for item i and examinee j.

In our study, we applied the *proportion correct greater than zero* (*P+>0* %) method. For assessments using the multiple-choice format, it is worth noticing that the probability of a correct answer is indeed greater than zero due to random guesses being taken into account. Specifically, for an item with five possible answers, the probability of a correct answer is about 0.2. In cases where examinees do not choose from a set of options but have to construct their own answers, the probability of randomly selecting the correct answer is zero. To determine the threshold *P+>0* %, the responses are sorted in increasing order of the time taken to respond. The threshold is defined as the shortest response time at which the first correct answer is obtained (Goldhammer et al., 2016).

2.5. Variables

The variables included in the analysis were those that have been found to influence test-taking effort in previous studies. These factors are discussed separately at the item and test-taker levels. The testing conditions were constant, so test situation-related factors were not included in the analysis. Table 1 presents the characteristics of the variables included in the analysis.

2.5.1. Item-related variables

Item position. To examine the effect of item position, the sequence number of items within the test was coded as an independent variable. Item position ranged from one to ten.

Item difficulty. In this study, item difficulty was calculated by dividing the correct responses for a given item by the total responses, so the higher the value, the easier the question. Item difficulty ranged between 0.273 and 0.824.

Table 1
Variables included in the analysis.

Variables	Level	Description	Values	Measurement
Disengaged response	Item	Outcome variable Disengaged responses for a given item and test-taker	0, 1	Dichotomous
Item position	Item	The sequence number of items within the test	1–10	Scale
Item difficulty	Item	Rate of correct responses for a given item out of total responses	0–1	Scale
Gender	Person	Demographic predictor variable representing students' gender.	0 = male 1 = female	Dichotomous
Mother's education	Person	Demographic predictor variable representing mothers' education.	0 = ISCED 0–1 1 = ISCED 2 2 = ISCED 3–5 3 = ISCED 6–8	Ordinal
Entrance score	Person	Students' entrance score	280–500	Scale
Working memory capacity	Person	Students' visual memory capacity	0–16	Scale
Self-reported effort	Person	Students' self-reported effort, SRE	1–5	Scale

Notes: Item = Level 1; Person = Level 2.

2.5.2. Person-related variables

Gender. Student's self-reported gender was coded as a dichotomous variable to examine the effect of gender (male = 0; female = 1).

Family background. Family background was represented by mother's education level. To our knowledge, the effect of mother's education on test-taking effort has not been investigated, but several studies suggest that it has an effect on academic performance (Csapó & Molnár, 2017; Rodríguez-Hernández et al., 2020). We therefore included it in the analysis. Mother's education was coded as: 0 = ISCED 0–1; 1 = ISCED 2; 2 = ISCED 3–5; 3 = ISCED 6–8.

Entrance score. In Hungary, the entrance score is partly based on academic results and partly on the results of the Matura examinations. The entrance score was included as a continuous variable and ranged from 280 to 500, with a mean of 399.52 (SD = 47.45).

Working memory. In this study, working memory capacity was measured using visual memory tasks, ranging from 0 to 16, with a mean of 9.98 (SD = 3.17).

Self-reported effort. A Likert-scale questionnaire related to students' effort ranges from 1 to 5, with a mean of 4.31 (SD = 0.93).

2.6. Data analysis: multilevel modeling

Multilevel data means that data structures are "nested". In multilevel modeling, variables can be identified at any level of the hierarchy. The lowest level (Level 1) is typically the level of individuals. Therefore, in educational research, we mostly investigate students who attend different classes or schools. In hierarchical data structures, the individual observations are usually not independent. For example, pupils at the same school are generally more similar to each other compared to other students, due to the selection processes and the impact of the school. As a result, traditional statistical methods are biased, which can be addressed by multilevel modeling (Hox et al., 2017).

