UNIVERSITY OF SZEGED

DOCTORAL SCHOOL OF EDUCATION

INFORMATION AND COMMUNICATION TECHNOLOGIES IN EDUCATION

SALEH AHMAD ALRABABAH

PHD DISSERTATION

MEASURING COMPLEX PROBLEM SOLVING IN JORDANIAN HIGHER EDUCATION: FEASIBILITY, CONSTRUCT VALIDITY AND LOGFILE-BASED BEHAVIOURAL PATTERN ANALYSES

SUPERVISOR:

GYÖNGYVÉR MOLNÁR DSC

PROFESSOR OF EDUCATION

SZEGED, HUNGARY

2022

AKCNOWLEDGMENT

Firstly, I would like to thank my supervisor Prof. Gyöngyvér Molnár for providing me with the wonderful opportunity to do my PhD study under her supervision. The words are not enough to thank her for all the support and patience in guiding me through my PhD studies.

Studying abroad is not always an easy thing to do. Thanks for all people who made this possible. A special thanks to my wife for encouraging me. Your support by having more responsibilities toward our children and family was priceless. My children (Hala, Malek, and Mohammad), I love you.

My sincere appreciation goes to the Doctoral School of Education at the University of Szeged. To my colleague, having you from every piece in this world was a golden chance to know more about the planet that we share.

Thank you, mom and dad, for always encouraging and supporting me and for your prayers. Your wisdom and advice guide me to achieve what I have done. I'm proud to be your son. Special Thanks to my brothers and sisters. You always give me the power for this life.

List of Abbreviations

aBIC	adjusted Bayesian Information Criterion
AIC	Akaike Information Criterion
ALCP	Average Latent Class Probabilities
APA	American Psychological Association
AR	Augmented Reality
BIC	Bayesian Information Criterion
CB	Computer-Based
CBA	Computer-Based Assessment
CFI	Comparative Fit Index
COMPED	Computers in Education Study
CPS	Complex Problem Solving
df	Degrees of Freedom
DM	Data Mining
EDM	Educational Data Mining
ICILS	International Computer and Information Literacy Study
ICT	Information and Communications Technology
IEA	International Association for the Evaluation of Educational Achievement
IT	Information Technology
KAC	Knowledge Acquisition

KAP	Knowledge Application
LCA	Latent Classes Analysis
LMS	Learning Management System
LSA	Large-Scale Assessments
MOL	Math Online
MR	Mixed Reality
NAEP	National Assessment of Educational Progress
PB	Paper-Based
PIAAC	International Assessment of Adult Competencies
PIRLS	The International Reading Literacy Study
PISACPS	Programme for International Student Assessment Complex Problem Solving
CPS	Complex Problem Solving
PP	Paper-and-Pencil
RMSEA	Root Mean Square Error of Approximation
RWD	Reading and Writing Direction
SBAC	Smarter Balanced Assessment Consortium
SITES	Second Information Technology in Education Study
TBA	Technology Based Assessment
TIMSS	The Trends in International Mathematics and Science Study
TLI	Tucker Lewis Index
VOTAT	Vary-One-Thing-At-A-Time-Strategy

VR Virtual Reality

WLSMV Weighted Least Squares Mean and Variance adjusted

WOL Writing Online

Contents

Publication list for this study-based dissertation

Paper 1

Alrababah, S. A. & Molnár, G. (2021). The Evolution of Technology-based Assessment: Past, Present, and Future. *International Journal of Learning Technology*. 16(2), 134–157.

Paper 2

Alrababah, S. A. & Molnár, G. (2021). Analysing Contextual Data in Educational Context: Educational Data Mining and Logfile Analyses. *Journal of Critical Reviews*, 8(1), 261–273.

Paper 3

Alrababah, S. A., Wu, H., & Molnár, G. (2022). Measuring Complex Problem-Solving in Jordan: Feasibility, Construct Validity and Behaviour Pattern Analyses. *SAGE open.* (Submitted).

Alrababah, S., Wu, H., & Molnár, G. (2022). Measuring Complex Problem-Solving in Jordan: Feasibility, Construct Validity and Behaviour Pattern Analyses. *Advance*. Preprint. https://doi.org/10.31124/advance.20272437.v1

Paper 4

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) e08775.

Additional publication:

Alrababah, S. A. & Molnár, G. (2021, June). Assessing Complex Problem Solving in Jordanian Higher Education Context. Paper presented at inclusive excellence and inclusive universities Conference. Pecs, Hungary.

Alrababah, S. A. & Molnár, G. (2021, November). Jordanian students' test-taking behaviour and problem-solving achievement. Paper presented on the 21st Conference on educational sciences, Szeged, Hungary.

Alrababah, S. Molnár, G. (2019, November). The developmental tendencies of educational assessment: from traditional to third-generation computer-based assessment. Paper presented at XIX. Országos Neveléstudományi Konferencia, Pecs, Hungary

Alrababah, S., Wu, H. & Molnár, G. (2020, December). The Efficacy of Students' Problem-Solving Strategies in Higher Education in Jordan: Log File Analyses. Paper presented at the EARLI SIG 27 Conference. Antwerp, Belgium.

Alrababah, S. Molnár, G. (2022, April). Measuring complex problem-solving in Jordanian higher education: The effect of demographic, cognitive and affective factors on students' achievement. Paper accepted in CEA 2022 conference, Szeged, Hungary.

1

INTRODUCTION

Technology has been used in assessment since the beginning of using computers in various domains and education. It has provided many advantages, possibilities, and challenges in the field of educational assessment (Pásztor-Kovács et al., 2021). By means of technology, teachers and educational administrations could have developed new policies which fit on a higher level to the expectations of the 21st century, e. g. measuring 21st-century skills even in international large-scale assessments (see e.g. Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA) creative or collaborative problem-solving module; see e.g. OECD, 2014a; Griffin et al., 2012).

The use of technology in the assessment leads to improve the possibilities, the efficacy, the validity, and the quality of assessment and offers numerous advantages over traditional assessments (Alrababah & Molnár, 2021), such as automatic item development, automatic scoring (Becker, 2004; Csapó et al., 2014; Dikli, 2006; Mitchell et al., 2002; Valenti et al., 2003), and reducing costs (Bennett, 2003; Christakoudis et al., 2011; Wise & Plake, 1990). In educational assessment, it provides the basis for innovations, e.g., measuring new constructs, using new item types (Dörner & Funke, 2017). Information and communication technologies, especially computers, had an immense impact on the development of educational assessment from quantitative and qualitative points of view. New science has emerged in the field of assessment, which focuses not only on the analyses of the actual answer and achievement data, but more deeply on the analyses of the contextual data collected during the data collection beyond the students' actual answers. Log file analyses and educational data mining have become state-of-the-art educational assessment analyses attracting increasing research interest. They make it possible to answer research questions that could not be answered through traditional assessment techniques (Molnár & Csapó, 2018).

The growing field of educational data mining uses data mining techniques and methods for searching for different patterns in the recorded, basically unstructured contextual data for analyzing and extracting hidden information about students' actions or test-taking behaviour for a more deeply and better understanding of the examined phenomenon. Educational Data Mining (EDM) has become a significant aspect of analysis, the basis of further developments in educational research and practice (Dahiya, 2018).

On the one hand, logfile analyses (e.g., time-on-task, number of clicking, navigation within the test), based on structured data files can also provide information that is not available with traditional assessment techniques and contribute to a better understanding of the examined

phenomenon. By means of logfile analyses, we can get data about students' test-taking behaviour, e.g. applied exploration strategies while solving complex problems (see e.g. Molnár & Csapó, 2018), or we can analyze the relations of time-on-task and achievement data. Logfile analysis also offers many challenges (e.g. how to make sense of the amount of data extracted) and possibilities (e.g. cover unhidden pattern of the educational phenomenon under examination) in the field of educational assessment (Stadler et al., 2020).

Problem-solving is one of the most often assessed reasoning skills in large-scale educational assessment projects. It is considered one of the most essential skills in the 21st century (Krieger, et al., 2021). Its assessment provides a reasonable basis for introducing how educational assessment techniques developed from traditional paper-and-pencil to computer-based assessment and how the type of research questions and used problem types varied by the changing possibilities in the field of assessment (Wu & Molnár, 2021).

In the present research project, we explore the feasibility and the potential for using computer-based assessment for assessing 21st century skills in Jordan. More specifically, we decided to assess students' 21st century skills during their higher education studies using most of the advantages of computer-based assessment. Complex Problem Solving (CPS) proved to be a good candidate for such a role. Because the test of CPS contains tasks including multimedia elements, requiring interaction (not only clicking on a radio button or entering a text in a textbox) of the test taker with the problem scenarios. It offers great possibilities for monitoring students' test-taking behaviour over time on task or the number of clicks via logged data. CPS as a construct involves knowledge acquisition (KAC) and knowledge application (KAP), which are basic learning elements. CPS allows us to investigate how knowledge is acquired in a new problem situation (KAC) and then applied to actually solve a problem (KAP) in an uncertain situation, which is independent of domain-specific content (Greiff et al., 2013). CPS is, by its nature, an important educational outcome in the twenty-first century (Krieger et al., 2021). Understanding how students acquire knowledge and then applying it has become essential because it highly predicts educational achievement (Schweizer et al., 2013).

CPS has been widely assessed in large-scale international assessments (see OECD, 2014b). However, not all of the countries which participated in the 2012 PISA cycle took part in assessing problem-solving. Only a few countries from the Middle East chose it as an international option. Jordan, the country under investigation, did not. As a result, the current study is likely to be the first to report Jordanian students' CPS skills. Despite the extensive

usage of CPS in international samples, little attention has been paid to analyzing its measurement invariance across cultures and nations.

Structure of the dissertation

The theoretical part of the dissertation (first and second papers) investigates the developmental trends in technology-based assessment in an educational context and highlights how technology-based assessment has reshaped the purpose of educational assessment and the way we think about it. Developments in technology-based assessment stretch back three decades. Around the turn of the millennium, studies centred on computer-based and paper-and-pencil test comparability to ascertain the effect of delivery medium on students' test achievement. A systematic review of media studies was conducted to detect these effects. We present the developmental trends in EDM techniques and logfile analysis in the educational context and their contribution to understanding the contextual data collected beyond the particular response data. We conduct a comparison analysis based on the Scopus database to show the developmental trends by year and domain. Then we shed light on measuring complex problem-solving in the educational context and its methods with different approaches. Finally, the applications of computer-generated logfile analyses in the domain of complex problem solving were investigated. The theoretical studies contain of two journal articles:

- Alrababah, S. A. & Molnár, G. (2021). The Evolution of Technology-based Assessment: Past, Present, and Future. *International Journal of Learning Technology*, 16(2), 134–157.
- Alrababah, S. A. & Molnár, G. (2021). Analysing Contextual Data in Educational Context: Educational Data Mining and Logfile Analyses. *Journal of Critical Reviews*, 8(1), 261–273.

Despite the importance of Complex problem-solving (CPS), we have no knowledge of its measurability, development, or comparability in Arab countries, with a short history of computer-based assessment. We fill this niche and beyond monitoring the applicability of third-generation innovative tests in a Jordanian higher educational context, we run international research to understand the behavioural differences in students' test-taking and problem-solving behaviour in case of European students (Hungarian) and Arab students (Jordanian). The results provide important insights into cross-cultural differences in test-taking behaviour and hidden behavioural patterns of students coming from Arab and European countries as they solve computer-based complex problems and contribute to an understanding of how students from

different educational contexts behave while solving tests, especially, technology-based complex problems.

Papers 3-4 introduce the empirical studies conducted within the confines of the PhD research. In all data collection, CPS was measured using the MicroDYN approach (Greiff & Funke, 2017) and the online eDia platform was used to administer the test (Molnár & Csapó, 2019). Paper 4 presents the main research questions, methods and results of the pilot study. The main aim of the pilot study was to test the applicability of technology-based assessment, especially the feasibility of using innovative third-generation tests in the Jordanian higher education context, where the use of technology in assessment has less attention; and validating a third-generation online test of complex problem-solving in higher education. We also investigated students' test-taking and problem-solving behaviours while working on complex problems in a digital environment using both directly collected answer data and logfile analyses. As a result, this study investigated the role of strategic exploration, various problem solving, and test taking behaviour in CPS success by using log file data to visualize and quantify Arabic students' problem solving behaviour in six CPS problems of varying difficulty and characteristics. The results of this study have been submitted for publication to SAGE Open:

Alrababah, S. A., Wu, H., & Molnár, G. (2022). Measuring Complex Problem-Solving in Jordan: Feasibility, Construct Validity and Behaviour Pattern Analyses. SAGE Open. (Submitted).

The results of the cross-cultural comparison study are introduced in paper 4. This study analyzes behavioural and overall performance data in CPS from two different countries with very different cultures: Jordan and Hungary. First, we monitored measurement invariance of CPS (i.e., MicroDYN) across Jordanian and Hungarian context. Then, in three steps, we examined the nature of the developmental differences. First, we used the traditional scoring method to focus on students' actual answer data. Second, we gained insight into what high- and low-achieving students did during the problem-solving process. Specifically, how motivated they were, as seen by how much effort they had shown during the test administration (number of clicks) and how much time they had spent on the problems. Third, we discovered different problem-solving profiles in both countries using logfiles and a behaviour pattern-finding algorithm. We compared students' behavioural features based on their class profiles and final scores. The results of the cross-national comparison study have been published in form of a journal article:

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 (e08775)

Finally, the sixth part of the dissertation consists of the conclusions derived from the discussions of the findings of the studies. It also includes the recommendations and the suggestions for future research and reveals the limitations of the studies.

References

Alrababah, S., & Molnár, G. (2021). Analyzing contextual data in the educational context: educational data mining and logfile analyses. *Journal of Critical Reviews, 8*(1), 261–273. doi: 10.31838/jcr.08.01.31

Becker, J. (2004). Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT). Unpublished PhD dissertation. Freie Universität, Berlin.

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Princeton, NJ: Educational Testing Service.

Christakoudis, C., Androulakis, G. S., & Zagouras, C. (2011). Prepare items for large scale computer based assessment: Case study for teachers' certification on basic computer skills. *Procedia-Social and Behavioral Sciences, 29*, 1189–1198.

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school-readiness and reasoning skills. *Journal of Educational Psychology, 106*(2), 639–650.

Dahiya, V. (2018). A survey on educational data mining. International Journal of Research in Humanities, Arts and Literature, 6(5), 23-30.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1).

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8*:1153. doi: 10.3389/fpsyg.2017.01153

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. doi: 10.1037/a0031856

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In *Assessment and teaching of 21st century skills* (pp. 1–15). Dordrecht (Netherlands): Springer.

Krieger, F., Stadler, M., Bühner, M., Fischer, F., & Greiff, S. (2021). Assessing complex problem-solving skills in under 20 minutes. *Psychological Test Adaptation and Development*. https://doi.org/10.1027/2698-1866/a000009. Retrieved March 13, 2022 from https://psycnet.apa.org/fulltext/2021-79444-001.pdf

Mitchell, T., Russel, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerized marking of free-text responses. In M. Danson (Ed.), *Proceedings of the Sixth*

International Computer Assisted Assessment Conference. Loughborouh: Loughboroug University. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1884/1/Mitchell_t1.pdf

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Log-file analyses. *Frontiers in Psychology*, *9*, 302.

Molnár, G., & Csapó, B. (2019). How to make learning visible through technology: The eDia-online diagnostic assessment system. In *Proceedings of the 11th International Conference on Computer Supported Education* (CSEDU 2019) - Volume 2, pages 122–131. ISBN: 978-989-758-367-4.

Organisation for Economic Co-operation and Development. (OECD) (2014a). *Results: creative problem solving - students' skills in tackling real-life problems* (Vol. V). Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (OECD) (2014b). *PISA 2012 Technical Report*. Paris: OECD

Pásztor-Kovács, A., Pásztor, A., & Molnár, G. (2021). Measuring collaborative problem solving: research agenda and assessment instrument. *Interactive Learning Environments*, 1–21.

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, *24*, 42–52.

Stadler, M., Hofer, S. & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior*, *111*. https://doi.org/10.1016/j.chb.2020.106442

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*(1), 319–330.

Wise, S. L., & Plake, B. S. (1990). Computer-based testing in higher education. *Measurement and evaluation in counseling and development*.23(1), 3–10.

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*, 1–24. https://doi.org/10.1007/s10212-020-00516-y

2

THE EVOLUTION OF TECHNOLOGY-BASED ASSESSMENT: PAST, PRESENT, AND FUTURE

This article available as:

- Alrababah, S. A. & Molnár, G. (2021). The Evolution of Technology-based Assessment: Past, Present, and Future. *International Journal of Learning Technology*, 16(2), 134–157.

Abstract:

This paper presents developmental trends in technology-based assessment in an educational context and highlights how technology-based assessment has reshaped the purpose of educational assessment and the way we think about it. Developments in technology-based assessment stretch back three decades. Around the turn of the millennium, studies centred on computer-based and paper-and-pencil test comparability to ascertain the effect of delivery medium on students' test achievement. A systematic review of media studies was conducted to detect these effects; the results were varied. Recent work has focused on logfile analysis, educational data mining and learning analytics. Developments in IT have made it possible to design different assessments, thus boosting the number of ways students can demonstrate their skills and abilities. Parallel to these advances, the focus of technology-based assessment has shifted from an individual and summative approach to one which is cooperative, diagnostic and more learning-centred to implement efficient testing for personalised learning.

Keywords: Information and communications technology; ICT; computer-based assessment; CBA; personalization of instruction; time on task; media comparison studies

Introduction

Paper-based (PB) testing, which falls under 'traditional assessment', has played a key role in educational assessment. Its possibilities are greatly restricted compared to technology-based assessment (TBA). TBA covers all forms of assessment which are delivered and marked with the aid of technology, that is, via the most commonly used computers [computer-based assessment (CBA)] or other electronic tools and devices (Kuzmina, 2010). In other words, through TBA there is an interaction between the student and the technology used. We are aware that computers play a dominant role in TBA because of its versatility. We have thus decided to use these terms as synonyms in the study.

If CBA is delivered online, which is the main focus of the present discussion, the benefits increase significantly, mostly building on the possibilities of automatic scoring and feedback. Other forms of TBA and CBA (e.g., optical mark readers for multiple-choice tests) are excluded from the main discussion.

Traditional paper-and-pencil (PP) tests are usually fixed tests; thus, every student receives the same items and tasks in the same order during data collection, independent of ability level. The most crucial disadvantages of PP tests are the long feedback time, the restricted suitability of test design, including difficulty, and the use of a limited range of item types.

The use of technology in assessment may lead to improved assessment, thus offering numerous advantages (e.g., automatic item generation, presenting dynamic stimuli and automatic scoring; Becker, 2004; Csapó et al., 2014; Dikli, 2006; Mitchell et al., 2002; Valenti et al., 2003), cutting costs (e.g., delivery, distributing results and evaluating answers; Bennett, 2003; Christakoudiset al., 2011; Wise and Plake, 1990) and laying the groundwork for new innovations (e.g., measuring new constructs and using new item types; Dörner and Funke, 2017; Pachler et al., 2010) in educational assessment. The possibilities, advantages and challenges of TBA are growing in accordance with the level of application (e.g., item development, delivery, scoring and feedback), type of technology (e.g., desktop computer, touchscreen tablets and eye-tracking technologies), methodology used (e.g., fixed testing or adaptive testing), delivery (e.g., internet-based, local server delivery and delivery on removable media), scoring (e.g., automatic, computer-based (CB), but not automatic, human scoring; item-level scoring based on the actual answer of the students or logfile and process data

analyses based on the actions of the students), item types (e.g., traditional multiple-choice or state-of-the-art third-generation innovative item types, including interactivity), domains (e.g., domains can be assessed using traditional methods, such as reading fixed texts, or domains requiring TBA, such as reading digital and printed texts) and the technological conditions of the assessment. Through technology, teachers and educational authorities and managers can develop new policies that truly meet the expectations of the 21st century (Shatunova et al., 2019), e.g., measuring 21st century skills [i.e., critical thinking, problem-solving, creativity, collaboration (teamwork), learning to learn, entrepreneurship and information literacy (Binkley et al., 2012; Redecker et al., 2010)] even on international large-scale assessments (LSA) [see e.g., the OECD Programme for International Student Assessment (PISA) creative or collaborative problem-solving module; Griffin et al., 2012; OECD, 2014).

Information and communications technologies, especially computers, have had an immense impact on the development of educational assessment not only from a quantitative perspective, but also from a qualitative one. New science has emerged in educational assessment, which focuses not only on an analysis of the actual answer and achievement data, but more deeply on an analysis of contextual data gathered during data collection beyond the actual answers provided by the students. Logfile analysis, educational data mining and learning analytics (Csapó et al., 2014; Johnson et al., 2016; Wise, 2019) have become the state of the art in educational assessment analysis and attracted increasing research interest. They make it possible to answer research questions that would be unanswerable using traditional assessment techniques.

To sum up, this paper presents a systematic literature review of the different qualitative or quantitative stages in the development of TBA, from the first use to the latest developments, including a systematic analysis of the media effect and media comparison studies on students' performance using the same test (or measuring the same construct) in different media. We also present and discuss the impact of large-scale international assessments on the evolution of TBA and the challenges of TBA developments for the future.

Research questions

We posited the following research questions on developmental trends in CBA:

RQ1 What role does technology play in educational assessment?

RQ2 Do large-scale international assessments have on effect on the evolution of TBA? If so, what is the nature of this effect?

RQ3 Are PP and CB test results comparable?

RQ4 What is required for the application of CBA among kindergarten children and its systematic integration into everyday school practice?

RQ5 How can an advanced use of the advantages and possibilities of TBA promote a shift in the aim of assessment from effective summative testing to personalised learning?

