UNIVERSITY OF SZEGED
DOCTORAL SCHOOL OF EDUCATION

RÓBERT CSÁNYI

TEST-TAKING ENGAGEMENT IN LOW-STAKES CONTEXT: AN EDUCATIONAL DATA SCIENCE APPROACH

PHD DISSERTATION

SUPERVISOR:
PROF. DR. GYÖNGYVÉR MOLNÁR
PROFESSOR OF EDUCATION

SZEGED
2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Gyöngyvér Molnár, for her unwavering support, invaluable guidance and insightful feedback throughout my doctoral journey. Her expertise and dedication have played a major role in shaping my research and academic development. I am truly fortunate to have had the opportunity to work under her mentorship.

Moreover, my grateful thanks go to the Doctoral School of Education of the University of Szeged for providing the stimulating academic environment and necessary resources that made this research possible. The faculty, staff and fellow doctoral candidates have all contributed to a vibrant and supportive community that has enriched my academic experience. Outstanding among the DS community was Prof. Benő Csapó, who encouraged us to carry out good research and publish in good journals. Although he is no longer with us, his support was a great inspiration to me and I am grateful to learned from him.

On a personal note, I am deeply grateful to my wife, dr. Katalin Horváth, for her endless encouragement, understanding and patience during this exhausting process. Her strong belief in me has been a constant source of strength and motivation. I am thankful to my parents, my brother, my children and my extended family for their love and support. Their faith in my abilities has been the cornerstone of my perseverance and determination.

This dissertation is as much a testament to the collective support and encouragement of my loved ones as it is the culmination of my academic efforts. To everyone who has stood by me, I offer my heartfelt thanks.

ABBREVIATIONS

aBIC	adjusted Bayesian information criterion
AIC	Akaike information criterion
ANOVA	Analysis of Variance
BIC	Bayesian information criterion
CBA	computer-based assessment
CHC	Cattell–Horn–Carroll theory
CPS	complex problem solving
DEFF	design effect
EDS	educational data science
df	degrees of freedom
g	general intelligence
ICC	intraclass correlation coefficient
ICT	information and communications technology
ISCED	International Standard Classification of Education
L-M-R	Lo–Mendell–Rubin test
NT	normative threshold
OECD	The Organization for Economic Co-operation and Development
P+>0%	proportion correct greater than zero threshold
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
RTE	response time effort
RTF	response time fidelity
SB	solution behavior
SD	standard deviation
SE	standardized error
SRE	self-reported effort
TBA	technology-based assessment
TOT	time on task
VOTAT	vary-one-thing-at-a-time strategy
WM	working memory

ABSTRACT

The research conducted within this dissertation aims to explore and comparatively investigate real-world behavioral outcomes of test-taking engagement in an interactive test environment through self-report and analysis of log and process data. This dissertation brings together three empirical studies on this issue.

The instrument used in this research consisted of 10 problems of different complexity requiring interaction to solve, which were suitable for measuring the exploration strategies of different problems. The problems, based on the MicroDYN model, were administered to first-year university students in a low-stakes testing context using the eDia online assessment platform. The exploration of the problems and the interactions used to solve them were recorded in log and process data format, and the effectiveness of different exploration and learning strategies and their impact on problem-solving performance were examined by analyzing these behavioral log data. In addition to self-report questions embedded in different parts of the test, test-taking effort was monitored by the time spent exploring and solving problems and the number of interactions.

In the research presented in the first paper, we measured students' test-taking effort using different methods and determined the optimal procedure to diagnose test-taking effort. The results suggest that the number of clicks plays an important role in predicting performance in interactive problem-solving tasks. The responses to the self-report questionnaire did not fully reflect the actual test-taking behavior of the participants. A maximum effort was not required to achieve good results, but only a certain amount.

The second study investigated item- and person-level factors that influence test-taking disengagement. For tasks administering later and for more difficult tasks, the proportion of disengaged responses increased. The proportion of disengaged responses was higher among women. Individuals with lower admission scores, lower working memory capacity and lower self-reported effort also had higher rates of disengaged responses.

In the third study, we investigated the role of test-taking effort in the knowledge acquisition through exploration behavior. Latent profile analysis of labeled behavioral data to monitor the effectiveness of the exploration strategy identified four groups. The degree of test-taking effort differed between groups and decreased to various degrees during testing.

Our results suggest that successful problem solvers put in enough time and effort to solve problems. A sufficient amount of effort does not guarantee a successful outcome, but success is not possible without it. Therefore, practitioners should place considerable emphasis on using methods that improve students' test-taking effort.

Table of contents

1. INTRODUCTION

The digital transformation of education has enabled the massive collection of student data, fostering the rise of educational data science (EDS) – an interdisciplinary field that combines education and data science to better understand and support learning. One key innovation is the use of process data, which captures not only students' responses but also their behavior during computer-based assessments (Provasnik, 2021).

A critical challenge in this context is interpreting test-taking behavior, especially in low-stakes assessments where students may not fully engage. Disengagement can distort assessment results and mask students' true abilities (Wise et al., 2014). Despite increasing interest in test-taking engagement and digital assessment, limited research has addressed how behavioral indicators – such as response time or strategy use – relate to complex problem-solving (CPS) performance.

This dissertation examines test-taking engagement on CPS tasks using process data from digital assessments administered to first-year university students. Its aims are to: (1) identify valid behavioral indicators of engagement, (2) examine factors influencing disengagement, and (3) investigate how engagement affects problem-solving strategies and outcomes. The research applies person-centered analysis, self-report measures, multilevel modelling and latent class analysis to uncover patterns in test-taking behavior.

The novelty of this work lies in its integration of self-assessment and behavioral data within digital assessments, offering new insights into the dynamics of student effort and performance. The dissertation comprises three empirical studies, each focusing on a different aspect of test-taking behavior, followed by a synthesis of findings and implications for research and practice.

Test-taking engagement

Students' performance on cognitive tests can be influenced by a variety of affective factors in addition to their actual knowledge and skills, including their test-taking engagement (Wise et al., 2014; Wolgast et al., 2020). Various studies have shown that disengaged students have significantly lower test performance than their engaged peers (Penk et al., 2014; Silm et al., 2020; Wise et al., 2021). The stakes of the tests have a significant influence on test-takers' engagement: as the stakes decrease, the level of test-taking engagement drops, thus potentially affecting test-takers' performance (Finn, 2015; Rios, 2021; Schüttpelz-Brauns et al., 2018; and Wise & Kong, 2005). On the contrary, there is also some research that suggests that maximum effort is not necessary to get valid test results, but rather that a certain level of effort is required (e.g. Gignac et al., 2019).

A frequently used theoretical framework to interpret test-taking engagement is the *expectancy-value theory* (Eccles & Wigfield, 2002; Wise & DeMars, 2005). According to the model, the level of motivation is determined by the expectation of performance and the value of the test (see Figure 1). 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*, i.e. 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-takers put 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).

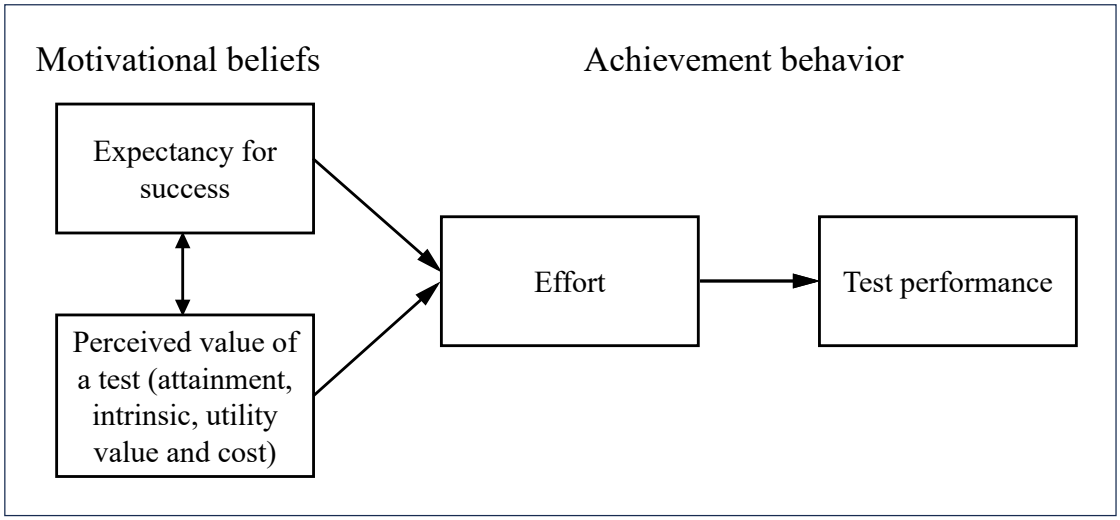


Figure 1. Expectancy-value theory (adapted from Penk & Richter (2017))

Test-taking engagement is influenced by a number of factors (e.g. Rios & Soland, 2022). These factors are related to the items, the test situation and the test-takers. Research suggests that *item-related factors* are (a) item position (e.g. Attali, 2016), (b) item difficulty (e.g. Lindner et al., 2017), (c) item type (e.g. Guo et al., 2022), (d) item length (e.g. Setzer et al., 2013) and (e) illustrations (Lindner, 2020). *Test situation related factors* are (a) stakes of the test (e.g. Wise et al., 2014), (b) time of testing (e.g. Wise et al., 2024), (c) motivational instructions (Liu et al., 2015) and (d) monetary incentives (Wise & DeMars, 2005). *Person-related factors* are (a) ability level (e.g. Kong et al., 2007), (b) working memory capacity (e.g. Lindner et al., 2019), (c) educational attainment (Goldhammer et al., 2016), (d) gender (Wise & DeMars, 2010), (e) age (Rios & Guo, 2020), (f) ethnicity (Soland, 2018) and (g) native language (Rios & Soland,

2022). Understanding these factors is essential because modifying them can have a significant impact on test-taking effort.

There are several methods for measuring test-taking effort. In the early phase, *self-report questionnaires* were applied, which generally measured the components of test-taking effort on a Likert scale. In general, students need to rate their effort in completing the test after they have finished it. This approach assumes that test-taking effort is constant, while many studies have found that it tends to decrease during the test (e.g. Attali, 2016; Penk & Richter, 2017) due to mental fatigue. Despite their usability, a major limitation is that they are subjective and it is not possible to know how honest the test-taker's responses were (Wise & Kong, 2005). The widespread diffusion of technology-based assessments allowed the development of *response time-based methods*. Response time is the time the examinee spends on a task from the time the task is presented until the "next" button is pressed. The basic assumption of response time-based methods is that disengaged examinees spend less time on tasks and therefore respond faster than their engaged counterparts (Wise & Kong, 2005). The main advantages of these methods are that they measure the actual behavior of the examinees rather than their perceptions, they do not require extra work for them, and the change in effort can be tracked from item to item (Wise & Ma, 2012). Using response time-based methods, we define a threshold of a certain way, and if the response time is shorter than the threshold, we identify the response as disengaged; if it is longer, we identify it as engaged (Wise & Kong, 2005). This is an important issue to identify the appropriate threshold (e.g. Bulut et al., 2023).

Test-taking behavior has an important role in test performance. Various studies examined the topic, but in a minority of cases they used a person-centered approach, looking at the individual test-taking behavior of learners. Research used cluster analysis to examine the behavior of students. The number of clusters and the characteristics of the students in each cluster are dependent on a number of factors, including the task, the sample and the variables included in the analysis (Goldhammer et al., 2017; Lundgren & Eklöf, 2020; Stenlund et al., 2018).

Technology-based assessment

Educational assessment in the digital age has gone through a major transformation, driven by the rapid development of information and communication technologies (ICT). While traditional paper-pencil tests have long dominated summative assessment practices focusing on factual knowledge (Molnár & Csapó, 2019) new social and educational demands require the assessment of complex skills such as problem solving, collaboration and digital literacy that

cannot be effectively assessed with traditional instruments (Redecker & Johannessen, 2013; Goldhammer et al., 2013).

Technology-based assessment (TBA) goes far beyond the mere digitalization of traditional tests. Leveraging the possibilities of technology to create interactive, dynamic and learner-centered assessment environments. Compared to paper-pencil testing, TBA provides a range of advantages, such as adaptive testing, multimodal item formats, embedded simulations and real-time feedback (Fink et al., 2023). These assessments can also collect process data - such as response time, number of interactions and navigation pathways - that provide insights into learners' cognitive and behavioral processes (Goldhammer et al., 2013; Greiff et al., 2012). This process-oriented approach supports both summative objectives and both diagnostic and formative learning processes.

Formative assessment, a central component of 21st-century education, aims to improve learning by identifying students' current understanding and providing feedback that helps them progress (Redecker & Johannessen, 2013). In digital environments, technology can enhance the feedback loop by delivering immediate, personalized, and actionable insights to learners and educators alike (Shute, 2008). This integration of assessment and instruction reinforces learning outcomes and promotes self-regulation and motivation.

Educational assessment has gone through a fundamental paradigm shift, moving beyond the mere digitalization of traditional tests to the integration of technology, which is fundamentally redefining what and how we assess. The progression of technology-based assessment can be conceptualized in four broad phases. The first phase (computer-based assessment, CBA) involves the digital administration of conventional assessments, where paper-based tests are transferred to digital platforms, allowing for automated scoring and faster data processing. This stage offers limited benefits of the technology. The second phase can be termed a technology-based assessment (TBA), because the testing was not only carried out on a personal computer, but also on other technological devices (e.g. laptop, tablet). In this phase, it started to measure constructs that could not be assessed with paper-pencil tests (e.g. complex problem-solving). The third stage assessment steps further and often includes interactive tasks or simulations and is integrated into the teaching-learning process. In addition, the data collected during the learning process provides real-time insights not only into learners' achievements but also into their learning behavior. The current and most advanced stage is beginning to evolve with the embedding of artificial intelligence in the evaluation process (Foster & Piacentini, 2023; O'Leary et al., 2018; Molnár & Csapó, 2019; Fink et al., 2023; Redecker & Johannessen, 2013).

In summary, technology-based assessment enables the measurement of complex skills, improves the quality and immediacy of feedback, and provides process-level insights into students' engagement and behavior. These features make TBA an effective tool for understanding and supporting learning in contemporary education systems.

Educational data science

The remarkable development and widespread availability of IT tools in the late 20th and 21st centuries have made it possible to use technology in the teaching-learning process on a daily basis. The use of technology has the potential to encompass the whole educational process, covering all levels of education from pre-school through higher education to adult learning. One of the biggest benefits of using technology in education is the possibility to move beyond the traditional, frontal, teacher-centered "one size fits all" approach and to make personalized learning possible (Csányi et al., 2024; Molnár, 2021).

Nowadays, there is so large amount of data available in the field of education that educational research is becoming a data-intensive field, using data science methods and techniques (Daniel, 2019). Irizarry (2020) proposed a definition that "*data science* is an umbrella term to describe the entire complex and multistep processes used to extract value from data". Data science provides a structure and principles that support the extraction of information and knowledge from data (Daniel, 2019). Similar to the previous definition, it is also suggested to use the term *educational data science* (EDS) (Cerezo et al., 2024; Peña-Ayala, 2023). EDS in a narrower definition is the application of statistics and IT tools in the field of education. In a broader definition, however, it refers to a number of new, non-traditional quantitative methods applied to educational problems, often using novel data (McFarland et al., 2021). Before the advent of the EDS, quantitative educational research was mainly based on administrative data, such as students' socio-economic background and educational achievement, but the EDS has opened up new possibilities.

