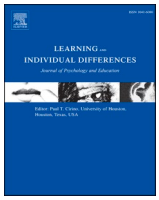
Learning and Individual Differences 106 (2023) 102340

Contents lists available at ScienceDirect

Learning and Individual Differences

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



How do test-takers rate their effort? A comparative analysis of self-report and log file data

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

[a] Doctoral School of Learning and Instruction, University of Szeged, Szeged, Hungary
[b] Institute of Education, University of Szeged, MTA-SZTE Digital Learning Technologies Research Group, Szeged, Hungary

ARTICLE INFO

Keywords:
Test-taking effort
Logfile analyses
Time on task
Number of clicks
K-means clustering

ABSTRACT

The present study investigates students' test-taking effort by integrating and comparing traditional self-report questionnaire data and students' test-taking behavior, based on log data analyses. Previous studies have shown that different methods often lead to different results. A computer-based measure of complex problem-solving in uncertain situations was used to minimize the influence of factual knowledge on test performance. K-means cluster analysis was used to build groups of students differing in test-taking effort, resulting in 3 distinct groups. The correlation between students' test-taking effort and test performance proved to be weaker based on the self-reported questionnaire data than on their actual test-taking behavior. Both the self-report questionnaire and the log data showed a decrease in test-taking effort during the test. The number of clicks played the largest role in predicting performance. Results suggest that (1) self-report questionnaire data are not consistent with students' actual test-taking behavior and (2) it's not necessary to make the maximum effort to obtain valid test results, but a certain level of effort is needed.

Educational relevance statement

In the implementation of effective personalised education, smart education, an increasingly important role is played by the accurate, fast and valid diagnostic of students' ability level. As for educational relevance, we stated that:

(1) Self-reported data are not always consistent with students' actual test-taking behavior, therefore log data-based methods are more appropriate than self-report questionnaires to investigate test-taking effort. For problem-solving tasks, the P+ > 0 % method performed better.

(2) For problem-solving tasks, the number of clicks plays the largest role in predicting performance. Using the number of clicks may increase the validity of response time-based methods.

(3) There is not necessary to make the maximum effort to obtain valid test results but rather to reach a certain level of effort.

1. Introduction

Students' cognitive test performance is not only determined by their actual knowledge and skills (Wolgast, Schmidt, & Ranger, 2020) but it is also potentially influenced by a variety of affective factors. The stakes of the tests can significantly affect the validity of the results: as the stakes decrease, the level of test-taking motivation drops (Wise, Ma, & Theaker, 2014). In one of the most prominent large-scale international studies – beyond intelligence and prior test achievement – 1–29 % of the variance of students' mathematical test results could be explained by their test-taking motivation (Kriegbaum, Jansen, & Spinath, 2014). In addition, according to Wise and DeMars (2005), unmotivated students scored more than half a standard deviation lower on tests than their motivated peers. This is supported by research results from (Finn, 2015; Schüttpelz-Brauns et al., 2018 and Wise & Kong, 2005), which indicated higher performance among more motivated test-takers. On the contrary, according to Gignac, Bartulovich, and Salleo (2019) it is not necessary to make the maximum effort or to have a very high level test-taking motivation to obtain valid test results, but it is rather needed to reach a certain level of effort.

From a methodological point of view, recent studies of test-taking effort generally use a single method design (Silm, Pedaste, & Täht, 2020). They generally administer a cognitive test and a self-report questionnaire at the end of the test, assuming a valid self-evaluation of test-taking effort and a constant value of this throughout the test.

* Corresponding author.
E-mail addresses: csanyi.robert@edu.u-szeged.hu (R. Csányi), gymolnar@edpsy.u-szeged.hu (G. Molnár).

https://doi.org/10.1016/j.lindif.2023.102340
Received 31 July 2022; Received in revised form 3 July 2023; Accepted 6 July 2023
Available online 18 July 2023

Only a few studies have examined students' actual test-taking behavior by using questions between test items or examined students' actual test-taking behavior by using log data, assuming that test-taking effort is not a constant value specific to a student, but can vary during a single test (Goldhammer, Naumann, Rölke, Stelter, & Tóth, 2017; Lundgren & Eklöf, 2020; Qiao & Jiao, 2018). Varying test-taking effort can present new challenges, since test-taking performance will reflect an unknown amount of the tested construct in this case and can harm the validity of the results.

We went further and first applied both approaches for measuring test-taking effort on students' test performance (self-report questionnaire and log data-based analyses) and secondly monitored its changing nature throughout the test-taking process in a cognitive domain where factual knowledge does not matter. We measured test-taking effort, raising questions at several time points during the test-taking process and analyzing log data using a longitudinal perspective to investigate students' test-taking effort across different test-taking profiles. According to our best knowledge, multiple method and test-taking profile analysis has not yet been integrated in the literature.

2. Literature review

2.1. Test-taking effort

A commonly used approach to interpreting test-taking motivation is expectancy-value theory (Eccles & Wigfield, 2002; Wise & DeMars, 2005). According to the model, the level of motivation is determined by the expectation of the performance and the value of the test. Test-takers' expectations are determined by (1) their perception of their own *abilities* and (2) the *difficulty* of the tasks. Values consist of four components: (1) *attainment value*, i.e., the importance of the test; (2) *intrinsic value*, defined by the enjoyment of engaging in the task; (3) *utility value*, i.e., how the task is related to future goals; and (4) *cost*, defined by the negative aspect of the task (e.g., time spent on the task or test anxiety). Test-taking motivation is manifested in the effort that the test-taker puts into completing the test. Test-taking effort is "the amount of resources that a test-taker uses in trying to achieve the best possible score on a specific test" (Lundgren & Eklöf, 2020).

There are several methods for measuring test-taking effort, which can be grouped into three categories: (1) self-assessment/self-report questionnaires, (2) response time-based approaches, and (3) model-based analyses. Self-report questionnaires are based on the test-takers' own judgements, the response time approach uses log data from computer-based tests, and the model-based approach can use both.

Self-report questionnaires are the longest-standing and most widely used means of measuring test-taking effort, typically measuring the components of test-taking effort on a Likert scale. In the simplest, most widely used case, students receive questions about their test-taking effort at the end of the test after completing the very last task. This approach assumes that students' test-taking effort is static, i.e., that it does not change during the test-taking process. This approach does not make it possible to monitor the dynamism involved in the process (Silm et al., 2020). It is also possible to monitor the change or even constancy of test-taking effort by answering questions about the current level of effort at the beginning of the test, between test items, and afterwards (Penk & Richter, 2017). The latter research design also provides an opportunity to monitor changes in test-taking effort. The simplest of the self-report questionnaires is the *Effort thermometer*, also used on the PISA survey, which only measures test-taking effort compared to previous personal experience (Butler & Adams, 2007). At the other end of the scale there is the *Online Motivation Questionnaire*, which contains seven subscales and 32 items (Crombach, Boekaerts, & Voeten, 2003). An important advantage of self-report questionnaires is that they are relatively easy to use and can even be implemented in traditional paper-and-pencil testing. However, a limitation of this approach is its subjectivity, as we have no knowledge of the degree of sincerity of the test-takers in their answers (Wise & Kong, 2005).