Early studies in TBA

Using technology in assessment started in the 1920s when Sidney L. Presses designed a machine for testing (Alruwais et al., 2018; Skinner, 1958). 1935 saw the first attempt to use a test scoring machine, the IBM model 805, to test millions of Americans in a type of objective test (Khoshsima & Hashemi, 2017). In the 1970s and 1980s, new computer systems were launched in language testing for purposes (test design, test construction, tryout, delivery, management, scoring, analysis and interpretation, and reporting) beyond simple test scoring (Fulcher, 2000).

The next major development took place in the 1990s, with the focus on the applicability of a broad range of technologies from the most common to the cutting edge (Baker & Mayer, 1999). In recent decades, educational assessment has represented one of the most dynamically developing areas in education; as a result, CBA has become part of large-scale international assessments.

In the early studies of this implementation process, the focus was on the comparability of traditional (PP or face-to-face) and computer-based (CB) test results, or media comparison studies. In media comparison studies, researchers compare the test results of students tested with one medium versus those of – in an ideal case, the same – students tested with another medium using the same test or at least measuring the same construct. It is challenging to conduct valid media comparison research because of difficulties in ensuring that the results are only influenced by the test medium.

Most types of traditional items, such as multiple-choice items, could easily be transferred to a CB assessment platform. The common research question among these studies was the following: whether traditionally administered test results are equivalent to those of CB tests using the same questions and item formats for determining score equivalence (Kuzmina, 2010).

The Guidelines for Computer-Based Tests and Interpretations published by the American Psychological Association (APA) in 1986 specified score equivalence between CB and PP tests. They concluded that

(1) the rank order of the test scores in PP and in CB mode was approximately the same,

(2) the means, standard deviations and shapes of the distribution curves were also nearly the same, at least after rescaling and transforming the data (APA, 1986; Kuzmina, 2010).

In parallel with this issue and building on the results of the different media studies, a great deal of research highlighted the significance and benefits of TBA over traditional paper-based testing.

The effect of large-scale national and international assessments on the evolution of TBA

Around the turn of the millennium, large-scale international assessments [e.g. the National Assessment of Educational Progress (NAEP) and Programme for International Student Assessment of the OECD (PISA)] were conducted to capitalize on CB delivery and implement TBA (Csapó, Ainley, Bennett, Latour, & Law, 2012; OECD, 2010) with the aim of replacing traditional face-to-face and PP testing. One of the hot topics of this period was a comparison of the results of PP and CB assessments for the same construct (Kingston, 2008; Wang, Jiao, Young, Brooks, & Olson, 2008).

Csapó and Molnár (2019) summarized the role of large-scale international assessment in the development of TBA. They argued that the OECD PISA assessments have had an impact on the development of TBA in two major ways: they have advanced the technological infrastructure, and they have tested the preparedness of different countries for the assessments. In PISA the first CBA took place in 2006, when the Computer-Based Assessment of Science was an optional domain (OECD, 2010). Only three countries took part in the data collection (Denmark, Iceland and Korea), but this research served as good practice for future assessments. Three years later, an assessment of digital reading was an extra optional domain in PISA. The research design made it possible to compare the results in PP and digital reading (OECD, 2011). In the following PISA cycle, assessments for reading and mathematics as well as creative problem-solving as an innovative domain were offered in CB delivery mode (OECD, 2013, 2014). This assessment has had a huge impact on the development of CBA and has resulted in a complete shift from PP to CB testing in PISA (OECD, 2016); thus, in 2015, the transition of PISA to CBA was complete, with all the assessments being administered via computer.

The Trends in International Mathematics and Science Study (TIMSS) of the International Association for the Evaluation of Educational Achievement (IEA) is an international comparative study measuring fourth and eighth graders' achievement in mathematics and science as a continuation of IEA's previous studies conducted from the 1960s through the 1980s. Since 1995, with a four-year assessment cycle, TIMSS has assessed student achievement using PP methods on six occasions – in 1999 (eighth grade only), 2003, 2007, 2011 and 2015 (Mullis & Martin, 2017). In the 2019 assessment cycle, TIMSS shifted to CBA and was called eTIMSS with expanded problem-solving and inquiry tasks and novel item types, including drag and drop, sorting and drop-down menu input types. Just around half of the 65 TIMSS countries used eTIMSS in 2019, while the remainder administered TIMSS with the PP format. The shift from traditional PP administration to a fully CBA expanded the coverage of the TIMSS assessment frameworks (Fishbein, Martin, Mullis, & Foy, 2018).

The International Reading Literacy Study (PIRLS) is an assessment of reading comprehension in the fourth grade, which was developed by the IEA and has been conducted every five years since 2001. PIRLS provides information on trends in reading literacy achievement among students in countries that have participated in the assessment cycles. PIRLS was expanded in 2016 to include ePIRLS – an innovative assessment of online reading. ePIRLS is a CBA that uses an engaging, simulated Internet environment to present students with authentic school-like assignments involving social studies and science topics (Mullis, Martin, Foy, & Hooper, 2017).

The IEA has long been concerned with the use of information and communications technology (ICT) in education. The first IEA study in this field was the Computers in Education Study (COMPED) conducted in 1989 and 1992, followed by IEA's Second Information Technology in Education Study (SITES) Module 1 in 1998–99 and Module 2 in 2001 and 2006, which assessed ICT goals and practices in education and the infrastructure in twenty-six countries (Fraillon, Ainley, Schulz, Duckworth, & Friedman, 2019). In 2013, the first cycle of the International Computer and Information Literacy Study (ICILS) was conducted, collecting data in 21 education systems. It investigated how students in Grade 8 in these countries developed the ICT literacy skills that would enable them to participate in the digital world. It researched the differences within and between participating education systems and the relationship of achievement to learning environment and student background. ICILS 2018 also included the computational thinking domain as a process of working out exactly how computers can assist people in solving problems (Fraillon et al., 2019).

The National Assessment of Educational Progress (NAEP) in the USA is one of the first large-scale online assessments in the world. President Barack Obama (2009)[1] said that "I'm calling on our nation's governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess twenty-first century skills like problem-solving and critical thinking and entrepreneurship and creativity". This reflects a trend toward the use of novel methods and techniques in assessment. The NAEP started in 1969. The largest nationwide, continuous, representative assessment in the USA, it focuses on what students know and can do in various subject areas. At the turn of the millennium, a project was designed to explore the use of technology, especially the use of the computer, as a tool to enhance the quality and efficiency of educational assessments, particularly the NAEP. In 2001, the Math Online (MOL) study was the first field investigation; it was followed by the Writing Online (WOL) project in 2002 and the problem-solving in technology-rich environments project in 2003. It investigated how CBA can be used to measure skills that cannot be measured with a PP test (Beller, 2013). In the second stage of development in 2009, almost ten years later, interactive computer tasks were administered in science. 2011 saw the launch of a CB writing assessment, with scenario-based tasks following in 2014. From 2017, the NAEP assessment was fully computerized.

Another national assessment in the USA, the Smarter Balanced Assessment Consortium (SBAC), began in 2014. It tested students using computer-adaptive technology that tailors questions to students based on their answers to previous questions. The SBAC continued to use one test at the end of the year for accountability purposes but created a series of interim tests to inform students, teachers and parents as to whether students are on track (SBAC, 2016). Table 1. summarizes the year of the transition to CBA among large-scale assessments from the NAEP in 2001 to the TIMSS in 2019.

[1] https://www.cbsnews.com/news/obamas-remarks-on-education/

Table 1.

From PP to CB: The transition year for the main large-scale assessments

Large-scale assessment	Start of transition from PP to CB	Transition completed
NAEP	2001	2017
PISA	2006	2019
ICILS (started as computer-based)	2013	2013
SBAC (started as computer-based)	2014	2014
PIRLS	2016	n.d. (2021 – both versions in parallel)
TIMSS	2019	n.d.

Note. n.d.: no date is given

Media comparison studies: CBA vs. PP assessments

Over the past two decades, various media studies have been carried out to determine the effect of delivery medium on students' test achievement (Oz & Ozturan, 2018). We conducted a review of these studies (see Table 2) to obtain a comprehensive overview of the main results in the Google Scholar database. As a first step, we defined the keywords, all connected to the topic of media comparison studies. These studies evaluate the comparability issues (e.g. validity, reliability, objectivity, advantages, costs and effect on test results) of different delivery modes, that is online testing, face-to-face testing and PP testing. We used the following terms separately during a Google Scholar search: media study in computer-based assessment; paper-based vs. computer-based assessment; technology-based assessment/paper-based assessment; computer-based assessment/paper-based assessment; technology-based assessment/paper-and-pencil assessment; and comparison between paper-based assessment and computer-based assessment. As a second filter, we only focused on studies where the same construct was assessed in both modes, CBA and PP, and established after the turn of the millennium. Table 2. summarizes these studies according to age level and sample size, field of study, country and main results.

Table 2.

CBA and PP assessment of the same construct

Researchers	Age level	Sample size	Field of study	Country	Main results
Clariana & Wallace (2002)	Under-graduates	105	Business courses	USA	Students' achievement in the CB environment was significantly higher than that of PP mode.
Choi, Kim, & Boo (2003)	Under-graduates	971	English language proficiency	South Korea	A significant difference between the results in the two modes supports comparability between them.
Bodmann & Robinson (2004)	Under-graduates	55	Web-based course management system	USA	There was no significant difference.
Higgins, Russell, & Hoffmann (2005)	Fourth grade	219	Reading comprehen-sion	USA	There were no statistically significant differences.
Horkay, Bennett, Allen, Kaplan, & Yan (2006)	Eighth grade	1308	Writing assessment	USA	There were no differences in students' writing skills in the two media.
Schatz & Putz (2006)	Under-graduates	30	Management and assessment of sports-related concussion	USA	Significant but modest correlations were found between the modes.
Akdemir & Oğuz (2008)	Under-graduates	47	Educational Measurement course	Turkey	Test scores were not different for the CB and PP tests.
Csapó, Molnár, & Tóth (2009)	Fifth graders (11 years old)	5000	Mathematics and reading comprehen-sion	Hungary	Participants' achievement was lower in CB testing than in the PP format.
Karadeniz (2009)	Under-graduates	38	Computer Hardware and Micropro-cessors course	Turkey	A significant difference was found in the scores in favour of TBA.

Al-Amri (2009)	University medical students	167	English reading tests	Saudi Arabia	Students' achievement in PP mode was significantly better than in CB.
Blazek & Forbey (2011)	Undergraduate students	387	Psychopathology test	USA	There were some significant differences in favour of CB.
Cagiltay & Zalp-Yaman (2013)	First-year engineering students	209	Chemistry course	Turkey	There was no significant performance difference between PP and CB.
Mojarrad, Hemmati, Jafari Gohar, & Sadeghi (2013)	8 to 12 years	66	Reading comprehension assessments in English as a foreign language	Iran	The quantity of reading comprehension did not differ considerably.
Csapó et al. (2014)	First-grade children	364–435	Inductive reasoning	Hungary	PP and CB tests measured pupils inductive reasoning skills very similarly, not only at the overall test level, but at the item level as well.
Logan (2015)	Grade 6	804	Mathematics	Singapore	There were no statistically significant differences.
Hensley (2015)	Grades 4–5	155	Mathematics	USA	There was no difference found in performance on PP and CB tests based on overall performance in mathematics.
Retnawati (2015)	Adults	600	Test of English proficiency	Indonesia	The reliability between the scores for the CB and PP tests was almost the same.
Khoshsima & Hashemi (2017)	Undergraduate students	228	Language knowledge and proficiency	Iran	Test-takers' scores were not different in CB and PP mode.
Hakim (2017)	Foundation year students from the English	200	English language proficiency	Saudi Arabia	There were statistically significant differences between test results in PP and CB mode, with

	Language Institute				participants in CB performing better.
Hardcastle, Herrmann-Abell, & DeBoer (2017)	Elementary, middle and high school	34,068	Science	USA	Performance varied with different test modes according to students' age level.
Garas & Hassan (2018)	University level	78	Financial accounting courses	United Arab Emirates	There was no statistically significant difference between the students' PP and CB scores.
Fishbein et al. (2018)	Fourth and eighth grades	16,894	Maths and science	International	There was an overall mode effect.

A meta-analysis of these studies shows various results on the effect of media on students' test scores, i.e. on students' achievement. More specifically, some of these results demonstrated a significant difference between the two testing modes in favour of CB mode (e.g. Blazek et al., 2011; Clariana & Wallace, 2002; Hakim, 2017; Karadeniz, 2009), while others found the opposite result of participants performing better in PP mode (e.g. Al-Amri, 2009; Csapó et al., 2009). Still other studies reported no significant differences in the two testing modes (e.g. Akdemir & Oğuz, 2008; Bodmann & Robinson, 2004; Cagiltay & Zalp-Yaman, 2013; Garas & Hassan, 2018; Hensley, 2015; Higgins et al., 2005; Horkay et al., 2006; Khoshsima & Hashemi, 2017; Logan, 2015; Mojarrad et al., 2013; Retnawati, 2015).

Beyond the actual test scores, some of the media studies also investigated participants' perceptions, attitudes and opinions with regard to the two-delivery medium. Donovan, Mader and Shinsky (2007) explored students' opinions on the application of computer-based assessment (CBA) instead of PP testing. According to the results of the survey-based study, 88.4% of the students preferred CBA to PP. Llamas-Nistal, Fernández-Iglesias, González-Tato and Mikic-Fonte (2013) confirmed this result, with 43 students out of 52 choosing online testing over traditional assessment methods. Tubaishat, Bhatti and El-Qawasmeh (2006) conducted a study at university level. 59% of the students at the University of Jordan and 50% of the students at Zayed University in the United Arab Emirates liked online exams better than PP exams. Barros (2018) confirmed these findings; that is, students unequivocally preferred CB tests over PP tests.

To sum up, the differences between PP and CB test performance among secondary students and undergraduate students have been widely studied and well documented; however, there is still a gap. Very few studies have focused on the comparability issues of traditional and CB testing among kindergarten children and primary students. Most of the latest media comparison or media effect studies among secondary students have indicated that PP and CB testing are comparable and that students prefer CB tests to PP testing. Based on the few studies focusing on primary students, we can conclude that existing differences decrease over time as computers become widely accessible at schools (Csapó et al., 2014; Mayrath, Clarke-Midura, & Robinson, 2012) and thus test mode effects should no longer represent an issue (Way, Davis, & Fitzpatrick, 2006), at least among secondary students.

Increased effectiveness and advantages of CBA

The development, spread and accessibility of technology offer extraordinary opportunities for the improvement of educational assessment. For example, CBA facilitates highly efficient data collection and more exact, more varied testing procedures to measure more complex skills and abilities and administer more realistic, application-oriented tasks in more authentic testing environments than those of PP assessments (Beller, 2013; Bennett, 2002; Breiter, Groß, & Stauke, 2013; Bridgeman, 2010; Christakoudis et al., 2011; Csapó et al., 2012; Farcot & Latour, 2009; Kikis, 2010; Martin, 2010; Martin & Binkley, 2009; Moe, 2010; Ripley, 2010; van Lent, 2010). Its increased effectiveness and advantages can be observed on every level of assessment:

1) *The costs of testing.* Among the benefits of late proliferation are the lower costs compared to PP assessment. The following activities are necessary for each PP testing session: item writing, proofreading, task editing and test assembly; preparation for printing and printing/copying; test delivery: packing, shipping and distribution; and data collection, collecting the tests, shipping, evaluation, coding, data recording, data cleaning, running the analysis, writing feedback and storing the tests. Each activity has its own cost implications. In the case of CBA, we do not need to print, copy, pack, ship, evaluate, code or record the data. Thus, the costs of data collection can be greatly reduced (Bennett, 2003; Choi & Tinkler, 2002; Christakoudis et al., 2011; Csapó et al., 2012; Csapó, Molnár, & Tóth, 2008; Peak, 2005; Rose, Hess, Hörhold, Brähler, & Klapp, 1999; Valenti, Neri, & Cucchiarelli, 2003; Wise & Plake, 1990). An analysis of the costs of testing showed that even two-thirds of documentation costs can be saved through CBA (Rose et al., 1999). Based on Farcot and Latour's (2009) cost analysis, the initial costs of

PP testing prove to be the lowest. However, this type of testing can only remain competitive in the long run if one does not need to produce many tasks and the complexity of the tasks can be low. As the number of required tasks and their complexity increase, CBA will be a more economical and sustainable method. In sum, the costs of CBA drop significantly in the medium and long term (Bennett, 2003; Choi & Tinkler, 2002; Farcot & Latour, 2009; Kuzmina, 2010; Peak, 2005).

2) *The speed and safety of test administration and data flow*. CBA makes data processing faster and easier (Csapó et al., 2012). It is safer to maintain test-taking security with user names and passwords (Kuzmina, 2010; Marriott & Teoh, 2012). The possibility of selecting questions at random or using adaptive techniques reduces cheating, thus improving safety and providing more objectivity (Marriott & Teoh, 2012). Moreover, an adaptive test algorithm allows a more precise (lower measurement error) or less time-consuming (with the same level of measurement error) assessment of levels of knowledge, skills and abilities (Frey, 2007; Jodoin, Zenisky, & Hambleton, 2006).

3) *The option of providing immediate feedback on completion of testing* (Becker, 2004; Csapó et al., 2014; Dikli, 2006; Mitchell et al., 2002; Valenti, Neri, & Cucchiarelli, 2003) increases the efficiency of the assessment by making it possible to measure even sudden improvement among students with diagnosed atypical development; that is, it paves the way for individualized diagnostic testing beyond the predominantly summative approach (Kettler, 2011; Redecker & Johannessen, 2013; Van der Kleij, Eggen, Timmers, & Veldkamp, 2012).

4) *Indicators of test goodness and efficiency.* The behaviour of the tests – that is, the generalizability of the results, the validity of the construct measured, and the objectivity of data collection and evaluation – is characterized by three indicators: reliability, validity and objectivity. These are assured when the test scores, i.e. the achievement of the students, only depend on the students' level of knowledge and skills, independent of any other factors, such as the circumstances of the data collection and the harshness of the test scorer. With technology, the level of standardization of testing conditions can be significantly boosted, thus ruling out the uncertainty of the human factor. That is, CBA promotes an increase in the indicators of test goodness (Csapó et al., 2014; Jurecka & Hartig, 2007; Marriott & Teoh, 2012; Ridgway & McCusker, 2003). We can thus achieve

improved efficiency and greater measurement precision in the assessment domains already established (Csapó et al., 2014).

5) *Options for measuring new constructs.* CBA has paved the way for the development and use of new, more complex and innovative item types beyond the more traditional first-generation CB items (e.g. multiple choice; Alruwais et al., 2018). With multimedia elements, second-generation items made it possible to create more real-life problems and a more standardized testing environment (e.g. everybody listening to the same voice) than first-generation items. Finally, third-generation tests (Greiff, 2012; Greiff, Wüstenberg, & Funke, 2012; Ripley, Harding, Redif, Ridgway, & Tafler, 2009), including interaction, simulations and cooperation, facilitated the measurement of construct, a feature which would be impossible with traditional assessments that rely on standard item formats (e.g. Complex Problem-Solving (CPS); see Dörner and Funke, 2017; Greiff et al., 2012; in PISA 2012, it was called Creative Problem-Solving). With second- and third-generation tests, we can replicate complex, real-life situations and use authentic tasks, interactions, dynamism, virtual worlds and collaboration within the test to measure even more complex, 21st-century skills (Pachler et al., 2010; Ridgway, McCusker, & Pead, 2004), thus increasing the quality of educational assessment.

6) *Student motivation towards testing changes* (Meijer, 2010; Sim & Horton, 2005). Technology allows creative task presentation through innovative item development opportunities (Pachler et al., 2010; Strain-Seymour, Way, & Dolan, 2009), thus raising the motivation and enjoyment level of the assessment in a way that would have been impracticable in the PP environment. CBA can provide test environments that are similar to entertainment activities (Ridgway et al., 2004).

7) *Effective tools for logging and analysing contextual data* (e.g. time on task and number of student attempts to modify solutions; Csapó et al., 2014), *not only observed variables.* Logfile analysis, educational data mining and learning analytics offer new indicators beyond traditional test results, thus making it possible to conduct a more thorough analysis of the student's behaviour and the structure of the knowledge, skills and abilities measured.

Challenges and drawbacks of using TBA

Despite the many advantages TBA and CBA offer educational researchers, they also face several challenges that call for further research and also involve some drawbacks. Drawbacks

of TBA can be viewed as bigger challenges for the future, thus requiring further developments and researches in the field of educational assessment.

In most cases, the basic technological solutions are already available at the student and/or school level, but – as we have seen in the situation generated by COVID 19 worldwide, even at the international level – their useful integration and application in everyday school practice are limited and require further development. This integration is strongly hindered by issues of diversity, connectivity, and lack of systematicity and compatibility.

There exists no fit for all approaches to TBA. Different assessment needs require different technological conditions, that is, the same solution cannot optimally serve every possible assessment scenario (Csapó et al., 2014). Beyond the proper infrastructure (Alruwais et al., 2018), different problems arise when TBA is used e.g. for high-stakes/low-stakes testing, large-scale/small-scale data collection, standardized/unstandardized assessment, fixed/adaptive testing, summative/formative/diagnostic assessment, using more traditional/innovative item types, replacing traditional PB assessment/launching assessment of skills related to the digital word, placing students in testing centres/in the classroom environment/at home, assessing kindergarten children/primary students/secondary students and students' familiarity/lack of familiarity with TBA. Independent of the aim, place, type and methods of assessment, validity still remains an important issue.

8. Latest developments

The latest developments in the CBA revolution in the educational context highlight two points: first, we have seen a shift from summative to formative and diagnostic assessments, which better reflect students' learning needs, facilitate understanding and provide students with immediate feedback; second, logfile analysis, educational data mining and learning analytics have contributed significantly to an understanding of the phenomenon under examination and expanded the possibilities not only from a quantitative perspective, but also from a qualitative one.