The use of modern technologies in education allows the gathering of contextual data (time on task, jumps back and forth, eye movements, clicks, etc.) that are unimaginable in traditional teaching and assessment systems (Tóth et al., 2017). Based on the analysis of contextual data, it is possible to personalize education even further, thus improving the effectiveness of education. Instructors can receive objective feedback on the teaching-learning process, which can be used to analyze students' learning and behavior (Romero & Ventura, 2020). The automatic gathering of data on students' behavior, often referred to as *log files* or *process data*, has become increasingly common in digital-based assessment. These information can help determine student engagement, improve test design, identify construct inferences (Anghel et al., 2024), reducing costs (e.g. Wise,

2019) and allow automatic item development, automatic scoring (e.g. Csapó et al., 2014). According to Provasnik (2021) the two terms overlap, but they are not synonymous. He declares, that "logfiles are everything captured in digital-based assessment (DBA)", while "process data are the empirical data that reflect the process of working on a test question". Based on this definition, log files are a subset of process data. Process data may include information that is not commonly recorded as part of the DBA, such as eye-tracking information, magnetic resonance imaging (MRI) or observations by interviewers.

The application of new methods is the other important feature of EDS, thanks to the use of machine learning. Machine learning algorithms can be divided into four broad categories: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning (Peña-Ayala, 2023). Supervised learning creates a function that models an input variable onto an output variable based on input-output pattern pairs. In this case, we label the input and output variables (e.g. regressions, deep neural networks, k-nearest neighbors, decision trees). Unsupervised learning analyses unlabeled data sets without human assistance based on similarity in the population, e.g. clustering. Semi-supervised learning uses both labelled and unlabelled data, e.g., classification, clustering. In reinforcement learning, the algorithm trains itself through repeated trial and error iterations as it interacts with the environment (Cady, 2017; McFarland et al., 2021; Peña-Ayala, 2023).

Complex problem-solving

A *problem* can be defined as "whenever there is a gap between where you are now and where you want to be, and you don't know how to find a way to cross that gap" (Hayes, 1981). Solving a problem requires both understanding the nature of the problem and finding ways to bridge the gaps. *Simple problems* are static situations in which all information is available to the problem-solvers, who have to find a solution in this unchanging environment (Fischer et al., 2015). Problems where part of the information is not available at the beginning of the problem situation, and where the situation is dynamically changing and has many interrelated elements, are termed *complex problems* (Greiff et al., 2018). Frensch & Funke (1995) defined complex problem-solving (CPS) as "to overcome barriers between a given state and a desired goal state by means of behavioral and/or cognitive, multistep activities" (p. 37). In the literature, there are a number of synonyms for CPS, such as *dynamic, interactive and creative problem solving* (Csapó & Funke, 2017). Based on previous research, CPS is a specific ability, which has high predictive power for job performance and academic performance (Greiff et al., 2012; Wüstenberg et al., 2012). The importance of the CPS is demonstrated by the fact that it is considered one of the 21st

century skills (e.g. Csapó & Funke, 2017; Rios et al., 2020) and is therefore included in both the Programme for International Student Assessment (PISA), where it was assessed in 2012, and Programme for the International Assessment of Adult Competencies (PIAAC) assessments, where it was included in both assessment cycles (Csapó & Funke, 2017; Greiff et al., 2017).

The process of carrying out CPS tasks can be divided into two phases: (1) knowledge acquisition and (2) knowledge application phases (Greiff et al., 2013), which are characteristic of the MicroDYN approach, a well-established framework for assessing complex problem-solving through computer-based simulations. During the knowledge acquisition phase, students interact with the simulated system, understand the structure of the problem and represent the newly generated knowledge in the form of a mental model. In the knowledge application phase, they apply the new knowledge and the representation of the problem structure in a targeted way to solve a given problem (Wüstenberg et al., 2014). An important feature of CPS tasks is that they do not require prior content knowledge, as they focus on domain general processes (Csapó & Funke, 2017; Frensch & Funke, 1995; Funke, 2001).

CPS tasks can be solved using a variety of methods, but the most successful and therefore most studied strategy is VOTAT (vary-one-thing-at-a-time) strategy (Molnár & Csapó, 2018). Within this approach, problem solvers are presented with a system comprising input variables that can be manipulated and output variables that can be observed. Using the VOTAT strategy, the problem solver changes only one input variable at a time while keeping the others constant, allowing them to isolate and identify the effect of that single variable on the output. The exploration part of the two phases dominates the whole problem-solving process, because if it fails, it means the whole process fails, even if after the first phase the learners are given the correct problem structure, while, if successful, the likelihood of successfully solving the problem increases significantly (Molnár & Greiff, 2023). This is the reason why the dissertation focuses on monitoring and analysing the activities in the first phase of problem solving.

The VOTAT strategy was used by the majority of researchers as an input variable for the person-centred approach, as it is one of the most important indicators of high CPS performance (Greiff et al., 2015; Molnár & Csapó, 2018; Molnár et al., 2022). The instrument is interactive, i.e. to solve the problem it is necessary to interact with the tasks, which provides more potential for reproducing and analysing the behavior of the participants during testing. Previous research has grouped students according to the extent to which they use the strategy. Research investigated different age groups in different educational contexts, identifying three to six latent groups (Greiff et al., 2018; Molnár & Csapó, 2018; Molnár, 2022; Molnár et al., 2022; Wu & Molnár, 2021). If the same age group is studied in the same educational context, the latent structure is stable, i.e.

the same latent classes are obtained (Molnár, 2017; Molnár, 2022; Molnár et al., 2022). Thus, research suggests that the number and characteristics of latent classes depend on the age group and the educational context.

Aims and methods of the research

The primary aim of this dissertation is to investigate test-taking engagement in technology-based assessments through the lens of educational data science. Based on three empirical studies, it examines how process data and response patterns can provide insights into students' engagement during assessments. The dissertation seeks to enhance the validity and interpretability of assessment results in digital environments.

The dissertation addresses three main research gaps in the study of test-taking engagement in technology-based assessments. First, while both self-reported and response time-based indicators are commonly used to measure engagement, few studies have examined these approaches simultaneously or systematically compared their outcomes. Second, although previous research has identified various predictors of test-taking engagement, the results have been partly contradictory and many studies have ignored the multilevel structure of the data, limiting the robustness of their conclusions. Third, while students' problem-solving behavior, especially when using the VOTAT strategy, has been widely studied, the role of test-taking engagement in these behaviors has not been taken into account. By addressing these gaps, the dissertation contributes to a more comprehensive understanding of engagement in low-stakes digital assessment contexts.

The research used a complex problem-solving test based on the MicroDYN approach. These tasks focus on fictitious situations and are thus independent of the influence of prior school learning (Funke, 2014). The assessment was administered via the eDia system, which allowed the collection of process data (response time, number of interactions). The research sample consisted of first-year students at the University of Szeged, and the participation was voluntary, and they received one academic credit for completing the tests (see the Ethics Approval in the Appendix).

In the first study, students' test-taking effort was examined by integrating and comparing traditional self-report questionnaire data and students' test-taking behavior, based on the analysis of process data. Previous research has developed several methods based on process data to identify disengaged responses. These methods produce different results on the same sample (e.g. Goldhammer et al., 2016), so we aimed to investigate the optimal method. We analyzed the relation between test performance, time on task, number of clicks and test-taking effort. Finally, previous research suggest that the test-taking profile of students depends on a number of factors,

and studies have found different results. The number and characteristics of groups vary depending on the task, the sample and the variables included in the analysis. Therefore, we used k-means cluster analysis to group students based on the data generated during the test completion and to examine the characteristics of the clusters. These findings have provided the basis for further investigation of how individual and item-related factors contribute to test-taking behavior.

The second study examined the item- and person-level factors that influence test-taking disengagement. 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. As a consequence, multilevel modelling was used to identify item- and person-level factors that influence test-taking disengagement. Building on these results, the third study focused on how test-taking effort is related to learners' behavioral patterns in complex problem-solving tasks.

The aim of the third study was to investigate the role of test-taking effort in knowledge acquisition via problem exploration behavior used in complex problem-solving environment. Several studies have investigated how students can be classified using their problem-solving strategy. Previous research has produced different results and has not always been able to explain the different characteristics of different groups. Therefore, we investigated the role of test-taking effort in the problem-solving strategy used, specifically in the knowledge acquisition phase of the problem-solving process. Students' exploration behavior was coded based on the VOTAT strategy, and latent class analysis was used to identify students' behavioral and learning profiles. Then the extent and variation of test-taking effort in the different groups was examined. This allowed a deeper understanding of the interaction between test-taking engagement and problem-solving behavior in low-stakes digital assessment contexts.

The present dissertation contributes to a deeper understanding of test-taking engagement in technology-based assessments by combining process data with traditional self-report measures, and by applying innovative analytical approaches. Despite the increasing availability of process data in technology-based testing, relatively few studies have systematically compared behavioral and self-reported indicators of disengagement. Furthermore, the research addresses methodological gaps by accounting for the multilevel structure of assessment data and exploring under-investigated predictors such as working memory. The third study adds to the field by integrating test-taking effort into the analysis of complex problem-solving behavior, an aspect largely ignored in previous research. Together, these studies offer a novel, data-driven perspective

on how students engage with assessment tasks, and provide valuable insights for improving the validity and interpretation of test results in educational measurement.

This dissertation is presented in a *study-based format*. The main part of the dissertation contains three empirical studies published in D1 and Q1 journals:

(1) Csányi, R., & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences*, *106*, 102340. https://doi.org/10.1016/j.lindif.2023.102340

(2) Csányi, R., & Molnár, G. (2024). Item- and person-level factors in test-taking disengagement: Multilevel modelling in a low-stakes context. *International Journal of Educational Research Open*, *7*, 100373. https://doi.org/10.1016/j.ijedro.2024.100373

(3) Csányi, R., & Molnár, G. (2025). Looking beyond students' exploration and learning strategies: The role of test-taking effort in complex problem-solving. *Intelligence*, *109*, 101907. https://doi.org/10.1016/j.intell.2025.101907

References

Anghel, E., Khorramdel, L., & von Davier, M. (2024). The use of process data in large-scale assessments: a literature review. *Large-Scale Assessments in Education*, *12*. https://doi.org/10.1186/s40536-024-00202-1

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

Bulut, O., Gorgun, G., Wongvorachan, T., & Tan, B. (2023). Rapid guessing in low-stakes assessments: Finding the optimal response time threshold with random search and genetic algorithm. *Algorithms*, *16*(2). https://doi.org/10.3390/a16020089

Cady, F. (2017). Machine learning overview. In *The data science handbook* (pp. 87–91). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781119092919.ch6

Cerezo, R., Lara, J.-A., Azevedo, R., & Romero, C. (2024). Reviewing the differences between learning analytics and educational data mining: Towards educational data science. *Computers in Human Behavior*, *154*, 108155. https://doi.org/https://doi.org/10.1016/j.chb.2024.108155

Csányi, R., Lőrincz, M. M., & Molnár, G. (2024). Egy felnőttképzési MOOC-programon részt vevők aktivitási adatainak elemzése. *Információs Társadalom*, *24*(1), 34. https://doi.org/10.22503/inftars.xxiv.2024.1.2

Csapó, B., & Funke, J. (Eds.). (2017). *The nature of problem solving: Using research to inspire 21st century learning*. OECD Publishing. https://doi.org/10.1787/9789264273955-en

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology*, *106*(3), 639–650. https://doi.org/10.1037/a0035756

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. In *British Journal of Educational Technology* (Vol. 50, Issue 1, pp. 101–113). Blackwell Publishing Ltd. https://doi.org/10.1111/bjet.12595

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

Fink, A., Spoden, C., & Frey, A. (2023). Determinants of higher education teachers' intention to use technology-based exams. *Education and Information Technologies*, *28*(6), 6485–6513. https://doi.org/10.1007/s10639-022-11435-4

Fischer, A., Greiff, S., Wüstenberg, S., Fleischer, J., Buchwald, F., & Funke, J. (2015). Assessing analytic and interactive aspects of problem solving competency. *Learning and Individual Differences*, *39*, 172–179. https://doi.org/10.1016/j.lindif.2015.02.008

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills*. OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Frensch, P., & Funke, J. (1995). Complex problem solving — The European perspective. In *Learning to Solve Complex Scientific Problems*.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning*, *7*(1), 69–89. https://doi.org/10.1080/13546780042000046

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

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

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

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills. *European Journal of Psychological Assessment*, *29*(4), 263–275. https://doi.org/10.1027/1015-5759/a000153

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

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., Scheiter, K., Scherer, R., Borgonovi, F., Britt, A., Graesser, A., Kitajima, M., & Rouet, J.-F. (2017). *Adaptive problem solving: Moving towards a new assessment domain in the second cycle of PIAAC* (OECD Education Working Papers, Vol. 156). https://doi.org/10.1787/90fde2f4-en

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers and Education*, *91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, *36*(3), 189–213. https://doi.org/10.1177/0146621612439620

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

Hayes, J. R. (1981). *The complete problem solver*. Franklin Institute Press.

Irizarry, R. A. (2020). The Role of Academia in Data Science Education. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.dd363929

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

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), e0180149. https://doi.org/10.1371/journal.pone.0180149

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

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

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

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

McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education Data Science: Past, Present, Future. *AERA Open*, *7*. https://doi.org/10.1177/23328584211052055

Molnár, G. (2017). A problémamegoldó és tanulási stratégiák változása 11 és 19 éves kor között: Logfile elemzések. *Magyar Pedagógia*, *117*(2), 221–238. https://doi.org/10.17670/MPed.2017.2.221

Molnár, G. (2021). Az IKT szerepe a felsőoktatás megújításában. *Magyar Tudomány*, *182*(11), 1488–1501.

Molnár, G. (2022). How to make different thinking profiles visible through technology: the potential for log file analysis and learning analytics. In M. Virvou, G. A. Tsihrintzis, L. H. Tsoukalas, & L. C. Jain (Eds.), *Advances in Artificial Intelligence-based Technologies. Learning and Analytics in Intelligent Systems* (pp. 125–145). Springer. https://doi.org/10.1007/978-3-030-80571-5_9

Molnár, G., Alrababah, S. A., & Greiff, S. (2022). How we explore, interpret, and solve complex problems: A cross-national study of problem-solving processes. *Heliyon*, *8*(1). https://doi.org/10.1016/j.heliyon.2022.e08775

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

Molnár, G., & Csapó, B. (2019). How to make learning visible through technology: The eDia-online diagnostic assessment system. In H. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *CSEDU 2019. Proceedings of the 11th International Conference on Computer Supported Education. Volume 2.* (pp. 122–131). Heraklion, Crete: Scitepress.