Methods based on log data emerged in parallel with the spread of computer-based assessments. Log data are computer-generated records (logs) that are connected to users' activity. These methods are mostly based on response time, which is the time the test-taker spends on a given task from the time the task is administered until its completion (i.e., when they click on the "next" button). A traditional paper-and-pencil test only enables test-takers' answers to be evaluated. If computer-based assessment is used, a great deal of contextual data (clicks, time spent on tasks, jumping back and forth, eye movements, etc.) can be recorded that used to be unimaginable with traditional paper-and-pencil assessment systems, and their analysis can reveal deeper relationships (Tóth, Rölke, Goldhammer, & Barkow, 2017). Response time-based methods are based on the assumption that participants with low test-taking effort spend less time completing tasks and therefore respond more rapidly than those with higher levels of motivation (Wise & Kong, 2005). Time spent on tasks may be supplemented with other data, such as number of clicks and type of clicks. Similarly, a lower number of clicks also indicates lower levels of motivation (Sahin & Colvin, 2020). Response time-based methods have several advantages over self-report questionnaires. Test-taking effort can be measured without intervention. No extra work is imposed on the examinee. In addition, measurement is based on test-takers' real behavior, not on their judgements. It will therefore be less biased. Changes in motivation can be tracked much more accurately because response time data are available for each item, not just at specific moments in time (Wise & Kong, 2005). In response time-based methods, a threshold time must be defined. If the response time is shorter than the threshold, the response is assumed not to be motivated (Wise & Kong, 2005). The simplest and longest-established solution involves a *constant threshold,* that means using a given, predefined threshold for each item. A more sophisticated solution entails *item-specific thresholds*. These are defined based on the assumption that the minimum time required to complete each item is different for each item. While test-takers can quickly solve a simple arithmetic problem, reading, interpreting, and solving a complex problem-solving task take much more time (Goldhammer, Martens, Christoph, & Lüdtke, 2016). This means that the threshold is not the same for all items but can differ item by item, task by task.

The model-based approach is based on the following assumption: the pattern of motivated test-takers' responses is related to the difficulty of the items. The approach is based on tools used within item response theory. The response pattern of test-takers is compared with a theoretical model: if there is a poor fit, it indicates non-normal behavior. The main advantage of the model-based approach is that it is based on the observation of test-takers' performance on the test, not on their self-assessment. Therefore, the bias may be lower. One drawback is that the abnormal pattern may be caused not only by unmotivated responses, but also by other factors, such as cheating and lucky guesses. Another important limitation is that it cannot characterize the level of motivation item by item. It only provides a global picture, making it a less common method for assessing test-taking effort (Wise & Smith, 2016). In this study, we integrated and compared the results obtained by applying the most frequently used methods: self-report questionnaires and response time-based methods.

2.2. Relation between test-taking effort and test performance

Previous research has shown a positive correlation between test-taking effort and test performance. Most research has examined test-taking effort with only one method; there have been relatively few studies that have applied multiple methods simultaneously on the same sample. Wise and Kong (2005) administered low-stakes, computer-based assessment tests to college freshmen ($N = 472$). Performance showed a higher correlation with response time effort ($r = 0.54$) than with self-reported effort ($r = 0.34$). Rios, Liu, and Bridgeman (2014) conducted research with volunteer college seniors ($N = 132$). They used

a computer-based achievement test that assessed critical thinking, reading, writing and mathematics. Test performance also demonstrated a higher correlation ($r = 0.67$) with response time effort than with self-reported effort ($r = 0.58$). Silm et al. (2020) conducted a meta-analytic review of the relationship between performance and test-taking effort. It encompassed 104 articles, most of which examined the test-taking effort using a single method. Test performance showed a higher correlation ($r = 0.72$) with response time effort than with self-reported effort ($r = 0.33$). These findings suggest that these two types of measures could be markedly different.

Examining the correlation between test-taking effort and test performance on the full sample provides a comprehensive picture of the relationship, but the details remain hidden. By examining the clustered parts of the sample, we can get a more accurate picture of the details. Hofverberg, Eklöf, and Lindfors (2022) investigated the PISA 2015 assessment of scientific literacy and performed a latent profile analysis which produced four student profiles. Highly motivated and interested students with sophisticated beliefs achieved the best results. This is contradicted by Lundgren and Eklöf (2020) who examined one problem-solving task and performed a cluster analysis. They found in the case of students who completed the task, level of effort was in a weak negative correlation with test performance. In addition, students in the low-effort cluster who solved the task were the highest performers. Together, these studies indicate that further studies are needed to explore the details.

Time spent on tasks and number of clicks are two indicators of test-taking effort in the literature. Research results on the relationship between time spent on tasks and test performance are not consistent. According to Wise and Kong (2005), there was a positive correlation between total time spent on tasks and test performance. Other research has produced similar findings on problem-solving tasks. Better planning of problem-solving, which takes more time, led to better solutions (AlZoubi, Fossati, Di Eugenio, Green, & CHEN, 2013; Eichmann, Greiff, Naumann, Brandhuber, & Goldhammer, 2020). In contrast, Greiff, Niepel, Scherer, and Martin (2016) found that too much time spent on problem-solving tasks was linked to lower test scores. While measuring the time spent on tasks makes sense for any type of task, measuring number of clicks only makes sense for tasks that require more interaction to complete. Previous research found a positive correlation between number of clicks and test performance (Eichmann et al., 2020; Goldhammer et al., 2014).

2.3. Changes in test-taking effort within the same testing session

In most of the self-report questionnaire-based research, test-taking effort has typically been measured only once during a testing session. However, multiple measurements during a testing session provide an opportunity to track changes in test-taking effort. Various studies have been carried out in which a self-report questionnaire was completed several times during the test and it was found that test-taking motivation fell (Barry, Horst, Finney, Brown, & Kopp, 2010; Penk & Richter, 2017; Wolgast et al., 2020). Log data-based methods provide an opportunity to measure test-taking efforts more than a few times and during each item. A decrease in test-taking effort has been supported by a number of log data-based and model-based studies (Attali, 2016; Goldhammer et al., 2016; Nuutila, Tapola, Tuominen, Molnár, & Niemivirta, 2021; Penk & Richter, 2017; Wise, Pastor, & Kong, 2009).

The changes are well explained by the process model of self-control depletion (Inzlicht, Schmeichel, & Macrae, 2014). The model proposes that people want to achieve an optimal balance between “have-to” and “want-to” goals. “Have-to” goals refer to labor-intensive tasks which are necessary to achieve long-term goals. In contrast, “want-to” goals refer to leisure activities that we like to do. After working hard within a particular time, motivation shifts from “have-to” goals towards “want-to” goals. This is supported by Lindner, Nagy, and Retelsdorf (2018) on changes in 1840 apprentices' state self-control capacity and their motivational test-taking effort. Test-takers repeatedly rated their state self-control capacity and test-taking effort during a 140-min. achievement test in mathematics and science. Researchers found drops in state self-control capacity correlated with drops in test-taking effort over the course of time using growth curve analyses. In addition, they also found that trait self-control helped to keep state self-control capacity and test-taking effort at a higher level during the test. Lindner, Lindner, and Retelsdorf (2019) investigated changes in students' self-control capacity and exhaustion during a learning session and also after three testing sessions. In the course of the four sessions, they found that decreasing self-control capacity was related to increasing exhaustion. In another study, Lindner and Retelsdorf (2019) found that students who report high self-control depletion during a test of English as a foreign language were less motivated to work on a subsequent test. They also reported more distracting thoughts, their performance was lower, and they felt more depleted at the end of the testing session. In summary, these results showed that focusing attention during a testing session while inhibiting task-irrelevant thoughts and/or emotions requires self-control. This can lead to mental fatigue that is closely related to changes in test-taking effort.

Apart from mental fatigue, there are other factors that affect changes in test-taking effort. Barry and Finney (2016) investigated test-taking effort during a low-stakes, three-hour testing session ($N = 683$). The first test was a difficult cognitive test, followed by non-cognitive and affective measurements. Self-reported test-taking effort increased linearly during the first four tests and decreased from test 4 to test 5. This means that students' test-taking effort was the lowest on the first test, which was the longest and most difficult, cognitive test. This is consistent with previous findings in low-stakes contexts in which test-takers made more of an effort on less difficult tests than on more demanding ones (DeMars, 2000; Wise, 2006). Other research has indicated that test-takers put more effort into completing a test that matched their abilities, i.e., one with tasks that were neither too difficult nor too easy (Asseburg & Frey, 2013).