8.1. From efficient testing to personalized learning: Integrating assessment into teaching by means of technology

There is no longer any question as to whether we can develop authentic, real-life, complex, high-quality tests. At the same time, summative test results have limited usefulness with regard to personalized intervention and student-level feedback in general (Csapó & Molnár, 2019).

They are often used for accountability purposes, causing negative effects in testing, such as test coaching (teaching for testing) and test score inflation (see e.g. Koretz, 2018). These effects can have a harmful influence on school climate and teacher stress (Saeki, Segool, Pendergast, & von der Embse, 2018). However, this does not mean that testing is harmful. We must change the purpose of assessment from a rather summative to a more learning-centred, personalized approach, where testing meets the individual needs of students through a frequent, low-stakes assessment combined with prompt and proper feedback about their level of knowledge, skills and abilities (Umami, 2018). Formative and diagnostic CB testing helps personalize learning with effective, adaptive learning and instruction programs (Grant & Basye, 2014). Teachers can use assessment platforms and programs to assess student performance before, during and after learning, which can be used to identify domains of weakness and strength and to promote directed personalized instruction (Grant & Basye, 2014). With TBA, teachers are no longer limited to standardized, yearly, summative exams or periodical, summative classroom tests. They have the opportunity to provide feedback at every step of the learning process and to use these regular assessments to measure the progress of educational objectives for individual students (Cole, 2008). Regular feedback enables teachers to tailor instruction and to aid in students' development more effectively by supplying more frequent information to parents on their children's learning progress (Grant & Basye, 2014).

TBA also makes it possible to fit the difficulty level of the tasks to the ability level of the students by giving students more difficult or less challenging questions. Through this adaptive approach, both the motivation level of the students and the information extracted during testing can be increased and the measurement error decreased.

8.2. Options in logfile analysis, educational data mining and learning analytics are increasing

Contextual information plays a significant role in educational assessment, contributes to a deeper understanding of the phenomenon under examination and can provide answers to research questions which could not be answered with traditional assessment techniques. Traditional assessment methods supply the researcher with very few indicators, such as test scores (quantitative) or subjective feedback (qualitative) from students on the testing/training session. Technology makes it possible to log, collect and analyse students' behaviour during the testing/learning session (e.g. the time needed to execute the task, the number of student attempts to adjust solutions, and the location and number of clicks made by students during the

task and during the test) and thus to quantify even qualitative developmental differences to better understand the fine mechanisms of the phenomenon under examination. However, logfile data are collected more often than they are analysed (Bruckman, 2006).

Table 3 summarizes the number of publications in Scopus as of 2011 with these keywords (phrases) restricted and filtered to these domains and illustrates the ever growing importance and role of logfile analysis in the social sciences and psychology, including time on task, learning analytics, educational data mining and big data. The keywords were used separately, filtered for the domains of the social sciences and psychology and resulting in 60 hits for the year. Based on the results, we can conclude that the history of the analysis of all kinds of log data dates back to 2010. In the last ten years, the number of publications focusing on an analysis of logged data has grown immensely. The most often used state-of-the-art terms are educational data mining, learning analytics and big data.

Table 3.

Search results in Scopus for keywords filtered for the social sciences and psychology (6 Dec. 2019)

Keywords	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Logfile or log-file analysis	4	11	8	4	9	11	13	11	11	11
Time on task	27	27	45	45	50	43	49	67	66	60
Educational data mining	8	12	24	22	38	52	68	87	88	124
Learning analytics	0	5	39	65	144	188	287	330	357	410
Big data	6	7	44	178	426	892	1485	1520	1993	1647

In the following, we only focus on papers containing the phrases “logfile analysis” or “log-file analysis” with the results of those papers illustrating how it is possible to use this type of analysis to quantify qualitative developmental differences to learn more about the phenomenon under examination beyond the score data. These papers also use state-of-the-art analysis (e.g. latent profile analysis) in most cases and go far beyond the possibilities of classical test theory (which is often used in time-on-task analyses). Several of them focus on students’ problem-solving behaviour on third-generation tests (e.g. Greiff et al., 2018; Greiff, Krkovic, &

Hautamäki, 2015; Greiff, Niepel, Scherer, & Martin, 2016; Herde, Wüstenberg, & Greiff, 2016), with a similarly focused paper beyond Scopus found through Google Scholar. As a result of the analyses, qualitatively different exploration strategies have been defined in a complex problem-solving environment (Greiff et al., 2018). It has been confirmed that using a theoretically effective strategy does not always result in high performance and that awareness also plays an influential role in problem-solving. The analyses have identified qualitatively different problem-solving class profiles. The most interesting group is that of rapid learners. These students start out as non-performers in their exploration behaviour in the first problem-solving scenarios but show a rapid learning curve and reach the same high level of exploration behaviour by the end of the test as proficient explorers. However, their final score is exactly the same as those who are high performers on the easiest problems, but low performers on the complex ones, with no so-called intermediate strategy users identified. Generally, the analyses have expanded the scope of previous studies and made it possible to detect a central component of children's scientific reasoning and problem-solving behaviour.

These opportunities and research results are expected to revolutionize education. We are thus able to predict what types of activities would be most beneficial for different students, contributing significantly to the personalization of education (Wise, 2019). According to Johnson et al. (2016), learning analytics is one of the most significant developments of the 21st century. Score-based data and analyses from previous educational research have provided opportunities for post-correction, intervention and modification (e.g. improvement and refinement of tests), with almost all of these data and analyses being output-oriented. Learning analytics enables us not only to confirm that a particular learning unit has been mastered, but also to monitor the learning activity in real time. Based on these data, both computer-controlled and human-driven techniques can be used to better tailor education to the needs of learners, thus moving away from a one-size-fits-all approach (Wise, 2019).

9. Perspectives in the present and challenges for the future

Different areas can be distinguished by discussing the perspectives on educational assessment based on the developments and experiences of the last twenty years. In our view, these developments can enhance the efficiency and efficacy of assessment, thus maximizing students' engagement, motivation and learning (Adesope & Rud, 2019) if they are used not for its own sake (Gonski et al., 2018), but in an integrated and combined way that provides links between assessment, teaching and learning (Neumann et al., 2019).

Innovative technologies combine to form an integrated multi-sensory interactive application to present information to students and thus offer exciting opportunities to increase the efficiency of assessments that are more useful for teachers and more supportive, motivating and effective for students (Gonski, 2018; Koomen & Zoanetti, 2018). However, the real advantage of these technologies, such as touchscreens, augmented reality (AR), virtual reality (VR), mixed reality (MR), robots and behavioural monitoring (e.g. voice recognition, eye gaze, face recognition and touchless user interface) can be effectively used if they are linked to the proper assessment, and educational and developmental theories and methods. However, ways, models and theories must be devised to adapt these technologies to the human mind, including how we learn, and experimental research evidence is needed to determine which instructional features maximize learning outcomes and promote learning processes (Adesope & Rud, 2019). The systematic introduction and application of TBA in everyday school practice, including TBA, using the most common technologies (as we saw in the quarantine situation worldwide because of COVID 19) or even emerging ones, require further research and provide new challenges for educational researchers.

New learning and assessment theories and the reconceptualization of research are needed – integrating models on multimedia learning, machine learning, learning analytics, educational data mining, knowledge representation, developmental psychology and assessment, including visualization of the results to support human learning (Bottou, 2014; Markauskaite, 2010; Martin & Sherin, 2013; Mayer, 2014) – to maximize the use and possibilities of these tools to enhance and facilitate students' learning instead of merely summarizing the current state of their knowledge based on the answer data given, which has been in the focus of educational assessment in the last 20 years. TBA can provide (1) fine-grained, process-oriented data, which can open up new possibilities to understand how we learn (Kramer & Benson, 2013) and thus (2) knowledge which supports personalized learning with constructive feedback. The ability to use available tools calls for new assessment theories (e.g. a more detailed analysis of logfiles and process data beyond the commonly used latent profile and time-on-task analysis). Developments in TBA are moving toward intelligent systems that facilitate students' personalized learning and monitor their emotional and cognitive status, where continuous diagnostic adaptive assessment techniques provide a challenging multimedia learning environment for the user.

The possibilities are becoming almost unlimited; however, implementing them in everyday school practice requires a great deal of research, development and time. As an example of the

very long implementation process, in the 1960s, Rasch published the Rasch model, the well-known and broadly used one-parameter item response theory model. This largely established the basis for adaptive testing, a special form of CBA that is adaptive to each test-taker's ability level. Empirical studies in the 1980s (e.g. Weiss & Kingsbury, 1984) proved that computer-adaptive testing is more effective, reduces testing time without deteriorating measurement precision and strongly increases test-takers' motivation compared to fixed tests, that is, tests comprising the same items for everybody. It took almost 40 years between demonstrating empirical evidence for the effectiveness of the Rasch model and applying it in the most prominent large-scale assessment, OECD PISA. (Please note that PISA was launched in 2000; that is, in the history of PISA, it took almost 20 years.)

10. Limitations

Limitations of the study include the sampling procedure. We restricted the sample to the large research databases on Google Scholar and Scopus. In other words, papers, dissertations and documents which are not indexed in Google Scholar or Scopus were excluded from the analyses. In addition, searches in Scopus were filtered further for the social sciences and psychology; that is, papers which are not indexed in these domains were also excluded from the analyses. We focused on the most prominent, mostly international large-scale assessments (LSA) and excluded other research developments by analysing the effect of LSAs on TBA.

11. Discussion and conclusion

The ICT revolution has reshaped society, required new competences, and opened up new possibilities and challenges in educational assessment. Measuring and developing 21st-century skills (Borodina, Sibgatullina, & Gizatullina, 2019) requires new assessment which goes beyond testing knowledge and provides prompt, meaningful feedback for learners and teachers as well. Traditional assessment methods are sorely lacking in this regard.

The development encompasses three main steps which lead to ever growing possibilities in educational assessment. First-generation CB tests looked very similar to traditional PP testing, but already used several advantages of CBA (e.g. feedback time and delivery mode). Second-generation CBA includes multimedia elements and makes adaptive testing possible. While employing third-generation tasks, even very complex constructs can be measured (e.g. 21st-century skills) by activating interaction, simulation, cooperation and dynamically changing items. To sum up, technology plays an important role in the development of educational

assessment (RQ1), and we observed a significant effect of large-scale international assessments on the evolution of TBA (RQ2).

A number of media studies were conducted around the turn of the millennium, when CBA emerged as a real alternative to PP testing even in large-scale assessments. The results were divergent because of the different samples, knowledge, skills and abilities assessed, and item formats used, but the eventual differences between PP and CB delivery mode and students' test performance have been widely studied and well documented. The latest studies have clearly indicated that PP and CB tests are comparable. Some of these results demonstrated a significant difference between the two testing modes in favour of CB mode, while others found the opposite result. Still other studies reported no significant differences in the two testing modes. If there are differences, they decrease over time as computers become widely accessible with students preferring CB tests to PP testing. Thus, with test mode effects no longer an issue, we can concentrate on the further possibilities of the new technologies in educational assessment (RQ3).

The use of technology has greatly improved the efficiency of testing procedures: it speeds up data collection, supports real-time automatic scoring, accelerates data processing, facilitates immediate feedback and revolutionizes the whole process of assessment, including innovative task presentation (for a detailed discussion of technological issues, see Csapó et al., 2012). Also, it provides new opportunities in item and test development. Beyond these options, technology makes it possible to store and analyse contextual data. This new approach is often called educational data mining, logfile analysis or learning analytics, each representing a slightly different form of analysis. Because of the many advantages, the most important assessments in the near future will probably be administered in a technological environment; however, there is still a need for further research and development on the application of CBA among kindergarten children and its systematic integration into everyday school practice (RQ4).

This trend is explicitly noticeable in the most prominent international large-scale summative assessments (e.g. IEA TIMSS and PIRLS; OECD PISA). In the last few years, taking advantage of one of the greatest possibilities of CBA, automatic feedback, there has been an emphasis on individualized diagnostic assessment beyond the mainly summative approach, thus using the power of prompt, proper feedback to personalize learning and instruction (Shatunova et al., 2019). That is, there is a need for an advanced use of the advantages and

possibilities of TBA in the learning process to shift the aim of assessment from effective summative testing to personalized learning (RQ5).

Undoubtedly, CBA will replace PP at all levels of testing – summative or formative, low- or high-stakes – and offers new opportunities in assessment (e.g. online diagnostic assessment, adaptive testing, embedded assessment, measuring new constructs and learning more about students' test-taking behaviour by analysing logfiles). The technology further expands the possibilities not only from a quantitative perspective, but also from a qualitative one, thus strengthening the use of CBA (Csapó et al., 2012).

References

Adesope, O. O. & Rud, A. G. (2019) Maximizing the affordances of contemporary technologies in education: Promises and possibilities. In O. O. Adesope & A. G. Rud (Eds.), *Contemporary technologies in education* (pp. 1–16). Cham: Springer Nature.

Akdemir, O., & Oguz, A. (2008). Computer-based testing: An alternative for the assessment of Turkish undergraduate students. *Computers & Education, 51*(3), 1198–1204. https://doi.org/10.1016/j.compedu.2007.11.007.

Al-Amri, S. S. (2009). *Computer-based testing vs paper-based testing: Establishing the comparability of reading tests through the evolution of a new comparability model in a Saudi EFL context* (Doctoral dissertation, The University of Essex).

Alruwais, N., Wills, G., & Wald, M. (2018). Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology*, *8*(1), 34–37. https://doi.org/10.18178/ijiet.2018.8.1.1008

American Psychological Association. Committee on Professional Standards, American Psychological Association. Board of Scientific Affairs. Committee on Psychological Tests, & Assessment. [APA] (1986). *Guidelines for computer-based tests and interpretations*. The American Psychological Association.

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in human behavior*, *15*(3–4), 269–282.

Barros, J. P. (2018, March). Students' perceptions of paper-based vs. computer-based testing in an introductory programming course. *10th International Conference on Computer Supported Education, CSEDU 2018* (Vol. 2, pp. 303–308). SciTePress. https://doi.org/10.5220/0006794203030308

Becker, J. (2004). Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT). Unpublished PhD dissertation. Freie Universität, Berlin.

Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. In von Davier, M., E. Gonzalez, I. Kirsch, K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 25–45). Dordrecht (Netherlands): Springer.

Bennett, R. E. (2002). Using electronic assessment to measure student performance: Online testing. *State Education Standard*, *3*(3), 23–29.

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Princeton, NJ: Educational Testing Service.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). *Defining twenty-first century skills*. *In Assessment and teaching of 21st century skills* (pp. 17–66). Springer, Dordrecht.

Blazek, N. L., & Forbey, J. D. (2011). A comparison of validity rates between paper-and-pencil and computerized testing with the MMPI-2. *Assessment*, *18*(1), 63–66. https://doi.org/10.1177/1073191110381718

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51–60. https://doi.org/10.2190/GRQQ-YT0F-7LKB-F033

Borodina, T., Sibgatullina, A., & Gizatullina, A. (2019). Developing creative tfhinking in future teachers as a topical issue of higher education. *Journal of Social Studies Education Research*, *10*(4), 226–245.

Bottou, L. (2014). From machine learning to machine reasoning: An essay. *Machine Learning*, *94*, 133–149.

Breiter, A., Groß, L. M., & Stauke, E. (2013). Computer-based large-scale assessments in Germany.In D. Passey, A. Breiter, & A. Visscher (Eds.), *Next generation of information technology in educational management* (pp. 41–54). Berlin, Heidelberg: Springer.

Bridgeman, B. (2010). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 39–44). Brussels: European Communities.

Bruckman, A. (2006). Analysis of log file data to understand behavior and learning in an online community. In J. Weiss, J. Nolan, J. Hunsinger, & P. Trifonas (Eds.), *The international handbook of virtual learning environments* (pp. 1449–1465). Dordrecht (Netherlands): Springer.

Cagiltay, N., & Ozalp - Yaman, S. (2013). How can we get benefits of computer-based testing in engineering education. *Computer Applications in Engineering Education*, 21(2), 287–293.

Choi, I-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20, 295–20. https://doi.org/10.1191/0265532203lt258oa.

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer based assessment in a K-12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Christakoudis, C., Androulakis, G. S., & Zagouras, C. (2011). Prepare items for large scale computer based assessment: Case study for teachers' certification on basic computer skills. *Procedia-Social and Behavioral Sciences, 29*, 1189–1198.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602. https://doi.org/10.1111/1467-8535.00294.

Cole, R. W. (2008). *Educating everybody's children: Diverse teaching strategies for diverse learners*. ASCD.

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology*, *10*, 1522. https://doi.org/10.3389/fpsyg.2019.01522

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century* skills (pp. 143–230). New York: Springer. https://doi.org/10.1007/978-94-007-2324-5_4

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school-readiness and reasoning skills. *Journal of Educational Psychology, 106*(2), 639–650.

Csapó, B., Molnár, G., & Tóth, K. (2009). Comparing paper-and-pencil and online assessment of reasoning skills: A pilot study for introducing TAO in large-scale assessment in Hungary. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 113–118). Luxemburg: Office for Official Publications of the European Communities.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Donovan, J., Mader, C., & Shinsky, J. (2007). Online vs. traditional course evaluation formats: Student perceptions. *Journal of Interactive Online Learning*, *6*(3), 158–180.

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in psychology*, *8*, 1153.

Farcot, M., & Latour, T. (2009). Transitioning to computer-based assessments: A question of costs. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 108–16). Brussels: European Communities.

Fishbein, B., Martin, M. O., Mullis, I. V., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education*, *6*(1), 11.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Springer.

Frey, A. (2007). Adaptives Testen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Testkonstruktion* (pp. 261–278). Berlin, Heidelberg: Springer.

Fulcher, G. (2000). Computers in language testing. In P. Brett & G. Moterram (eds.*), A special interest in computers: Learning and teaching with information and communications technologies* (pp. 93–107). Manchester: IATEFL Publications.

Garas, S., & Hassan, M. (2018). Student performance on computer-based tests versus paper-based tests in introductory financial accounting: UAE evidence. *Academy of Accounting and Financial Studies Journal*.

Gonski, D. et al. (2018). *Through growth to achievement. Report of the review to achieve educational excellence in Australian schools*. Australian Government. Retrieved from https://docs.education.gov.au/system/files/doc/other/662684_tgta_accessible_final_0.pdf

Grant, P., & Basye, D. (2014). *Personalized learning: A guide for engaging students with technology*. International Society for Technology in Education.

Greiff, S. (2012). Assessment and theory in complex problem solving: A continuing contradiction. *Journal of Educatio nal and Developmental Psychology*, *2*, 49–56.

Greiff, S., Krkovic, K., & Hautamäki, J. (2015). The prediction of problem-solving assessed via microworlds. *European Journal of Psychological Assessment, 32*, 298–306.

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 & Education*, *126*, 248–263.

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 files. *Computers in Human Behavior*, *61*, 36–46.

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

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In *Assessment and teaching of 21st century skills* (pp. 1–15). Dordrecht (Netherlands): Springer.

Hakim, B. M. (2017). Comparative study on validity of paper-based test and computer-based test in the context of educational and psychological assessment among Arab students. *International Journal of English Linguistics, 8*(2), 85–91. http://doi.org/10.5539/ijel.v8n2p85 .

Hardcastle, J., Herrmann-Abell, C. F., & DeBoer, G. E. (2017). Comparing student performance on paper-and-pencil and computer-based tests. *Grantee Submission*. San Antonio, TX.

Hensley, K. K. (2015). Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement. PhD (Doctor of Philosophy) thesis, University of Iowa, 2015. https://doi.org/10.17077/etd.ireseh1q

Herde, C. N., Wüstenberg, S., & Greiff, S. (2016). Assessment of complex problem solving: What we know and what we don't know. *Applied Measurement in Education, 29*(4): 265–277.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation of reading test performance. *The Journal of Technology, Learning and Assessment*, *3*(4).

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B. A., & Yan, F. (2006). Does it matter if I take my writing test on computer?: An empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment*, *5*(2). Retrieved 23 November 2019, from https://ejournals.bc.edu/index.php/jtla/article/view/1641

Jodoin, M., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203–220.

Johnson, L., Becker, S. A., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). *NMC horizon report: 2016 higher education edition* (pp. 1-50). The New Media Consortium.

Jurecka, A., & Hartig, J. (2007). Computer- und Netzbasiertes Assessment. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 37–48). Berlin, Bonn: Bundesministerium für Bildung und Forschung.

Karadeniz, S. (2009). The impacts of paper, web and mobile based assessment on students' achievement and perceptions. *Scientific Research and Essay*, *4*(10), 984–991. Retrieved November 18, 2019 from http://www.academicjournals.org/app/webroot/article/article1380547300_Karadeniz.pdf .

Kettler, R. J. (2011). Computer-based screening for the new modified alternate assessment. *Journal of Psychoeducational Assessment, 29*(1), 3–13.

Khoshsima, H., & Hashemi Toroujeni, S. M. (2017). Comparability of computer-based testing and paper-based testing: Testing mode effect, testing mode order, computer attitudes and testing mode preference. *International Journal of Computer (IJC), 24*(1), 80–99.

Kikis, K. (2010). Reflections on paper-and-pencil tests to eAssessments: Narrow and broadband paths to 21st century challenges. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 99–103). Brussels: European Communities.

Kingston, N. M. (2008). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education*, *22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Koretz, D. (2018). Moving beyond the Failure of Test-Based Accountability. *American Educator*, *41*(4), 22–26.

Koomen, M., & Zoanetti, N. (2018). Strategic planning tools for large-scale technology-based assessments. *Assessment in Education: Principles, Policy & Practice*, *25*, 200–223. doi: 10.1080/0969594X.2016.1173013

Kuzmina, I. P. (2010). Computer-based testing: advantages and disadvantages. *Вісник Національного технічного університету України Київський політехнічний інститут. Філософія. Психологія. Педагогіка*, (1), 192–196. [Bulletin of the

National Technical University of Ukraine Kyiv Polytechnic Institute. Philosophy. Psychology. Pedagogy.]