Molnár, G., & Greiff, S. (2023). Understanding transitions in complex problem-solving: Why we succeed and where we fail. *Thinking Skills and Creativity*, *50*. https://doi.org/10.1016/j.tsc.2023.101408

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

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education*, *53*(2), 160–175. https://doi.org/10.1111/ejed.12271

Peña-Ayala, A. (2023). Educational data science: An "Umbrella term" or an emergent domain? In A. Peña-Ayala (Ed.), *Educational data science: Essentials, approaches, and tendencies* (pp. 95–147). Springer Singapore. https://doi.org/10.1007/978-981-99-0026-8_3

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

Provasnik, S. (2021). Process data, the new frontier for assessment development: rich new soil or a quixotic quest? *Large-Scale Assessments in Education*, *9*(1). https://doi.org/10.1186/s40536-020-00092-z

Redecker, C., & Johannessen, Ø. (2013). Changing Assessment-Towards a New Assessment Paradigm Using ICT. *European Journal of Education*, *48*(1). http://c4lpt.co.uk/top-100-tools-2012/;

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., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements. *Educational Researcher*, 1–10. https://doi.org/10.3102/0013189X19890600

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

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1–21. https://doi.org/10.1002/widm.1355

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

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

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.

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. (2019). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry*, *10*(1), 21–33. https://doi.org/10.1080/20004508.2018.1490127

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., Kuhfeld, M. R., & Lindner, M. A. (2024). Don't test after lunch: The relationship between disengagement and the time of day that low-stakes testing occurs. *Applied Measurement in Education*, 1–15. https://doi.org/10.1080/08957347.2024.2311925

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. In N. Kingston & A. Clark (Eds.), *Test Fraud: Statistical Detection and Methodology* (Issue January 2014, pp. 175–185). Routledge.

Wu, H., & Molnár, G. (2021). Logfile analyses of successful and unsuccessful strategy use in complex problem-solving: a cross-national comparison study. *European Journal of Psychology of Education*, *36*(4), 1009–1032. https://doi.org/10.1007/s10212-020-00516-y

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving - More than reasoning? *Intelligence*, *40*(1), 1–14. https://doi.org/10.1016/j.intell.2011.11.003

Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge and Learning*, *19*(1–2), 127–146. https://doi.org/10.1007/s10758-014-9222-8

2. PUBLISHED PAPERS

Paper 1: How do test-takers rate their effort? A comparative analysis of self-report and log file data

Published as:

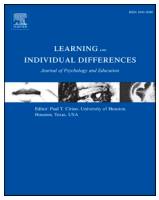
Csányi, R., & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences*, *106*, 102340. https://doi.org/10.1016/j.lindif.2023.102340

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

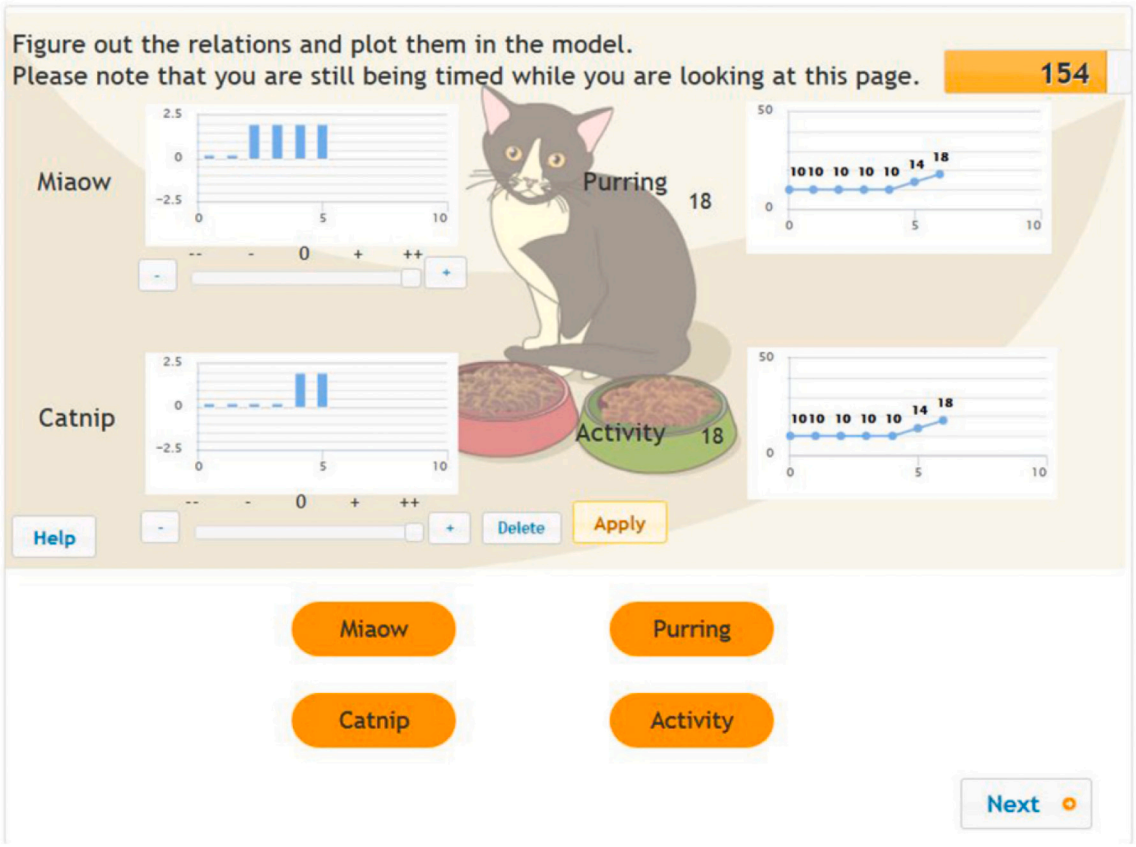


Fig. 1. The first stage of interactive problem-solving: exploring the relationship between two input and two output variables.

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

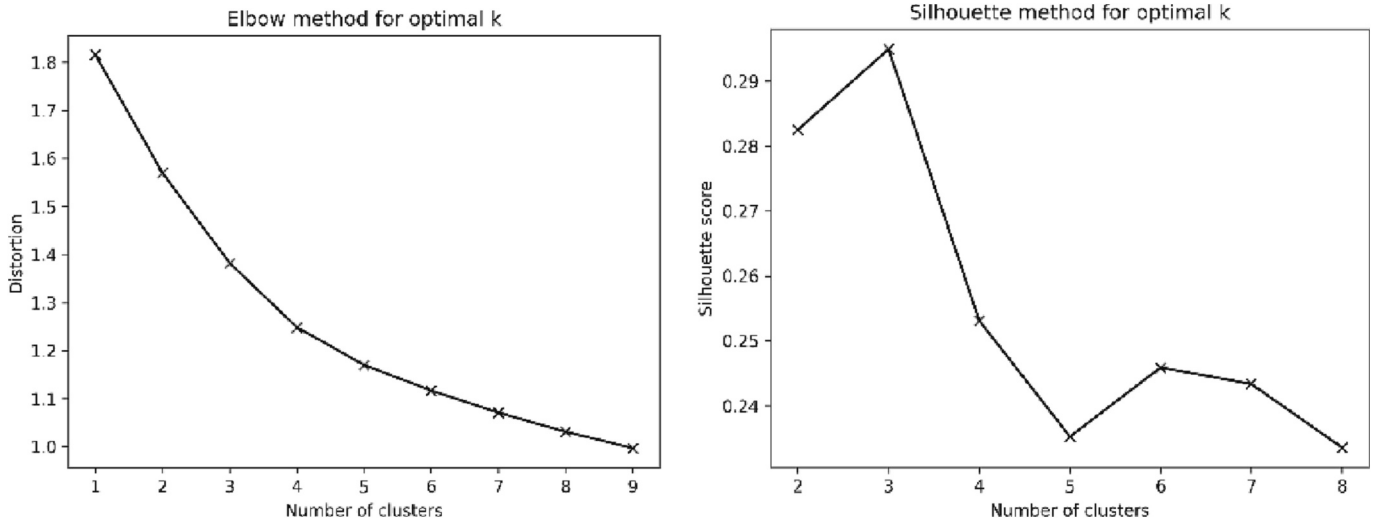


Fig. 2. Optimal cluster numbers for the elbow and silhouette methods.

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

Paper 2: Item- and person-level factors in test-taking disengagement: Multilevel modelling in a low-stakes context

Published as:

Csányi, R., & Molnár, G. (2024). Item- and person-level factors in test-taking disengagement: Multilevel modelling in a low-stakes context. *International Journal of Educational Research Open*, *7*, 100373. https://doi.org/10.1016/j.ijedro.2024.100373

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

Paper 3: Looking beyond students' exploration and learning strategies: The role of test-taking effort in complex problem-solving

Published as:

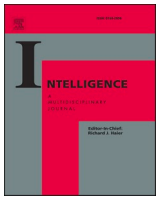
Csányi, R., & Molnár, G. (2025). Looking beyond students' exploration and learning strategies: The role of test-taking effort in complex problem-solving. *Intelligence*, *109*, 101907. https://doi.org/10.1016/j.intell.2025.101907

Intelligence 109 (2025) 101907

Contents lists available at ScienceDirect

Intelligence

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



Looking beyond students' exploration and learning strategies: The role of test-taking effort in complex problem-solving

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

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

ARTICLE INFO

Keywords:
Complex problem-solving
Test-taking effort
Exploration profiles
Knowledge acquisition
Latent profile analysis
Process data

ABSTRACT

The aim of this study was to investigate the role of test-taking effort in knowledge acquisition via problem exploration behaviour used in complex problem-solving (CPS) environments. The sample consisted of undergraduate students just starting their university studies ($n = 1748$). MicroDYN-based tasks with different levels of complexity were administered via the eDia online platform. Students' exploration behaviour was coded based on the VOTAT (vary-one-thing-at-a-time) strategy, and latent class analysis was used to identify students' behavioural and learning profiles. We identified four profiles: rapid learners, non-performers, proficient explorers and ineffective learners. Students' test-taking effort was measured based on the time they spent on the tasks. Results suggest a strong relation between VOTAT strategy use and test-taking effort. Rapid learners and proficient explorers displayed the greatest test-taking effort, followed by ineffective learners and non-performers. The results provide a new interpretation of previous analyses of the knowledge acquisition phase in CPS.

1. Introduction

Education in the 21st century faces many new challenges. One of the most significant of these is how to prepare students for an increasingly uncertain, volatile, complex and ambiguous world. In earlier times, the knowledge that teachers passed on to their students lasted a lifetime. Today, teachers need to prepare students for jobs that do not yet exist and for economic and social changes that have not been experienced before. This has also changed the demand for what students need to be taught. Instead of the hegemony and exclusivity of disciplinary knowledge, the applicability of knowledge and the ability to generate new knowledge have come to the forefront, thus boosting the role of skills that are generally required in all aspects of life. These skills are known by various labels, e.g. 21st-century skills, transversal skills and core skills. Among these skills, problem-solving is of particular importance, having become one of the measured domains in the most prestigious international large-scale educational assessments: PISA for students and PIAAC for adults.

General intelligence (*g*) is one of the most widely used constructs to predict cognitive performance (Wüstenberg et al., 2012). In general, tests that are highly correlated with *g* have strong predictive power for life outcomes, but the prediction is not perfect (Coyle and Greiff, 2021). Based on previous research, complex problem-solving (CPS) is a specific ability that differs conceptually and empirically from general intelligence (Greiff et al., 2013b; Wüstenberg et al., 2012) and has high predictive power for job performance and academic performance. CPS is incorporated into comprehensive models of human intelligence, such as the Cattell–Horn–Carroll (CHC) theory (Coyle and Greiff, 2021; McGrew, 2009).

Valid and reliable educational assessment provides information on students' learning and performance. However, students' cognitive test performance is influenced by a variety of affective factors, including test-taking disengagement (Wise et al., 2014). For low-stakes tests, the impact can be significant; according to Rios and Soland (2022), more than half of the test-takers on the 2018 PISA assessment showed disengagement on at least 10 % of the items. Previous research has clustered students according to their problem-solving strategies, but, to the best of our knowledge, none of these studies have taken test-taking effort into account. We fill this niche and analyse students' exploration strategies combined with their test-taking effort.

* Corresponding author at: Doctoral School of Education, University of Szeged, Petőfi sgt. 32–34., H–6722 Szeged, Hungary.
E-mail addresses: csanyi.robert@edu.u-szeged.hu (R. Csányi), gymolnar@edpsy.u-szeged.hu (G. Molnár).

https://doi.org/10.1016/j.intell.2025.101907
Received 5 October 2024; Received in revised form 30 December 2024; Accepted 25 February 2025

1.1. Complex problem-solving: definition, processes and problem-solving strategies

A *problem* occurs "whenever there is a gap between where you are now and where you want to be, and you don't know how to find a way to cross that gap" (Hayes, 1981, p. i). To solve a problem, it is essential to understand the nature of the problem and to find a means to bridge the gap. There are numerous definitions of simple and complex problems in the literature. *Simple problems* are static situations in which all the information is given and the problem-solver is expected to find a solution in an unchanging environment (Fischer et al., 2015). Problems where not all the information is given at the onset of the problem situation and where the situation is dynamically changing and contains many interrelated elements are referred to as *complex problems* (Greiff et al., 2018). Frensch and Funke (1995) defined CPS as "to overcome barriers between a given state and a desired goal state by means of behavioural and/or cognitive, multistep activities" (p. 37). CPS is not only an extension of simple problem-solving; rather, it is structurally different. The terms *dynamic*, *interactive* and *creative problem-solving* are also used in the literature as synonyms for CPS (Csapó and Funke, 2017).

One of the most frequently used options for investigating CPS is the MicroDyn approach (Funke, 2014; Greiff and Funke, 2017), which is applied in this study. This approach is the optimal method to study how effectively students discover unknown, fictional problems where they cannot use and apply their prior school knowledge, and then, after having identified the problem, how effectively they can apply the newly acquired knowledge. MicroDYN tasks consist of up to three input variables and up to three output variables. The input and output variables can be related in different ways, and these relationships need to be explored by the problem-solver (Greiff et al., 2013c). Each task has a fictitious cover story, with the names of the input and output variables also being fictitious (e.g. Miaow as the name for a cat food).

The process of completing CPS tasks can be broken down into two phases: the (1) knowledge acquisition and (2) knowledge application phases (Greiff et al., 2013c). Knowledge acquisition involves students' exploration of the problem space based on their interaction with the simulated system and depicting their understanding of the problem structure, the newly generated knowledge in the form of a mental model. Completing the tasks only requires participants to identify causal structures using active experimentation and then using those causal structures (Funke, 2014). Learning in the process of system exploration is a prerequisite for successful system operation (Wüstenberg et al., 2012) because problem-solvers have to learn how variables are related. Acquiring knowledge about an unfamiliar, complex system is a good indicator of a person's ability to learn. Adapting flexibly to new circumstances, understanding the situation and applying new information are strongly linked to learning and applying learned content (Greiff et al., 2013a). Knowledge application is the process of applying this new knowledge and the representation of the problem structure in a targeted way towards solving the problem (Wüstenberg et al., 2014). CPS tasks do not require previously mastered content knowledge, as they focus on domain-general processes (Csapó and Funke, 2017; Frensch and Funke, 1995; Funke, 2001).