2.4. Test-taking profiles

Test-taking behavior has an important role in test performance. It has been investigated in a number of studies, typically characterizing students' average behavior patterns, but only a few studies have classified students' individual test-taking behavior.

Stenlund, Lyrén, and Eklöf (2018) examined test-taking behavior in a high-stakes context with the Swedish Scholastic Assessment Test among participants with an average age of 22 (SD = 6.6). They used a self-report questionnaire to measure motivation, test anxiety, and risk-taking behavior. Then, they used hierarchical cluster analysis and identified three clusters: (1) moderate risk-taker, (2) calm risk-taker, and (3) test-anxious risk-averse. They concluded that test anxiety and risk-taking played a major role in a high-stakes context and that students with a calm risk-taker profile (high level of risk-taking and relatively low levels of test anxiety and motivation) proved to be the best performers.

Goldhammer et al. (2017) investigated 17-year-old (SD = 0.78) German students' test-taking effort while completing ICT literacy items in a stimulating web-based environment. Six log data-based variables were analyzed which describe students' web search: (1) number of web page views, (2) number of different pages viewed, (3) time spent on the significant page, (4) percentage of time spent on the significant page to total time on task, (5) percentage of time spent on the home page to total time on task, and (6) total time on task. Two clusters of test-taking effort were identified with k-means cluster analysis. Members of Cluster 1 spent most of their time on the home page pre-selecting pages, successfully completing the task in 53.88 s. Cluster 2 participants spent more time evaluating sources of information on irrelevant websites, managing to do the task successfully in 87.94 s. The results showed that higher-ability students needed less effort to solve problems successfully in a technology-rich environment. At the same time, less skilled learners were also able to produce successful solutions by making a greater effort

to compensate for their lower skill levels.

Lundgren and Eklöf (2020) analyzed 3231 fifteen-year-old Scandinavian students' test-taking behavior on the PISA 2012 traffic problem-solving task, where students sought the shortest travel time route between two fictitious cities. Using k-means cluster analysis on log data, the researchers identified four clusters of test-taking effort: (1) high effort, (2) low effort, (3) medium effort, and (4) planner, in which test-takers spent a relatively long time before starting to perform actions. Qiao and Jiao (2018) compared data mining methods in the US sample of the same PISA 2012 survey. They concluded that k-means cluster analysis as a method had successfully been used to investigate test-taking behavior on computer-based tests.

To sum up, we can highlight that test-taking effort and the number of clusters describing its variety and characteristics are strongly dependent on numerous factors, including the task, the sample, and the variables included in the analysis.

2.5. Research purpose and questions

Partial or total lack of test-taking effort can harm the validity of the results (Rios, 2021) because if an incorrect answer is identified as the test-taker's failure to solve the problem, rather than being identified as unmotivated, it will affect the score obtained. In previous research, a number of log data-based methods have been developed to identify unmotivated responses. These methods produce different results on the same sample (Goldhammer et al., 2016). Generally, there is a positive correlation between test-taking effort and test performance, but the relationship is not so clear when examining clustered groups of test-takers (Lundgren & Eklöf, 2020). Additionally, several previous studies have shown that students' effort varies throughout a single cognitive test, and this in turn affects test performance (Penk & Richter, 2017). Finally, students' test-taking profiles depend on many factors, and different results were reached in the investigations. The number of groups and their characteristics vary depending on the task, the sample, and the variables included in the analysis.

Consequently, the present research aimed to investigate students' test-taking effort on the same sample by integrating and comparing self-report and log data-based methods, giving interactive tasks and situations in which already existing factual knowledge could not be used during the problem-solving process. Students' self-report effort was measured by asking students to rate their test-taking effort. Collected log data included number of clicks and time on task. We also used response time effort, which refers to the level of effort of a given test-taker. We decided to include this term because the response time and number of clicks are also measure of effort, but previous research (Gignac et al., 2019) stated that it is not necessary to make the maximum effort to obtain valid test results, but it is rather needed to reach a certain level of effort. In addition, Stenlund et al. (2018) found that the best performers were the calm risk-takers (high level of risk-taking and relatively low levels of test anxiety and motivation). It is concluded above that response time and number of clicks alone cannot be used to measure effort. Answers were sought to the following research questions:

RQ1: Which of the methods tested is the most appropriate and valid response time-based method for measuring students' test-taking effort in interactive, complex problem-solving situations?

RQ2: What is the relationship between self-reported effort, effort reflected by log data (time on task and number of clicks), response time effort and test performance?

RQ3: How does test-taking effort change as the test progresses based on self-report questionnaire and log data-based methods?

RQ4: Which test-taking effort profiles can students be classified into based on self-reported data, log data (time on task and number of clicks) and test performance?

3. Materials and methods

3.1. Participants

The sample consisted of undergraduate students just starting their studies at one of the largest universities in (masked for review) in autumn 2021. The university has twelve faculties (e.g., faculty of medicine, law, the humanities and social sciences and natural science), all of which were involved in the assessment. All full-time freshers were informed of the details via the university's learning management system. Students' participation was voluntary. They received one credit as an incentive for successfully completing the tests. Due to the administrative requirements of the university, they were assigned to a specific course, called Pursuing a Career. A total of 1748 students representing 46.2 % of the target population, participated in the study (mean age = 19.80, SD = 1.92), 53.0 % of them being female.

3.2. Data collection procedure

Both the cognitive tasks and the questionnaire items were administered via the eDia system (Csapó & Molnár, 2019). The assessment was carried out in a large computer room at the university learning and information center with up to 150 participants at a time. Test administration was supervised by PhD students who had previously been trained. The test was administered during the first three weeks of the semester. Two-hour sessions were offered to the students, who were asked to do other learning-related cognitive tests within the confines of the course in addition to the complex problem-solving test. At the beginning of the test, participants were provided with instructions on how to use the user interface and a warm-up task. After logging in to eDia, students had 60 min to do the tasks and complete the questionnaire. If a student had used the maximum time in all the problem-solving exercises (a total of 45 min.), they still had enough time for the questionnaire. After taking the test, they received immediate feedback on their average performance and detailed feedback a week later, including comparative data with their peers.

Ethical approval was not required based on the national and institutional guidelines as (1) the data collection was an integral part of the educational processes at the university, (2) participation was voluntary and (3) all of the students in the assessment had turned 18. Consequently, it was not required or possible to request and obtain written informed parental consent from the participants, but (4) all of the participants confirmed with signature that their data would be used for educational and research purposes at both faculty and university level.