In our research, item-level variables were included at Level 1 and student-level variables at Level 2 by fitting a two-level random-intercepts model. We did this by nesting the disengaged responses to item *i* within examinee *j* for Y_{ij}, which is a dichotomous variable where 1 = disengaged response and 0 = effortful response. The theoretical equations were as follows:

$$Level\ 1:\ Y_{ij} = \beta_{0j} + \varepsilon_{ij} \tag{2}$$

$$Level\ 2:\ \beta_{0j} = \gamma_{00} + \mu_{0j} \tag{3}$$

$$Combined:\ Y_{ij} = \gamma_{00} + \mu_{0j} + \varepsilon_{ij} \tag{4}$$

In Eq. (2), the disengagement of item *i* in student *j* (Y_{ij}) can be modelled as a function of the mean disengagement for student *j* (β_{0j}) plus a residual term that reflects individual item differences around the mean of student *j* (ε_{ij}). In Eq. (3), the mean disengagement for student *j* (β_{0j}) is modelled as a function of a grand-mean disengagement (γ_{00}) plus a student-specific deviation from the grand mean (μ_{0j}). Substituting Eq. (3) into Eq. (2) yields the combined multilevel equation (Hox et al., 2017; Peugh, 2010).

As a first step in the analysis, we built an empty model (with no predictor variables) with twofold objectives: first, to determine how much of the variation in the output variable is associated with item and person level, and, second, to decide whether multilevel modeling is really needed. Based on the model, we calculated the intraclass correlation coefficient (ICC) (Hox et al., 2017; Sommet & Morselli, 2021).

$$ICC = \frac{Between-cluster\ variance}{Total\ variance} = \frac{var\left(\mu_{0j}\right)}{var\left(\mu_{0j}\right) + var\left(\varepsilon_{ij}\right)} \tag{5}$$

As shown in the equation above, the ICC corresponds to the proportion of the variance between test-takers $var\left(\mu_{0j}\right)$ in the total variance $var\left(\mu_{0j}\right) + var\left(\varepsilon_{ij}\right)$. The ICC represents the degree of similarity of observations belonging to the same test-taker and can vary between 0 and 1. A value of 0 indicates that test-taking effort is completely independent of the test-takers: all test-takers put in the same amount of effort; that is, there is no difference between them. A value of 1 indicates perfect interdependence among test-takers. In this case, the observations are completely dependent on test-takers: a given test-taker exerts the same effort on all items; that is, there is no variation between items (Peugh, 2010; Sommet & Morselli, 2021). An ICC value of 0.01 can be interpreted as small homogeneity among test-takers, 0.05 as medium and 0.20 as high (Sommet & Morselli, 2021).

Another metric for deciding whether the multilevel model is justified is the design effect (DEFF):

$$DEFF = 1 + (n-1)\cdot ICC \tag{6}$$

where *n* is the average number of items (Sommet & Morselli, 2021). DEFF is a measure of how different a multilevel sample is from a simple random sample. DEFF can vary between 1 and *n*, from no difference to a maximum difference. When DEFF exceeds 1.5, the use of a hierarchical structure is reasonable (Lai & Kwok, 2015).

The random intercept model was the most appropriate after testing various models:

$$Y_{ij} = \gamma_{00} + \gamma_{01} mother\_edu_j + \gamma_{02} gender_j + \gamma_{03} entrance\_score_j + \gamma_{04} WM_j + \gamma_{05} SRE_j + \gamma_{10} item\_position_{ij} + \gamma_{20} item\_diff_{ij} + \mu_{0j} + \varepsilon_{ij} \tag{7}$$

where $mother\_edu_j$ represents mother's education, $gender_j$ is a dummy-coded variable of the examinee's gender, $entrance\_score_j$ is the student's entrance score, WM_j is the student's working memory, SRE_j is the student's self-reported effort, $item\_position_{ij}$ is the sequence number of items on the test, and $item\_diff_{ij}$ is the rate of correct responses for a given item out of total responses.

3. Results

3.1. Results for research question 1 (RQ1): how much of the variation in disengaged responses can be detected at the item and person levels?

The ICC value for the degree of similarity of observations for the same test-taker was 0.227, meaning that 22.7 % of the variance in test-taking disengagement occurs between students. DEFF = 3.043 was above 1.5, meaning that multilevel modeling was justified.

3.2. Results for research questions 2a and 2b (RQ2a and RQ2b): can we define item-level factors which result in disengaged responses? Which item-level factors are predictive of disengaged responses?

Two item-level predictors were included in the analysis to investigate influencing factors in test-taking disengagement. Both item-level predictors, item position and item difficulty, were shown to be significant for test-taking disengagement. There is a positive relation between item position and test-taking disengagement; that is, the later the items, the greater the disengagement. Due to the definition of item difficulty, it takes a lower value for more difficult items. This means that the value for disengagement is higher for the more difficult items (Table 2).