These problems can be solved using a number of methods (Molnár and Csapó, 2018), but the most commonly used, most successful and consequently most studied strategy is vary-one-thing-at-a-time (VOTAT). With the VOTAT strategy, the problem-solver changes only one input variable at a time, leaving the others unchanged. Thus, it is only possible for the effect of the modified variable to be responsible for later changes. Based on Molnár and Greiff (2023), the exploration part is the dominant one in the entire problem-solving process. If that fails, it implies the failure of the whole process – in both phases, even if we provide students with the correct problem structure after the first phase. If successful, the probability of successful problem-solving increases significantly; that is, a key part of the problem-solving process is mapping the problem and its correct interpretation and visualization. This is also the reason why the main focus of the present study is on monitoring and analysing the activities carried out in the first phase of problem-solving.

1.2. Problem exploration and learning profiles in the knowledge acquisition phase of the CPS process

A majority of researchers have used the VOTAT strategy as an input variable for the person-centred approach, as it is one of the most relevant indicators of high CPS performance (Greiff et al., 2015; Molnár et al., 2022; Molnár and Csapó, 2018). Previous research has identified different groups of students depending on the extent to which they used this special strategy.

Greiff et al. (2018) used latent class analysis to investigate the exploration strategies of students in Grades 6–8 and identified six qualitatively different classes: proficient explorers, intermediate explorers, low-performing explorers, rapid learners, emerging explorers and non-persistent explorers. Molnár and Csapó (2018) studied the exploration strategies of 3rd- to 12th-grade students in CPS. Three subsamples were created by grade, ranging from four to six latent classes. Among university students, Molnár (2022) identified four latent class profiles: (1) proficient explorers, (2) almost high performers on the easiest problems but low performers on the complex ones, (3) rapid learners and (4) low to intermediate performers on the easiest problems but non-performers on the complex ones. Wu and Molnár (2021) compared the exploration strategy profiles of Hungarian and Chinese 6th graders in a CPS environment. Both the Chinese and Hungarian samples fell into three qualitatively different exploration class profiles, but they showed different behavioural patterns. Molnár et al. (2022) investigated exploration strategies among Hungarian and Jordanian university students. They identified four latent classes in both samples, but there were significant differences between the two samples. To sum up, the number and characteristics of latent classes depend on students' age and the educational context. Where research examined the same age group (university students) in the same educational context (*Hungary*), the latent structure is stable; that is, the same latent classes are obtained (Molnár, 2017; Molnár, 2022; Molnár et al., 2022).

1.3. Test-taking effort

One of the most common models to explain test-taking effort (Anghel et al., 2024) is the *expectancy-value theory* (Eccles and Wigfield, 2002). According to this theory, test-takers' motivation is determined by expected performance and the value of the test. Test-takers' expectations are influenced by their own perception of their abilities and the difficulty of the tasks. Values are made up of four components: (1) attainment value, which is the importance of the test; (2) intrinsic value, which is measured by the enjoyment with which the task is completed; (3) utility value, which represents the relationship of the task to future goals; and (4) cost, that is, the negative attributes of the task, such as time spent on the task and anxiety about the tests. Test-taking effort is a manifestation of the motivation that the examinee puts into test completion.

According to previous research, when test results have no personal consequences for students, examinees tend to invest low test-taking effort into the tasks, resulting in underestimation of student performance (Michaelides and Ivanova, 2022; Penk and Richter, 2017; Rios et al., 2017). In a meta-analysis, Wise and DeMars (2005) found that test-takers in low-stakes test contexts scored on average 0.59 standard deviations (SD) lower than test-takers in high-stakes assessment contexts. As a result, both score-based conclusions and decisions about educational stakeholders (e.g. schools and education systems) can be biased (Rios, 2021).

There are different approaches to measure test-taking effort, one of the most common of which is the response time-based method, both in the literature and in practice (Rios et al., 2022), which has been made

possible by the spread of digital-based assessments. Response time is the time between an item being administered and the examinee's response. The principle behind response time-based methods is that reading and interpreting a problem is a basic prerequisite for solving it, followed by solving the problem itself. Consequently, participants who display low test-taking effort spend less time on the task (reading, interpreting and completing it) and therefore respond faster than those with higher motivation (Wise and Kong, 2005). *Response time-based methods* have several advantages over the self-report questionnaires traditionally used in the educational context, thus leading to their widespread application: (1) their use is unobtrusive; that is, the examinees do not perceive that their behaviour is being monitored; (2) they measure the actual behaviour of the examinees rather than their perceptions; (3) they do not require additional work on the part of the examinee; and (4) the change in effort can be monitored from item to item (e.g. Silm et al., 2020; Wise and Ma, 2012). The latter feature is of great importance because previous research has shown that test-taking effort can vary during the testing process (e.g., Attali, 2016; Penk and Richter, 2017). When we use response time-based methods, a threshold value must be defined for a given method. When the response time is shorter than the threshold, the response is identified as disengaged. When it is longer or equal, it is identified as engaged (Wise and Kong, 2005). The method was introduced using a constant threshold, i.e. a predefined threshold for each item (e.g. 3 or 5 s). The main limitation of the constant threshold is that the minimum time required to complete a task is not the same for each task, so identification can be biased. For this reason, item-specific thresholds have been introduced, meaning that the threshold varies from item to item (Goldhammer et al., 2016).

Another indicator of test-taking effort is *number of clicks*. The underlying assumption of this approach is similar to that of response time-based methods; that is, disengaged participants use fewer clicks than their engaged counterparts (Sahin and Colvin, 2020). This is supported by previous research showing a correlation between number of clicks and test performance (Arguel et al., 2019; Csányi and Molnár, 2023). A severe limitation of these methods is that they can only be used for constructed response tasks, where the students are expected to interact with the system, and cannot be used for selected response tasks, where a single click is all it takes to provide the response.

1.4. Research purpose and questions

Several studies have investigated how students can be classified using their problem-solving strategy. Previous research has produced different results and has not always been able to explain the different characteristics of different groups. Therefore, we investigated the role of test-taking effort in the problem-solving strategy used, specifically in the knowledge acquisition phase of the problem-solving process.

In our study, students were classified into qualitatively different classes based on their exploration behaviour in CPS contexts. The application of the principle of isolated variation was analysed through a series of CPS tasks using latent class analysis. Then, we examined the test-taking effort of these classes. To our knowledge, no similar study has been performed. These objectives were addressed via the following research questions:

RQ1: What are the exploration and learning profiles?
RQ1a: Which exploration and learning profiles can students be classified into based on their exploration behaviour in a CPS environment?
RQ1b: Do students with different exploration and learning profiles differ in time spent on task, number of clicks and test performance?
RQ1c: Is there a change in time on task and number of clicks among the students in the different exploration and learning profiles while solving the complex problems on the test?
RQ2: What is the role of test-taking effort in strategy use?
RQ2a: Do students in each exploration and learning profile differ in their response time-based test-taking effort?
RQ2b: Is there a change in the response time-based test-taking effort made by the students in each exploration and learning profile during the test?
RQ2c: Is there a variation in the proportion of disengaged test-takers between the different profiles?

2. Methods

2.1. Participants

The sample consisted of first-year full-time undergraduates enrolled in one of the largest (*Hungarian*) universities. The assessment took place just after the beginning of their studies. The university has twelve faculties (such as law, medicine, humanities and social sciences, and natural sciences), with all of them included in the project. Although participation was voluntary, the students who completed the test successfully earned a credit as an incentive. Students participating in the assessment were assigned to a specific course, Career Development, due to university administrative requirements. A total of 1748 students (46.2 % of the target population) participated in the study (mean age = 19.80 years, SD = 1.92 years), of whom 53.0 % were women.

2.2. The problem-solving tasks

A CPS test containing ten fictitious problems developed with the MicroDYN approach was used in the project. MicroDYN problems can be solved in a relatively short time, are not based on prior school knowledge (Funke, 2014; Greiff et al., 2013c), and are reliable and valid forms for assessing CPS (Greiff et al., 2013c; Greiff et al., 2018; Molnár and Csapó, 2018). Each of the ten tasks took less than five minutes to complete, so the total testing time, including the instructions and warm-up task, was less than one hour. The principle of using a number of short-duration tasks is valid because interacting with unfamiliar real-life problems is usually also short-duration (e.g. learning to use the remote control for a new air conditioner) (Greiff et al., 2012).

The tasks consisted of two different phases, knowledge acquisition and knowledge application (Greiff et al., 2013c). In the first phase of problem-solving, test-takers explored the relationships between input and output variables by freely interacting with the problem environment (see Fig. 1). They were expected to change the values of the input variables (e.g. two different types of cat food) and then observe the effect of the changes on the values of the output variables (cat's behaviour, purring and activity). Students were able to change the values for the input variables by clicking on the "++" (value = +2), "+" (value = +1), "–"(value = −1) or "– –"(value = −2) buttons or by using the slider. The time series of the values for the input variables within a given scenario were plotted on the graphs associated with each input variable. In addition to input and output variables, each scenario included a Help, Delete, Apply and Next button. The Delete button returns the system to its original state. The Apply button causes the system to execute the changes made on the input variables, which are displayed as a diagram of each output variable. The Next button navigates between the different phases within a MicroDYN scenario and the different MicroDYN scenarios. Based on the information obtained and interpreted, participants drew the relationship(s) between the input and output variables, that is, the mental model of the problem structure in the form of a concept map, which is shown at the bottom of the figure (Molnár and Csapó, 2018). In this phase, the time limit was 180 s, but the number of interactions was not limited. In the second part of the problem-solving process, students were asked to apply the knowledge they had acquired in the first phase of the problem-solving process by changing the values of the input variables to obtain predefined values for the output variables. In the second phase, they had a time limit of 90 s and were allowed a maximum

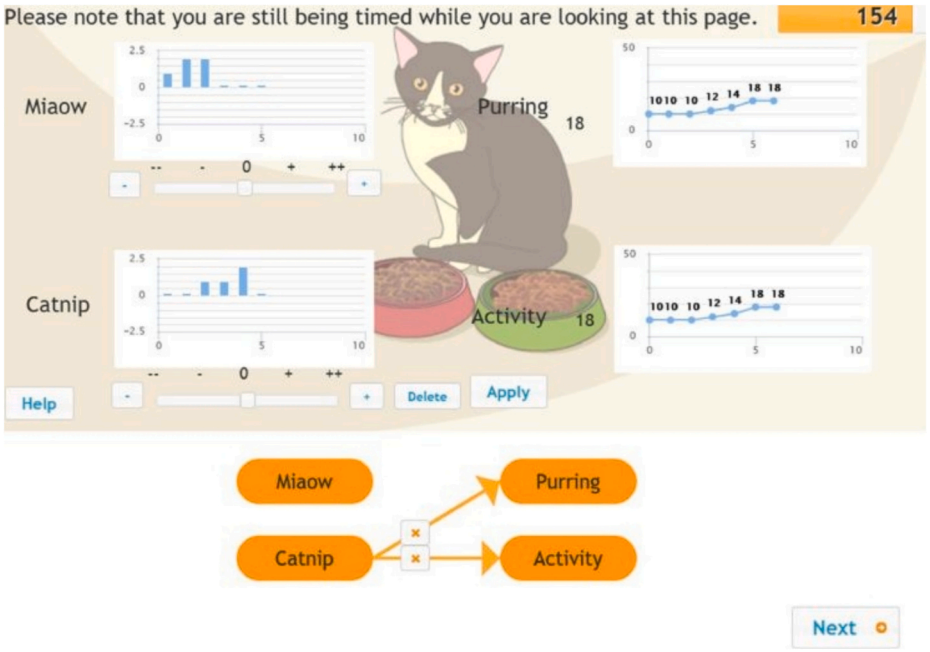


Fig. 1. The first stage of interactive problem-solving: Exploring the relationship between two input and two output variables.

of four trials. The problems were of increasing complexity, i.e. with an increasing number of relationships and input and output variables. For the easiest tasks, there were two input variables and one output variable with two relationships, while there were three input variables, three output variables and three relationships for the most difficult tasks. The reliability of the test was good ($\alpha = 0.88$).

The time spent on solving the problems and the number of clicks used in the problem-solving process are important indicators of test-taking effort (Csányi and Molnár, 2023; Ivanova and Michaelides, 2023). In the first phase of the problem-solving process, the number of possible clicks was not maximized and the maximum time was less limited. Therefore, both time and click data differentiated students more than in the second phase of the task-solving process. Consequently, the first phase was investigated in this study.

2.3. Procedures

2.3.1. Data collection

The assessment was administered via the eDia system (Csapó and Molnár, 2019) and took place in the main computer room of the university's information and learning centre. The testing procedure was supervised by invigilators. Students were allowed to choose their own schedules, so participants varied from 10 to 150 at each session. Those who enrolled in the assessment had two two-hour sessions to complete the CPS tests and other cognitive tests. At the beginning of the test, participants were introduced to the user interface and began a warm-up exercise. After logging into eDia, they had 60 min to complete all the tasks and questionnaires. They received immediate feedback on average performance after completing the test and detailed feedback one 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). In addition, (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, (4) they were informed in advance of the details via the university's learning management system, and (5) informed consent was obtained from all participants and they 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.

2.3.2. Scoring the performance indicators

In the first phase of the CPS process, we scored the visualized mental model of the problem structure, that is, the cognitive representation of the problem, dichotomously, which indicated the detected relations in the form of arrows between the variables presented on screen. A completely matching problem structure was assigned a score of 1 (e.g. on the "Cat" problem if students draw one arrow from the Catnip cat food to purring and another arrow from Catnip to movement but no other arrows); otherwise, the response was considered incorrect and earned a score of 0. Thus, across all ten problems administered on the CPS test, each student received ten binary (correct/incorrect) scores on their visualized cognitive mental map.

In the second phase, after the students received the correct representation of the mental model, the answers were marked correct ("1") if they managed to reach the given target values of the output variables within the given constraints (e.g. within a pre-specified number of steps; in the example, if students managed to increase the level of the cat's movement and purring from 10 to 21–23); otherwise, the solution was considered incorrect and assigned a score of 0. Thus, across all ten problems, each student received ten binary (correct/incorrect) scores on their knowledge use.

2.3.3. Scoring students' exploration based on log data; the principle of isolated variation

The test was structured in similar but increasingly complex scenarios according to the MicroDYN approach so that unknown situations could be explored using the same or almost the same exploration strategy. The change in exploration strategy, which can be reproduced from students' interactions, is a good indicator of their problem-solving effectiveness and therefore a good way of characterising their ability to learn in unfamiliar situations. Students' exploration strategies were analysed by assessing how well they used the principle of isolated variation to navigate problem environments. In each of the ten MicroDYN tasks, we evaluated their exploration methods during the first phase of the CPS process. We used the labelling procedure developed by Molnár and Csapó (2018). Students received 0 points if they failed to apply isolated variation to any input variables, 1 point if they applied isolated variation

to some but not all input variables and 2 points if they used isolated variation for all input variables. Consequently, each student received ten categorical scores, one for each CPS task, reflecting the degree to which they employed optimal exploration strategies.