3.3. The problem-solving tasks

We searched for a widely used and reliable instrument which excludes the effect of prior school learning (thus disregarding already existing attitudes towards different domains), yet provides learning opportunities with direct applications in various uncertain situations. Complex problem-solving, more specifically, the MicroDYN approach, was chosen, which involves dynamic problem-solving tasks which can be completed in a relatively short amount of time (Funke, 2014; Greiff et al., 2013). The MicroDYN approach has been shown to be reliable and valid for assessing complex problem-solving (Greiff et al., 2013; Greiff, Molnár, Martin, Zimmermann, & Csapó, 2018; Molnár & Csapó, 2018). The tasks consisted of two empirically distinguishable phases (Greiff et al., 2013), knowledge acquisition and knowledge application. In the first phase of the problem-solving process (see Fig. 1), test-takers explored the relationships between the input and output variables by freely interacting with the problem environment. In this phase, there was no limit to the number of interactions, but there was a time limit of 180 s. Based on the information obtained and interpreted, they drew (a) relationship(s) between the input and output variables (Molnár & Csapó, 2018) on a concept map presented on screen. In the second part of the

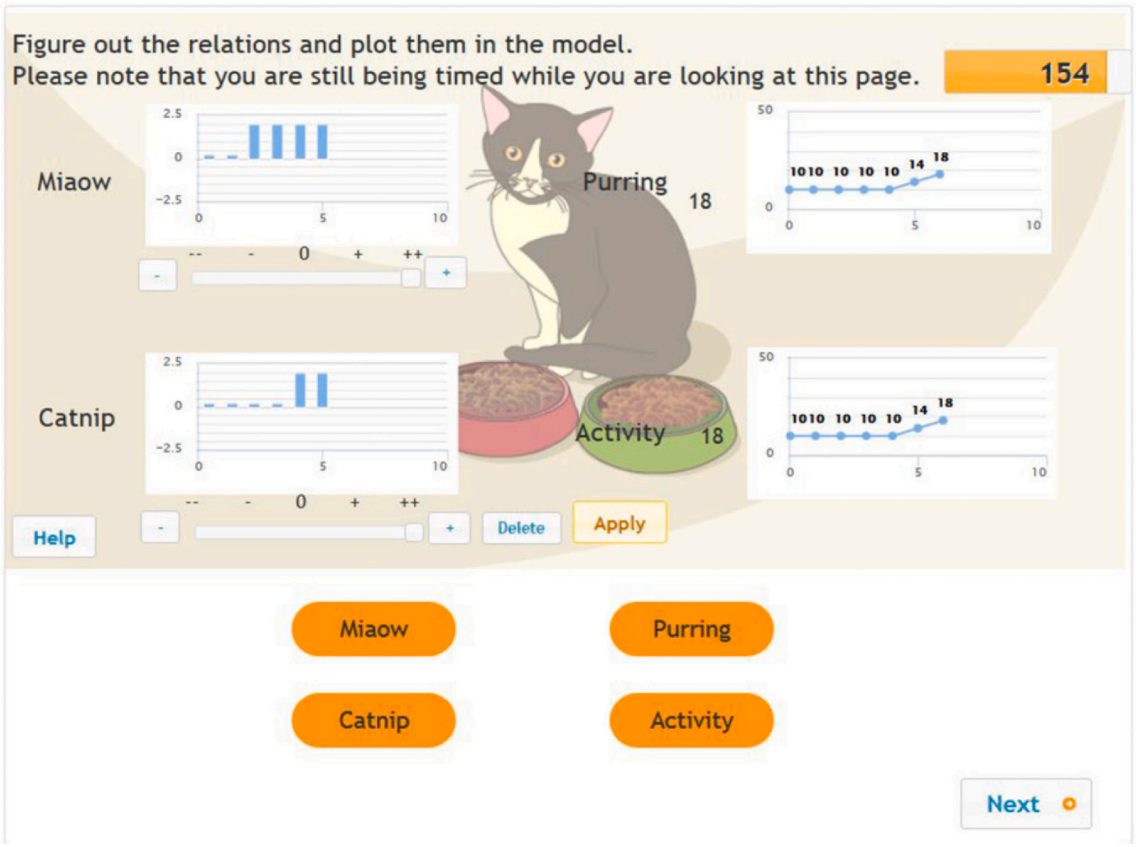


Fig. 1. The first stage of interactive problem-solving: exploring the relationship between two input and two output variables.

problem-solving process, they applied all this knowledge to operate the system, changing the values of the input variables to reach a given state of the problem environment. In the second phase of the test, they had a 90-s time limit, with a maximum of four trials, i.e., four applications of the different input variable settings. In the second phase of the example task presented in Fig. 1, the pre-determined values of purring and activity were adjusted in up to four steps by feeding the cat two different kinds of cat food in the right proportions. The test consisted of ten tasks of increasing complexity, i.e., an increasing number of input and output variables and an increasing number of relations. The reliability of the MicroDYN problems as a measure of knowledge acquisition and knowledge application was acceptable ($\alpha = 0.88$).

In this study, we focused on achievement and log data collected in the first phase of the problem-solving process, as the number of possible clicks – which can be a good indicator of test-taking effort – was not maximized and the maximum amount of time – which can also be an important indicator of test-taking effort – was also less restricted. Consequently, both the time and click data differentiated students to a larger extent than the log data collected in the second phase of the problem-solving process.

3.4. Data collected

Two different approaches were merged to measure test-taking effort, the self-report questionnaire-based design and the log data-based (time on task and number of clicks) approach. In addition, we also collected students' test performance data. Students were asked to rate their test-taking effort (*self-reported effort, SRE*) based on statements we had designed on a five-point Likert scale ("I worked on the tasks with full effort." 1: not true at all; 5: completely true). In order not only to obtain a static picture of the students' test-taking effort, but also to track changes in effort during the test process, we had the students complete the self-report questionnaire a total of six times during the cognitive test. The first one was done after the warm-up task, the next four times after every second problem scenario, and the final one after the last problem had been solved.

Two types of log data were included in the test-taking effort analysis, (1) time on task (TOT) and (2) number of clicks (CLICK). These represented how students behaved while completing the tasks. In response time-based methods, the indicator measured is time on task, meaning the time the test-taker spends on a task. An item-level threshold should also be defined for tasks with more items. If the response time for an item is less than the threshold, it is considered an unmotivated response. However, if it is greater than or equal to the threshold, it is considered a motivated response. Wise and Kong (2005) introduced the following relationship to measure the motivated or *solution behavior* (SB_{ij}) associated with item i and examinee j:

$$SB_{ij} = \begin{cases} 1, if\ RT_{ij} \geq T_i \\ 0, if\ RT_{ij} < T_i \end{cases} \quad (1)$$

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

Further, Wise and Kong (2005) introduced the term *response time effort* (*RTE*). RTE is the average motivated behavior for a given participant, i.e., the amount of effort invested. The RTE per examinee j is

$$RTE_j = \frac{\sum SB_{ij}}{k} \quad (2)$$

where k = number of items.

In our research, we investigated six different thresholds. We applied the two most commonly used *constant threshold* methods, the three-second (*3 s*) and five-second (*5 s*) thresholds (Wise & Kong, 2005). The *normative threshold* method (*NT10*) (Wise & Ma, 2012) sets the

threshold relative to the average time spent on tasks. The NT10 threshold is 10 % of the average time examinees spent on an item, up to a maximum of ten seconds. For example, if the average time on task for a given item is 38 s, the threshold for the item is 3.8 s. However, if the average time on task is 160 s, the threshold is 10s, instead of 16 s. Based on this rule, we also used the thresholds *NT15* and *NT20*. The *proportion correct greater than zero* (*P+ > 0 %*) method is used for constructed response items. For multiple-choice tests, the probability of a correct answer is greater than zero, even in the case of random guesses (e.g., 0.25 for a test with four answer options per item). In the case of constructed response items where test-takers are required to provide their own answers, the random chance of choosing the correct answer is zero. To determine the P+ > 0 % threshold, the proportion of correct answers within a given response time is calculated at one-second intervals. The responses are then sorted by size in ascending order of response time. The threshold is the shortest response time at which the proportion of correct responses is greater than zero (Goldhammer et al., 2016). For example, if the 4, 5, 7, and 8 s responses are incorrect for a given item after ordering and the first correct response is for the 9 s response time, this will be the threshold.