Table 2
Item-level predictors of disengaged responses.

Item-level predictors	Estimate	SE	p
Item position	0.004	0.001	< 0.001
Item difficulty	- 0.048	0.008	< 0.001

Note: Item = Level 1.

3.3. Results for research questions 3a and 3b (RQ3a and RQ3b): can we determine person-level factors which result in disengaged responses? Which person-level factors are predictive of disengaged responses?

To investigate influencing factors in test-taking disengagement, five person-level predictors were included in the analysis. Among the person-level predictors, mother's education had no significant effect on test-taking disengagement, but the effects of gender, entrance score, working memory and self-reported effort were considered to be significant. Higher levels of disengagement were found among females, students who scored lower on the Matura exam and working memory tasks, and those who rated their effort lower (Table 3).

4. Discussion

The main aim of this study was to investigate item- and examinee-level predictors of test-taking disengagement. In the context of complex problem-solving assessment among first-year students, results suggest that test-taking disengagement is an existing issue because 11.7 % of test-takers demonstrated test-taking disengagement in at least one case and 2.3 % of items were disengaged. The percentage of disengaged responses was relatively low (Lee & Chen, 2011; Rios & Soland, 2022; Wise et al., 2021). The reason for this is presumably that, although it was a low-stakes test, it was administered at the start of participating students' university studies and they were curious about their strengths and weaknesses. It follows that item- and examinee-level predictors also showed low values.

Research question 1 (RQ1). *How much of the variation in disengaged responses can be detected at the item and person levels?*

The measure of variance between examinees, as represented by the ICC, is 0.227. This means that 22.7 % of the variance in test-taking disengagement was explained by variance between students. According to Sommet and Morselli (2021), this represents a high level of homogeneity among examinees. This high degree of homogeneity can be explained by the relatively low level of test-taking disengagement, with the majority of students (88.3 %) exhibiting completely engaged test-taking behavior.

Data on examinees' test-taking effort are hierarchically structured. The item-level responses are nested in the individual level of the test-takers and are logically interconnected. The multilevel nature of data is often ignored when analysing test-taking effort. Our results suggest that the use of multilevel modeling is warranted. Understanding the phenomenon of test-taking disengagement can be enhanced by analysing the impact of different levels of variables and their interactions.

Research questions 2a and 2b (RQ2a and RQ2b). *Can we define item-level factors which result in disengaged responses? Which item-level factors are predictive of disengaged responses?*

In our research, disengagement increased as *item position* increased. Various studies have investigated possible changes in motivation during testing and their impact on examinees' test-taking engagement. Test-taking motivation may increase or decrease during the test. Increases can be interpreted as flow (Csikszentmihalyi, 2014). In a low-stakes testing context, however, it is more likely that test-takers show a decrease in motivation (Attali, 2016; Nuutila et al., 2021; Penk & Richter, 2017; Wise et al., 2009). The decline is well interpreted by the process model of self-control depletion (Inzlicht et al., 2014). According to the model, people want to reach an optimal balance between "have-to" and "want-to" goals. "Have-to" goals refer to duties that must be performed. In contrast, "want-to" goals refer to relaxing activities that we like to do. After hard work over a period of time, motivation changes from "have-to" goals to "want-to" goals. This model is supported by various research. Lindner et al. (2018) investigated changes of state self-control capacity and test-taking effort during a test. The researchers observed that a decrease in state self-control capacity correlated with a decrease in test-taking effort over the course of the test. In another study, decreased self-control capacity among students during testing was associated with increased fatigue (Lindner et al., 2019). In a different study, Lindner and Retelsdorf (2019) found that students who reported high self-control depletion on a given test were less motivated to work on the next test. These results suggested that focusing attention during testing requires self-control, which can lead to mental fatigue, which is closely related to changes in test-taking effort.