2.3.4. Identifying test-taking effort

Response time was included in the analysis of test-taking effort. Analysis of time spent on tasks can be complemented by analysis of other behavioural data, such as number and type of clicks. Previous research has found that a lower number of clicks, similar to time data, indicates a lower motivation level (Sahin and Colvin, 2020). This is a less widely used method, as it is only applicable to interactive tasks and not appropriate for certain types of test items, such as multiple-choice items.

In response time-based methods, the measured indicator is the time the test-taker spends on a task. If the response time for an item is less than a specified threshold, it is classified as a disengaged response. However, if it is greater than or equal to the threshold, it is classified as an engaged response. Wise and Kong (2005) introduced the following relationship to measure the engaged or *solution behaviour* (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 and RT_{ij} = response time for item i and examinee j.

We applied the *proportion correct greater than zero* ($P + > 0$ %) method in our study to determine the threshold. On multiple-choice tests, the probability of a correct answer is greater than zero because of random guessing. For example, for a four-choice problem, the probability of a correct answer is around 0.25. For constructed-response tests and for tests where there are many combinations of response options to choose from, the probability of randomly selecting the correct answer is approximately zero. To determine the $P + > 0$ % threshold, the responses are sorted in ascending order as a function of response time. The threshold is the shortest time taken to arrive at the first correct answer (Goldhammer et al., 2016).

Based on Eq. (1), Wise and Kong (2005) introduced the term *response time effort* (*RTE*). RTE is the average engaged behaviour for a given participant, that is, 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. RTE scores range from 0 to 1. The lower the RTE score, the lower the effort exerted by the examinee. Examinees with RTE < 0.90 are considered disengaged (Kong et al., 2007; Wise and Ma, 2012).

An item-specific approach to the characterisation of test-taking effort is called *response time fidelity* (RTF) (Wise, 2006). RTF is the average engaged behaviour for a given item, i.e. the ratio of the sum of engaged responses to the total number of examinees. The RTF for a given item i is

$$RTF_i = \frac{\sum SB_{ij}}{n}, \quad (3)$$

where n = number of examinees. RTF scores range from 0 to 1. The higher the RTF value, the more effort examinees exerted in completing the item. Overall, while RTE scores measure the effort exerted by individual examinees on all items, RTF scores measure the effort exerted by all examinees on particular items (Wise, 2006).

2.4. Data analysis

A latent profile analysis was performed to identify the latent classes of the problem-solvers based on their exploration behaviour (Collins and Lanza, 2010). Several criteria were used to determine the number of latent groups: these include relative fit indices, such as the AIC (Akaike information criterion), BIC (Bayesian information criterion) and aBIC (adjusted Bayesian information criterion). For all three criteria, lower values indicate a better model fit (Dziak et al., 2012). Entropy can be used to determine how homogeneous the groups are, that is, how accurately we can classify students into the right groups. Finally, we used the Lo–Mendell–Rubin adjusted likelihood ratio test to compare the model with n number of latent classes with the model with $n–1$ number of latent classes (Lo et al., 2001). A significant p-value ($p < .05$) indicates that model $n–1$ is rejected in favour of the model with n groups, as the model currently being tested is a better fit than the previous model (Muthén and Muthén, 2012).

We employed one-way ANOVA to examine the difference between time on task, number of clicks, test score and test-taking effort by exploration and learning profile. RTF was compared using repeated measures ANOVA to examine the change in test-taking effort.

3. Results

3.1. Results for research question 1a (RQ1a): which exploration and learning profiles can students be classified into based on their exploration behaviour in a CPS environment?

We used absolute and relative fit indices beyond the entropy-based approach and the Lo–Mendell–Rubin (LMR) test to empirically investigate the number of latent class profiles. We also considered interpretability in relation to theoretical assumptions and the probabilities of class attribution when determining the number of latent classes. Our findings include solutions ranging from two to five latent classes.

With regard to absolute fit, we consulted the Pearson χ^2 statistic and the likelihood ratio χ^2 statistic provided by Mplus. As a general rule, the higher the number of latent classes, the smaller the χ^2 values and the better the absolute fit. Moreover, a χ^2/df ratio below 2 is usually considered an indication of an adequate absolute fit in large samples. This condition was met for all solutions (see Table 1).

With regard to relative fit, the information theory criteria were used. The AIC, BIC and aBIC generally indicated a continuous decrease in a growing number of latent classes. Both the AIC and aBIC were lowest for the solution comprising four classes, whereas the BIC indicated that the three-class solution was most appropriate. The likelihood ratio statistical test (Lo–Mendell–Rubin adjusted likelihood ratio test) showed the best model fit for the 4-class solution. The entropy-based criterion reached the maximum values for the 2-class solutions, but it was also significant for the 3- and 4-class solutions. The entropy index for the 4-class model showed that 95 % of the first-year students were accurately categorized based on their class membership (Table 1). Based on the relative fit, absolute fit and entropy values as well as the LMR test, a distinction can be made between four qualitatively different groups by exploring CPS problems at university level (see Fig. 2).

The characteristics of the four groups, which can be described by distinct exploration and learning profiles, are shown in Fig. 2. The students in Class 1 (6.8 %) had only limited success in solving the simpler problems at the beginning of the test, but they had learned to manage even the most complex systems by the end. This is why they can be called rapid learners. The participants in Class 2, the non-performers (3.7 %), had less success in solving the simpler problems but struggled with the more complex ones. Those who fell into Class 3, the proficient explorers (85.2 %), had the highest level of learning efficiency. Those in Class 4, the ineffective learners (4.3 %), were able to improve and learn while solving the simpler tasks but could not grasp the more complex and sophisticated systems.

Table 1
Absolute fit, relative fit and entropy-based fit indices for the two- to five-class solutions.

Number of latent classes	*Likelihood ratio X^2 (df)*	*Pearson X^2 (df)*	*$X^2/df < 2$ for both*	*AIC*	*BIC*	*aBIC*	*Entropy*	*L–M–R test*	*p*
2	882.91 (58929)	3950.75 (58929)	yes	9177.44	9401.55	9271.30	0.990	4648.84	0.000
3	695.14 (58923)	2391.09 (58923)	yes	8763.69	9102.59	8905.62	0.949	452.86	0.000
4	513.49 (58891)	1997.261 (58891)	yes	**8656.91**	**9110.61**	**8846.92**	**0.951**	**147.83**	**0.044**
5	574.30 (58879)	2263.25 (58879)	yes	9638.36	9206.85	8876.45	0.959	60.16	1.000

Notes: AIC = Akaike information criterion; BIC = Bayesian information criterion; aBIC = adjusted Bayesian information criterion; L–M–R test = Lo–Mendell–Rubin adjusted likelihood ratio test. The best fitting model solution is in bold.

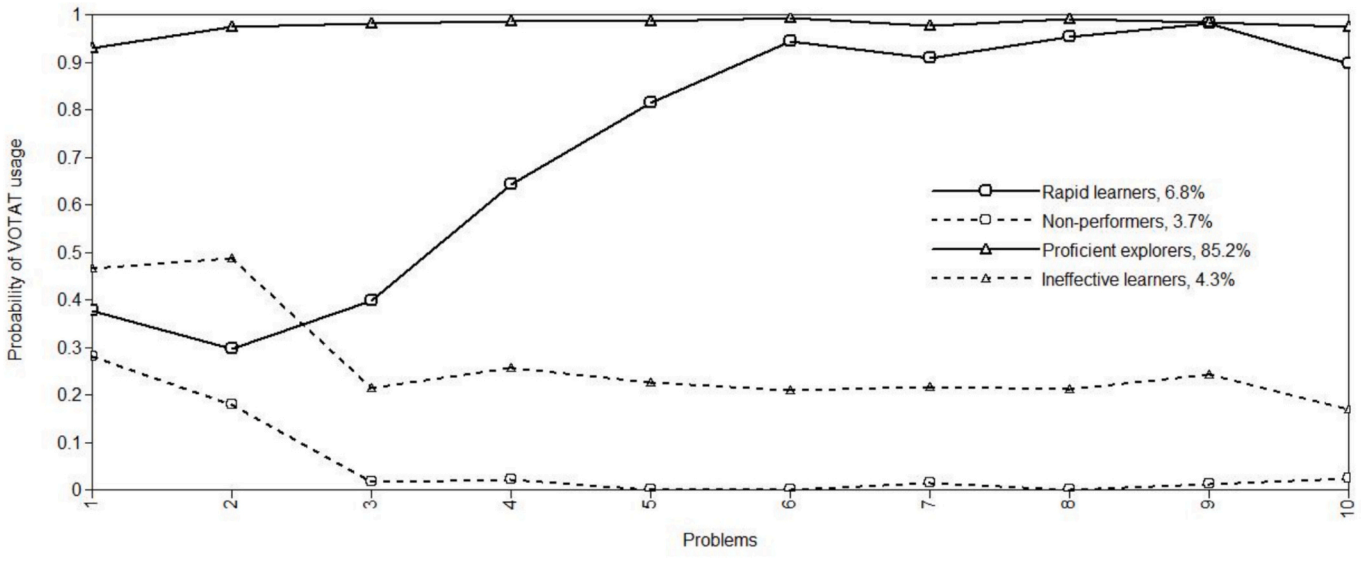


Fig. 2. Latent profile analysis based on the students' exploration and learning strategies.

3.2. Results for research question 1b (RQ1b): do students with different exploration and learning profiles differ in time spent on task, number of clicks and test performance?

We investigated time on task, number of clicks and test score using one-way ANOVA. In Fig. 3, the means for the variables are presented by class, and Table 2 shows the means and standard deviations of the variables. The rapid learners used a medium number of clicks for a long time and earned a medium score. The non-performers had few interactions for a short time and achieved low scores. The proficient explorers engaged in many interactions for a moderately long time and achieved the best performance. The ineffective learners used a medium amount of clicks over a long period of time and achieved a low score.

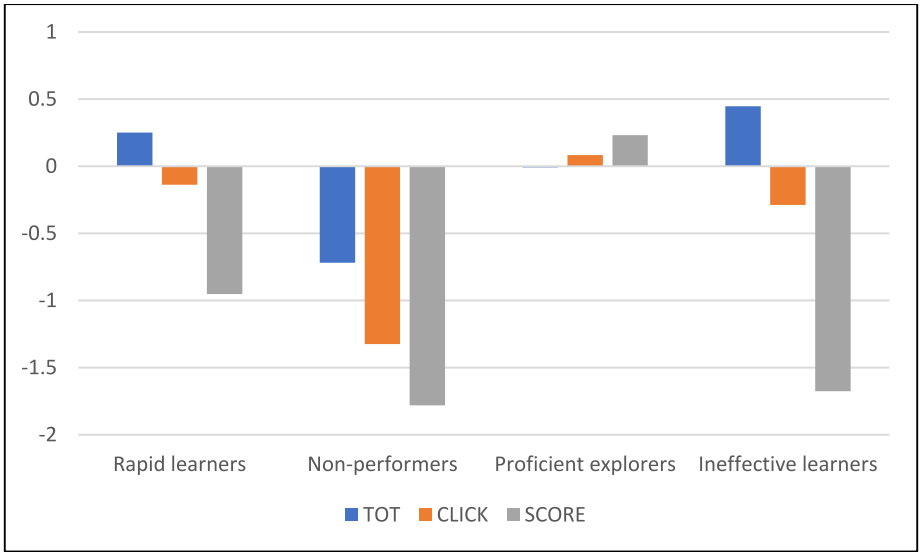


Fig. 3. Students' exploration and learning profiles based on time on task (TOT), number of clicks (CLICK) and test score (SCORE). Z scores are used instead of the original means.

Table 2
Features of the students' exploration and learning profiles generated from time on task, number of clicks and test score.

Variables	Class 1: Rapid learners ($N = 112$)		Class 2: Non-performers ($N = 67$)		Class 3: Proficient explorers ($N = 1497$)		Class 4: Ineffective learners ($N = 72$)		F*	Sig. different classes**
	M	SD	M	SD	M	SD	M	SD		
TOT	627.51	228.70	432.52	249.93	575.51	188.79	666.96	272.18	19.26	{2} < {3} < {1,4}
CLICK	51.91	19.59	25.37	20.25	56.87	21.46	48.54	25.95	49.10	{2} < {1,4} < {1,3}
SCORE	3.48	2.55	1.16	1.19	6.79	2.32	1.46	1.46	301.43	{2,4} < {1} < {3}

Notes: *All F-values are significant at the $p < .001$ level. **The "<" sign indicates the direction of the significant difference ($p < .05$) for significantly different classes. The comparison column between classes shows the significantly differentiated classes according to the Dunnett T3 test.

3.3. Results for research question 1c (RQ1c): is there a change in time on task and number of clicks among the students in the different exploration and learning profiles while solving the complex problems on the test?

We investigated the time on task item by item using one-way ANOVA. In Fig. 4, the means for time on task are presented by class, and Table 3 shows the means and standard deviations. In the first part of the test (Tasks 1–6), time on task was decreased for all the profiles. Time on task was similar for the non-performers and proficient explorers, as well as for the rapid learners and ineffective learners. The students in the latter two profiles spent more time completing the tasks. In the second part of the test (Tasks 7–10), the non-performers' time on task decreased further, while the proficient explorers began to spend more time on the tasks and matched and even exceeded the time taken by the rapid learners and ineffective learners.

We investigated the number of clicks item by item using one-way ANOVA. In Fig. 5, the mean number of clicks is presented by class, and Table 4 shows the means and standard deviations. The number of interactions for non-performers was much lower than for the others. It decreased in the first part of the test and then remained the same after that. The number of clicks for the other three profiles was the same and decreased in the first part of the test. In the second part of the test, the number of clicks for the proficient explorers increased considerably on the more difficult tasks. There was a smaller increase for the rapid learners and ineffective learners.

3.4. Results for research question 2a (RQ2a): do the students in each exploration and learning profile differ in their response time-based test-taking effort?

ANOVA showed a significant difference in the response time effort for the different exploration and learning profiles (Table 5). Based on the Dunnett T3 test, the non-performers had the lowest RTE, while RTE was higher for the ineffective learners and the highest for the rapid learners and proficient explorers. There was no significant difference between the latter two classes.

3.5. Results for research question 2b (RQ2b): is there a change in the response time-based test-taking effort made by the students in each exploration and learning profile during the test?

Repeated measures ANOVA showed a significant decrease of test-taking effort for all the exploration and learning profiles but at different rates (Fig. 6). Based on eta squared, which is a measure of effect size, the smallest decrease occurred among the proficient explorers (Wilk's $\lambda = 0.94$, $F(7, 1490) = 13.40$, $p < .001$, $\eta^2 = 0.06$). The decrease was slightly higher for the rapid learners (Wilk's $\lambda = 0.82$, $F(6, 106) = 3.96$, $p = .001$, $\eta^2 = 0.18$), much larger for the ineffective learners (Wilk's $\lambda = 0.60$, $F(7, 63) = 6.10$, $p < .001$, $\eta^2 = 0.40$) and the largest for the non-performers (Wilk's $\lambda = 0.43$, $F(9, 58) = 8.71$, $p < .001$, $\eta^2 = 0.58$).