Llamas-Nistal, M., Fernández-Iglesias, M. J., González-Tato, J., & Mikic-Fonte, F. A. (2013). Blended e-assessment: Migrating classical exams to the digital world. *Computers & Education*, *62*, 72–87.

Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal*, *27*, 423–441. https://doi.org/10.1007/s13394-015-0143-1

Markauskaite, L. (2010). Digital media, technologies and scholarship: Some shapes of eResearch in educational inquiry. *The Australian Educational Researcher*, *37*(4), 79–101.

Marriott, P., & Teoh, L. (2012). ICT for assessment and feedback on undergraduate accounting modules. *The Higher Education Academy. Retrieved from http://www. heacademy. ac.uk/resources/detail/disciplines/finance-and-accounting/using-ICT-in-assessment-and-feedback*.

Martin, R. (2010). Utilising the potential of computer delivered surveys in assessing scientific Literacy. In F. Scheuermann & J. Bjornsson (Eds.), *The Transition to Computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 172–177). Brussels: European Communities.

Martin, R., & Binkley, M. (2009). Gender differences in cognitive tests: A consequence of gender-dependent preferences for specific information presentation formats? In F. Scheuermann & J. Bjórnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 75–82). Luxembourg: Office for Official Publications of the European Communities.

Martin, T., & Sherin, B. (2013). Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences*, *22*(4), 511–520.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York: Cambridge University Press.

Mayrath, M., Clarke-Midura, J., & Robinson, D. (2012). Introduction to technology-based assessments for 21st century skills. *Technology-based assessments for 21st century skills*, 1-11.

Mitchell, T., Russel, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerized marking of free-text responses. In M. Danson (Ed.), *Proceedings of the Sixth*

International Computer Assisted Assessment Conference. Loughborouh: Loughboroug University. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1884/1/Mitchell_t1.pdf

Moe, E. (2010). Introducing large-scale computerized assessment – Lessons learned and future challenges. In F. Scheuermann & J. Bjórnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 51–56). Luxembourg: Office for Official Publications of the European

Mojarrad, H., Hemmati, F., Jafari Gohar, M., & Sadeghi, A. (2013). Computer-based assessment (CBA) vs. Paper/pencil-based assessment (PPBA): An investigation into the performance and attitude of Iranian EFL learners' reading comprehension. *International Journal of Language Learning and Applied Linguistics World*, *4*(4), 418–428.

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Amsterdam: The Netherlands.

Mullis, I. V., Martin, M. O., Foy, P., & Hooper, M. (2017). ePIRLS 2016: International Results in Online Informational Reading. *International Association for the Evaluation of Educational Achievement*.

Neumann, M. M., Anthony, J. L., Erazo, N. A., & Neumann, D. L. (2019). Assessment and technology: Mapping future directions in the early childhood classroom. *Frontiers in Education, 4*(116), doi: 10.3389/feduc.2019.00116

Organisation for Economic Co-operation and Development. (OECD) (2010). PISA 2012 field trial problem solving framework. Paris: OECD.

OECD. (2011). Education at a glance 2011: OECD indicators (p. 497). Paris.

OECD. (2013). PISA 2015 Draft collaborative problem solving framework. Paris: OECD Publishing.

OECD. (2016). Measuring student knowledge and skills. The PISA 2000 assessment of reading, mathematical and scientific literacy. Paris: OECD. https://doi.org/10.1787/9789264255425-en

OECD (2014). PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V). Paris: OECD. https://doi.org/10.1787/9789264208070-5-en

Oz, H., & Ozturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies, 14*(1), 67.

Pachler, N., Daly, C., Mor, Y., & Mellar, H. (2010). Formative e-assessment: Practitioner cases. *Computers & Education, 54*(3), 715–721. https://doi.org/10.1016/j.compedu.2009.09.032

Peak, P. (2005). *Recent trends in comparability studies*. Pearson educational measurement. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/5FC04F5A-E79D- 45FE-8484-07AACAE2DA75/0/TrendsCompStudies_rr0505.pdf

Quansah, F. (2018). Traditional or performance assessment: What is the right way in assessing learners. *Research on Humanities and Social Sciences, 8*(1), 21–24.

Redecker, C., Ala-Mutka, K., & Punie, Y. (2010). Learning 2.0: The impact of social media on learning in Europe. Policy brief. JRC Scientific and Technical Report. EUR JRC56958 EN, Retrieved from http://bit. ly/cljlpq

Redecker, C., & Johannessen, Ø. (2013). Changing assessment – Towards a new assessment paradigm using ICT. *European Journal of Education, 48*(1), 79–96.

Retnawati, H. (2015). The comparison of accuracy scores on the paper and pencil testing vs. computer-based testing. *Turkish Online Journal of Educational Technology–TOJET, 14*(4), 135–142.

Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education, 10*(3), 309–328.

Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment*. ⟨hal-00190440⟩. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.189.5286&rep=rep1&type=pdf

Ripley, M. (2010). Transformational Computer-based Testing. In F. Scheuermann & J. Bjórnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 92–98). Luxembourg: Office for Official Publications of the European Communities.

Ripley, M., Harding, R., Redif, H., Ridgway, J. & Tafler, J. (2009). Review of Advanced e-Assessment Techniques (RAeAT) Final Report. Joint Information Systems Committee.

Rose, M., Hess, V., Hörhold, M., Brähler, E., & Klapp, B. F. (1999). Mobile computergestützte psychometrische Diagnostik. Ökonomische Vorteile und Ergebnisse zur

Teststabilität. *Psychotherapie Psychosomatik Medizinische Psychologie, 49*, 202–207.

Saeki, E., Segool, N., Pendergast, L., & von der Embse, N. (2018). The influence of test-based accountability policies on early elementary teachers: School climate, environmental stress, and teacher stress. *Psychology in the Schools, 55*(4), 391–403.

Schatz, P., & Putz, B. O. (2006). Cross-validation of measures used for computer-based assessment of concussion. *Applied Neuropsychology*, *13*(3), 151–159. DOI: 10.1207/s15324826an1303_2

Shatunova, O., Anisimova, T., Sabirova, F., & Kalimullina, O. (2019). STEAM as an Innovative Educational Technology. *Journal of Social Studies Education Research*, *10*(2), 131–144.

Skinner, B. F. (1958). Teaching machines. *Science*, *128*(3330), 969–977.

Smarter Balanced Assessment Consortium. (SBAC) (2016). Smarter Balanced Assessment Consortium: 2014–15 tTechnical Report. Los Angeles.

Strain-Seymour, E., Way, W., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Research Report. Iowa City, IA: Pearson Education.

Tubaishat, A., Bhatti, A., & El-Qawasmeh, E. (2006). ICT experiences in two different Middle Eastern universities. *Issues in informing science & information technology*, *3*. https://doi.org/10.28945/922

Umami, I. (2018). Moderating influence of curriculum, pedagogy, and assessment practices on learning outcomes in Indonesian secondary education. *Journal of Social Studies Education Research*, *9*(1), 60-75.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research, 2*(1), 319–330.

Van Lent, G. (2010). Risks and benefits of CBT versus PBT in high-stakes testing. In F. Scheuermann & J. Bjornsson (Eds.), *The Transition to Computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 83–91). Brussels: European Communities.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing

mode effects. *Educational and Psychological Measurement*, 68, 5–24. https://doi.org/10.1177/0013164407305592

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. Annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal Educational Measurement, 21*, 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x. [Google Scholar]

Wise, A. F. (2019). Learning analytics: Using data-informed decision-making to improve teaching and learning. In O. Adesope & A. G. Rudd (Eds.), *Contemporary technologies in education: Maximizing student engagement, motivation, and learning* (pp. 119–143). New York: Palgrave Macmillan.

Wise, S. L., & Plake, B. S. (1990). Computer-based testing in higher education. *Measurement and evaluation in counseling and development, 23*(1), 3–10.

3

ANALYZING CONTEXTUAL DATA IN EDUCATIONAL CONTEXT: EDUCATIONAL DATA MINING AND LOGFILE ANALYSIS

This article available as:

Alrababah, S. A. & Molnár, G. (2021). Analysing Contextual Data in Educational Context: Educational Data Mining and Logfile Analyses. *Journal of Critical Reviews*, 8(1), 261–273.

Abstract:

In recent decades, attention has been focused on analyzing contextual data in educational context using educational data mining (EDM), which is a process of employing data mining techniques to transform initial data collected through educational systems into meaningful information. Logfile analysis entails analyzing behavioral processes, time-on-task, and the sequence of actions captured in logfiles and thus introduces novel methods to the analysis of the instruction and learning process as well as to educational assessment. This research paper aims to present the developmental trends in EDM techniques and logfile analysis in educational context and their contribution to a better understanding of the contextual data, collected beyond the particular response data. We conducted a comparison analysis based on the Scopus database to show developmental trends by year and domain. According to the results, (1) research interest in this field has grown immensely in the last few years; that is, EDM is an emerging discipline. In addition, (2) EDM and logfile analysis examine earlier hidden information to provide explanations of students' learning and testing behaviour from a new perspective, thus broadening our understanding of students' behavior, interests, learning processes, motivational aspects, and test results and the reason for their learning outcomes.

Keywords: data mining, educational data mining, contextual data, logfile analyses

Introduction

Improvements in the computer and data sciences have created new methods and possibilities from which education can benefit. Data mining (DM) is a crucial data analysis methodology that has been used in many domains successfully (Tanimoto, 2007). Recently, the expansion of computer-based testing and training in education (Molnár & Csapó, 2019a, 2019b; Mousa & Molnár, 2019) has made information available (Molnár & Csapó, 2018; Tanimoto, 2007; Wu, & Molnár, 2019), which can provide important indicators of students' learning processes but were hidden by traditional assessment and training methods. Nowadays, both educational data mining (EDM) and logfile analysis have become a state-of-the art area of educational assessment, as they can build the basis for new learning theories and new knowledge about how students' learn, behave, complete, and interact with the tasks and problems administered to them (see e.g. Al-Kabi, Shannaq, & Alsmadi, 2011; Chen & Chen, 2009; Guo, Deane, van Rijn, Zhang, & Bennett, 2018; Minaei, Kortemeyer, & Punch, 2004; Mylonas, Tzouveli, & Kollias, 2004; Romero,Ventura, & García, 2008; Xing, Guo, Petakovic, & Goggins, 2015; Zaiane & Luo, 2001; Zhang, Bennett, Deane, & van Rijn, 2019). Based on the literature in the field of assessment, the most commonly employed techniques are time-on-task analysis (see e.g. Alzoubi, Fossati, Di Eugenio, Green & Chen, 2013; Goldhammer, Naumann, Stelter, Tóth, Rölke, & Klieme, 2014; Greiff, Niepel, Scherer, & Martin, 2016; Naumann, 2019; Zoanetti & Griffin, 2017) and analysis of students' exploration and problem-solving behavior in connection with complex problems developed through the MicroDyn approach (see e.g. Goldhammer & Barkow, 2017; Greiff, Wüstenberg, & Avvisati, 2015; Molnár & Csapó, 2018; Tóth, Rölke, Wu, & Molnár, 2019). The possibilities are unlimited, but new methods and new models are needed to use these possibilities in research and educational practice (Molnár & Csapó, 2019b).

This paper summarizes and evaluates the main studies in this field in so as to present a theoretical framework for using contextual data in educational context. It visualizes how contextual data analyses have been utilized in different contexts over the years. The basis for the meta-analysis consisted of papers available in Scopus databases. This study is expected to provide both researchers and educators with information about the possibilities of logfile analysis and highlight the importance of using and analyzing contextual data for a deeper understanding of learning processes. Furthermore, the paper presents how logfile analysis has been used in the educational process with outstanding results, especially in assessment. It has

been employed in exploring the relationship between time-on-task and students' performance, exploring students' strategies and behaviors on problem-solving tests as well.

1. Developmental trends and challenges in EDM

1.1. Data mining

DM is a multidisciplinary field which is considered to be a branch of computer science. It is regarded as an exploratory process, but it could be used for confirmatory investigations (Cheng, 2017). Algarni (2016) views DM as a logical step of the knowledge discovery in database operation. DM methods have links to artificial intelligence, machine learning, statistics, and computer science (Dunham, 2006). However, it is somehow different from other search and analysis methods because it is exploratory, while other analyses are more confirmatory. Hidden patterns can be uncovered with a combination of an explicit knowledge rule, domain knowledge, and advanced analytical skills. As a consequence, the detected patterns and trends can assist organizations with meaningful information and guide their decision-making (Kiron, Shockley, Kruschwitz, Finch, & Haydock, 2011).

DM is a useful artificial intelligence tool that has the potential to uncover helpful information by analyzing data from many dimensions, classifying the information, and summarizing the specified relationships in the database (Algarni, 2016). Afterward, this information assistance makes or improves decisions. In DM solutions, algorithms can be employed to achieve the required results. A case in point is clustering algorithms that recognize modes and group data into varying groups. The data in each group vary from high to less consistent. Based on this, the findings can assist in creating a better decision model, while multiple algorithms implement one solution as they can conduct separate functions. For instance, using a regression tree method makes it possible to obtain financial predictions or association norms to conduct a market analysis (Algarni, 2016).

Nowadays, a significant amount of data in databases surpasses the human ability to extract and analyze the most useful information without the aid of new analysis techniques (Cheng, 2017). Knowledge discovery is the process of significant extraction of involved, unidentified, and potentially meaningful information from a big database (Kiron et al., 2011).

Data mining employed in knowledge discovery has revealed patterns (Algarni, 2016). The precise discovery of patterns through DM is affected by various factors, such as sample size, data fairness, and endorsement of the knowledge involved. All of them affect the degree of certainty that is suited to determining patterns (Dunham, 2006). Even though DM uncovers various patterns in databases, some of them can be interesting from the point of view of learning. It is also crucial to examine the extent of confidence in a given pattern when rating the validity of the results. Loops can take place between any two steps in the process, an area which calls for more iteration (Algarni, 2016).

In using data mining techniques, the following steps are required: (1) developing an understanding of the domain under examination, building relevant prior knowledge, and defining the objective of the analysis; (2) understanding the structure of the target dataset by focusing on the subset of data/variables which are targeted in the analysis; (3) cleaning and pre-processing databases by eliminating noise, developing methods for dealing with missing data, and accounting for time sequence details and documented changes; (4) the data reduction and projection stage, e.g. reduction of dimensionality or methods of transformation; (5) selecting the best fitting DM techniques according to what we call Knowledge Discovery goals; (6) fitting DM algorithms to the dataset to identify patterns; (7) deriving meaningful patterns from a specific form or collection of representations; and finally, before reporting, (8) explaining these mined and detected patterns and/or returning to previous steps for further iteration (Algarni, 2016).

1.2. Educational data mining (EDM)

EDM has become greatly important in research interests and possibilities recently (Amrieh, Hamtini, & Aljarah, 2016). It is a growing field which uses DM techniques and methods to analyze and extract hidden and unknown knowledge from educational data (Amrieh et al., 2016; Baker & Yacef, 2009; Dahiya, 2018; Romero et al., 2008). Silva and Fonseca (2017) defined EDM as a process of transforming initial data collected through educational systems into meaningful information that can be utilized to make decisions and answer research questions. Further issues, such as time, serialization, and context, also play a significant role in analyzing educational contextual data (see e.g. time-on-task; Fatima, Fatima, & Prasad, 2015). To sum up, educational data mining is a multidisciplinary field, including and combining knowledge from a number of domains, such as information retrieval, recommender systems,

visual data analytics, data visualization, social network analyses, cognitive psychology, and psychometrics (Cheng, 2017), machine instruction and learning (Baker & Yacef, 2009), and process mining (Juhaňák, Zounek, & Rohlíková, 2019).

The knowledge discovered by EDM through DM techniques and complex analysis (Amrieh et al., 2016) can assist educational researchers in developing new learning and teaching techniques, understanding learners' behavior, and designing more effective, motivating, and supportive learning environments, that is, enhancing the learning process. It can assist in producing high-quality research results (Amrieh et al., 2016). Dahiya (2018) divided the objectives of EDM into three categories: (1) educational or academic objectives (e.g. designing educational content); (2) administrative or management objectives (e.g. the maintenance of educational infrastructure), and (3) commercial or the market objectives (e.g. capturing the market in terms of enrolments). EDM also employs unique ways, methods, and techniques to address relevant educational problems and questions (Cheng, 2017), such as predicting students' performance (Miguéis, Freitas, Garcia, & Silva, 2018) and predicting students' academic failures (Arora, Singhal & Bansal, 2014; Costa, Fonseca, Santana, Araújo & Rego, 2017; Manhaes, da Cruz, & Zimbrão, 2014).

1.3. Source of the data used in EDM

A database, which can build the basis for educational data mining analysis, can be collected from different sources, such as computer-based assessments (Molnár & Csapó, 2018), web-based education systems, and online learning management systems, such as Moodle (Silva & Fonseca, 2017), educational repositories, or traditional surveys (Amrieh et al., 2016). The structures of these databases differ greatly, that is, the "one size fits all" approach cannot be applied in these types of analyses. Every database is unique and requires "personalized" methods.

1.4. Methods

DM is a powerful technology with great capacity to map student's behavior beyond particular answer data (Silva & Fonseca, 2017) by developing methods to explore large data sets collected in educational settings in order to gain a better understanding of students' behavior, motivation, interests, and results. For educational issues and problems, EDM uses DM techniques and tools

to discover unknown relationships and patterns in large-scale data. These techniques and tools involve machine learning methods, mathematical algorithms, and statistical models, such as clustering and decision trees. They also detect information within the data which queries and reports based exclusively on response data cannot effectively detect (Amrieh et al., 2016; Silva & Fonseca, 2017).

1.5. The use of educational data mining to evaluate students' behavior and actions

EDM is utilized to analyze students' behavior to detect the significantly different ways and learning behavior patterns of students in learning management systems (Kularbphettong, 2017). Evaluating activities and understanding behavior patterns offer new developmental perspectives (Rodrigues, Isotani, & Zarate, 2018).

Romero, Ventura, and García (2008) summarized the benefits of DM techniques and methods in connection with a management system course at the University of Cordoba, Spain. A total of 438 students took part in the research. EDM techniques assisted in classifying students and detecting the sources of any contradiction values from student activities.

Zaiane and Luo (2001) attempted to analyse 395 university students' activities in Britain and Canada to identify typical behavioral patterns from access logs and activities used by the students. They extracted meaningful behavioral patterns based on weblog analysis, which assisted teachers in their students' evaluation and pointed out students' needs toward improving their learning outcomes.

Minaei et al. (2004) analyzed data extracted from an online educational system developed at Michigan State University with Computer-Assisted Personalized Approach to detect and define profiles for different students based on their problem-solving behavior and to develop and recommend decision-making strategies that best fit the different profiles. Association rules were used to detect patterns and cluster profiles.

Chen, Chen, and Liu (2007) assessed the learning outcomes of students by analyzing data on their interactions and access to different educational media. The aim was to provide better insights for teachers into the main factors affecting students' learning outcomes. To achieve these aims, Grey's relational analyses, clustering techniques, fuzzy association rule techniques, and fuzzy inference algorithms were used. The information extracted by these algorithms aided

in developing better educational strategy plans and designing more effective educational materials.

There are no definitive techniques to assess performance during the e-learning process; typically, the final outcomes are given (Mylonas et al., 2004). Chen and Chen (2009) designed an online system using statistical association analyses and fuzzy clustering methods, among other techniques. The main aim was to develop a system which uses formative assessment techniques to profile students and personalize pedagogical materials. This system allows teachers and researchers to detect factors that impact students' learning for better curriculum developent and to find, fit, and personalize the best teaching strategy. The results confirm the importance of performing inferences in the individual performance of students and to evaluate educational strategies recommended to reduce students' failing and dropping out.

Miguéis et al. (2018) utilized EDM techniques to predict graduating students' overall academic performance. Barros, Neto, Plácido, Silva, and Guedes (2019) also used EDM techniques (e.g. decision tree, neural networks, and Balanced Bagging) to predict dropout rates in higher education in Brazil.

Process mining is among the basic EDM techniques. It is attracting increasing research attention (Juhaňák, Zounek, & Rohlíková, 2019; Reimann & Yacef, 2013). Reimann, Markauskaite, and Bannert (2014) focused on process mining in data-intensive research methods from the perspective of methodological challenges.

Schoor and Bannert (2012) analyzed empirical studies using process mining techniques in the context of computer-supported collaborative learning to map social organizational processes. They concluded that these methods are helpful to gain insights into the process of learning and recommended them for further analyses. Bannert, Reimann, and Sonnenberg (2014) used a process mining technique in self-regulated learning to analyze qualitative data collected through a think-aloud protocol. Sedrakyan, De Weerdt, and Snoeck (2016) employed process mining methods in connection with complex problem-solving data to detect students' learning behavior patterns and to link identified profiles to students' learning outcomes. They concluded that the use of such methods was highly beneficial to monitor and analyze cognitive learning processes. In the virtual learning environment, Vidal, Azquez-Barreiros, Lama, and Mucientes (2016) made use of these techniques to analyze log event recordings of students' and teachers' behavior.

Romero, Cerezo, Bogarín, and Anchez-Santill (2016) applied process mining techniques in the analysis of data collected via the Learning Management System (LMS) Moodle, while Papamitsiou and Economides (2016) focused on quiz-taking behavior through a process mining approach by identifying patterns of guessing behavior during testing. They suggested that process mining can provide new opportunities in modeling and detecting various types of students' quiz-taking and guessing behavior in online learning and test environments.

The application and development of DM in education are limited and fairly late compared to other fields, e.g. the life sciences and business administration. Indeed, EDM faces many challenges. One of these arises from the specific attributes of data (Silva & Fonseca, 2017). However, growth can be observed in the amount of educational data, thus opening doors for EDM analysis. Nowadays, EDM has become a significant aspect of analysis, the basis of further developments in educational research and practice (Dahiya, 2018).