Several studies have examined the effects of *item difficulty* on test-taking disengagement. Rios and Guo (2020) examined critical thinking in four countries on a 45-minute computer-based assessment consisting of 26 multiple-choice items. Examinees showed higher levels of disengagement for items with higher perceived difficulty. Analysing data from the Canadian sample of the PIAAC Cycle 1, Goldhammer et al. (2017) demonstrated a positive effect between item difficulty and test-taking disengagement. Barry and Finney (2016) investigated test-taking effort in low-stakes contexts across five consecutive tests. The first difficult cognitive test was followed by non-cognitive and affective measures. Self-reported test-taking effort was lowest for the first test, which was the longest and most difficult test. A plausible reason for this tendency is that, because of the low probability of success, examinees may tend to become unmotivated when they are faced with difficult tasks (Schunk et al., 2008). According to other research, test-takers put more effort into completing a test that matches their abilities, that is, one that is neither too difficult nor too easy (Asseburg & Frey, 2013). This can be explained by the flow, as tasks that are too easy are not challenging and tasks that are too difficult are too challenging (Csikszentmihalyi, 2014). Our results were in line with the research, with the proportion of disengaged responses increasing as the test progressed.

The increase in test-taking disengagement during the test and higher disengagement on more difficult items implies that more attention should be paid to developing low-stakes tests. Research has identified a number of interventions that can be used to motivate academically unmotivated students. Rios (2021) classified these factors into four main categories: (1) modifying test design, (2) providing feedback, (3) modifying test relevance and (4) providing external incentives. Test design can be modified by presenting test-takers illustrations (Lindner et al., 2017), as well as tasks that are moderately difficult (Pools & Monseur, 2021), not too mentally taxing (DeMars, 2000) and intrinsically interesting (Attali & Arieli-Attali, 2015). Giving feedback increases test-takers' motivation if it is timely and relevant (Wise & DeMars, 2005). The relevance of tests can be modified by increasing the stakes of the test, but this can also lead to cheating and anxiety (Wise & DeMars, 2005). Another approach is for invigilators to make students aware of the institutional importance of test performance (Liu et al., 2015). In a meta-analysis of data from 53 studies, Rios (2021) observed that the use of financial incentives has the greatest impact on increasing motivation.

Research questions 3a and 3b (RQ3a and RQ3b). *Can we determine person-level factors which result in disengaged responses? Which person-level factors are predictive of disengaged responses?*

Various research results indicate that males show greater test-taking disengagement than females. Wise and DeMars (2010) examined

Table 3
Person-level predictors of disengaged responses.

Person-level predictors	Estimate	SE	p
Gender	0.013	0.004	0.002
Mother's education = 0	- 0.008	0.059	0.886
Mother's education = 1	0.029	0.017	0.090
Mother's education = 2	0.011	0.012	0.365
Mother's education = 3	0.013	0.012	0.293
Entrance score	- 0.001	< 0.001	< 0.001
Working memory	- 0.002	0.001	0.014
Self-reported effort	- 0.011	0.002	< 0.001

Note: Person = Level 2.

test-taking efforts among first- and second-year university students using a low-stakes oral communication test. Test-taking disengagement was greater for male students than for their female peers in both grades. In a critical thinking assessment, males demonstrated higher rates of disengagement than females (Rios & Guo, 2020). According to DeMars et al. (2013), the gender gap is not evenly distributed. More men are at the low end of the effort scale than at the higher end; that is, more men with extremely low effort were found among test-takers, while there was not such a difference in effort levels among women.

Not all studies have shown a significant relationship between *gender* and test-taking disengagement. Lindner et al. (2019) investigated test-taking effort on a scientific literacy test among fifth- and sixth-grade students in Germany. The link between gender and test-taking effort was not significant. Wise et al. (2009) employed a natural world assessment test to assess the quantitative and scientific reasoning proficiencies of university students in a low-stakes context. No significant relationship was found between gender and test-taking effort in this study either. In the PIAAC Cycle 1 sample, males and females did not differ significantly in disengagement in numeracy and problem-solving, but disengaged responses in literacy were slightly higher for males (Goldhammer et al., 2016). In our research, females demonstrated higher levels of disengagement than males. A possible reason for this is that males are generally better at problem-solving than females (e.g. Csapó & Molnár, 2017) and are therefore better suited to these tasks.

A fundamental question is whether there is a relationship between *academic ability* and test-taking disengagement. The results are mixed and of substantial practical importance. In order to investigate this question, many studies have compared the total proportion of disengaged responses (response time effort; RTE) and ability measurement (such as SAT score and GPA). According to most studies, test-taking disengagement is unrelated to ability scores (Kong et al., 2007; Rios et al., 2014; Wise & DeMars, 2005; Wise & Kong, 2005), but some studies have reached different conclusions.