The variation between the student profiles is shown in Table 6. At the beginning of the test, the test-taking effort is almost similar between the students' exploration and learning profiles, with more differences as the test progresses. Test-taking effort is greatest for the rapid learners and proficient problem-solvers, with no significant difference between them.

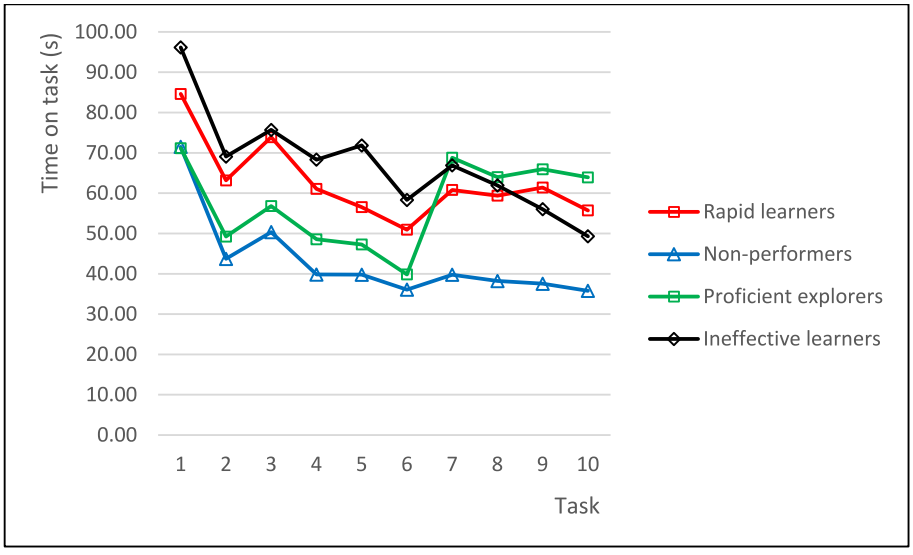


Fig. 4. Mean time on task for each task by exploration and learning profile.

Table 3
Time on task for each task by exploration and learning profile.

Tasks	Class 1: Rapid learners ($N = 112$)		Class 2: Non-performers ($N = 67$)		Class 3: Proficient explorers ($N = 1497$)		Class 4: Ineffective learners ($N = 72$)		F*	Sig. different classes**
	M	SD	M	SD	M	SD	M	SD		
T1	84.62	40.81	71.48	41.76	71.15	36.47	96.14	43.76	14.12	{2,3} < {1, 4}
T2	63.17	34.93	43.72	31.57	49.23	28.32	69.07	42.80	18.31	{2, 3} < {1, 4}
T3	73.89	39.67	50.34	37.77	56.81	29.15	75.68	43.88	19.60	{2, 3} < {1, 4}
T4	61.05	32.88	39.82	32.62	48.57	25.99	68.32	42.68	20.90	{2, 3} < {1, 4}
T5	56.54	34.98	39.78	33.19	47.28	25.44	71.85	43.18	23.66	{2, 3} < {1} < {4}
T6	50.93	32.38	36.08	27.90	39.83	19.90	58.31	43.15	23.45	{2, 3} < {1, 4}
T7	60.80	34.11	39.76	31.51	68.79	34.62	66.89	44.88	16.06	{2} < {1, 3, 4}
T8	59.38	33.12	38.22	33.86	63.98	31.09	61.92	41.29	14.46	{2} < {1, 3, 4}
T9	61.40	36.66	37.55	36.05	65.94	32.62	56.01	36.94	17.44	{2} < {1, 3, 4}
T10	55.72	33.24	35.78	35.01	63.93	29.63	49.30	39.59	24.02	{2} < {1, 4} < {3}

Notes: *All F-values are significant at the $p < .001$ level. **The "<" sign indicates the direction of the significant difference ($p < .05$) for significantly different classes. The comparison column between classes shows the significantly differentiated classes according to the Dunnett T3 test.

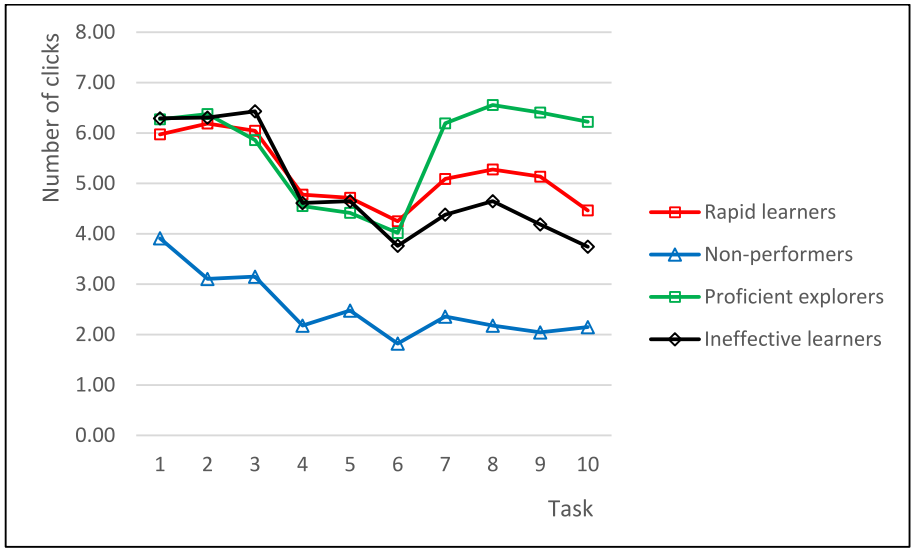


Fig. 5. Mean number of clicks for each task by exploration and learning profile.

Table 4
Number of clicks for each task by exploration and learning profile.

Tasks	Class 1: Rapid learners ($N = 112$)		Class 2: Non-performers ($N = 67$)		Class 3: Proficient explorers ($N = 1497$)		Class 4: Ineffective learners ($N = 72$)		F*	Sig. different classes**
	M	SD	M	SD	M	SD	M	SD		
T1	5.97	5.21	3.91	3.94	6.27	3.95	6.29	4.62	7.31	{2} < {1, 3, 4}
T2	6.19	5.47	3.10	2.70	6.38	4.06	6.31	5.83	12.96	{2} < {1, 3, 4}
T3	6.05	3.38	3.15	3.85	5.86	3.36	6.43	5.96	13.76	{2} < {1, 3, 4}
T4	4.78	2.84	2.18	2.45	4.55	2.44	4.61	3.41	19.74	{2} < {1, 3, 4}
T5	4.71	3.94	2.48	3.35	4.42	2.59	4.65	3.73	11.31	{2} < {1, 3, 4}
T6	4.25	2.70	1.82	1.92	4.02	2.11	3.76	3.21	22.32	{2} < {1, 3, 4}
T7	5.09	3.33	2.36	2.66	6.19	3.65	4.38	3.23	31.39	{2} < {1, 4} < {3}
T8	5.28	3.22	2.18	2.72	6.56	3.64	4.65	3.96	39.64	{2} < {1, 4} < {3}
T9	5.13	3.01	2.05	2.35	6.41	3.59	4.19	3.95	42.91	{2} < {1, 4} < {3}
T10	4.46	2.55	2.15	3.00	6.22	3.45	3.74	4.09	47.27	{2} < {1, 4} < {3}

Notes: *All F-values are significant at the $p < .001$ level. **The "<" sign indicates the direction of the significant difference ($p < .05$) for significantly different classes. The comparison column between classes shows the significantly differentiated classes according to the Dunnett T3 test.

It is medium for the ineffective learners and least for the non-performers.

3.6. Results for research question 2c (RQ2c): is there a variation in the proportion of disengaged test-takers between the different profiles?

A chi-square test of independence was performed to examine the

Table 5
Response time effort by exploration and learning profile.

Variable	Class 1: Rapid learners ($N = 112$)		Class 2: Non-performers ($N = 67$)		Class 3: Proficient explorers ($N = 1497$)		Class 4: Ineffective learners ($N = 72$)		F*	Sig. different classes**
	M	SD	M	SD	M	SD	M	SD		
Response time effort	0.98	0.05	0.74	0.25	0.99	0.04	0.90	0.13	334.06	{2} < {4} < {1,3}

Notes: *All F-values are significant at the $p < .001$ level. **The "<" sign indicates the direction of the significant difference ($p < .05$) for significantly different classes. The comparison column between classes shows the significantly differentiated classes according to the Dunnett T3 test.

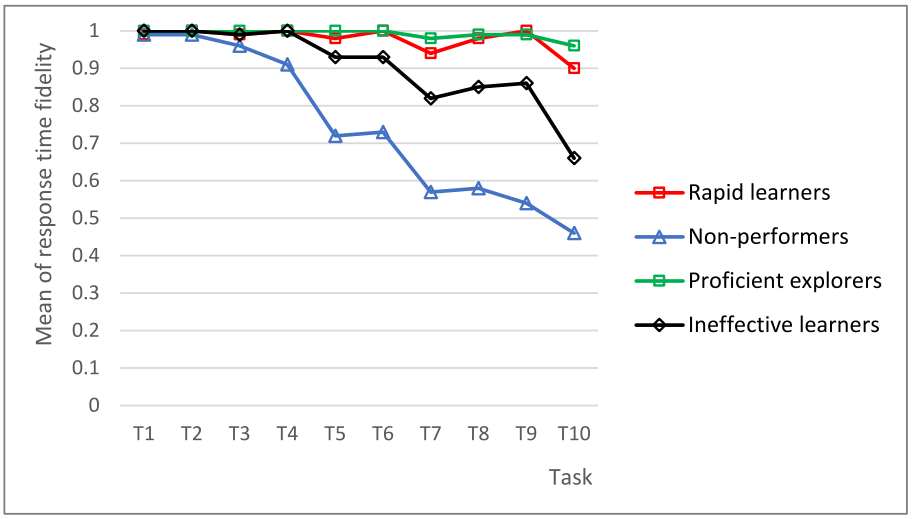


Fig. 6. Change of response time fidelity for each task by exploration and learning profile.

Table 6
Response time fidelity for each task by exploration and learning profile.

Tasks	Class 1: Rapid learners ($N = 112$)		Class 2: Non-performers ($N = 67$)		Class 3: Proficient explorers ($N = 1497$)		Class 4: Ineffective learners ($N = 72$)		F*	Sig. different classes**
	M	SD	M	SD	M	SD	M	SD		
T1	0.99	0.09	0.99	0.12	1.00	0.00	1.00	0.00	6.34	{1,2} < {1, 3, 4}
T2	1.00	0.00	0.99	0.12	1.00	0.00	1.00	0.00	8.47	{2} < {1, 3, 4}
T3	0.99	0.09	0.96	0.21	1.00	0.00	0.99	0.12	17.15	{2} < {1, 3, 4}
T4	1.00	0.00	0.91	0.29	1.00	0.03	1.00	0.00	45.88	{2} < {1, 3, 4}
T5	0.98	0.13	0.72	0.45	1.00	0.03	0.93	0.26	146.71	{2} < {4} < {1,3}
T6	1.00	0.00	0.73	0.45	1.00	0.03	0.93	0.26	150.07	{2} < {4} < {1, 3}
T7	0.94	0.24	0.57	0.50	0.98	0.15	0.82	0.39	110.02	{2} < {4} < {1, 3}
T8	0.98	0.13	0.58	0.50	0.99	0.10	0.85	0.36	161.62	{2} < {4} < {1, 3}
T9	1.00	0.00	0.54	0.50	0.99	0.12	0.86	0.35	179.62	{2} < {4} < {1, 3}
T10	0.90	0.30	0.46	0.50	0.96	0.21	0.66	0.48	114.62	{2} < {4} < {1, 3}

Notes: *All F-values are significant at the $p < .001$ level. **The "<" sign indicates the direction of the significant difference ($p < .05$) for significantly different classes. The comparison column between classes shows the significantly differentiated classes according to the Dunnett T3 test.

relation between exploration and learning profiles and the test-taking disengagement. The relation between these variables was significant, $X^2(3, N = 1748) = 510.90$, $p < .001$. More than half (67.72 %) of the non-performers and nearly a third (29.17 %) of the ineffective learners were disengaged, while the proportion was negligible among the rapid learners and proficient explorers, 1.79 % and 1.67 %, respectively (Table 7).

4. Discussion

The main aim of this study was to investigate the role of test-taking effort in students' exploration and learning strategies in a CPS environment. A number of studies have examined students' problem-solving strategies, but, to the best of our knowledge, no research has examined the role of test-taking effort in this process. In this study, we quantified qualitative behavioural differences in the students' problem-solving behaviour using cutting-edge analyses of process data. Based on the students' VOTAT strategy use, we determined groups of students and the characteristics of these groups, as well as examining their test-taking effort.

4.1. Describing the exploration and learning profiles

We used a person-centred approach to investigate the students'

Table 7
Engaged and disengaged test-takers by exploration and learning profile.

Exploration and learning profiles	Unit	Engaged test-takers	Disengaged test-takers	Total
Rapid learners	Count	110	2	112
	% within row	98.21 %	1.79 %	100 %
Non-performers	Count	29	38	67
	% within row	43.28 %	56.72 %	100 %
Proficient explorers	Count	1472	25	1497
	% within row	98.33 %	1.67 %	100 %
Ineffective learners	Count	51	21	72
	% within row	70.83 %	29.17 %	100 %
Total	Count	1662	86	1748
	% within row	95.08 %	4.92 %	100 %

exploration and learning strategies. Applying a latent class analysis of their use of the VOTAT strategy, we identified four groups of students. After identifying the different profiles, we wanted to better understand the differences between them and examine their test-taking behaviour. Previous research has shown that time on task and number of clicks indicate exerted effort. We therefore analysed these variables and the students' test performance based on their latent class membership.

The vast majority of the participants (85.2 %) were proficient explorers. They consistently used optimal exploration strategies throughout the test and were the most successful in interpreting the information extracted from the problem environment. They engaged in the most interactions over a moderately long period of time and thus performed the best.

The rapid learners (6.8 %) displayed low to medium performance in their exploration behaviour at the beginning of the test, after which they learned rapidly and started to use high-quality exploration strategies. In the second half of the test, they performed at almost the same level of exploration behaviour as the proficient explorers. Their overall test-taking behaviour was characterised by interacting with the problem environment as often as the proficient explorers over a longer period of time. Based on their test-taking and exploration behaviour, they would be expected to have a test result approaching that of the proficient explorers, but it was significantly lower. To possibly explain this, the test-taking behaviour needs to be examined at item level. In the first part of the test, the rapid learners used as many clicks as the proficient explorers but over a longer period of time. By the end of this phase, their exploration behaviour was almost the same level as the proficient explorers. In the second part of the test, as the tasks became more difficult, the rapid learners spent the same amount of time on the tasks as the proficient explorers but clicked less. Therefore, they probably performed worse than the proficient explorers because they had not yet fully applied the VOTAT strategy in the first part of the test and because they had fewer interactions with the problem environment in the second part of the test. If we only examine their test-level performance, they could be described as moderate problem-solvers, but their test-taking and exploration behaviour considerably refines their characteristics.