2. Developmental trends and challenges in logfile analysis

The emergence of computers in the psychological and educational context has enabled us to analyze the behavioral processes and sequence of actions through the information stored in logfiles (Greiff et al., 2015). Analysis of such data as the interaction between the user and the computer system dates back to 1967 (Al-Kabi et al., 2011).

In the context of computer-based assessment, student interactions during tasks are recorded easily to produce logfiles. These records typically describe keystrokes and mouse events (cursor movement, clicking, dropping, typing, dragging, etc.). Therefore, each separate process is typically logged using a timestamp or a similar occurrence time (Zoanetti & Griffin, 2017). The logfiles can be used as a direct source of information containing data on each behavioral action for each student (Wu & Molnar, 2019). The challenge here is how to make sense of this massive amount of data (Zoanetti & Griffin, 2017). Contextual data introduces new possibilities and raises new research questions for a better understanding of the educational phenomenon under examination (Molnár & Csapó, 2019).

There are many challenges in working with logfiles across a number of problem domains, such as event distribution and data size. A wide range of logfile formats, and different transport and storage methods exist. Usually, it is not possible to manually scan all of the logfile data to identify patterns, anomalies, or different tendencies in the data (Green, 2015).

2.1. A review of empirical studies using time-on-task analysis

Time-on-task can be defined as the amount of time spent on task completion. It involves the time spent identifying the task, the time spent completing it, and the time spent entering the solution. On tasks that require participants to take multiple steps and/or to interact with the problem environment, we can define time-on-task in smaller steps, that is, the time between each click.

There are two different approaches to modeling time-on-task. Time-on-task can be taken as an indicator of the latent construct or it can be employed in explanatory models as a predictor to explain differences in performance (Goldhammer et al., 2014).

Nowadays, most studies focus on the second approach and test the effect of time-on-task on the academic performance of students in a computer-based learning environment (Lee, 2018). Lustria (2007) argued that interactivity can significantly affect comprehension as well as attitudes towards health Web sites, that is undergraduate students who spent more time using interactive websites on health-related information achieved significantly higher on a related achievement test. Louw, Muller, and Tredoux (2008) analyzed the predictive power of different variables, such as previously existing mathematics knowledge, computer access outside of school, time spent on computers both outside and inside of school, confidence in using information technology, computer literacy, degree of enjoyment in learning mathematics, intention to study after school, motivation toward mathematics, parental encouragement, language used at home, and time spent with the computer-based tutoring system employed in the study. According to their research findings, time spent in the computer-based tutoring environment was the most influential predictive factor for academic success among participating students.

Krause, Stark, and Mandl (2009) studied the learning behavior of 137 undergraduate students engaged in studying statistics course in a computer-based learning environment. They found that time-on-task significantly correlated with the students' learning outcomes. Macfadyen and Dawson (2010) analyzed log data for a learning management system. They concluded that the number of log-ins and time spent in the learning management system explain more than 30% of the variance of the final grade earned by the participating undergraduate students. Cho and

Shen (2013) confirmed this result and reported that time-on-task logged in a learning management system, along with effort regulation, predicted students' academic achievement.

Landers and Landers (2015) focused on the predictive power of time-on-task in a game-based learning environment for academic achievement. They reported that undergraduate students engaged in learning in a game-based learning environment spent much more time on the school material than their peers working in a non-game-based learning environment. The additional time improved their academic performance significantly. These studies unanimously confirmed that time-on-task in a computer-based learning environment is a significant predictor for academic success.

In the field of educational assessment until the 21st century, the main focus was almost exclusively on the realization of reliable and valid tests and test-level achievement. With traditional assessment, it was not necessary to collect contextual data, which could be important indicators of behavioral processes that lead to students' final performance. Contextual data can be recorded via technology-based assessment systems, which can help the researcher to extract descriptors of theoretical importance to the task completion process (Goldhammer et al., 2014) and time spent on the task (or part of the task) during testing (Zoanetti & Griffin, 2017). Previous research has indicated that it is a challenging task to define the predictive power of time-on-task to task performance (Naumann, 2019). Spending too much time on complex problem-solving tasks was associated with poor performance (Greiff et al., 2016). In contrast, Alzoubi et al. (2013) observed that if students spent more time on a task on a problem-solving test, their performance became significantly better. Therefore, we would expect that a longer time spent on a problem-solving task results in higher achievement, meaning a longer time allows for longer planning and better planned solutions. Notably, according to the research results, for weak problem-solvers, spending more time on a task may be helpful in compensating for the lack in reading or computer usage (Goldhammer et al., 2014). That is, the key question is if students' attitudes affect their time in completing a task, which may affect their performance and skills in the target domain (Naumann, 2019).

Naumann (2019) examined three variables in digital reading tasks, comprehension skills, reading enjoyment, and reading strategies, which can predict students' performance. It also focused on the time students spend on digital reading tasks with varying difficulty. It analyzed computer-based assessment data retrieved in PISA 2009. Two indicators were built by the researcher in connection with time-on-task: the exact time spent by students on a task and the

average time students spent on each task. The study found positive correlations between enjoyment of reading, comprehension skills, and reading strategies knowledge, and concluded that students with high comprehension skills, enjoyment of reading, and reading strategies were better at dealing with task difficulty with respect to their time usage than those who were less skilled in these areas.

Goldhammer et al. (2014) analyzed the relations between time-on-task and students' achievement in reading and problem-solving. They used the (N=1020) International Assessment of Adult Competencies (PIAAC) computerized reading and problem-solving test data collected in Germany. The confirmed results for item difficulty, namely, time-on-task, increased with item difficulty. In problem-solving, time-on-task correlated positively with item difficulty, while with reading items, which called for more routine processing, time spent on tasks negatively correlated with achievement. They concluded that there is no common interpretation between time-on-task and achievement.

Scherer, Greiff, and Hautamäki (2015) distinguished CPS time-on-task and CPS ability with a positive correlation. According to their findings, cognitive and motivational factors have different predictive power on CPS time-on-task and CPS ability. As the developmental level of CPS skills is not only influenced by that of thinking skills (Goldhammer et al., 2014), the development of both cognitive and affective factors can be seen as a significant educational goal. "Understanding CPS time-on-task and CPS ability is therefore crucial for making progress in the conceptualization and assessment of CPS within and beyond educational context" (Scherer, Greiff, & Hautamäki, 2015, p. 47).

2.2. Analyzing learning processes

A unique challenge is to understand the hundreds of pieces of information that students may produce when participating in complex assessments. Questions such as (1) how to define these different pieces of information, (2) what makes sense, and (3) how to combine them into an evidence-based approach may have meaning in connection with auto-judging, fluency, or motivation, generally speaking, in educational context (Csapó, Ainley, Bennett, Latour & Lawet, 2012). Using logfile analysis can lend these kinds of data meaning and importance, and makes it possible to interpret them (Molnár & Csapó, 2018). One of the advantages of this

broad shift is that a large amount of the observable behavior is stored in logfiles created by the computer and can be accessed to provide additional information about students' behaviour, how they completed the tasks, and how they behaved during testing, thus offering more detailed information beyond the actual test scores. Since computers are available for assessment, the potential for this almost unlimited amount of information has been widely praised (Eichmann, Goldhammer, Greiff, Brandhuber, & Naumann, 2020).

Assessment and learning can be integrated through direct feedback and customized interventions from an educational science perspective. The emergence of different generations of computerized tests has finally led to intelligent measurement, a description of the overall integration of behavioral processes to assess students' skills while students engage in learning using computers. This vision is consistent with a description of logfile-driven integration of learning and computer-based assessment (Greiff et al., 2016). Log data analysis is considered a new approach for the assessment of learning to learn from a self-regulative and motivational perspective (Vainikainen, 2019).

Logfile analysis has introduced numerous modern methods for analyzing learning processes (Molnar & Csapó, 2018). One of them was used by Zhang et al. (2019). They recorded data on students' keystroke logs extracted from writing processes to monitor gender-level differences. The results indicated a female advantage in writing essays and highlighted the gender-level differences in the writing process. In the same line, Guo et al. (2018) also used keystroke logs to distinguish processes used by students in essay composition.

3. Growing research interest in logfile analysis and educational data mining: analyses based on papers indexed by Scopus

The sample for these analyses consisted of papers published between 1966 and 2019 and indexed by Scopus. We used the following keywords during the filtering process: contextual data, log file analysis, data mining, and educational data mining. To detect the significance of contextual data analysis in educational research, we investigated the history and main topics in logfile analysis and educational data mining that have been tackled.

We met 167,563 records, distributed as follows according to the keywords contextual data (1249), log file or log-file or logfile analysis (405) (with different spelling), data mining

(163,496), and educational data mining (2008). The huge number of studies confirmed that there is great research interest in working with contextual data.

Figure 1 shows the almost linear trend of how the number of data mining papers in Scopus has risen in the last twenty-five years. In 1995, there were one hundred such publications, while that number was 15,466 in 2019 (please note that this data only refers to the Scopus databases). It must also be noted that the very first study was published in 1966.

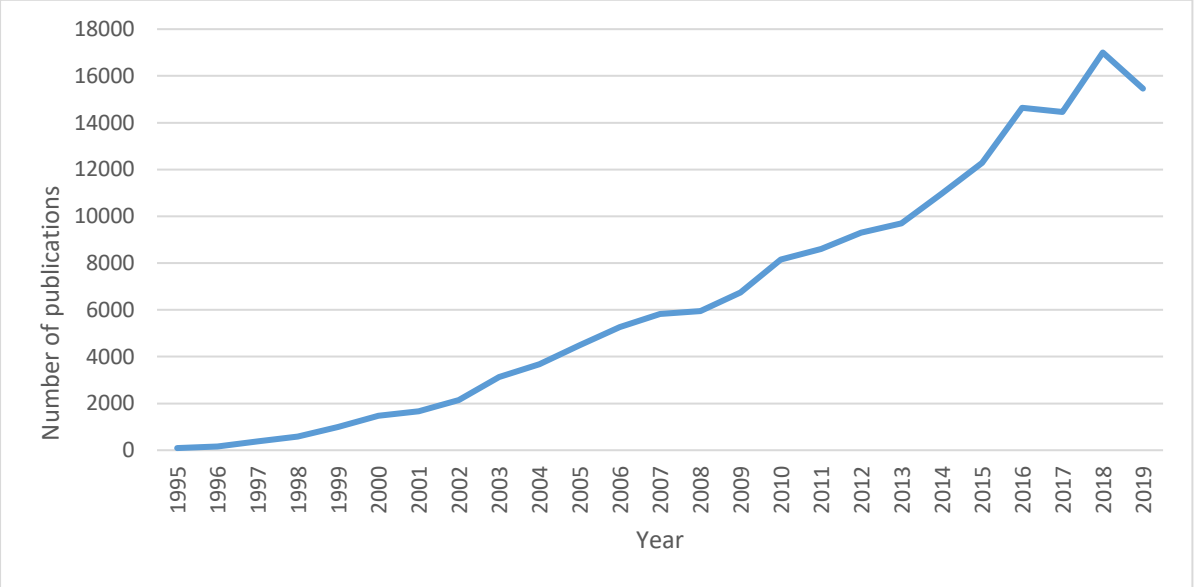

Figure 1. Number of data mining papers published annually between 1995 and 2019

As a next step, we conducted research filtered for the different subject areas defined in Scopus. Figure 2 shows the frequency distribution for data mining papers, but in the different fields. As we expected, computer science proved to be the top field in data mining research (113,045), followed by engineering (46,501) and mathematics (35,978). The social sciences, including education, are among the top ten domains; however, the cumulative number of papers published beyond these three domains does not reach that published in computer science alone. Clearly, it is still rare for these techniques to be applied, as they are basically under development.

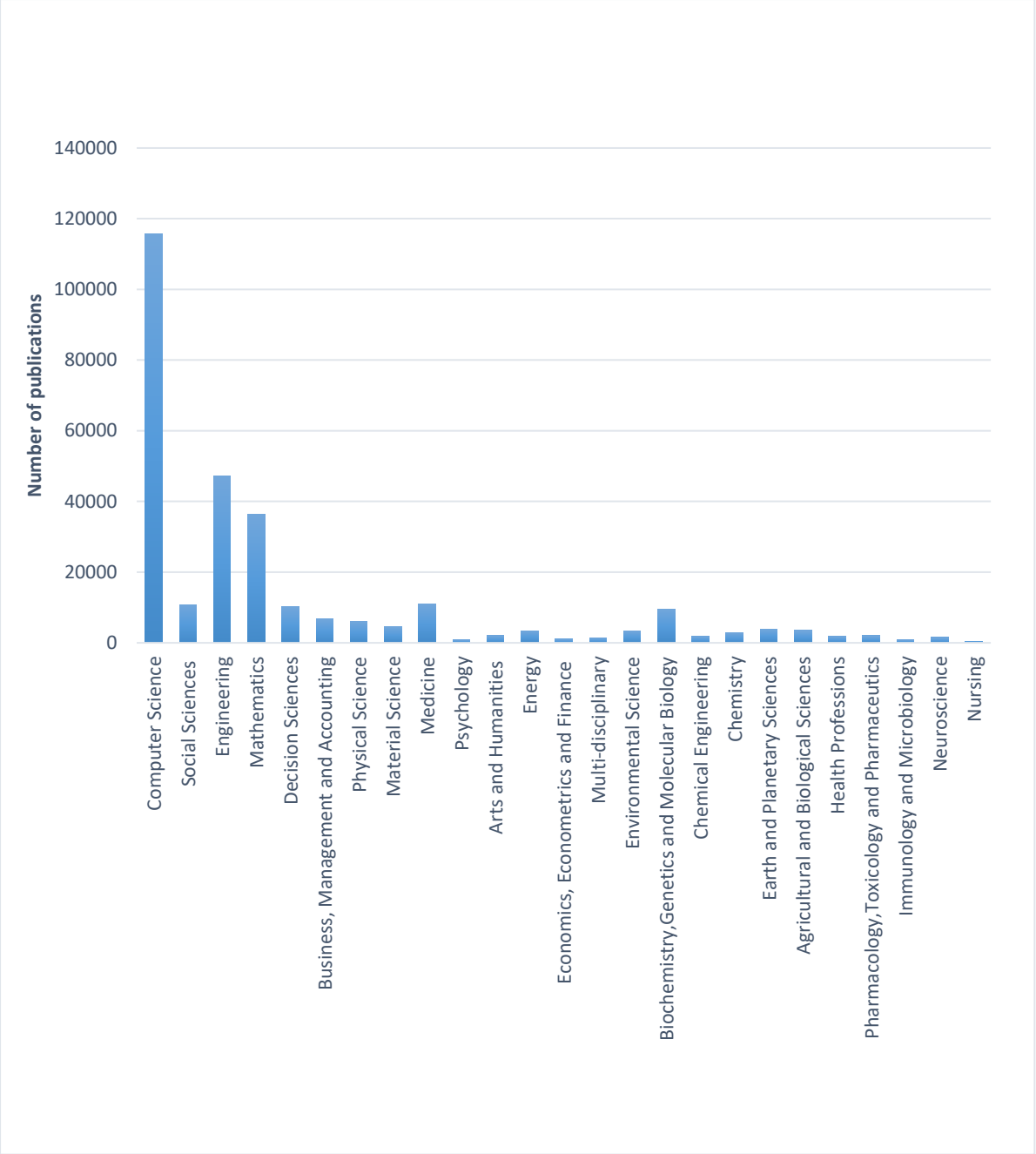

Figure 2. The frequency distribution of data mining studies in different domains

Figure 3 shows the result of the third analysis. It presents how keywords, and thus research interest, have changed in data mining in the last 25 years. The terms contextual data (139 records) and logfile analysis (29 records) received less, but still increasing attention in Scopus compared to educational data mining (314 records). In contrast, the term data mining still tops the list. In 2019 alone, there were 15,466 data mining publications in Scopus.

The most rapidly growing area under the umbrella term data mining was educational data mining. Significant changes happened in 2005, 2008, and 2012. Between 1995 and 2005, the number of publications was almost the same year by year, but they started to increase in 2005. Until 2008, the term logfile was used more often than educational data mining. This may be caused by the effect of the large-scale assessment programs, which became computer-based and made great use of recorded logfiles. Another important change occurred in 2012, when educational data mining became the leading term.

To sum up, data mining became an important method in education, since technology is broadly used in educational assessment (Molnár & Csapó 2019) and learning (especially in the time of COVID 19), which are among the main sources of the contextual data.

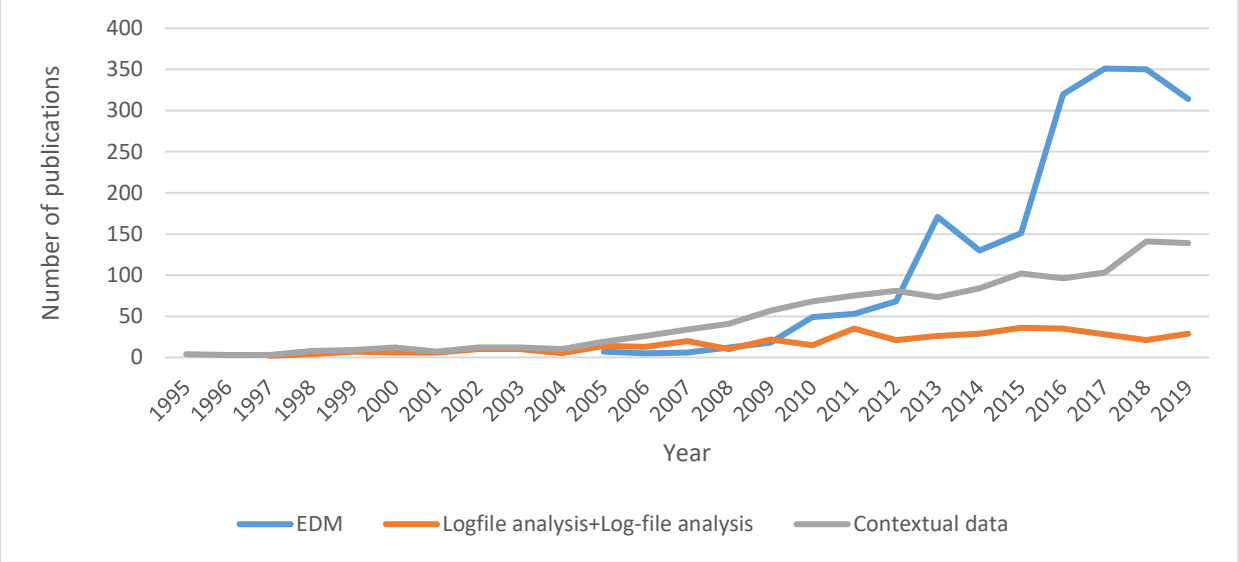

Figure 3. The ever growing research interest in EDM, logfile analysis, and contextual data

Figure 4 shows the result of the fourth analysis of the application level of educational data mining techniques, contextual data, or logfile data analysis in the different domains. The term contextual data was first used in 1981, logfile analysis dates back to 1995, and the latest term has proved to be educational data mining (2000). Interestingly, this most recent term (1700 records) more often occurs in papers published in computer science (641 times) than in the social sciences (279). This confirms our previous statement that educational data mining techniques are still under development and that their real application is still ahead of us. This trend was also observed for the other two keywords (logfile analysis: computer science: 249; social sciences: 93; contextual data: computer science: 641; social sciences: 279).

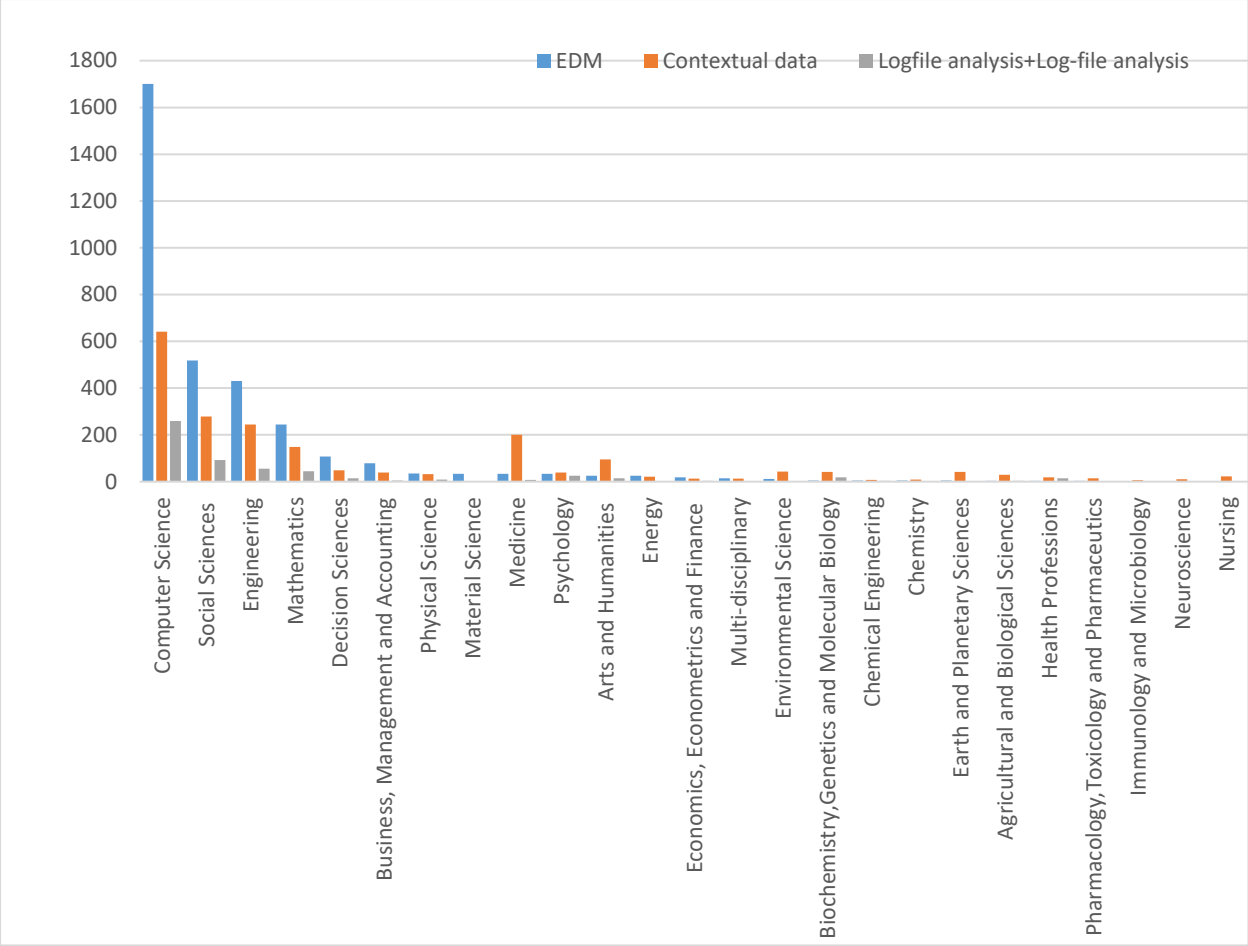

Figure 4. The frequency distribution of studies involving EDM, contextual data, or logfile analysis in different domains

4. Discussion and conclusion

The present research paper has aimed to summarize and evaluate the different definitions and approaches in educational data mining and logfile analysis. It has revealed that educational data mining is still strongly related to computer science in its use of data mining techniques, which are considered a useful artificial intelligence tool. Logfile analysis is also strongly tied to that field; however, since the emergence and spread of computers in psychological and educational research, working with structured databases has made it possible to analyze the behavioral processes captured during data collection and stored in logfiles. EDM research has extracted and analyzed hidden data to evaluate students' behaviours and actions with a focus on the following issues:

- Classifying students and detecting sources of any incongruous values from student activities (Romero, Ventura, & García, 2008).
- Assisting teachers in evaluating their students so as to improve their performance (Zaiane & Luo,2001).
- Detecting changes in students' performance during the teaching and learning process (Schoor & Bannert, 2012).
- Helping teachers to design and develop good educational strategy plans and digital school materials (Chen & Chen, 2009).
- Predicting the likelihood that students will fail or drop out (Neto, Plácido, Silva, & Guedes, 2019).
- Analyzing event logs based on students' and teachers' behaviors in digital learning environments (Lama & Mucientes, 2016).
- Detecting students' quiz-taking behavior (Papamitsiou & Economides, 2016).