Rios et al. (2017)b) investigated a 108-item university-level ETS Proficiency Profile test that assesses critical thinking, mathematics, reading and writing among first-year students ($n = 1322$). They employed five threshold methods (3 s, NT15, NT20, NT25 and visual inspection) and found that motivated students' SAT scores were significantly higher than those of unmotivated peers with every method. Effect size varied between $d = 0.34$ and $d = 0.51$ depending on the method. Wise et al. (2009) used a multiple-choice test to assess the quantitative and scientific reasoning proficiencies of university students in a low-stakes context. A lower proportion of higher-ability students' responses were disengaged. Deribo et al. (2021) employed multiple-choice and complex multiple-choice items to assess ICT literacy among young adults ($N = 4960$) and showed that lower-ability examinees tend to be disengaged more frequently.

The practical significance of the question raised above is how to address disengaged behavior. A widely used method to deal with disengaged responses is motivation filtering, where either disengaged responses or all data from disengaged test-takers are deleted, leaving only engaged data in the sample and only taking these into account. Rios et al. (2017)b) developed the term *response-level filtering* to refer to the former type of motivation filtering and *examinee-level filtering* to refer to the latter. In the case of examinee-level filtering, a person can be classified as unmotivated if the percentage of disengaged responses exceeds a predefined threshold, usually 10 % (Wise & Kong, 2005). Examinee-level filtering is based on the assumption that disengaged response behavior is unrelated to test-takers' true ability. If this assumption is not correct, then deletion of respondents of higher or lower abilities will lead to bias (Rios et al., 2017). In our research, lower-ability examinees exhibited higher test-taking disengagement, suggesting that there is a relation between academic ability and test-taking disengagement. This implies that item-level filtering should be preferred to examinee-level filtering.

Previous research indicates that *working memory capacity* is crucial to students' problem-solving performance (Bull & Lee, 2014; Lindner et al., 2017). Lindner et al. (2019) found that working memory capacity significantly influenced test-taking disengagement in a low-stakes testing context. Our research yielded similar results: examinees with higher working memory capacity had lower test-taking disengagement. Research has demonstrated that working memory capacity is not fixed and can be improved in various ways (e.g. Brady et al., 2016). Students' working memory enhancement may be a good method to increase test-taking effort.

Most studies have used one method (self-reported or time-on-task-based) to examine test-taking effort. There are relatively few studies that have used both methods simultaneously on the same sample. Time-on-task-based effort showed significant correlations with self-reported effort (Rios et al., 2014; Silm et al., 2020; Wise & Kong, 2005). In our research, students who rated effort higher demonstrated lower levels of test-taking disengagement, which is consistent with the research. Self-reported questionnaires tend to provide the big picture, while response time-based methods enable item-by-item tracking of test-taking effort (Wise & Ma, 2012). This implies that if digital-based testing is applied, response time-based methods are preferable.

5. Limitations

Our study has a number of limitations. One was that the test consisted entirely of interactive problem-solving tasks. Research has found that subject matter has an effect on test-taking disengagement, so it is conceivable that we would obtain different results for different subject matter. Another important limitation is that convenience sampling was used at university level and the sample consisted exclusively of first-year university students who were willing to participate in the study. A further limitation is that test-taking disengagement was investigated in the knowledge acquisition phase, whereas this phase is not applicable to most tests, which mainly involve the knowledge application phase. The final limitation is that the test was carried out in a low-stakes context but with a relatively low proportion of disengaged responses.

6. Conclusions

The main objective of our research was to examine item- and person-level factors that influence test-taking disengagement, as the research has been contradictory as regards a number of factors. Multilevel modeling allows these factors to be identified more precisely. Among the predictors, item-level factors are remarkable because they can be changed to influence the motivation of examinees to do a test. Tests that are too long and items that are too difficult will lead to higher test-taking disengagement. Among the person-level factors, test-taking disengagement was predicted by gender, entrance score, working memory and self-reported effort.

As for the educational implications, the entrance score is of particular importance, as it is a proxy for academic ability. The method of dealing with disengaged responses is essentially determined by whether there is a relationship between academic ability and test-taking disengagement. Our research suggests that there is indeed such a relationship, with lower-ability examinees showing greater test-taking disengagement. According to our research, item-level filtering should be preferred to examinee-level filtering.