3.5. Data analysis

In order to compare the log data-based methods, we calculated the proportion of responses rated as unmotivated and it was compared by task and method. Validation criteria were used to select the optimal log data-based methods. A valid indicator should aptly separate unmotivated responses from motivated ones. This is based on the assumption that motivated responses should be more likely to be among the correct responses than unmotivated ones (Goldhammer et al., 2016).

To examine the relationship between self-reported effort, response time effort, time on task, number of clicks and test performance, we applied Pearson correlation, furthermore to compare the correlations, we used Steiger's Z method (Steiger, 1980).

In order to examine the change in self-reported test-taking effort, the mean for each measurement time was taken and compared using Repeated Measures ANOVA. In case of log data-based test-taking effort, we took the solution behavior (SB) scores for each task and also compared them using Repeated Measures ANOVA also.

K-means cluster analysis was used to construct the student groups to identify students' test-taking effort profiles. K-means clustering had successfully been used in previous studies of test-taking behavior on computer-based assessments (Goldhammer et al., 2017; Lundgren & Eklöf, 2020; Qiao & Jiao, 2018). In order to prevent bias caused by different scales, we used *Z*-score standardization. One of the most important issues when performing cluster analysis is to determine the optimal number of clusters. According to Shi et al. (2021), one of the most commonly used methods is the *elbow method.* It enables us to find the optimal cluster number at the maximum change in slope of the plotted values. The disadvantage of this method is that it is difficult to read this value from the graph in many cases, so it is not possible to clearly determine the optimal number of clusters. For this reason, the *silhouette method* was used in the analyses, where the optimal cluster number can be identified from the maximum for the silhouette value (Shi et al., 2021). Fig. 2 shows a comparative analysis of the use of the elbow and silhouette methods. Using the elbow method, it is more difficult to read the breakpoint where the change in slope is the greatest. This is because there are several major breakpoints: for cluster numbers 3, 4, and 5, and it is difficult to select the largest of these. Based on the silhouette method, it can clearly be identified that the maximum for the silhouette value is 3, so this value is the optimal number of clusters.

4. Results

4.1. Results for research question 1 (RQ1): Which of the methods tested is the most appropriate and valid response time-based method for measuring students' test-taking effort in interactive, complex problem-solving situations?

A comparison of the log data-based methods used to measure test-taking effort is shown in Table 1. Average rate of unmotivated responses varies across methods, ranging from 0.0 % to 2.3 %. Method 3 s identified the fewest responses as unmotivated, while the P+ > 0 % method identified the most. There are more significant differences at the individual task level. For Task 10, method 3 s identified 0.1 % of the responses as unmotivated, while method P+ > 0 % identified 7.8 % of them as such. In addition to the proportion of unmotivated responses, Table 1 also shows the proportion of correct responses for each task, indicating that the tasks became more difficult towards the end of the test.

4.1.1. Validation criteria

Table 2 shows how the proportion of correct answers classified as motivated and unmotivated changed for the six methods (3 s, 5 s, NT10, NT15, NT20, and P+ > 0 %). For each method, the proportion of correct responses was obtained by averaging the results of the responses to the tasks. For the P+ > 0 % method, the proportion of unmotivated correct answers should be zero due to the principle of the method. The results

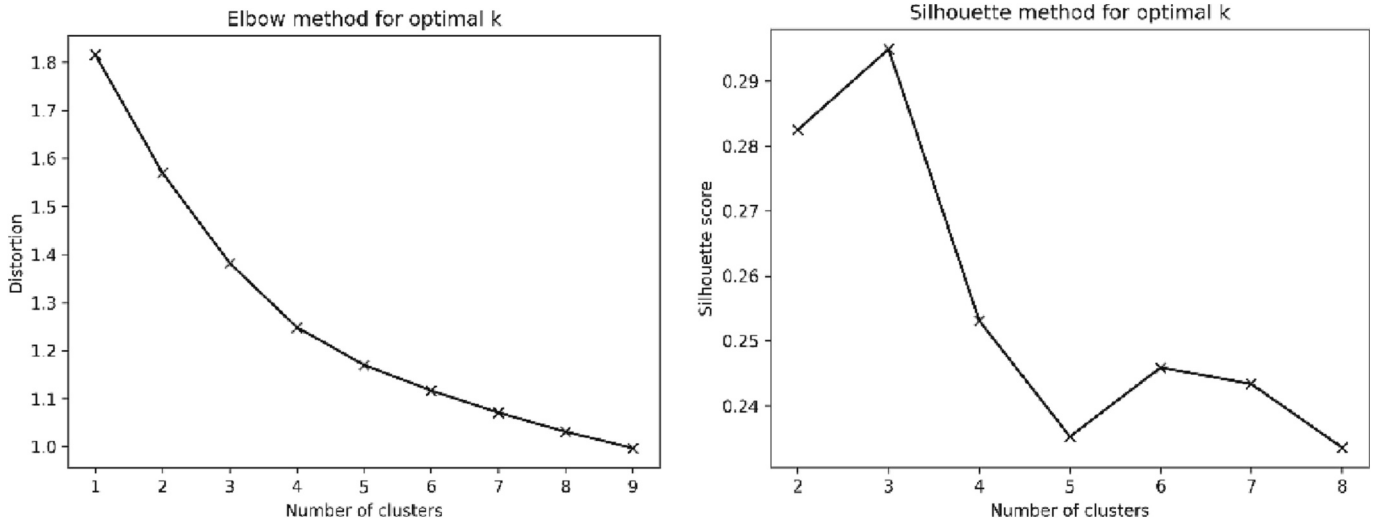


Fig. 2. Optimal cluster numbers for the elbow and silhouette methods.

Table 1
Percentage of unmotivated responses by task and method, and percentage of correct responses.

Methods	Percentage of unmotivated responses per task										
	1	2	3	4	5	6	7	8	9	10	Mean
3 s	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0
5 s	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.2	0.1
NT10	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.2	0.1	0.5	0.1
NT15	0.1	0.1	0.1	0.2	0.1	0.1	0.5	0.5	0.6	1.0	0.3
NT20	0.2	0.5	0.2	0.2	0.3	0.4	1.0	0.9	1.1	1.3	0.6
P+ > 0 %	0.1	0.1	0.3	0.4	1.5	1.4	4.6	3.2	3.5	7.8	2.3
Proportion of correct responses	78.9	79.9	81.7	82.4	77.5	80.2	27.3	38.2	35.4	32.7	61.4

Notes: 3 s: three-second threshold; 5 s: five-second threshold; NT10: normative threshold 10; NT15: normative threshold 15; NT20: normative threshold 20; P+ > 0 %: proportion correct greater than zero threshold.

Table 2
The proportion of motivated and unmotivated correct responses and their differences between each method.

Methods	Proportion of correct responses – motivated	Proportion of correct responses – unmotivated	Difference
3 s	0.61	0.00	0.61
5 s	0.61	0.00	0.61
NT10	0.62	0.00	0.62
NT15	0.62	0.02	0.60
NT20	0.62	0.03	0.59
P+ > 0 %	0.63	0.00	0.63

Notes: 3 s: three-second threshold; 5 s: five-second threshold; NT10: normative threshold 10; NT15: normative threshold 15; NT20: normative threshold 20; P+ > 0 %: proportion correct greater than zero threshold.

show that the P+ > 0 % method produces the greatest difference between the proportion of motivated and unmotivated correct answers; hence it is the method that best separates motivated from unmotivated answers. For this reason, further analyses were performed with this method.

4.2. Results for research question 2 (RQ2): What is the relationship between self-reported effort, effort reflected by log data (time on task and number of clicks), response time effort and test performance?

Table 3 shows descriptive statistics of the variables included in the analyses: time on task (TOT), number of clicks (CLICK), score achieved (SCORE), self-reported effort (SRE) and response time effort (RTE).