Molnár and Csapó (2019) stated that essential educational assessments will be administered in the near future via a technological environment, providing the option to use the numerous advantages of technology-based assessment, including the methods and tools developed from educational data mining and logfile analysis. This study aims to offer researchers and educators information about developmental trends in analyzing contextual data in educational context through an evaluation of the number and distribution of publications in one of the most prominent publication databases, Scopus, in the field of data mining, more specifically, educational data mining, contextual data analysis, and logfile analysis.

The interest in contextual data, especially in educational context, could already be observed in the last century; however, use of the term data mining dates back to 1966. In the 1980s, contextual data was the relevant term, while logfile analysis as a keyword started to be used in the 1990s. The latest term is educational data mining, which has been common since 2000. Despite its educational context, it is still often found in computer science publications, indicating that the methods and algorithms are still under development and not yet at a stage of widespread application.

This study confirms that providing additional indicators during the educational process, especially in assessment, draws researchers' interest to the new domain, thus supporting the

educational process with an abundance of indicators. Indeed, using technology in learning has come into its own as a research field.

The acknowledgment

Preparation of this article was funded by OTKA K135727.

References

Algarni, A. (2016). Data mining in education. *International Journal of Advanced Computer Science and Applications, 7*(6), 456–461.

Al-Kabi, M., Shannaq, M., & Alsmadi, I. (2011). A comparative study of Web usage and searches at Yarmouk University and Jordan. *Abhath Al-Yarmouk: Basic Science & Engineering, 20,* 1–21.

Alzoubi, O., Fossati, D., Di Eugenio, B., Green, N., & Chen, L. (2013). Predicting students' performance and problem solving behavior from iList log data. In *ICCE 2013, 21st International Conference on Computers in Education*.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119–136.

Arora, Y., Singhal, A., & Bansal, A. (2014). Prediction & Warning: a method to improve student's performance. *ACM SIGSOFT Software Engineering Notes, 39*(1), 1–5.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analyzing patterns and strategies in students' self-regulated learning. *Metacognition and learning, 9*(2), 161–185.

Barros, T. M., Neto, S., Plácido, A., Silva, I., & Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. *Education Sciences, 9*(4), 275.

Chen, C. M., & Chen, M. C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education, 52*(1), 256–273.

Chen, C. M., Chen, Y. Y., & Liu, C. Y. (2007). Learning performance assessment approach using web-based learning portfolios for e-learning systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 37*(6), 1349–1359.

Cheng, J. (2017). Data-mining research in education. *arXiv preprint arXiv:1703.10117*. Retrived from https://arxiv.org/pdf/1703.10117.pdf.

Cho, M.-H., & Shen, D. (2013). Self-regulation in online learning. *Distance Education, 34*(3), 290–301.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st Century skills* (pp. 143–230). New York: Springer.

Dahiya, V. (2018). A survey on educational data mining. *International Journal of Research in Humanities, Arts and Literature*, *6*(5), 23–30.

Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*. Pearson Education India.

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology.* Advance online publication

Fatima, D., Fatima, S., & Prasad, A. K. (2015). A survey on research work in educational data mining. *IOSR Journal of Computer Engineering (IOSR-JCE), 17*(2), 43–49.

Funke, J. (2010). Complex problem solving: A case for complex cognition?. *Cognitive processing, 11*(2), 133–142.

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.

Green, B. (2015). Data mining log file streams for the detection of anomalies (Doctoral dissertation, Auckland University of Technology).

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 files. *Computers in Human Behavior*, *61*, 36–46.

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.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement, 55*(2), 194–216.

Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior, 92*, 496–506.

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2011). Analytics: The widening divide. *MIT Sloan Management Review, 53*(3), 1–22.

Krause, U. M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and instruction, 19*(2), 158–170.

Kularbphettong, K. (2017). Analysis of students' behavior based on educational data mining. In *Proceedings of the Computational Methods in Systems and Software* (pp. 167–172). Springer, Cham.

Landers, R. N., & Landers, A. K. (2015). An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance. *Simulation & Gaming, 45*(6), 769–785.

Lee, Y. (2018). Effect of uninterrupted time-on-task on students' success in Massive Open Online Courses (MOOCs). *Computers in Human Behavior, 86*, 174–180.

Louw, J., Muller, J., & Tredoux, C. (2008). Time-on-task, technology and mathematics achievement. *Evaluation and Program Planning, 31*(1), 41–50.

Lustria, M. L. A. (2007). Can interactivity make a difference? Effects of interactivity on the comprehension of and attitudes toward online health content. *Journal of the American Society for Information Science and Technology, 58*(6), 766–776.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588–599.

Manhães, L. M. B., da Cruz, S. M. S., & Zimbrão, G. (2014). WAVE: an architecture for predicting dropout in undergraduate courses using EDM. In *Proceedings of the 29th annual acm symposium on applied computing* (pp. 243–247).

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36–51.

Minaei-Bidgoli, B., Kortemeyer, G., & Punch, W. (2004). Association analysis for an online education system. In *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004* (pp. 504–509). IEEE.

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Log-file analyses. *Frontiers in Psychology, 9*, 302.

Molnár, G., & Csapó, B. (2019). How to Make Learning Visible through Technology: The eDia-Online Diagnostic Assessment System. In *Proceedings of the 11th International Conference on Computer Supported Education* (CSEDU 2019) - Volume 2, pages 122–131. ISBN: 978-989-758-367-4.

Molnár, G., & Csapó, B. (2019). Technology-Based Diagnostic Assessments for Identifying Early Mathematical Learning Difficulties. In *International Handbook of Mathematical Learning Difficulties* (pp. 683–707). Springer, Cham.

Mousa, M., & Molnár, G. (2019). Applying Computer-based Testing in Palestine: Assessing Fourth and Fifth Graders Inductive Reasoning. *Journal of Studies in Education, 9*(3), 1–13.

Mylonas, P., Tzouveli, P., & Kollias, S. (2004). Towards a personalized e-learning scheme for teachers. In IEEE *International Conference on Advanced Learning Technologies* Proceedings (pp. 560–564). IEEE.

Naumann, J. (2019). The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment. *Frontiers in Psychology, 10*, 1429.

Papamitsiou, Z., & Economides, A. A. (2016, July). Process mining of interactions during computer-based testing for detecting and modelling guessing behavior. In *International Conference on Learning and Collaboration Technologies* (pp. 437–449). Springer, Cham.

Reimann, P., & Yacef, K. (2013). Using process mining for understanding learning. In R. Luckin, S. Puntambekar, P. Goodyear, B. Grabowski, J. Underwood, & N. Winters (Eds.), *Handbook of design in educational technology* (pp. 472–481). New York: Routledge.

Reimann, P., Markauskaite, L., & Bannert, M. (2014). e-Research and learning theory: What do sequence and process mining methods contribute. *British Journal of Educational Technology*, *45*(3), 528–540.

Rodrigues, M. W., Isotani, S., & Zarate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, *35*(6), 1701–1717.

Romero, C., Cerezo, R., Bogarín, A., & Sanchez-Santill_an, M. (2016). Educational process mining: A tutorial and case study using moodle data sets. In S. ElAtia, D. Ipperciel, & O. R. Zaïane (Eds.), *Data mining and learning analytics: Applications in educational research* (pp. 3–28). Hoboken, NJ: John Wiley & Sons.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368–384.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50.

Schoor, C., & Bannert, M. (2012). Exploring regulatory processes during a computer supported collaborative learning task using process mining. *Computers in Human Behavior*, *28*(4), 1321–1331.

Sedrakyan, G., DeWeerdt, J., & Snoeck, M. (2016). Process-mining enabled feedback: "Tell me what I did wrong" vs. "tell me how to do it right". *Computers in Human Behavior*, *57*, 352–376.

Silva, C., & Fonseca, J. (2017). Educational data mining: a literature review. In *Europe and MENA Cooperation advances in information and communication technologies* (pp. 87–94). Springer, Cham.

Tanimoto, S. L. (2007). Improving the prospects for educational data mining. *In Proceedings of the complete online proceedings of the workshop on data mining for user modelling, 11th International conference on user modeling* (UM 2007) (pp. 106–110).

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. 193–210). Paris: OECD.

Vainikainen, M. P. (2019). Assessing learning to learn in the 2020S – Theory, methods, and practice. Paper presented at 17th Conference on Educational Assessment, Szeged, Hungary.

Vidal, J. C., Vazquez-Barreiros, B., Lama, M., & Mucientes, M. (2016). Recompiling learning processes from event logs. *Knowledge-Based Systems, 100*, 160–174.

Wu, H., & Molnár, G. (2019). Cross-national Differences in Students' Exploration Strategies in a Computer-Simulated Interactive Problem-Solving Environment: Log-file Analyses. Paper presented at Symposium at 17th Conference on Educational Assessment. Szeged, Hungary.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47*, 168–181.

Zaiane, O. R., & Luo, J. (2001). Towards evaluating learners' behaviour in a web-based distance learning environment. In *Proceedings IEEE International Conference on Advanced Learning Technologies* (pp. 357–360). IEEE.

Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice, 38*(2), 1–13.

Zoanetti, N., & Griffin, P. (2017). Log-file data as indicators for problem-solving processes. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 177–191). Paris: OECD.

4

MEASURING COMPLEX PROBLEM SOLVING IN JORDAN: FEASIBLITY, CONSTRUCT AND BEHAVIOUR PATTERN ANALYSES

This article available as:

Alrababah, S., Wu, H. & Molnár, G. (2022): Measuring Complex Problem-Solving in Jordan: Feasibility, Construct Validity and Behaviour Pattern Analyses. Advance. Preprint. https://doi.org/10.31124/advance.20272437.v1

Alrababah, S. A, Wu, H. and Molnár, G. (2022). Measuring Complex Problem-Solving in Jordan: Feasibility, Construct Validity and Behaviour Pattern Analyses. *SAGE Open*. Submitted.

Abstract:

In the 21st century, complex problem-solving (CPS) serves as a key indicator of educational achievement. However, the elements of successful complex problem-solving have not yet been fully explored. This study investigates the role of strategic exploration and different problem-solving and test-taking behaviours in CPS success, using logfile data to visualize and quantify students' problem-solving behaviour on ten CPS problems with different levels of difficulty and characteristics. Additionally, in the present study, we go beyond the limits of most studies that focus on students' problem-solving behaviour pattern analyses in European cultures and education systems to examine Arabic students' CPS behaviour. Results show that students in the Arabic school system interpret CPS problems the same way. That is, we confirmed the two-dimensional model of CPS, indicating the processes of knowledge acquisition and knowledge application as separate dimensions during the problem-solving process. Large differences were identified in the test-taking behaviour of students in terms of the efficacy of their exploration strategy. We identified four latent classes based on the students' exploration strategy behaviour. The study thus leads to a better understanding of how students solve problems and behave during the problem-solving process in uncertain situations.

Keywords: Complex problem-solving, Logfile analysis, Test-taking behaviour, Higher education, Exploration strategy

1. Introduction

Nowadays, schools should prepare their students for jobs and technologies that do not yet exist and solve problems that have never been faced before (OECD, 2018) to succeed in this new world. Those prospects represent novel needs in higher education and have led to a growing interest in assessment instruments that cover a broader area of competencies than traditional domain-specific skills and disciplinary knowledge (Molnár & Csapó, 2018). These assessment instruments can be used to measure students' 21st-century skills.

The present study focuses on problem-solving, especially complex problem-solving (CPS), in an Arabic higher education environment. We assess the suitability of education programmes in terms of the development of students' 21st-century skills in Jordan, obtaining more knowledge about the factors and mechanisms that constitute a successful complex problem-solver, while considering the problem-solving behaviour of students socialized in different cultures. This study therefore investigates different factors, such as the role of strategic exploration and test-taking behaviour in CPS success, using logfile data to visualize and quantify students' problem-solving behaviour in ten CPS problems with different levels of difficulty and characteristics. Our paper is among the first to study the feasibility and validity of an interactive, innovative third-generation computer-based test, such as the globally examined CPS assessments (see e.g. Csapó & Funke, 2017; Dörner & Funke, 2017; Molnár & Csapó, 2018; Molnár, Greiff, Wüstenberg, & Fischer, 2017; OECD, 2014a; Wüstenberg et al., 2014; Wu & Molnár, 2021) in Jordanian higher education. We investigated whether Jordanian students interpret problems the same way as students in other, mostly European, countries (Greiff et al., 2013; Greiff, Wüstenberg, & Avvisati, 2015; OECD, 2014a; Wüstenberg et al., 2014), where most CPS studies have been carried out (Molnár, Alrababah, & Greiff, 2022). Beyond students' CPS performance, computer-based assessment makes it possible to monitor additional test-taking behavioural actions, such as mouse clicks, time-on-task and problem-solving strategy (Gnaldi, Bacci, Kunze, & Greiff, 2020). These sorts of information have the potential to provide policymakers and researchers with valuable insights into students' CPS skills and offer new ways to assist them in optimizing their cognitive capacity (Wu & Molnár, 2021). Such information is still missing from the evaluation and development of different education systems with increasing cultural diversity.

2. Theoretical background

2.1 Complex problem-solving and its assessment

The classical view defines problem-solving as a step-by-step process, which is passive, reproductive and domain-general, mostly based on trial and error (Greiff, Wüstenberg, Molnár, Fischer, Funke, & Csapó, 2013). In contrast, the Gestalt view considers problem-solving as a productive and active process, where insight, reorganisation and functional fixedness play an important role (Baadte & Müller, 2010). The development of the information-processing approach and Newell and Simon's problem space theory has opened the door to new directions in research. North American research has typically focused on examining the development of expertise in separate domains, while most of the research in Europe has concentrated on the problem-solving processes of complex, unknown problems with the help of computerized scenarios. Reeff, Zabal, and Blech (2006) defined problem-solving as guided thinking and action in situations with no routine solution. Eichmann, Goldhammer, Greiff, Brandhuber, and Naumann (2020) distinguished analytical and interactive problem-solving according to the interactive nature of the problem scenario.

In analytical (static) problem-solving environments, both the problem and the related information are static. That is, there are no changes during the problem-solving process, with all the relevant information being presented at the beginning of the problem-solving process (Greiff, Wüstenberg, Molnár, Fischer, Funke, & Csapó, 2013).

Complex problem-solving (CPS) requires a sequence of complex cognitive processes or continuous activities (Funke, 2010). Previous research has recognised two different approaches to measuring CPS (Buchner 1995; Funke, 2014):

(1) Computer-simulated microworlds, which have a large number of variables like real-life problems. For example, the well-known microworld scenario "Lohhausen", which consists of nearly 2000 associated variables (Dörner et al., 1983). This approach results in highly complex problems with high-level similarities to real-world problems. However, (a) their application requires a very long testing time, and (b) they fail to employ common theoretical frameworks to produce comparable problems in a systematic way (Funke, 2001; Funke & Frensch, 2007). In addition, (c) participants' performance is influenced by many other factors, such as prior knowledge about the problem context, not only their problem-solving skills (Greiff et al., 2015). Finally, (d) the majority of microworld-based problems consist of a few items or many interconnected items, both harming instrument reliability (Greiff et al., 2015).

(2) Simplified, artificial, but still complex problems that follow specific construction rules. Most (though not all) of the characteristics of a complex system are present in minimal complex systems (dynamic, complex and intransparent; see Funke, 1991). A minimal complex system has a low number of variables and relations, resulting in reduced testing time compared to the highly complex and challenging microworlds. The MicroDYN approach falls into this category (Greiff & Funke, 2009; Greiff et al., 2012, Schweizer et al., 2013). It applies a number of independent "fake" scenarios to prevent the influence of participants' previous knowledge (Greiff et al., 2015), it uses only a few variables – that is, problems are easy to scale – and it is widely accepted among problem-solving assessments (see e.g. Csapó & Molnár, 2017; Greiff et al., 2015a; Greiff & Wüstenberg, 2014; Mustafić et al., 2019; OECD, 2014). However, there are limitations to generalization to consider as regards problems of minimal complexity in comparing real-life problems because variables cannot be selectively controlled in a real-life context in most cases (Funke, 2021).

The focus of the present study is on complex problem-solving, especially the MicroDYN approach (Funke, 2014), measured in a computerized and interactive environment. According to the theoretical understanding, complex problem-solving (CPS) in the MicroDYN approach is a two-dimensional construct (Funke, 2001; Greiff et al., 2012; Greiff et al., 2013; Leutner, Wirth, Klieme, & Funke, 2005), consisting of knowledge acquisition and knowledge application. In the first phase of the problem-solving process, the problem-solver needs to acquire knowledge in uncertain situations (knowledge acquisition), while in the second phase, this newly acquired knowledge must be applied in a goal-directed way toward the problem solution (Funke, 2001; Greiff et al., 2018; Novick & Bassok, 2005). In a real-life setting, these two processes are related and take place at the same time. However, in an assessment situation, they are usually separated.

2.2. The role of strategic exploration in problem-solving

Exploring and generating effective information represent the secret to solving a problem successfully. According to Wittmann and Hattrup (2004), "riskier strategies [create] a learning environment with greater opportunities to discover and master the rules and boundaries [of the problem]" (p. 406). Thus, there may be differences in the efficacy of the exploration strategies when gathering information about a problem (Wu & Molnár, 2021). Problem-solvers are supposed to explore the problem environment by acquiring knowledge during strategic exploration (Fisher, Greiff, & Funke, 2012). The development and implementation of strategic

exploration are central actions of the problem-solving process (Wüstenberg, Greiff, & Funke, 2012). Problem-solving success in MicroDYN scenarios, which are simplifications and simulations of real-world problems, is also affected by the adoption and application of strategic exploration. In these artificial problem situations, the isolated variation strategy has been the most frequently discussed exploration strategy (it is often called the vary-one-thing-at-a-time strategy; VOTAT; Vollmeyer, Burns, & Holyoak, 1996). Using the VOTAT strategy, the problem-solver directly detects the effects of a single variable at a time by manipulating a given variable in a systematic way, while keeping the other variables unchanged, i.e. in the neutral position (Molnár & Csapó, 2018). According to previous studies, participants who know how to apply VOTAT are more likely to achieve better on problem-solving tasks (Greiff, Molnár, Martin, Zimmermann, & Csapó, 2018), particularly in minimal complex systems (Fischer et al., 2012). According to Lotz, Scherer, Greiff, and Sparfeldt (2017), effective use of VOTAT correlates with higher levels of intelligence, and successful exploration behaviour may lead to better results in problem-solving (Wu & Molnár, 2021).

VOTAT is among the most effective exploration strategies in most problem-solving environments (Lotz et al., 2017; Wu & Molnár, 2021), and it is the most effective in minimal complex systems (such as MicroDYN). Based on Greiff et al. (2018) and Molnár and Csapó (2018), we have discerned and quantified three types of exploration strategies in each of the problem scenarios in the present analyses: (1) No VOTAT (no VOTAT trial was applied); (2) partial VOTAT (VOTAT trials were used for some but not all of the variables in a given problem scenario); (3) full VOTAT (VOTAT trials were applied for all of the variables in a given CPS scenario) (see Greiff et al., 2018; Molnár & Csapó, 2018; Wu & Molnár, 2021).