Due to the test design, we were not able to include all moderators of interest in our analysis. Among the factors not investigated, item type is worth considering in future research. Studies have found that test-taking effort is higher for selected response tasks than for constructed response tasks (e.g. DeMars, 2000). However, there are many types of selected-response tasks that have not been extensively studied in relation to test-taking effort.

CRediT authorship contribution statement

Róbert Csányi: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Gyöngyvér Molnár:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was prepared with the professional support of the Doctoral Student Scholarship Program of the Co-operative Doctoral Program of the Ministry of Innovation and Technology financed from the National Research, Development and Innovation Fund. This research was supported by a Hungarian National Research, Development and Innovation Fund grant (under the OTKA K135727 funding scheme), by the Hungarian Academy of Sciences Research Programme for Public Education Development grant (KOZOKT2021–16) and by the Humanities and Social Sciences Cluster of the center of Excellence for Interdisciplinary Research, Development and Innovation of the University of Szeged. GM is a member of the Digital Learning Technologies Incubation Research Group.

References

Akyol, P., Krishna, K., & Wang, J. (2021). Taking PISA seriously: How accurate are low-stakes exams? *Journal of Labor Research, 42*(2), 184–243. https://doi.org/10.1007/s12122-021-09317-8

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*, 92–104.

Attali, Y. (2016). Effort in low-stakes assessments: What does it take to perform as well as in a high-stakes setting? *Educational and Psychological Measurement, 76*(6), 1045–1058. https://doi.org/10.1177/0013164416634789

Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance? *Computers and Education, 83*, 57–63. https://doi.org/10.1016/j.compedu.2014.12.012

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*(1), 46–64. https://doi.org/10.1080/08957347.2015.1102914

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences of the United States of America, 113*(27), 7459–7464. https://doi.org/10.1073/pnas.1520027113

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*, 2309–2344.

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

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives, 8*(1), 36–41. https://doi.org/10.1111/cdep.12059

Csapó, B., & Molnár, G. (2017). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.02022

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01522

Csikszentmihalyi, M. (2014). *Flow and the foundations of positive psychology*. Netherlands: Springer. https://doi.org/10.1007/978-94-017-9088-8

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3

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

Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement, 58*(2), 281–303. https://doi.org/10.1111/jedm.12290

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

Eitel, A., Endres, T., & Renkl, A. (2020). Self-management as a bridge between cognitive load and self-regulated learning: The illustrative case of seductive details. *Educational Psychology Review, 32*(4), 1073–1087. https://doi.org/10.1007/s10648-020-09559-5

Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Frontiers in Psychology, 5*, 1–3. https://doi.org/10.3389/fpsyg.2014.00739

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. In *OECD education working papers, 133*. https://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 Assessments in Education, 5*(1). https://doi.org/10.1186/s40536-017-0051-9

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers and Education, 126*(February), 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts — Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*(2), 364–379. https://doi.org/10.1037/a0031856

Guo, H., Rios, J. A., Ling, G., Wang, Z., Gu, L., Yang, Z., et al. (2022). Influence of selected-response format variants on test characteristics and test-taking effort: An empirical study. *ETS Research Report Series, 2022*(1), 1–20. https://doi.org/10.1002/ets2.12345

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd Edition). Routledge.

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences, 18*(3), 127–133. https://doi.org/10.1016/j.tics.2013.12.009

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

Lai, M. H. C., & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *Journal of Experimental Education, 83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Lau, A., Swerdzewski, P. J., Jones, A., Anderson, R., & Markle, R. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*(3), 196–217. https://doi.org/10.1353/jge.0.0045

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379. http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/06_Lee.pdf.

Lindner, C., Lindner, M. A., & Retelsdorf, J. (2019). Die 5-item-skala zur messung der momentan verfügbaren selbstkontrollkapazität (SMS-5) im lern- und leistungskontext. *Diagnostica, 65*(4), 228–242. https://doi.org/10.1026/0012-1924/a000230

Lindner, C., Nagy, G., Ramos Arhuis, W. A., & Retelsdorf, J. (2017). A new perspective on the interplay between self-control and cognitive performance: Modeling progressive depletion patterns. *PloS One, 12*(6), Article e0180149. https://doi.org/10.1371/journal.pone.0180149

Lindner, C., Nagy, G., & Retelsdorf, J. (2018). The need for self-control in achievement tests: Changes in students' state self-control capacity and effort investment. *Social Psychology of Education, 21*(5), 1113–1131. https://doi.org/10.1007/s11218-018-9455-9