The ineffective learners (4.3 %) used the VOTAT strategy to a moderate extent at the beginning of the test and then to a low extent from the third part of the test onwards. They had a medium amount of interaction over a long period of time and achieved a low result on the test. Their test-taking behaviour is similar to that of the rapid learners, not only overall, but also at the item level. The essential difference is that their test performance was much lower. A possible explanation is the low efficiency of their exploration strategy use.

A minority of the learners (3.7 %) were non-performers, who made very limited use of the VOTAT strategy throughout the test. They interacted little in a short period of time. Their results were similar to those of the ineffective learners, while their test-taking and exploration behaviour differed.

Our research confirmed research findings reported by Molnár and Greiff (2023). Efficient test-taking behaviour was contradictory for low- and high-complexity tasks. Effective explorers tended to engage in the same number of interactions in a shorter time than less effective ones on tasks of low and medium complexity. However, in completing more complex tasks, they interacted most with the problem environment and spent more time doing that. For the less efficient explorers, there was little variation in the time spent on the task and the number of clicks, regardless of the complexity of the task.

These exploration profiles, or parts of them, have been identified by various previous studies that have investigated university students. Molnár (2022) identified four latent class exploration profiles among university students, three of which (rapid learners, proficient explorers and low to intermediate performers on the easiest problems but non-performers on the complex ones) correspond to the profiles obtained in our research. In another study, Molnár et al. (2022) compared the exploration strategies used by Hungarian and Jordanian university students. In both samples, four latent classes were observed, but there were significant differences between them. Three of the four Hungarian latent classes (rapid learners, proficient explorers and non-performing explorers) were similar to those in our study. Our research confirmed previous findings by examining the exploration strategies used by students of similar age groups in similar educational contexts. Overall, students' test-taking and exploration behaviour, however, does not fully explain their performance on the test, which is the reason we included test-taking effort in the analysis.

4.2. *The role of test-taking effort in strategy usage*

After identifying the different exploration and learning profiles, we wanted to better understand the differences between them and thus examined test-taking effort. The rapid learners and proficient explorers showed the greatest test-taking effort. Their response time effort was close to 1.0, indicating a high level of engagement. The effort made by the ineffective learners was significantly lower, with the lowest being that of the non-performers.

In order to obtain a more accurate understanding of the test-taking effort of the profiles, we examined the change in it. At the beginning of the test, the RTF of all four profiles was very high, almost the same, near or equal to 1.0. The rapid learners and proficient explorers had high RTF scores throughout the test, with only a minimal decrease, to 0.9 and 0.96, respectively, by the end. For the ineffective learners, the RTF started to decrease from the middle of the test and reached 0.66 by the end. For the non-performers, the RTF decreased sharply from one third of the test onwards and reached 0.46 by the end. Our results are consistent with several previous studies demonstrating that test-taking effort decreases as the test progresses (e.g. Lindner et al., 2017; Wise, 2006). The novelty of our research is that we examined change by profile, rather than overall, and that we showed that it varies significantly across profiles. While there was hardly any decrease for the rapid learners and proficient explorers, it was substantial for the ineffective learners and non-performers, thus presumably influencing the use or non-use of an effective exploration strategy.

We have determined the proportion of disengaged students to obtain a more accurate picture of the role of test-taking effort in the profiles. The proportion of disengaged students was negligible for the rapid learners and proficient explorers, only 1.79 % and 1.67 %, respectively. However, nearly a third (29.17 %) of the ineffective learners and more than half (56.72 %) of the non-performers were disengaged. Therefore, the rapid learners and proficient explorers were fundamentally engaged in completing the tasks, so the difference in performance between them is presumably caused by the difference in strategy use in the first part of the test and the difference in the number of interactions on more complex problems in the second part of the test. In contrast, the rates of

disengaged behaviour were significantly high among the ineffective learners and even more among the non-performers, which strongly influenced their performance. Students in these profiles did not demonstrate the use of exploration strategies of which they could actually be capable but rather less effective because of their lack of engagement. This raises the question of whether they are underperforming because they are disengaged or they are disengaged because they are underperforming. Various studies have investigated the relationship between ability levels and test-taking effort, and the results are contradictory. Some studies suggest no link between them (Kong et al., 2007; Rios et al., 2014; Wise and DeMars, 2005), while others suggest that there is indeed a relation (Csányi and Molnár, 2024; Deribo et al., 2021). This question should be investigated in further research. Our results are consistent with a number of previous studies that show that test-taking disengagement significantly affects test performance for a proportion of examinees (e.g. Finn, 2015; Kriegbaum et al., 2014; Schüttpelz-Brauns et al., 2018). In addition, excluding disengaged responses/examinees from the sample increases the validity of the test (e.g. Rios and Deng, 2021; Sahin and Colvin, 2020).

5. Limitations

There are some important limitations to this study that need consideration. One important limitation is that non-representative convenience sampling was used, with participants being exclusively first-year students at one university. Participation was not compulsory, and only students who were willing to participate were included in the study. Another limitation of the study is that in order to minimize the uncontrollable effects of prior knowledge, the tasks included fictitious cover stories; that is, the results are necessarily applicable to all kinds of everyday, complex and dynamic real-world problems. There is no information about the intelligence profile of the students related to the four classes, so this has not been examined. The final limitation is that the test was conducted in a low-stakes context. Thus, the results cannot be generalized.

6. Conclusions and implications

The main aim of our study was to investigate the role of test-taking effort in students' exploration strategies. A number of prior studies have investigated students' exploration strategies in CPS and the role of test-taking effort in low-stakes contexts, but, to our knowledge, these have not been studied in tandem. The results shed new light on the interpretation of students' exploration strategy usage in the problem-solving process.

As for the educational implications, understanding the characteristics of students' problem-solving behaviour also provides valuable input for designing appropriate training tasks and training students to become better problem-solvers. The proficient explorers invested great effort in completing the tasks by applying the VOTAT strategy consistently, so, in their case, no intervention other than practice is needed. The non-performers and ineffective learners invested low effort in completing the tasks, and their VOTAT strategy use was also at the low level. This calls for both improved strategy use and increased effort. The rapid learners are an interesting class because they used the VOTAT strategy less than the ineffective learners did at the beginning of the test, but they learned it while completing the tasks. They exerted a great deal of effort into the completion of the tasks and thus performed better than the ineffective learners. This also suggests the role of test-taking effort during the problem-solving process.

Our results confirm findings that successful problem-solvers invest enough time and effort into solving problems (e.g. Csányi and Molnár, 2023). The appropriate amount of effort is no guarantee of a successful outcome, but success is also not possible without it. Therefore, practitioners need to place great emphasis on using methods that improve students' test-taking effort.

CRediT authorship contribution statement

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

Declaration of competing interest

The authors declare no conflicts of interest.

Data availability

Data will be made available on request.

Acknowledgements

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), by the University of Szeged Open Access Fund 7577 and by the Centre of Excellence for Interdisciplinary Research, Development and Innovation of the University of Szeged.

References

Anghel, E., Khorramdel, L., & von Davier, M. (2024). The use of process data in large-scale assessments: A literature review. *Large-Scale Assessments in Education, 12*. https://doi.org/10.1186/s40536-024-00202-1

Arguel, A., Lockyer, L., Chai, K., Pachman, M., & Lipp, O. V. (2019). Puzzle-solving activity as an indicator of epistemic confusion. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00163

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

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. John Wiley & Sons, Inc.. https://doi.org/10.1002/9780470567333

Coyle, T. R., & Greiff, S. (2021). The future of intelligence: The role of specific abilities. *Intelligence, 88*. https://doi.org/10.1016/j.intell.2021.101549

Csányi, R., & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences, 106*, Article 102340. https://doi.org/10.1016/j.lindif.2023.102340

Csányi, R., & Molnár, G. (2024). Item- and person-level factors in test-taking disengagement: Multilevel modelling in a low-stakes context. *International Journal of Educational Research Open, 7*, Article 100373. https://doi.org/10.1016/j.ijedro.2024.100373

Csapó, B., & Funke, J. (Eds.). (2017). *The nature of problem solving: Using research to inspire 21st century learning*. OECD Publishing. https://doi.org/10.1787/9789264273955-en.

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

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

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). Sensitivity and specificity of information criteria. In *Technical Report Series #12-119*. Issue Retrieved from https://www.methodology.psu.edu/fil.

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

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

Fischer, A., Greiff, S., Wüstenberg, S., Fleischer, J., Buchwald, F., & Funke, J. (2015). Assessing analytic and interactive aspects of problem solving competency. *Learning and Individual Differences, 39*, 172–179. https://doi.org/10.1016/j.lindif.2015.02.008

Frensch, P., & Funke, J. (1995). Complex problem solving — The European perspective. In *Learning to solve complex scientific problems*.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning, 7*(1), 69–89. https://doi.org/10.1080/13546780042000046

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. *OECD Education Working Papers, 133*. https://doi.org/10.1787/5jlzfl6fhxs2-en

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013a). A multitrait-multimethod study of assessment instruments for complex problem solving. *Intelligence, 41*(5), 579–596. https://doi.org/10.1016/j.intell.2013.07.012

Greiff, S., & Funke, J. (2017). Interactive problem solving: Exploring the potential of minimal complex systems. In B. Csapó, & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning*. OECD Publishing. https://doi.org/10.1787/9789264273955-en.

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 in Education, 126*(February), 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers in Education, 91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*(3), 189–213. https://doi.org/10.1177/0146621612439620

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013b). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development, 61*(3), 407–421. https://doi.org/10.1007/s11423-013-9301-x

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013c). 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

Hayes, J. R. (1981). *The complete problem solver*. Franklin Institute Press.

Ivanova, M. G., & Michaelides, M. P. (2023). Measuring test-taking effort on constructed-response items with item response time and number of actions. *Practical Assessment, Research and Evaluation, 28*(15).

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

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., 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

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767–778. https://doi.org/10.1093/biomet/88.3.767

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

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. (2017). A problémamegoldó és tanulási stratégiák változása 11 és 19 éves kor között: Logfile elemzések. *Magyar Pedagógia, 117*(2), 221–238. https://doi.org/10.17670/MPed.2017.2.221

Molnár, G. (2022). How to make different thinking profiles visible through technology: The potential for log file analysis and learning analytics. In M. Virvou, G. A. Tsihrintzis, L. H. Tsoukalas, & L. C. Jain (Eds.), *Advances in artificial intelligence-based technologies. Learning and analytics in intelligent systems* (pp. 125–145). Springer. https://doi.org/10.1007/978-3-030-80571-5_9.

Molnár, G., Alrababah, S. A., & Greiff, S. (2022). How we explore, interpret, and solve complex problems: A cross-national study of problem-solving processes. *Heliyon, 8*(1). https://doi.org/10.1016/j.heliyon.2022.e08775

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

Molnár, G., & Greiff, S. (2023). Understanding transitions in complex problem-solving: Why we succeed and where we fail. *Thinking Skills and Creativity, 50*. https://doi.org/10.1016/j.tsc.2023.101408

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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

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., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education, 9*(1). https://doi.org/10.1186/s40536-021-00110-8

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment, 27*, 1–18. https://doi.org/10.1080/10627197.2022.2110465

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

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education, 8*(5), 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*(101). https://doi.org/10.1186/s12909-018-1196-0

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

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. In *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, January, 2014*, 175–185.

Wu, H., & Molnár, G. (2021). Logfile analyses of successful and unsuccessful strategy use in complex problem-solving: A cross-national comparison study. *European Journal of Psychology of Education, 36*(4), 1009–1032. https://doi.org/10.1007/s10212-020-00516-y

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving – More than reasoning? *Intelligence, 40*(1), 1–14. https://doi.org/10.1016/j.intell.2011.11.003

Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge and Learning, 19*(1–2), 127–146. https://doi.org/10.1007/s10758-014-9222-8

3. DISCUSSION

The aim of the research conducted in this dissertation is to explore and comparatively investigate test-taking behavior in an interactive test environment through self-report and process data analysis. This chapter discusses and summarizes the results of three empirical studies that contribute to a deeper understanding of test-taking engagement from different perspectives. The first study focuses on comparing self-report and behavioral indicators of engagement, the second uses multilevel modeling to examine person- and item-level predictors, and the third examines how engagement influences students' exploration strategies in problem-solving tasks. The aim of this chapter is to integrate these findings, interpret them on the basis of existing research, and highlight their theoretical and practical implications.

The first study focused on measuring test-taking effort, a construct most often captured using self-report or behavioral indicators such as response time. There are several operationalizations of response time-based measures, of which we examined six to determine the most appropriate for our research. Validation criteria was used to select the optimal response time-based method (Goldhammer et al., 2016). The optimal method is the one that best separates correct responses identified as disengaged from those identified as engaged. Among the methods, the *proportion correct greater than zero* P+>0% method was found to be the most accurate based on the validation criterion used, which is consistent with results reported in Goldhammer et al. (2016). To determine a threshold of P+>0%, the responses and their corresponding response times are sorted in ascending order of response time. The threshold is the shortest response time at which the proportion of correct responses is greater than zero. Previous studies have investigated different methods based on response time and found that item-specific thresholds result in higher accuracy than constant thresholds (e.g. Wise & Ma, 2012). Our study supports this finding and found that a relatively rarely used method proved to be the most accurate for problem-solving tasks. To ensure the validity of the tests, it is important to use the proper method to identify disengaged responses.

The results also highlighted important differences between self-reported and behavioral measures of effort. Although both were positively correlated with test performance, the correlation was significantly stronger for the response time-based effort (r = .37) than for self-reported effort (r = .10). These results suggest that self-reported and behavioral indicators may capture different constructs of engagement - a finding supported by previous meta-analyses Silm et al. (2020).

While a multiple-choice test can be solved by clicking on the correct option, the complex problem-solving tasks used in the test require interactions with the test environment, and the students have to explore possible relationships between variables. There was a high correlation between response time and the number of interactions ($r = .62$, $p < .01$), meaning that if someone put in a lot of effort, they needed more time. Students who completed a sufficient number of interactions were able to achieve high scores on the test. Test performance was significantly correlated with the number of interactions, but not with the response time. This highlights the task-specific importance of the different indicators: in CPS contexts, success depends primarily on active interaction and systematic exploration, and response time becomes a function of these.

To further investigate the effort of test-takers, cluster analysis was used to identify distinct groups of students. The response time and the number of interactions are indicators of the effort exerted in completing the test. The results show that test-taking behavior was not consistent with self-reported data for all clusters, i.e. participants' responses did not fully reflect their actual testing behavior, indicating limitations of self-report questionnaires.

If we analyze groups of students (clusters) rather than the whole population, it gives a more accurate insight into the details. A positive correlation between test-taking effort and test performance is obtained for all students, but this represents only the overall picture. A more accurate insight is obtained by examining the relationships for each cluster. Students in some clusters achieve the same results with less effort than students in other clusters. This suggests that good results do not require maximum effort, but only a certain amount of effort. This is in line with the results of Gignac et al. (2019). Goldhammer et al. (2017) found that higher-ability students needed less effort to solve problems successfully, which is also consistent with our results.

Study 1 established the foundation for subsequent studies by providing validated indicators of test-taking effort that take into account the unique requirements of CPS tasks. It also highlighted the limitations of self-report methods and emphasized the need for analysis of process data to better understand engagement behavior.