Table 4 displays the correlation coefficients between variables. The correlation between time on task and number of clicks was found to be the strongest ($r = 0.62$, $p < .01$). Self-reported effort showed a significantly lower correlation ($Z = 8.73$, $p < .01$) with performance ($r = 0.11$, $p < .01$) than log data -based one ($r = 0.37$, $p < .01$). Number of clicks demonstrated a significant correlation with performance ($r = 0.32$, $p < .01$), but there was no correlation between time on task and performance.

We used multiple regression to highlight the role of the independent variables, the results of which are shown in Table 5. The table shows the individually and cumulatively explained variances of the independent variables. Number of clicks and response time effort predicted performance to the greatest extent. In comparison, they have a predictive power which is higher by one order of magnitude than self-reported effort and time on task.

Table 3
Descriptive statistics of the variables.

Variables	Minimum	Maximum	Mean	SD
TOT	90	1610	577.13	201.28
CLICK	0	188	55.00	22.37
SCORE	0	10	6.14	2.80
SRE	1.00	5.00	4.28	0.92
RTE P+ > 0 %	0.10	1.00	0.98	0.08

Notes: TOT: total time on task; CLICK: total number of clicks; SCORE: test score; SRE: self-reported effort; RTE P+ > 0 %: response time effort based on proportion correct greater than zero method.

Table 4
Correlation between variables.

Variables	Correlation between variables			
	SRE	TOT	CLICK	SCORE
SRE	–			
TOT	0.09**	–		
CLICK	0.07**	0.62**	–	
SCORE	0.10**	−0.01	0.32**	–
RTE P+ > 0 %	0.13**	0.30**	0.32**	0.37**

Notes: **p < .01; SRE: self-reported effort; TOT: total time on task; CLICK: total number of clicks; SCORE: test score; RTE P+ > 0 %: response time effort based on proportion correct greater than zero threshold.

Table 5
Results of multiple regression analysis for test score as a dependent variable.

Independent variables	r	b	r·b·100	p
SRE	0.11	0.07	0.69	0.001
TOT	−0.01	−0.40	0.42	<0.001
CLICK	0.32	0.45	14.33	<0.001
RTE P+ > 0 %	0.37	0.34	12.49	<0.001
Total variance explained			27.93	

Notes: $N = 1748$; $F(1747) = 169.02$ $p < .001$; r: Pearson correlation; b: standardized regression coefficient; r·b·100: explained variance.

4.3. Results for research question 3 (RQ3): How does test-taking effort change as the test progresses based on self-report questionnaire and log data-based methods?

4.3.1. Change in test-taking effort based on self-report questionnaire

As the test progressed, a significant difference in test-taking effort was observed (Wilk's $\lambda = 0.91$, $F(5, 1740) = 36.70$, $p < .001$, $\eta^2 = 0.10$). The Bonferroni adjusted pairwise tests identified which measurement time points were significantly different, suggesting that test-taking effort fell significantly as the test progressed (Table 6). The following significantly distinct measurement time points were observed {1} > {2} > {3, 4} > {5, 6}.

4.3.2. Change in test-taking effort based on log data

As the test progressed, a significant difference in test-taking effort was observed (Wilk's $\lambda = 0.91$, $F(9, 1737) = 19.13$, $p < .001$, $\eta^2 = 0.09$). The Bonferroni adjusted pairwise tests identified which measurement time points were significantly different, suggesting that test-taking effort decreased significantly as the test progressed (Table 7). The following significantly distinct measurement time points were observed: {1, 2, 3,

Table 6
Change in test-taking effort based on self-report questionnaire.

Measuring time	SRE		ANOVA		Sig. of different times measured*
	M	SD	F	p	
1.	4.45	0.86	36.70	< 0.001	{1} > {2} > {3,4} > {5,6}
2.	4.36	0.97			
3.	4.30	1.01			
4.	4.31	1.02			
5.	4.26	1.07			
6.	4.22	1.08			

Notes: *The figures in the comparison column refer to the results of the measurement times ($p < .05$). SRE: self-reported effort.

Table 7
Change in test-taking effort based on P+ > 0 % method.

Measuring time	RTE		ANOVA		Sig. of different times measured*
	M	SD	F	p	
1	1.00	0.03	19.13	<0.001	{1, 2, 3, 4} > {5, 6} >> {7, 8, 9} > {10}
2	1.00	0.02			
3	1.00	0.05			
4	1.00	0.06			
5	0.98	0.12			
6	0.99	0.12			
7	0.95	0.21			
8	0.97	0.18			
9	0.97	0.18			
10	0.92	0.27			

Notes: *The figures in the comparison column refer to the results for the measurement times (p < .05). RTE: response time effort.

4} > {5, 6} > {7, 8, 9} > {10}.

4.4. Results for research question 4 (RQ4): Which test-taking effort profiles can students be classified into based on self-reported data, log data (time on task and number of clicks) and test performance?

For the cluster analysis, the time on task, number of clicks, test score and self-reported effort values were taken into account. In Fig. 3, the means for the standardized values (*Z*-scores) of the variables are presented by cluster, and Table 8 shows the means and standard deviations of the variables. The results of the analysis of variance show that there is a significant difference between the three clusters ($p < .001$). The F-values show that there are differences between the means for the clusters mostly by number of clicks and least by effort. In post hoc analyses of variance, i.e., analysis to examine differences between clusters, the variances are not homogeneous, so a Dunnett-T3 test was used.

The first cluster (Cluster 1) is made up of 310 students, 18 % of the sample. They are characterized by low amount of clicks in a short time. There was no significant difference in the number of clicks and time spent on tasks as compared to students in Cluster 2. Of the three clusters, they achieved the worst results and rated their effort significantly lower than their peers in the other two clusters.

The second cluster (Cluster 2) comprises 1000 students (57 %). Students in this cluster clicked little in a short period of time; that is, they put low amount of effort into completing the tasks, just as the students in Cluster 1. They achieved good results; there was no significant difference in scores as compared to students in Cluster 3. They rated their effort the highest.

The third cluster (Cluster 3) consists of 438 students (25 %). The students in this cluster were the ones who spent the most time solving the problems and clicked the most times. Their results are similar to those of the students in the second cluster, but better than those in the first. They rated their effort lower than those in the second cluster.

5. Discussion

The main aim of this study was to investigate test-taking effort on the same sample with multiple methods. Most of the research examines test-taking effort according to a single principle – in very few of the papers we reviewed did we find research involving multiple methods used simultaneously. This is also supported by a meta-analysis by Silm et al. (2020), in which approximately 10 % of the studies reviewed used multiple approaches. The vast majority of them measured test-taking effort in a single way. In our research, we used a self-report questionnaire to measure test-taking effort as well as applying log data-based methods.

Research question 1 (RQ1): Which of the methods tested is the most appropriate and valid response time-based method for measuring students' test-taking effort in interactive, complex problem-solving situations?

There are several methods to specify the response time-based time threshold. In order to determine the most appropriate method for our research, we investigated six different threshold methods, two of them constant (3 s, 5 s) and four of them item-specific (NT10, NT15, NT20, P+ > 0 %).

A number of studies have examined the appropriateness of each threshold. Hauser and Kingsbury (2009) argued that the three-second threshold is inappropriate for items with a great deal of reading material. Wise and Ma (2012) compared two thresholds for multiple-choice items and found that the normative threshold performed better than

Fig. 3. Student profiles based on time spent on task, number of clicks, score achieved and self-reported effort (TOT: total time on task; CLICK: total number of clicks; SCORE: test score; SRE: self-reported effort).

Table 8
Features of the students' test-taking effort profiles generated from the log data, score and self-reported effort.