3. Aims

Nowadays, there is a positive attitude toward using technology in higher education in Jordan (Al-Khayat, 2017), but we do not have any proof of its feasibility and applicability, especially in the field of assessment. Thus, at the initial phase of the study, we had to test the feasibility and applicability of using innovative, interactive, third-generation computer-based tests in Jordan in a higher-educational context. We also explored students' test-taking and problem-solving behaviour while solving complex problems in a digital environment with both directly collected answer data and logfile analyses. We thus aim:

(Research Aim 1) to test the applicability of an interactive, innovative third-generation test, such as the CPS test, in a country where computer-based assessment has a relatively short history;

(Research Aim 2) to test the structure of the assessed construct (construct validity), that is, to test the underlying dimensionality of CPS measured in the Jordanian educational context, assuming – based on theory and international (mostly European) assessments – a measurement model consisting of two different but highly correlated dimensions or processes of problem-solving (i.e. knowledge application and knowledge acquisition);

(Research Aim 3) to discover and describe the type of strategic exploration used by Jordanian university students while solving CPS problems with different characteristics to understand the mechanism underlying successful CPS; and

(Research Aim 4) to detect the relationships between different types of test-taking and problem-solving behaviour and CPS performance to find new ways to assist students in optimizing their cognitive capacity.

4. Methods

4.1 Participants

The participants were undergraduate students (Mean_age=21.50, SD_age=3.03, N=195) from two Jordanian universities with 15 and 13 faculties, respectively. Students from two faculties took part in the assessment: Arts and Sciences.

4.2 Instruments

CPS was measured with a computer-based test developed within the MicroDYN approach (Greiff & Funke, 2017) and adapted into the Arabic style. In MicroDYN, problem environments consist of up to six variables with up to four different types of relations. The problems are embedded in fictitious cover stories, thus eliminating the influence of prior knowledge (for example, “When you get home in the evening, a young cat is lying on your doorstep. It is exhausted and can barely move. You decide to feed the cat. A neighbour gives you two kinds of cat food. Find the relation between the cat food and the cat’s movement/purring”).

The test consisted of six complex problems with different characteristics and different levels of complexity. On each MicroDYN problem, participants were first expected to explore the

structure of the problem scenario by freely operating the system during the knowledge acquisition phase, that is, by manipulating one or more input variables (displayed on the right side according to the Arabic style) for no more than three minutes (see Fig. 1), and then analyse their effects on the output variables (displayed on the left side according to the Arabic style). In parallel, within the 180 seconds of the knowledge acquisition phase, they were expected to visualize the detected relations by drawing lines between the variables on a concept map presented at the bottom of the screen (see Fig. 2). The history of the settings was shown on a graph linked to each input and output variable. In practice, each problem scenario has four buttons beyond the adjustment sliders and buttons for the input variables: Help, Apply, Reset and Next. By clicking on the Reset button, the participant has the option of deleting all the histories presented on the graphs and setting all the values back to their original values. Each input variable has five stages: +2 (++), +1 (+), 0, −1 (–) and −2 (– –), which can be set using the sliders or buttons (+ or –) next to the input variables. Their effects on the output variables can be tested by clicking on the Apply button. The changes in the output variables are presented in both numerical and graphic formats in the problem scenario. The Next button makes it possible to navigate between the MicroDYN scenarios and its different phases.

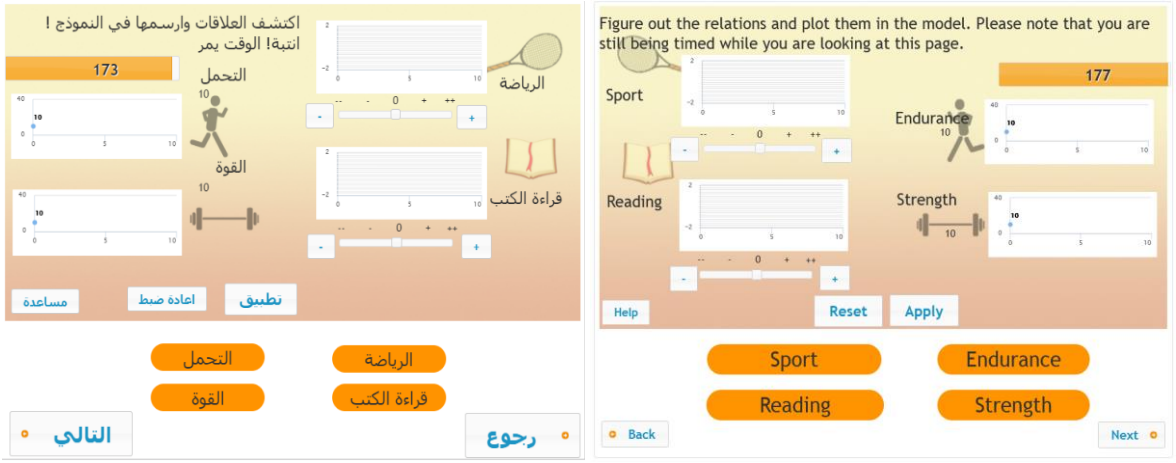

Fig. 1. Sample item from the Arabic-language version of the CPS test – Knowledge acquisition phase. In the example, the task is to find out about the effects of sport and reading on endurance and strength. The controllers of the input variables range from "- -" (value=-2) to "++" (value=+2). In the English version (to the right), they are presented on the left side of the problem environment and on the right side in the Arabic version (to the left). The model is shown at the bottom of the figure.

Fig. 2. Example of problem representation: Drawing relations on a concept map provided onscreen. The English version is provided to the right.

.

Second, in the knowledge application phase, students are expected to use the system in a goal-directed way to reach particular target values (e.g. a given level of movement/purring) of the output variables. To avoid item dependence in this phase, the right concept map is presented at the bottom of the screen. In this part of the problem-solving process, students have no more than 90 seconds and four trials (clicking four times on the Apply button) to solve the problem, that is, to reach the target values of the output variables. Fig. 3 provides a screenshot of the knowledge application phase for a problem with four variables (two input and two output variables) with two direct effects.

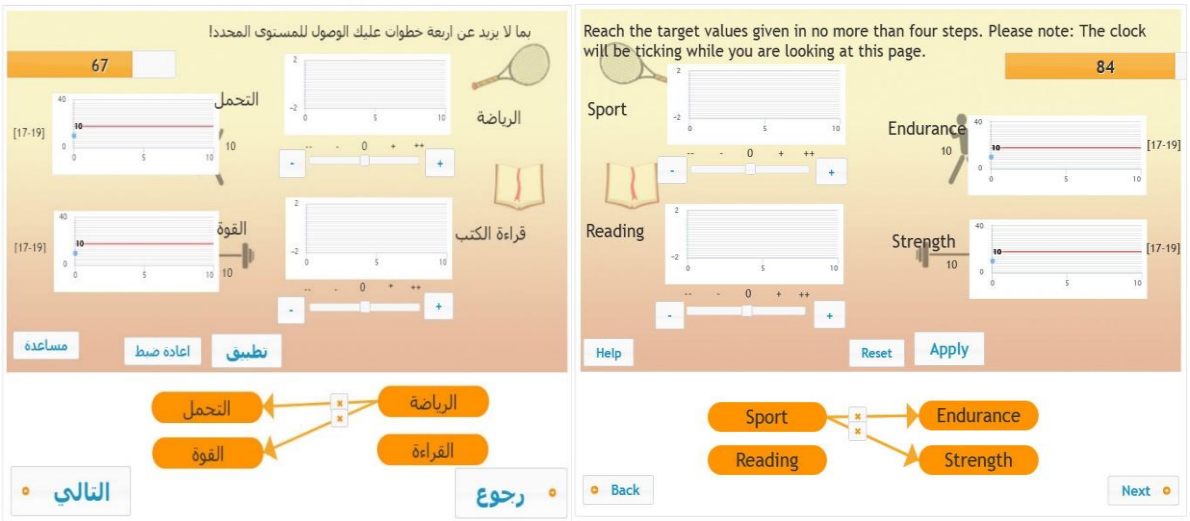

Fig.3. Screenshot of the MicroDYN task "Sport" - Knowledge application phase. The controllers of the input variables range from "- -" (value=-2) to "++" (value=+2). In the English version, they are presented on the left side of the problem environment (screenshot presented to the right) and on the right side in the Arabic version (screenshot presented to the left). The right concept map is presented at the bottom of the figure

The items were adapted from the European to the Arabic writing style by changing the direction of the items from left to right to make them suitable for the right-to-left reading and writing convention of the Arabic language (see Figs. 1 and 3) and by translating the instructions into Arabic. The complexity of the problem was scaled by the number of variables (input-output; 2-2, 3-2, 3-3) and the number (2-4) and type (direct or indirect) of relations. According to Beckmann, Birney, and Goode (2017), raising the number of both variables and relations will boost the difficulty of the CPS problems.

4.3 Procedure

Test administration. The eDia online assessment platform (Molnár & Csapó, 2019) was used for the test administration. The data collection lasted 45 minutes at each university's computer labs. As an achievement indicator, we applied the traditional scoring for both CPS phases (see e.g. Csapó & Molnár, 2017; Fischer et al., 2012; Molnár & Csapó, 2018):

Scoring the answers and labelling the logfiles. If the visualized relations matched the theoretical structure of the problem, students obtained a score of 1. Otherwise, the response was assigned 0 points (for the first phase). Further, if the problem-solver managed to achieve the target values of the output variables within the given time (90 min.) and trial frames (clicking on the Apply button four times), students earned another 1 point, or 0 points otherwise. Applying the traditional scoring, we generated databases for the analyses for Research Aims 1 and 2. Beyond the traditional scoring, students' activity during the problem-solving process was logged and coded based on Molnár and Csapó's (2018) mathematical model and labelling system, which had been developed based on the effectiveness of the strategy usage. Every trial was labelled in the databases. Students' problem-solving behaviour was defined in each problem situation separately by evaluating all of the trials executed within the same problem. If the problem-solving behaviour followed meaningful regularities, it was labelled as a strategy. Three categories were defined within the problem-solving strategies observed: (a) no VOTAT at all – which earned a score of 0 points; (b) partial VOTAT, when VOTAT was used only for some, but not for all of the input variables – which was assigned a score of 1 point; and (c) full VOTAT, when the VOTAT strategy was used for all the input variables – which garnered a score of 2 points. These scores provided the foundation for fulfilling Research Aims 3 and 4.

Data analyses. The descriptive analyses were executed by SPSS (Research Aim 1). Confirmatory factor analyses was used to test the underlying measurement model of complex problem-solving, assuming two different problem-solving processes, knowledge acquisition and application. These analyses were executed by MPlus (Research Aim 2). We have accepted the cut-off values suggested by Hu and Bentler (1999), who indicated that a CFI (Comparative Fit Index) and a TLI (Tucker–Lewis Index) value above .95 and a RMSEA (Root Mean Square Error of Approximation) below .06 indicate a good model fit. We used the preferred estimator for categorical variables, Weighted Least Squares Mean and Variance adjusted (WLSMV; Muthén & Muthén, 2010). Latent class analysis (LCA) was used for Research Aim 3 and was also executed by MPlus. LCA is a pattern-finding algorithm, which searches for latent classes which share similarly observed variables (Collins & Lanza, 2010). In this study, LCA was used to establish latent classes regarding students' problem-solving behaviour. The quality of the LCA was evaluated with the following fit indices: the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and adjusted Bayesian Information Criterion (aBIC). As regards these fit indices, lower values indicate a better model fit. Entropy was utilized to test the accuracy of the classification. The Lo–Mendell–Rubin Adjusted Likelihood Ratio was used to compare the LCA models with different numbers of latent classes (Lo et al., 2001).

5. Results

5.1. Results for Research Aim 1 focusing on testing the applicability of an interactive, innovative third-generation test, such as the CPS test, in a country where computer-based assessment has a short history

Using the traditional scoring for the CPS problems, the internal consistency of the test was high (Cronbach's alpha=.83). The phase-level reliabilities also proved to be good and acceptable (KAC (knowledge acquisition phase): .83; KAP (knowledge application phase): .65). The test proved to be difficult for the students (M=16.8%; SD=16.7% points), whose achievement was significantly higher in the knowledge acquisition phase (M=25.3%; SD=25.7% points) than in the knowledge application phase (M=8.1%; SD=13.0% points; t=10.2, p<.001). To sum up, using interactive, innovative, third-generation computer-based assessments is feasible and reliable in the Jordanian higher education context.

5.2 Results for Research Aim 2 focusing on the underlying dimensionality of CPS measured in the Jordanian educational context, assuming – based on theory and international assessments – a measurement model with two different problem-solving processes (i.e. knowledge application and knowledge acquisition)

The bivariate correlations between the two CPS processes, knowledge acquisition and knowledge application, proved to be medium (r=.45; see Table 1), indicating the measurement of different aspects of CPS.

Table 1. Test and phase level correlations

	KAC	*KAP*
KAC	1.00	
KAP	.453**	1.00
CPS	.925**	.758**

*Note: KAC: knowledge acquisition; KAP: knowledge application; CPS: complex problem-solving; **p<.01 level significant*

Confirmatory factor analyses indicated a good fit (see Table 2). A special χ2-difference test in Mplus (Muthén & Muthén, 2010) was carried out to compare the one- and two-dimensional models. This test revealed that the two-dimensional model fit the data significantly better (Chi-Square Test for Difference Testing=55.317, df=1, p<.001). Thus, we confirmed the theory and the earlier empirical results based on European and Asian data collections as regards CPS (Wu & Molnár, 2021). CPS is a two-dimensional construct, where the KAC and KAP processes can be distinguished empirically.

Table 2. Goodness of fit indices for testing dimensionality of CPS in Jordan

Model	χ^2	*df*	*p*	CFI	TLI	RMSEA
2-dimensional	234.938	169	.001	.965	.961	.045
1-dimensional	289.945	170	.001	.936	.929	.061

Note. df=degrees of freedom; CFI=Comparative Fit Index; TLI=Tucker–Lewis Index;

RMSEA=Root Mean Square Error of Approximation; WLSMV estimator was used in the analyses.

5.3. Results for Research Aim 3 to discover and describe the exploration behaviour of the Jordanian university students while solving computer-based CPS problems with different characteristics

Contrary to our expectations, based on the results for Research Aim 1, the percentage of theoretically effective strategy use was 56.5% for the more complex problems and 64.2% for the less complex ones (see Table 3).

Table 3. Percentage of theoretically effective and non-effective strategy use

Complexity of problem		Percentage (%)	
Number of input and output variables	Number of relations	Effective strategy use	Non-effective strategy use
2-2	2	64.2	35.8
3-2	3	59.8	40.2
3-3	4	56.5	43.5

A large percentage of the Jordanian students employed theoretically effective exploration strategies, including the VOTAT strategy, where the problem-solver manipulates only one input variable systematically while at the same time keeping the other variables unchanged to be able to test the direct effect of the input variables under investigation on the output variables during the problem-solving process. These manipulations allow direct monitoring of changes in output variables to demonstrate the impact of the variable just modified (Molnár & Csapó, 2018). Table 4 summarizes the percentage of no VOTAT, partial VOTAT and full VOTAT strategy users. Independently of problem complexity, a majority of the students applied the most effective exploration strategy during the problem-solving process, but, according to the results for Research Aim 1, they were unable to interpret its meaning. That is, at the very end, most of them failed to solve the problems properly.

Table 4. Percentage of no VOTAT, partial VOTAT and full VOTAT strategy use

Complexity of problem		No VOTAT (%)	Partial VOTAT (%)	Full VOTAT (%)
Number of input and output variables	Number of relations			
2-2	2	37.4	10.0	52.6
3-2	3	39.2	4.6	56.2
3-3	4	40.0	6.9	53.1

5.4. Results for Research Aim 4 to detect the relationships between different types of problem-solving behaviour and problem-solving performance

Only half (52.1%) of the students who applied a theoretically correct strategy made a correct decision as well, solving the easiest problems correctly. This ratio increased to 59.8% with the second sort of complexity before dropping slightly on the most complex problems. Note that the complexity of a problem was defined by the number of variables and the number of relations (Table 5).

Table 5. The ratio of high and low achievers among the theoretically effective strategy users during problem-solving

Complexity of problem		Frequency (%)	
Number of input and output variables	Number of relations	High achievement	Low achievement
2-2	2	52.1	47.9
3-2	3	59.8	40.2
3-3	4	57.4	42.6

Fig. 4 displays the ratio of high and low achievers among theoretically effective strategy users at the task level. The ratio for the effective strategy users to correctly solve an item was higher than 50% on most of the items (except on item 2). Compared to the relatively low performance for the overall sample (see Section 5.1), the theoretically effective strategy users showed a remarkably better performance.

Problem-solving performance among the theoretically effective strategy users suggests the guessing factor, which indicates a correct solution despite theoretically non-effective strategy usage. This also includes participants who recall a theoretically effective strategy but apply it

wrongly and then solve the problem (see Table 6). The guessing factor varied from 15.3% to 7.5%, from the least to the most complex tasks.

The guessing factor (indicating those who solve a problem without an effective strategy) showed the highest effectiveness on item 1. Item 1 is of the 2-2 type, which is the easiest. The effectiveness dropped for the rest of the items. Low achievement for the non-theoretically effective strategy users was very noticeable for all the CPS items.

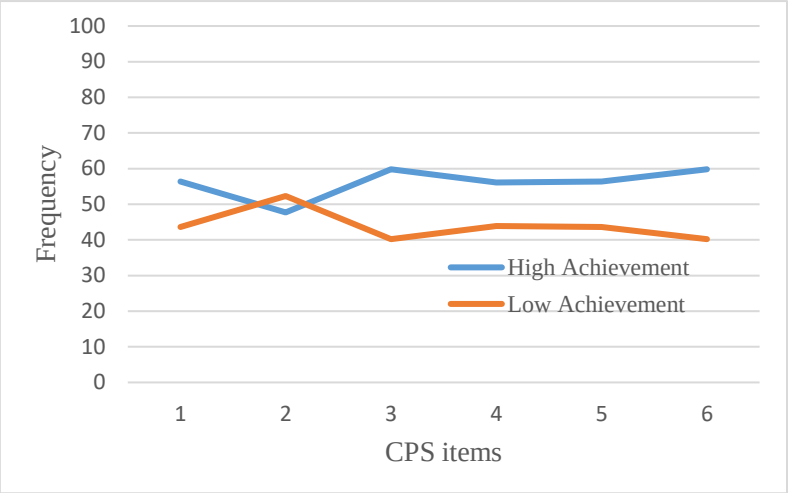

Fig. 4. Problem-solving performance among the theoretically effective strategy users

Table 6. Problem-solving effectiveness among the theoretically non-effective strategy users

Complexity of problem		Frequency (%)	
Number of input and output variables	Number of relations	High achievement	Low achievement
2-2	2	15.3	84.7
3-2	3	11.7	88.3
3-3	4	7.4	92.6

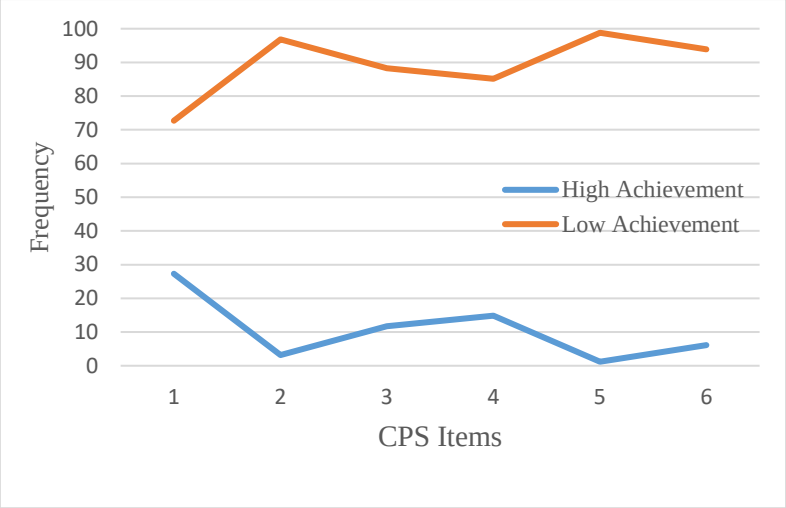

Fig. 5. Problem-solving performance among the theoretically non-effective strategy users

After analysing the performance of theoretically right and theoretically wrong strategy users, we went further to obtain a statistical model of students' problem-solving ability. First, using the tools of latent class analysis and log data to ascertain the use of VOTAT strategies based on students' exploration behaviour, we distinguished three qualitatively different VOTAT strategy users. The Akaike, Bayesian and adjusted Bayesian Information Criterion indices decreased with a growing number of latent classes up to the 4-class solutions. The entropy index reached its maximum value for the 2-class model. However, it was also high for the 3- and 4-class solutions. The Lo–Mendell–Rubin adjusted likelihood ratio test indicated the best model fit for the 3-class model, and it proved to be no longer significant for the 4-class model. Thus, we used the 3-class model – where 93% of the Jordanian students were accurately categorized – to distinguish three qualitatively different class profiles in the further analyses: 50.5% of these students were among the proficient strategy users, who consistently employed VOTAT strategies almost from the very first problem; 18.1% proved to be intermediate explorers, who used VOTAT strategies with lower but still intermediate frequency; and 31.4% were low-level strategy users, who barely made use of VOTAT strategies throughout the assessment process.

Table 7. Fit indices for latent class analyses monitoring students' problem-solving behaviour in uncertain situations

Number of latent classes	AIC	BIC	aBIC	Entropy	L–M–R test	p
2	1504	1585	1506	0.967	613	<.001
3	1446	1570	1449	0.931	83	<.05
4	1440	1607	1445	0.944	31	>.05

Table 8 indicates the problem-solving performance of all three classes of participants (low-level strategy users, intermediate explorers and expert explorers). The results indicate that all three classes of participants performed better on the easier items (the 2-1 and 2-2 types) than on the more complex problems (the 3-3 type). Furthermore, the results confirmed that VOTAT is the most effective strategy. Problem-solvers that used it had a higher chance to solve a problem correctly, with the exception that the intermediate explorers performed slightly worse than the low-level strategy users on the 3-3 problems.