Lindner, C., & Retelsdorf, J. (2019). Perceived — And not manipulated —Self-control depletion predicts students' achievement outcomes in foreign language assessments. *Educational Psychology, 40*(4), 490–508. https://doi.org/10.1080/01443410.2019.1661975

Lindner, M. A. (2020). Representational and decorative pictures in science and mathematics tests: Do they make a difference? *Learning and Instruction, 68*(September 2019), Article 101345. https://doi.org/10.1016/j.learninstruc.2020.101345

Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology, 51*, 482–492. https://doi.org/10.1016/j.cedpsych.2017.09.009

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology, 10*, 1–15. https://doi.org/10.3389/fpsyg.2019.01533

Lindner, M. A., Schult, J., & Mayer, R. E. (2020). A multimedia effect for multiple-choice and constructed-response test items. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000646. *February 2021*.

Liu, O. L., Bridgeman, B., & Adler, R. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*, 352–362. https://doi.org/10.3102/0013189X12459679

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment, 20*(2), 79–94. https://doi.org/10.1080/10627197.2015.1028618

Martin, S., Helmschrott, S., & Rammstedt, B. (2014). The use of respondent incentives in PIAAC: The field test experiment in Germany. *Methods, Data, Analyses, 8*(2), 223–242. https://doi.org/10.12758/mda.2014.009

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

Molnár, G. (2021). Challenges and developments in technology-based assessment: Possibilities in science education. *Europhysics News, 52*(2), 16–19. https://doi.org/10.1051/epn/2021202

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Logfile analyses. *Frontiers in Psychology, 9*(MAR), 1–17. https://doi.org/10.3389/fpsyg.2018.00302

Nuutila, K., Tapola, A., Tuominen, H., Molnár, G., & Niemivirta, M. (2021). Mutual relationships between the levels of and changes in interest, self-efficacy, and perceived difficulty during task engagement. *Learning and Individual Differences, 92* (December 2020). https://doi.org/10.1016/j.lindif.2021.102090

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-Scale Assessments in Education, 2*(1). https://doi.org/10.1186/s40536-014-0005-4

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability, 29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*(1), 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

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 Assessments in Education, 9*(1), 10. https://doi.org/10.1186/s40536-021-00104-6

Rios, J. A. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education, 34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741

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., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

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), 69–82. https://doi.org/10.1002/ir.20068

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

Rodríguez-Hernández, C. F., Cascallar, E., & Kyndt, E. (2020). Socio-economic status and academic performance in higher education: A systematic review. In *Educational research review, 29*. Elsevier Ltd. https://doi.org/10.1016/j.edurev.2019.100305

Rosenzweig, E., Wigfield, A., Eccles, J., Renninger, K., & Hidi, S. (2019). *The Cambridge handbook of motivation and learning* (pp. 617–644). https://doi.org/10.1017/9781316823279.026

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications*. Pearson/Merrill Prentice Hall.

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

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. (2018). Are achievement gap estimates biased by differential student test effort? *Teachers College Record, 120*(12), 1–26.

Sommet, N., & Morselli, D. (2021). Keep calm and learn multilevel linear modeling: A three-step procedure using SPSS, Stata, R, and Mplus. *International Review of Social Psychology, 34*(1). https://doi.org/10.5334/irsp.555

Wang, Q., Mousavi, A., Lu, C., & Gao, Y. (2023). Examining adults' behavioral patterns in a sequence of problem solving tasks in technology-rich environments. *Computers in Human Behavior, 147*. https://doi.org/10.1016/j.chb.2023.107852

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

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., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment, 15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment, 26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. *Paper Presented at.* In *Proceedings of the annual meeting of the national council on measurement in education* (pp. 1–24). *March*.

Wise, S.L., Ma, L., Kingsbury, G.G., & Hauser, C. (2010). An investigation of the relationship between time of testing and test-taking effort. *National Council on Measurement in Education, March*, 1–18.

Wise, S. L., Ma, L., & Theaker, R. A. (2014). Identifying non-effortful student behavior on adaptive tests: Implications for test fraud detection. In N. Kingston, & A. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 175–185). Routledge.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185–205. https://doi.org/10.1080/08957340902754650