Variables	Cluster 1 ($N = 404$)		Cluster 2 ($N = 929$)		Cluster 3 ($N = 415$)		F*	Sig. different clusters**
	M	SD	M	SD	M	SD		
1. TOT	−0.48	0.68	−0.39	0.57	1.24	0.94	914.87	{1,2} < {3}
2. CLICK	−0.45	0.73	−0.39	0.54	1.22	0.98	866.01	{1, 2} < {3}
3. SCORE	−0.37	1.13	0.04	0.94	0.17	0.97	28.93	{1} < {2,3}
4. SRE	−1.72	0.78	0.47	0.42	0.15	0.73	1665.15	{1} < {3} < {2}

Notes: *All F-values are significant at the $p < .001$ level. **For significantly different clusters, the "<" sign indicates the direction of the significant difference ($p < .05$). The comparison column between clusters shows the significantly differentiated clusters according to the Dunnett-T3 test. TOT: total time on task; CLICK: total number of clicks; SCORE: test score; SRE: self-reported effort.

the three-second threshold. Goldhammer et al. (2016) used data from the PIAAC (Round 1) survey, where the majority of the tasks are of the constructed response type. They investigated the impact of using different thresholds. Four thresholds were compared: two constant (three-second and five-second) and two item-specific (proportion correct greater than zero and visual inspection). They found that the proportion correct greater than zero method provided the most valid results.

We used Goldhammer et al. (2016) validation criteria to select the optimal log data-based method. The optimal method is the one that better separates unmotivated correct answers from motivated ones. Among the response time-based methods, the P+ > 0 % method was found to be the most accurate based on the validation criterion used. This finding is consistent with results reported in (Goldhammer et al., 2016). One possible reason for this is that here, too, there were constructed response answers, where the probability of a correct answer is close to zero when guessed.

An important domain of application for low-stakes tests is international large-scale assessments, where one of the subject areas is problem-solving. These assessments usually apply constant thresholds to identify unmotivated responses. In the PISA assessment, items that are not reached and rapid responses are also excluded from the analysis. If response time is less than five seconds, the response is identified as rapid guessing (Buchholz, Cignetti, & Piacentini, 2022). In the PIAAC assessment, only omitted responses could be considered unmotivated responses (but not rapid responses). If response time is less than five seconds on an item and the respondents only engage in 0–2 actions, the non-response is considered not attempted and therefore excluded from the analysis (Khorramdel, von Davier, Gonzalez, & Yamamoto, 2020). Previous studies examined different time-on-task-based methods and found that item-specific thresholds produce greater accuracy than constant thresholds (Goldhammer et al., 2016; Wise & Ma, 2012). Our study also supports this finding, and we found that a relatively rarely used method proved to be the most accurate. For international comparability, it is important to use the proper method to identify unmotivated responses. Further research is required to investigate different methods for large-scale, international assessments.

Research question 2 (RQ2): What is the relationship between self-reported effort, effort reflected by log data (time on task and number of clicks), response time effort and test performance?

We investigated the correlations between the variables. Significant correlations were found between the variables tested (self-reported effort, response time effort, time on task, number of clicks and test score) in all but one case. The self-reported effort has a significantly lower correlation with performance ($r = 0.10$) than response time effort ($r = 0.37$). Both values are significantly lower ($Z = 10.13$, $p < .01$ and $Z = 21.65$, $p < .01$, respectively) than the results of the meta-analysis conducted by Silm et al. (2020) ($r = 0.33$ and $r = 0.72$, respectively). Overall, the above data suggest that self-report effort and log data-based methods could be different.

Due to the nature of the interactive problem-solving exercises used on the test, the problems cannot be solved by heart. The test-takers must therefore test the possible relationships between variables in order to succeed. The correlation between time spent on the tasks and number of clicks was the strongest ($r = 0.62$, $p < .01$), meaning that if someone was making a great deal of effort, they needed more time. The students who were able to achieve high scores on the test were those who made the appropriate number of attempts on the tasks. Number of clicks significantly correlated with performance ($r = 0.32$), but time spent on tasks did not ($r = -0.01$). Supposedly, the high-ability problem-solvers were able to complete numerous trials in a short time, while for the low-ability problem-solvers it took much longer. However, the tasks could not be completed successfully with a very low number of attempts. The results indicate that for problem-solving tasks, the number of clicks plays the largest role in predicting performance. Previous research findings are not consistent on the relationship between time on task and test scores. Greiff et al. (2016) found that too much time spent on tasks was associated with lower test scores, but other researchers found a positive correlation between these two variables (AlZoubi et al., 2013; Eichmann et al., 2020; Wise & Kong, 2005).

Performance showed a higher correlation with number of clicks than time spent on tasks, thus possibly suggesting the need for further research. In case of interactive tasks not only too short response time can be an indicator of lack of motivation, but also too few clicks. Number of clicks may be a promising method to identify unmotivated test-takers. Sahin and Colvin (2020) supplemented response time with type of response behavior (e.g. clicks, keystrokes, and running a simulation) and total number of response behaviors (the sum of all clicks and keystrokes). The method is based on the assumption that not only time on task but also response actions are related to level of motivation. Therefore, if fewer response actions are measured, this indicates unmotivated behavior. The method yields more accurate results than only time-on-task-based methods for some cases, but no clear pattern was observed. This would be a fruitful area for further work.

Research question 3 (RQ3): How does test-taking effort change as the test progresses based on self-report questionnaire and log data-based methods?

Both the self-report questionnaire and log data-based methods show a significant decrease in test-taking effort, but the decrease is not fully consistent. The results are consistent with a number of previous studies, showing a decrease in test-taking effort as the test progresses (Lindner, Lüdtke, Grund, & Köller, 2017; Wise, 2006).

Decreasing test-taking effort implies that more attention needs to be paid to developing valid tests. Previous studies have demonstrated that adding representational pictures to text-based items improves students' test-taking motivation and test performance (Lindner, 2020; Lindner, Nagy, Ramos Arhuis, & Retelsdorf, 2017). These realistic schematic pictures illustrate important information supplied in the text but do not provide any additional information relevant to the solution beyond what is found in the text (Lindner, Nagy, et al., 2017). In contrast to representational pictures, seductive details in item stems are interesting and entertaining but task-irrelevant. Inhibiting the impulse to focus on seductive details requires high level of self-control capacity, which falls during testing (Eitel, Endres, & Renkl, 2020). Decreasing self-control capacity is linked to declining test-taking effort (Lindner et al., 2018; Lindner & Retelsdorf, 2019). Therefore, adding representational

pictures and reducing seductive details in test items lower mental fatigue effects and improve test-taking effort. Further research could usefully explore the joint effect of representational pictures and seductive details.

Research question 4 (RQ4): Which test-taking effort profiles can students be classified into based on self-reported data, log data (time on task and number of clicks) and test performance?

We identified groups of learners by defining learner test-taking effort profiles. By considering the variables noted above, we found that the optimal number of clusters was three. Students in the first cluster (Cluster 1) clicked just as little in a short period of time, as students in Cluster 2. Because time on task and number of clicks correspond to the effort invested in the tasks, they made little effort. They achieved the worst results, also rating their effort significantly lower than students in the other two clusters. Therefore, in this cluster, the self-reported data is consistent with the log data.

Students in the second cluster (Cluster 2) clicked little in a short amount of time; that is, they put little effort into completing the tasks. They achieved as good results as the students in Cluster 3 but rated their effort the highest. Participants in the third cluster (Cluster 3) clicked the most during the longest period when doing the tasks. Their results are similar to those of the second cluster, and they rated their effort lower than their peers in the second.