Building on the methodological foundations of the previous study, the second study used a multilevel modelling approach to examine predictors of disengaged behavior at both person and item levels. We considered the hierarchical structure of the data - item responses embedded in the person tested - which allowed for more precise estimation of measures of variance.

In our research, test-taking disengagement increased with the increase in item position and for more difficult items. These suggest that more attention should be paid to the development of low-stakes tests. Previous research has identified a number of interventions that can be used to

increase test-taking engagement. Rios (2021) categorized these factors into four main categories: (1) modifying the test design, (2) providing feedback, (3) modifying the relevance of the test, and (4) providing external incentives.

Among the person-level factors, test-taking disengagement was predicted by gender, entrance score, working memory capacity and self-reported effort. Among women, the percentage of disengaged responses was higher. People with lower entrance scores, lower working memory capacity and lower self-reported effort also had a higher proportion of disengaged responses. A fundamental question is whether there is a correlation between academic ability and test-taking disengagement. Previous results are mixed and of significant practical importance, i.e. how to deal with disengaged behavior.

Previous research suggests that working memory capacity is crucial for students' problem-solving performance (e.g. Lindner et al., 2017). Our study showed that examinees with higher working memory capacity had lower test-taking disengagement. However, research has demonstrated that working memory capacity is not fixed and can be developed in different ways (Brady et al., 2016). Developing students' working memory may be a good method to increase test-taking engagement.

Motivational filtering is a widely used method for dealing with disengaged responses and can be applied in two ways. It is possible to remove (1) disengaged responses or (2) all data from disengaged examinees and leave only the engaged data in the sample and only these are analyzed. Rios et al. (2017) developed the term response-level filtering to refer to the former type of motivation filtering and examinee-level filtering to refer to the latter. Examinee-level filtering is based on the assumption that disengaged response behavior is unrelated to examinees' true ability. If this assumption is not correct, then deleting examinees with higher or lower ability leads to bias (Rios et al., 2017). In our study, lower ability examinees showed higher disengagement, suggesting that there is a relationship between academic ability and test-taking disengagement. This implies that item-level filtering is preferable to examinee-level filtering.

Taken together, the first two studies show that test-taking disengagement is a complex phenomenon, influenced by cognitive, motivational and contextual factors. They also suggest that it is important to use sophisticated, psychometrically founded methods to identify disengaged behavior in digital assessments.

While the first two studies focused on the measurement and predictors of engagement, the third study addressed the functional role of test-taking engagement in shaping learning behavior. Using a person-centered approach, four latent profiles of learners were identified based on their use of the VOTAT (vary-one-thing-at-a-time) strategy: (1) rapid learners, (2) non-performers, (3)

proficient explorers and (4) ineffective learners. Previous research has shown that time on task and number of clicks are indicators of effort. Therefore, we analyzed these variables and students' test performance based on their latent class membership.

The majority of participants were proficient explorers, i.e. they consistently used optimal exploration strategies throughout the test and were the most successful in interpreting information extracted from the problem environment. They performed most of the interactions in the test over a moderately long period of time and thus performed best.

Rapid learners showed low to moderate performance in exploration behavior at the beginning of the test, then learned quickly and used high quality exploration strategies in the second half of the test. In the first part of the test, rapid learners interacted with the test environment as many times as proficient explorers, but for a longer period of time. In the second part of the test, as the tasks became more difficult, rapid learners spent the same amount of time on the tasks as proficient explorers, but interacted less. If we consider only their test performance, we could characterize them as average problem solvers, but their test-taking and exploration behavior fine-tunes their characteristics.

Ineffective learners used the VOTAT strategy at a medium level at the beginning of the test, and then quickly dropped to a low level. Over a long period of time, they interacted moderately and scored low on the test.

A minority of students were non-performers who made very limited use of the VOTAT strategy during the test. They had few interactions with the test environment in a short period of time, their results were similar to ineffective learners, but their test-taking and exploration behavior differed.

Our research is in line with the findings reported by Molnár & Greiff (2023). There is a difference between efficient test-taking behavior for low and high complexity tasks. For low and medium complexity tasks, efficient explorers are able to perform the same number of interactions in less time than their less efficient counterparts. In more complex tasks, however, efficient explorers performed the most interactions and spent more time on them. Less efficient explorers showed little change in response time and number of interactions, regardless of task complexity. It also underlines the importance of the number of interactions with the test environment in effective problem solving.

To better understand the different latent profiles, we examined the extent of test-taking effort and its change during the test. At the beginning of the test, the test-taking effort of all four profiles was very high. Test-taking effort for rapid learners and proficient explorers was high during the test and only minimally decreased at the end of the test. Test-taking effort of ineffective learners

declined from the middle of the test, while that of non-performers declined steeply from one-third of the test. Our results are consistent with a number of previous studies showing that test-taking effort decreases as the test progresses (e.g. Wise, 2006). The novelty of our research is that we examined change by profile rather than overall, and showed that change varies significantly by profile. While there was little decrease for rapid learners and proficient explorers, there was a significant decrease for ineffective learners and non-performers, presumably influencing the use or non-use of an effective exploration strategy.

A number of prior studies have investigated students' exploration strategies in CPS and the role of test-taking effort in low-stakes contexts, but, to our knowledge, these have not been studied in tandem. The results shed new light on the interpretation of students' exploration strategy usage in the problem-solving process. Our results suggest that successful problem solvers put in enough time and effort to solve problems. A sufficient amount of effort does not guarantee a successful outcome, but success is not possible without it. Therefore, practitioners should place considerable emphasis on using methods that improve students' test-taking effort.

Limitations

The dissertation has several limitations. One is that the test consisted exclusively of complex problem-solving items. The most commonly used multiple-choice items were therefore not investigated, nor were the tasks related to a specific subject, because the fictional context of the CPS test was unfamiliar to everyone. 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. A final limitation is that the test was in a low-stakes context.

Conclusions

The three studies are linked by their focus on testing in complex problem solving in low-stakes digital environment, while each adds a distinct layer of understanding. Study 1 addressed how to accurately measure engagement. Study 2 highlighted who tends to be disengaged and under what circumstances. Study 3 examined test-taking engagement in complex problem-solving in latent groups of students. Together, the studies provide a comprehensive picture of test-

taking engagement as a measurable construct and as a behavioral phenomenon with real-life consequences.

From a theoretical point of view, this research contributes to the discussion on the nature of participation in technology-based assessments. It is consistent with multidimensional models that conceptualize engagement as comprising behavioral, cognitive and emotional components, and operationalizes this construct using self-report and process data. It also extends the literature on test-taking engagement and the assessment of latent behavioral profiles in assessment contexts.

From a practical point of view, the results have implications for test design and implementation. Designers of digital assessments should consider integrating process indicators (e.g. response time, number of clicks) and item-level measures of effort to track engagement. Test developers should avoid over-relaying on self-report metrics and use adaptive testing strategies that reduce disengagement in later test phases. Educators should be cautious when interpreting performance data without consideration of engagement, especially in low-stakes contexts.

Finally, this research suggests that targeted interventions - such as training in strategic exploration or support for improving working memory - can help reduce disengagement and improve performance. Future work should test such interventions experimentally and examine their impact across different student populations and subject areas.

References

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

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

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. *OECD Education Working Papers*, *133*. https://doi.org/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

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), e0180149. https://doi.org/10.1371/journal.pone.0180149

Molnár, G., & Greiff, S. (2023). Understanding transitions in complex problem-solving: Why we succeed and where we fail. *Thinking Skills and Creativity*, *50*. https://doi.org/10.1016/j.tsc.2023.101408

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., 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

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

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., & 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.

SZEGEDI TUDOMÁNYEGYETEM
NEVELÉSTUDOMÁNYI DOKTORI ISKOLA
ETIKAI BIZOTTSÁGA

6722 SZEGED, Petőfi S sgt. 30-34.
Tel.: (62) 544163, 544032; Fax: (62) 420034

Ügyiratszám: 11/2023
Tárgy: kutatás-etikai engedély igazolása

ETIKAI ENGEDÉLY

… Molnár Gyöngyvér részére a 2023.08.10-én a „*Karrierépítés támogatása: a Szegedi Tudományegyetemen felvételt nyerő hallgatók felkészültségének monitorozása és támogatása*" c. kutatás tárgyában (kutatásvezető: Molnár Gyöngyvér, résztvevők: Habók Anita, Pásztor Attila, Magyar Andrea) benyújtott etikai kérelmet az SZTE Neveléstudományi Etikai Bizottság a kutatási terv és a kérelemhez csatolt kiegészítések áttanulmányozása alapján elbírálta, és a következő döntést hozta:

A bizottság a szakmai-etikai engedélyt **jóváhagyja**/~~nem hagyja jóvá~~.

INDOKLÁS:

A kutatás célja egy online értékelési rendszer kidolgozása, amely alkalmas a tanulási sikeresség szempontjából meghatározó jelentőségű tudás- és képességszintek felmérésére, illetve hozzájárulhat a lemorzsolódás csökkentéséhez. A kutatást az MTA Közoktatás-fejlesztési Kutatási Program, az SZTE IKIKK és OMIKK, valamint a kutatócsoport tagjainak esetleges további pályázatai támogatják. A kutatásban minden SZTE nappali képzésben frissen felvett hallgató megszólításra kerül. A részvétel önkéntes, a minta nagysága évenként változó, több ezer fő. Az adatgyűjtés módszere online teszt és kérdőív. A tesztek megoldása és a kérdőívek kitöltése az eDia online platformon keresztül történik. A tájékozott beleegyezést megvalósul. Nem vesznek részt a vizsgálatban 18 éven aluli személyek.
A kutatók biztosítják, hogy a résztvevők személyiségi jogai, testi és lelki egészsége ne sérüljön.
Mindezek alapján megállapítható, hogy a benyújtott kutatási terv a neveléstudományi és a tágabb értelemben vett társadalomtudományi humán kutatások szakmai-etikai kritériumainak megfelel.

Szeged, 2023. augusztus 23.

SZEGEDI TUDOMÁNYEGYETEM Nevelés-tudományi Doktori Iskola

Prof. Dr. Pikó Bettina
az Etikai Bizottság elnöke

AUTHOR'S PUBLICATION

No	Articles	Indexing
1.	Csányi, R., & Molnár, G. (2025). Looking beyond students' exploration and learning strategies: The role of test-taking effort in complex problem-solving. *Intelligence*, *109*, 101907. https://doi.org/10.1016/j.intell.2025.101907	SJR Q1
2.	Csányi, R., & Molnár, G. (2024). Item- and person-level factors in test-taking disengagement: Multilevel modelling in a low-stakes context. *International Journal of Educational Research Open*, *7*, 100373. https://doi.org/10.1016/j.ijedro.2024.100373	SJR Q1
3.	Csányi, R., & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences*, *106*, 102340. https://doi.org/10.1016/j.lindif.2023.102340	SJR D1
4.	Csányi, R., & Molnár, G. (2024). Az önértékelés buktatói: a tesztmegoldási erőfeszítés kérdőíves és logadatalapú elemzése. *Scientia et Securitas*, *5*(2), 86–95. https://doi.org/10.1556/112.2024.00167	
5.	Csányi, R., Lőrincz, M. M., & Molnár, G. (2024). Egy felnőttképzési MOOC-programon részt vevők aktivitási adatainak elemzése. *Információs Társadalom*, *24*(1), 34. https://doi.org/10.22503/inftars.xxiv.2024.1.2	SJR Q3
6	Csányi, R., & Molnár, G. (2021). A tesztmegoldási motiváció kérdőíves és logadat alapú mérésének összehasonlító elemzése alacsony téttel rendelkező interaktív problémamegoldó környezetben. *Magyar Pedagógia*, *121*(3), 281–307. https://doi.org/10.17670/mped.2021.3.281	
7.	Csányi, R., & Molnár, G. (2022). A tesztmegoldási motiváció szerepe az alacsony téttel rendelkező tesztek eredményeinek értékelésében. *Iskolakultúra*, *32*(1), 44–63. https://doi.org/10.14232/ISKKULT.2022.1.44	
	Conference papers	
8.	Csányi, R., & Molnár, G. (2024). A tesztmegoldási erőfeszítés mérése a feladattal töltött idő és a kattintások száma alapján alacsony téttel bíró teszteken. In A. Habók & M. T. Nagy (Eds.), *XX. Pedagógiai Értékelési Konferencia = 20th Conference on Educational Assessment* (pp. 67–67). Szegedi Tudományegyetem Neveléstudományi Doktori Iskola.	
9.	Csányi, R., & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. In *EARLI 2023. Education as a Hope in Uncertain Times: Book of Abstracts* (pp. 287–287).	

10. Csányi, R., Lőrincz, M., & Molnár, G. (2023). A lemorzsolódás csökkentésének kulcsa a korai beavatkozás: egy felnőttképzési IT MOOC résztvevőinek tanulási profilelemzése. In L. Kasik & Z. Gál (Eds.), *XIX. Pedagógiai Értékelési Konferencia = 19th Conference on Educational Assessment : PÉK 2023 = CEA 2023* (pp. 58–58). Szegedi Tudományegyetem Neveléstudományi Doktori Iskola.

11. Csányi, R., & Molnár, G. (2023). A tesztmegoldás során azonosított motiválatlan válaszok feladat- és személy szintű vizsgálata elsőéves egyetemisták körében. In A. Bajzáth, K. Csányi, & J. Győri (Eds.), *Elkötelezettség és rugalmasság: a neveléstudomány útjai az átalakuló világban: Absztraktkötet* (pp. 260–260). MTA Pedagógiai Tudományos Bizottság, ELTE Pedagógiai és Pszichológiai Kar (ELTE PPK).

12. Csányi, R., & Molnár, G. (2022). A tesztmegoldási erőfeszítés önértékelő kérdőíven alapuló és logadat alapú mérésének összehasonlító vizsgálata. In J. Steklács & Z. Molnár-Kovács (Eds.), *21. századi képességek, írásbeliség, esélyegyenlőség : XXII. Országos Neveléstudományi Konferencia. Absztraktkötet* (pp. 313–313). MTA Pedagógiai Tudományos Bizottság, Pécsi Tudományegyetem Bölcsészet- és Társadalomtudományi Kar, Neveléstudományi Intézet.

13. Csányi, R., & Molnár, G. (2022). A tesztmegoldási motiváció mérési lehetőségeinek összehasonlító elemzése alacsony téttel rendelkező tesztek esetén. In J. B. Fejes & A. Pásztor-Kovács (Eds.), *XVIII. Pedagógiai Értékelési Konferencia = 18th Conference on Educational Assessment: program és összefoglalók = program and abstracts* (p. 78). Szegedi Tudományegyetem, Neveléstudományi Doktori Iskola.

14. Csányi, R., & Molnár, G. (2021). A tesztmegoldási motiváció szerepe az alacsony téttel bíró tesztek eredményeinek értékelésében. In G. Molnár & E. Tóth (Eds.), *A neveléstudomány válaszai a jövő kihívásaira: XXI. Országos Neveléstudományi Konferencia Szeged, 2021. november 18-20.: program, előadás összefoglalók* (p. 690 pp. 522–522). MTA Pedagógiai Tudományos Bizottság, SZTE BTK Neveléstudományi Intézet.