Students in the second cluster achieved similar results in significantly less time and with fewer clicks than those in the third cluster. This suggests that participants in the second cluster have a higher ability level than those in the third cluster. The higher-ability students in Cluster 2 clicked significantly less in less time, while rating their effort higher than their lower-ability peers in Cluster 3, who clicked significantly more in more time. This suggests that the participants' responses do not fully reflect their real test-taking behavior, thus indicating the limitations of self-report questionnaires. The reasons for their answers not fully reflecting reality could be social expectations, which may lead some students to record what is expected of them when answering, not their real thoughts and feelings. It is also possible that the less capable participants in the third cluster, who generally require more effort to complete the tasks because of their weaker abilities, underestimated their effort on the test. Another possible explanation is the inadequate self-awareness and self-esteem of some students. The results show that the answers to the self-report questionnaire are not fully consistent with the respondents' actual test-taking behavior. This is also supported by Silm et al. (2020) meta-analysis which suggests that these two types of measures could be markedly different.

One of the advantages of cluster analysis is that it offers a more accurate insight into the details. For research question 2, we examined the relationship between test-taking effort and test performance, but the positive correlation only represents the big picture. Examining the behavior shown on the test, we found that only the performance of the students in Cluster 1 was consistent with their effort. The students in Cluster 2 achieved good results with medium effort, and those in Cluster 3 achieved similar results with a great deal of effort. This finding shows that a good result does not require maximum effort, only a certain amount. This supports Gignac et al., 2019) results, and is also consistent with the results of Stenlund et al. (2018), who found that the best performers have high level of risk-taking and relatively low level of motivation. Goldhammer et al. (2017) found that higher-ability students needed less effort to solve problems successfully, which is also consistent with our findings.

6. Limitations

Our study has several limitations. One is that the test consisted exclusively of interactive problem-solving items. For this reason, the same analyses could not be used on many other types of tests, e.g., a multiple-choice test, where the correct answer for each item can be provided with a single click. Another important limitation is that we used convenience sampling at the university level and that the sample consisted of only freshers; that is, we only involved first-year university students willing to take part in the study. A further limitation is that although test performance was not related to factual knowledge, the relationship to students' cognitive abilities and problem-solving skills was not investigated. An additional limitation is that we investigated the response time effort, total time, and number of clicks in the knowledge acquisition phase, whereas this phase does not exist on most tests, which mainly consist of the knowledge application phase. A final limitation is that the test was in a low-stakes context. Thus, the results cannot be generalized.

7. Conclusions

The main objective of our research was to compare the results of self-report questionnaire-based and log data-based measures of test-taking effort in a low-stakes situation. The correlation between test-taking effort and test performance proved to be weaker based on self-reported questionnaire data than on actual test-taking behavior. Results of k-means cluster analysis also suggested that self-report questionnaire data are not completely consistent with students' actual test-taking behavior. Both the self-report questionnaire responses and the log data showed a decrease in test-taking effort during the testing session, which contained increasingly difficult, interactive, complex problem-solving tasks developed with the same approach. The level of correlation between number of clicks and test score suggests that including number of clicks in response time-based analyses may be a useful direction for further research. As for the educational implications, we are confident that a better understanding of students' test-taking behavior will both help teachers identify individual differences and provide opportunities for increased validity of low-stakes tests.

Author contributions

RC and GM were actively involved in writing the article, from planning, research, and data analysis to the preparation of the final manuscript. Both authors have read and approved the published version of the manuscript.

Funding

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 and has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the OTKA K135727 funding scheme and supported by the Research Programme for Public Education Development of the Hungarian Academy of Sciences (KOZOKT2021-16).

Institutional review board statement

Ethical approval was not required for this study based on the national and institutional guidelines. The assessments which provided data for this study formed integral parts of the educational processes of the participating university. Participation was voluntary. All of the students in the assessment were over 18; that is, it was not required or possible to request and obtain written informed parental consent from the participants.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

CRediT authorship contribution statement

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

Declaration of competing interest

The authors declare no conflicts of interest.

Data availability statement

The data presented in this study are available on request from the corresponding author.

References

AlZoubi, O., Fossati, D., Di Eugenio, B., Green, N., & CHEN, L. (2013). Predicting Students' performance and problem solving behavior from iList log data. In *Proceedings of the 21st international conference on computers in education, ICCE 2013* (pp. 1–6).

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

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

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

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

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

Crombach, M. J., Boekaerts, M., & Voeten, M. J. M. (2003). Online measurement of appraisals of students faced with curricular tasks. *Educational and Psychological Measurement, 63*(1), 96–111. https://doi.org/10.1177/0013164402239319

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

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

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

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

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

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

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

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

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. In *, 133. OECD Education Working Papers* (pp. 0–67). https://doi.org/10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In *Competence assessment in education: Research, models and instruments* (pp. 407–425). https://doi.org/10.1007/978-3-319-50030-0

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716

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., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log fi les. *Computers in Human Behavior, 61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

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

Hauser, C., & Kingsbury, G. G. (2009). *Individual score validity in a modest-stakes adaptive educational testing setting* (The Annual Meeting of the National Council on Measurement in Education).

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

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

Khorramdel, L., von Davier, M., Gonzalez, E., & Yamamoto, K. (2020). Plausible values: Principles of item response theory and multiple imputations. In D. B. Maehler, & B. Rammstedt (Eds.), *Large-Scale Cognitive Assessment: Analyzing PIAAC Data* (pp. 27–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_3.

Kriegbaum, K., Jansen, M., & Spinath, B. (2014). Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. *Learning and Individual Differences, 43*, 140–148. https://doi.org/10.1016/j.lindif.2015.08.026

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

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

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., & 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

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology, 9*, 2231. https://doi.org/10.3389/fpsyg.2018.02231

Rios, J. A. (2021). *Improving test-taking effort in low-stakes group-based educational testing: A Meta-analysis of interventions* (pp. 1–22). March: Applied Measurement in Education. https://doi.org/10.1080/08957347.2021.1890741

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

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

Schüttpelz-Brauns, K., Kadmon, M., Kiessling, C., Karay, Y., Gestmann, M., & Kämmer, J. E. (2018). Identifying low test-taking effort during low-stakes tests with the new Test-taking Effort Short Scale (TESS) – Development and psychometrics. *BMC Medical Education, 18*(1), 101. https://doi.org/10.1186/s12909-018-1196-0

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking, 2021*(31). https://doi.org/10.1186/s13638-021-01910-w

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*(July 2019). https://doi.org/10.1016/j.edurev.2020.100335

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251.

Stenlund, T., Lyrén, P. E., & Eklöf, H. (2018). The successful test taker: exploring test-taking behavior profiles through cluster analysis. *European Journal of Psychology of Education, 33*(2), 403–417. https://doi.org/10.1007/s10212-017-0332-2

Tóth, K., Rölke, H., Goldhammer, F., & Barkow, I. (2017). Educational process mining: New possibilities for understanding students' problem-solving skills. In B. Csapó, & J. Funke (Eds.), *The nature of problem solving. Using research to inspire 21st century learning* (pp. 193–209). Paris: OECD. https://doi.org/10.1201/9781003160618-1.

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., & 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 the 2012 Annual Meeting of the National Council on Measurement in Education, March, 1–24.

Wise, S. L., Ma, L., & Theaker, R. A. (2014). Identifying non-effortful student behavior on adaptive tests: Implications for test fraud detection. *Test Fraud: Statistical Detection and Methodology*, 175–185. January 2014.

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

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In *Handbook of Human and Social Conditions in Assessment* (pp. 204–220).

Wolgast, A., Schmidt, N., & Ranger, J. (2020). Test-taking motivation in education students: Task battery order affected within-test-taker effort and importance. *Frontiers in Psychology, 11*, Article 559683. https://doi.org/10.3389/fpsyg.2020.559683