Table 8. Problem-solving performance for low-level strategy users, intermediate explorers and expert explorers

Latent class profiles	Frequency (%)					
	2-2 problems		2-3 problems		3-3 problems	
	High achievement	Low achievement	High achievement	Low achievement	High achievement	Low achievement
Low level strategy users	32.0	68.0	34.4	65.6	33.9	66.1
Intermediate explorers	33.8	66.2	35.3	64.7	33.3	66.7
Expert explorers	45.4	54.6	46.5	53.5	37.7	62.3

6. Discussion

Research Aim 1: To test the applicability of an interactive, innovative third-generation test, such as the CPS test, in a country where computer-based assessment has a short history.

In this study, we used logfile analysis to examine Jordanian undergraduate students' problem-solving behaviour. First, we monitored the feasibility and applicability of computer-based assessment in the Jordanian educational context. The internal consistency of the CPS tests was high, but the mean achievement was relatively low, indicating that it is difficult for the students to solve interactive problems. Based on all the descriptive results, we can conclude that computer-based assessment and innovative online tests are feasible and valid in Jordan at the level and in the context of higher education.

Research Aim 2: To test the structure of the assessed construct (construct validity), that is, to test the underlying dimensionality of CPS measured in the Jordanian educational context, assuming – based on theory and international (mostly European) assessments – a measurement model with two different problem-solving processes (i.e. knowledge application and knowledge acquisition).

The analyses of the structural stability of the measured construct confirmed earlier research results obtained in Europe (e.g. Funke, 2001; Wüstenberg, Greiff, & Funke, 2012) and Asia that CPS is a two-dimensional construct. The processes of the KAC and KAP phases can also be empirically distinguished in the Jordanian context. The bivariate correlation (r=.45) between KAC and KAP was consistent with earlier research results, which varied between r=.14 and r=.94 (Nicolay, Krieger, Stadler, Gobert, & Greiff, 2021). The reason for this wide range of correlation coefficients is the use of different problem-solving approaches and CPS assessments to measure KAC and KAP. Since CPS skills are a key competence for educational success, research results on CPS have important implications for filling the gap between students' ability to acquire and then apply that knowledge in uncertain situations, which has become highly important in the 21st century. CPS serves as a relevant showcase for addressing a crucial existing gap in modern-day education: the gap between students' ability to acquire knowledge and then apply this knowledge in uncertain situations, which is increasingly significant in the 21st century.

Research Aim 3: To discover and describe the type of strategic exploration used by Jordanian university students while solving CPS problems with different characteristics to understand the mechanism underlying successful CPS.

Logfile-based analyses have expanded the scope of previous studies on CPS, especially in the Arabic environment, and enabled us to identify key components of students' problem-solving skills: the way they explore and understand relatively simplified problems and the relationships within the problem. A large number of students showed systematic strategies but failed to solve the problem; that is, the use of a theoretically effective strategy does not always lead to high problem-solving achievement, a finding which confirms research results by de Jong and van Joolingen (1998), who claim that learners often have trouble understanding data. In contrast, we have detected another relatively large number of students who achieved high performance without collecting all the information necessary to be able to solve the problem correctly; that is, they applied a theoretically non-effective exploration strategy. Beyond guessing, it is more difficult to find a clear explanation for this discrepancy in students' problem-solving behaviour. The result is consistent with previous research (e.g. Greiff et al., 2015; Molnár & Csapó, 2018; Vollmeyer et al., 1996) that indicates that high performance is not always in line with the right kind of problem exploration and interpretation. To sum up, the use of a theoretically effective strategy does not always lead to high performance, and, in contrast, high performance does not always indicate the right kind of exploration and interpretation, i.e. the application of the right kind of problem-solving strategy.

Research Aim 4: To detect the relationships between different types of test-taking and problem-solving behaviour and CPS performance to find new ways to assist students in optimizing their cognitive capacity.

The analysis explored Jordanian students' problem-solving behaviours in greater depth, focusing on the type of problem exploration and helping us to understand the reasons behind discrepancies between the high ratio of theoretically right exploration behaviour, i.e. collecting information, and low problem-solving achievement. One possible explanation is that students did not provide the proper meaning for the information obtained during the first phase of the problem-solving process. Molnár and Csapó (2018) have shown that there is an inverse relation between problem complexity and the probability of strong problem-solving performance without the use of an effective problem-solving strategy. It is clear that the achievement for all participants on the easiest item (2-2) was not the best. Students' performance was better on

problems of medium complexity (2-3) because they had sufficient experience after solving the first type of problem. On more complex problems (3-3), students' performance declined, despite having sufficient experience in solving problems. As regards the increasing numbers of input variables, output variables and relations between them, the participants experienced greater difficulty (Beckmann et al., 2017). More analyses are required to detect the reasons for the large differences between the expertise level in the exploration and the lower achievement in the decisions made in problem-solving.

7. Limitations

The study is considered as a small-scale study with 195 participants from two Jordanian universities. Thus, it does not represent the entire university student population in Jordan. Hence, the results from this study are not generalizable. A bigger sample size from more universities and faculties is required to obtain a wider view of Jordanian students' problem-solving behaviour.

In addition, some participants suffered from the weakness of the internet during the test at peak intensity; all the students used the university system at the same time. This caused some difficulty in retaining access, as some sessions required a high-speed connection. Another limitation stems from the translation and adaptation of the items. Originally, the languages of the items were German and Hungarian. Then, both the Hungarian and German versions were translated into English. After validating the Hungarian, German and English versions, the test was translated into Arabic by specialist translators for distribution to the Jordanian students. Beyond translating the problem texts and instructions, we changed the direction of the test to suit the Arabic format, from right to left (the earlier versions of the test were produced in left-to-right format). We also changed tables, boxes, pictures and all the connecting elements.

The MicroDYN approach was used in the study to assess students' problem-solving abilities using an instrument which is valid and reliable for measurement purposes, but uses artificial problems, where the number of variables and relations are limited. Hence, the problem-solving behaviour observed in MicroDYN scenarios cannot be generalized to all types of problems we face in everyday life.

8. Conclusion and implications

The study points to the possibility and feasibility of problem-solving measurements in the Jordanian context. It highlights the importance of explicit development of problem-solving skills and problem-solving strategies as a means of applying knowledge in new contexts in higher education. The findings highlight the importance of developing instructional methods to improve students' CPS skills by enhancing their individual learning strategies. The results also suggest the need for further investigation to explore a larger representation of the relationships between students' cognitive skills and their behaviour in problem-solving situations. To sum up, the study has shed light on Jordanian students' problem-solving development from the perspective of their behaviour, thus providing a solid basis for further study in the Jordanian context.

9. Funding

This research was supported by the Hungarian National Research, Development and Innovation Fund (grant under the OTKA K135727 funding scheme) and the Hungarian Academy of Sciences (Research Programme for Public Education Development of the Hungarian Academy of Sciences, grant KOZOKT2021-16).

References

Al-Khayat, M. (2017). Students and instructors' attitudes toward computerized tests in business faculty at the main campus of Al-Blaqa Applied University. *Journal for Research-B (Humanities), 31*(11), 6.

Alquraan, M. F. (2012). Methods of assessing students' learning in higher education. *Education, Business and Society: Contemporary Middle Eastern* Issues. 5, 124–133

Beckmann, J. F., Birney, D. P., & Goode, N. (2017). Beyond psychometrics: the difference between difficult problem solving and complex problem solving. *Frontiers in psychology*, *8*, 1739.

Christakoudis, C., Androulakis, G. S., & Zagouras, C. (2011). Prepare items for large scale computer based assessment: Case study for teachers' certification on basic computer skills. *Procedia-Social and Behavioral Sciences, 29*, 1189–1198.

Collins, L. M., & Lanza, S. T. (2010). Latent class and latent transition analysis. In With applications in the social, behavioral, and health sciences. New York: Wiley

Csapó, B., & Funke, J. (2017). Epilogue. In B. Csapó and J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp.265–268). Paris: OECD Publishing.

Csapó, B., & Molnár, G. (2017). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Frontiers in Psychology*, 8, 2022. https://doi.org/10.3389/fpsyg.2017.02022

Csapó, B., & Molnár, G. (2017). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Frontiers in Psychology, 8*, 2022. https://doi.org/10.3389/fpsyg.2017.02022

Csapó, B., Lörincz, A., & Molnár, G. (2012). Innovative assessment technologies in educational games designed for young students. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning* (pp. 235–254). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-3546-4_13

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school-readiness and reasoning skills. *Journal of Educational Psychology, 106*(2), 639–650.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research, 68*(2), 179-201.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of experimental psychology: General*, *122*(3), 371.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Dörner, D., & Funke, J. (2017). Complex problem solving: what it is and what it is not. *Frontiers in Psychology*, 8, 1153.

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology*.

Fischer, A., Greiff, S. és Funke, J. (2012): The process of solving complex problems. *Journal of Problem Solving*, 4. 19–42.

Fukuda, K., & Vogel, E. K. (2009). Human variation in overriding attentional capture. *The Journal of Neuroscience*, *29*(27), 8726–8733.

Funke, J. (1991). Solving complex problems: exploration and control of complex systems. In R. J. Sternberg & P.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning*, 7(1), 69–89.

Funke, J. (2010). Complex problem solving: A case for complex cognition?. *Cognitive processing*, *11*(2), 133–142.

Funke, J. (2021). It requires more than intelligence to solve consequential world problems. *Journal of Intelligence*, *9*(3), 38.

Gnaldi, M., Bacci, S., Kunze, T., & Greiff, S. (2020). Students' complex problem solving profiles. *Psychometrika*, 85(2), 469–501.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving-the MicroDYN approach. In F. Scheuermann (Ed.), The transition to computer-based assessment-lessons learned from largescale surveys and implications for testing (pp. 157–163). Luxembourg: Office for Official Publications of the European Communities.

Greiff, S., & Funke, J. (2017). Interactive problem solving: Exploring the potential of minimal complex systems. In B. Csapó and J. Funke (Eds.), The nature of problem solving (pp. 93–105). Paris: OECD Publishing.

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 & Education*, 126, 248–263.

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 & Education*, 91, 92–105.

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

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.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In *Assessment and teaching of 21st century skills* (pp. 1–15). Dordrecht (Netherlands): Springer.

Harsel, Y., & Wales, R. (1987). Directional Preference in Problem Solving. *International Journal of Psychology, 22*(2), 195–206.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Leutner, D., Wirth, J., Klieme, E., & Funke, J. (2005). Ansätze zur Operationalisierung und deren Erprobung im Feldtest zu PISA 2000. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 21–36). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767–778.

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence, 64*, 98–112.

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, in press.

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Logfile analyses. *Frontiers in psychology*, 9, 302.

Molnár, G., & Csapó, B. (2019). How to make learning visible through technology: The eDia-online diagnostic assessment system. In Proceedings of the 11th International Conference on Computer Supported Education, Volume 2, 122–131.

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking skills and Creativity*, *9*, 35–45. https://doi.org/10.1016/j.tsc.2013.03.002

Molnár, G., Greiff, S., Wüstenberg, S. & Fischer, A. (2017). Empirical study of computer based assessment of domain-general dynamic problem solving skills.In B. Csapó and J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 123–143). Paris: OECD Publishing.

Mousa, M., & Molnár, G. (2019). The feasibility of computer-based testing in Palestine among lower primary school students: Assessing mouse skills and inductive reasoning. *Journal of Studies in Education*, *9*(2), Terjedelem-16. https://doi.org/10.5296/jse.v9i2.14517

Mousa, M., & Molnár, G. (2020). Computer-based training in math improves inductive reasoning of 9-to 11-year-old children. *Thinking Skills and Creativity*, *37*, 100687.

Mustafić, M., Yu, J., Stadler, M., Vainikainen, M. P., Bornstein, M. H., Putnick, D. L., & Greiff, S. (2019). Complex problem solving: Profiles and developmental paths revealed via latent transition analysis. *Developmental psychology*, *55*(10), 2090.

Muthén, L. K., & Muthén, B. O. (2010). Mplus user's guide. Los Angeles, CA: Muthén & Muthén.

Newell, A., & Simon, H. A. (1972). Human problem solving (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.

Nicolay, B., Krieger, F., Stadler, M., Gobert, J., & Greiff, S. (2021). Lost in transition–Learning analytics on the transfer from knowledge acquisition to knowledge application in complex problem solving. *Computers in Human Behavior*, *115*, 106594.

Novick, L. R., & Bassok, M. (2005). *Problem Solving*. Cambridge University Press**.**

OECD (2018). The future of education and skills. Education 2030. OECD.

OECD. (2014a). *Results: creative problem solving - students' skills in tackling real-life problems (Vol. V)*. Paris: OECD Publishing.

OECD. (2014b). *PISA 2012: technical report*. Paris: OECD Publishing.

Reeff, J. P., Zabal, A., & Blech, C. (2006). The assessment of problem-solving competencies: A draft version of a general framework. *German Institute for Adult Education (DIE).*

Román, A., El Fathi, A., & Santiago, J. (2013). Spatial biases in understanding descriptions of static scenes: the role of reading and writing direction. *Memory & cognition*, *41*(4), 588–599.

Schnotz, W., Baadte, C., & Müller, A. (2010). Creative thinking and problem solving with depictive and descriptive representations. In Use of representations in reasoning and problem solving (pp. 21-45). Routledge.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2), 257–285.

Vo, D. Csapó, B. (2020). Development of inductive reasoning in students across school grade levels. *Thinking Skills and Creativity*, 37, 100699.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. Cognitive Science, 20(1), 75–100.

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, 21(4), 393–409.

Wu, H., & Molnár, G. (2018). Interactive problem solving: assessment and relations to combinatorial and inductive reasoning. *Journal of Psychological and Educational Research, 26*(1), 90–105.

Wu, H., Molnár, G. (2021). Logfile analyses of successful and unsuccessful strategy use in complex problem-solving: a cross-national comparison study. *Eur J Psychol Educ*. https://doi.org/10.1007/s10212-020-00516-y

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

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, 29, 18–29.

Yang, Y., Majumdar, R., Li, H., Akçapinar, G., Flanagan, B., & Ogata, H. (2021). A framework to foster analysis skill for self-directed activities in data-rich environment. *Research and Practice in Technology Enhanced Learning*, 16(1), 1–17.

5

HOW WE EXPLORE, INTERPRET, AND SOLVE COMPLEX PROBLEMS: A CROSS-NATIONAL STUDY OF PROBLEM-SOLVING PROCESSES

This article available as:

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 (e08775)

Abstract

Complex problem-solving (CPS) is considered an important educational outcome in the 21st century. Despite its importance, we have no knowledge of its measurability, development, or comparability in Arab countries, where there is a short history of computer-based assessment. The results of the current study provide important insights into the international validity of CPS measurements and shed light on the different hidden behavioral patterns and test-taking behavior of Jordanian (N=431) and Hungarian (N=1844) students as they solve complex problems. CPS proved to be measurement-invariant in Jordan and Hungary among university students. Analyzing log data, we have identified large differences in students' test-taking behavior in terms of the effectiveness of their exploration strategy, time-on-task, and number of trials at an international level. Based on the students' exploration strategy behavior, we identified four latent classes in both samples. The tested process indicators proved to be non-invariant over the different latent profiles; that is, there are big differences in the role of the number of manipulations executed, time-on-task, and type of strategy used in actual problem-solving achievement between students that fall within different thinking profiles. This study contributes to an understanding of how students from different educational contexts behave while solving complex problems.

Keywords: International validity Process indicators Test-taking behavior Exploration strategies Latent class analysis Complex problem solving

1. Introduction

As a result of an increasingly interconnected global economy, students today will compete with each other not only on national labor markets but also on international ones. Highly skilled adults are more likely to be employed and have access to better-paying jobs than poorly skilled ones. That is, the aim of national education systems should be to equip students with internationally competitive knowledge and skills. Parallel to this issue has been a change in what is considered valuable knowledge. The role of declarative knowledge has decreased, and the value of the applicability of knowledge and that of new knowledge creation and innovation, that is, the role of thinking skills, have increased. In the 21st century, problem-solving represents one of the most cited, highlighted, and important skills on the labor market.

Complex problem-solving (CPS)[2] is among the most extensively studied areas among the problem-solving skills in educational context over the past few decades (Csapó & Molnár, 2017; Greiff et al., 2013; Greiff, Fischer, Stadler, & Wüstenberg, 2015; Dörner & Funke, 2017). It is a specific form of problem-solving (Funke, 2014), where the problem-solver needs to explore, understand, and control problem environments that are unknown, non-transparent in nature, and consisting of a number of interconnected elements (Buchner, 1995; Dörner, 1986; Funke, 2001; Wüstenberg, Greiff, & Funke, 2012). CPS tasks focus on domain-general processes and disregard the role of content knowledge and rote learning (see Funke, 2001; Funke & Frensch, 2007; Greiff, Wüstenberg, & Funke, 2012). When solving complex problems, the problem-solver is more effective when relying on abstract representation schemata by understanding the structure of the problems rather than based on specifically relevant school knowledge or example problems (see Holyoak, 1985; Klahr, Triona, & Williams, 2007).

CPS enables us to study, first, how knowledge is gathered in a new problem situation (i.e., knowledge acquisition) and, second, how this knowledge is applied to actually solve a problem (i.e., knowledge application), independently of domain-specific content (Greiff, Holt, & Funke, 2013; Moln'ar et al., 2013). By its nature, CPS is considered an important educational outcome in the 21st century. Since it strongly predicts educational achievement (Greiff et al., 2012; Schweizer, Wüstenberg, & Greiff, 2013), it has become essential to understand the fine

[1]As concerns terminology, please note that there are different labels for the subject under investigation in the literature (e.g., complex problem-solving, dynamic problem-solving, creative problem-solving, and interactive problem-solving, see Csapó & Funke, 2017). In the present paper, we use the most widely used modifier, complex.

mechanisms of CPS, especially to understand the reasons students' behavior lags behind the differing CPS performance to be able to design effective educational programs to improve it.

The enormous development and spread of computer-based assessment and analytical techniques (e.g., developments in structural equation modelling, in pattern finding techniques, and in process and logfile analyses) have made it possible to learn a great deal about the processes and specific features related to CPS in the last few years. In fact, a number of studies have confirmed the international usability of tests measuring CPS (Greiff et al., 2015; Molnár & Csapó, 2018; Mustafic et al., 2019; OECD, 2014a; Wu & Molnár, 2021).

In 2012, CPS was also assessed as a core marker of educational achievement in one of the most prominent international large-scale assessments, OECD PISA (OECD, 2014a), where 15-year-old students from 40 countries took part in the CPS data collection. Based on the PISA results (OECD, 2014a), we have some knowledge of how cultures differ in their problem-solving performance, but we know little about how underlying processes may differ. Güss et al. (2010) analyzed CPS processes based on cultural-psychological theories by investigating think-aloud protocols in five countries (Brazil, Germany, India, the Philippines, and the United States). Their results showed cross-national differences in all CPS steps, including knowledge acquisition and knowledge application.

Despite the relatively great attention paid to CPS (see Schoppek et al., 2018), we have no knowledge of its measurability, development, and comparability issues in countries in the Arab region, especially in Jordanian communities, where computer-based assessment has less of a history. Beyond the lower prevalence of computer-based tests, spatial biases of the Arabic language – which runs from right to left and not from left to right like European languages – can influence human behavior and can cause biases at both low-level perceptuo-motor skills and high-level conceptual representations (Román et al., 2015). Language has the potential to influence cognitive processes as it may direct attention to conceptual representations and distinctions that are encoded in a given language over others (Gleitman & Papafragou, 2005; Landau et al., 2010); it is a type of tool that influences human representational resources (Ünal & Papafragou, 2018). Thus, language-related factors can cause invariance in solving problems. In addition, cultural mindset can also influence problem-solving (Arieli & Sagiv, 2018). Members of individualist cultures (like the Hungarian culture; see Holicza, 2016) perform better on rule-based problems, whereas members of collectivist cultures (such as the Jordanian culture; Ourfali, 2015) solve context-based problems more easily (Arieli & Sagiv, 2018). For

members of individualist cultures, the task is more important than personal relationships (Al Suwaidi, 2008), while in-group goals take priority over personal goals in collectivist cultures (Schwartz & Bilsky, 1990), where group performance is more important than individual task performance (Hofstede & Hofstede, 2005) and consultative decision-making is preferred over autonomous decisions (Al Suwaidi, 2008). However, please note that both individualist and collectivist countries vary along a continuum of individualism and collectivism (Al Suwaidi, 2008). Educational national traditions and practices may also influence these issues. For example, according to Alkailani (2012), most university students in Jordan live with their families and enjoy full social and financial support from them; that is, they do not need to solve problems or make autonomous decisions, since they depend on their families and their decisions. All these cultural, national, linguistic, and educational factors may have a powerful effect on students' reasoning skills, resulting in large differences in performance among students. As an example of the continuum noted above, compare teaching methods and educational success based on international large-scale assessments of two largely collectivist countries: China and Jordan.

To sum up, we accept Triandis's (1994) argument about culture and the role of cultural differences in human behavior: "culture is to be seen as a web of significances that direct, guide and shape human action" (Triandis, 1994). Indeed, it is a complex phenomenon. Along these lines, it is important to study CPS processes in countries that fall within different cultures as outlined above – a topic that has so far been neglected in current research.

In the present paper, we address this shortcoming and analyze behavioral and overall performance data in CPS across two different countries that fall within different cultures: the Jordanian and Hungarian cultures. Specifically, after adapting the CPS problems to both languages, we analyze the measurement invariance of one of the most commonly-used CPS measures (i.e., MicroDYN) across Jordanian and Hungarian students in Research Question 1. Subsequently, we investigate the nature of the developmental differences in three steps. First, we focus on the concrete answer data of the students, using the traditional method for scoring the problems (Research Question 2). Second, we go deeper to reconstruct what high- and low-achieving students did during the problem-solving process, that is, how motivated they were, e.g., how much time they spent on the problems and how much effort they showed during the test administration (number of clicks) (Research Question 3). Third, using logfiles and a behavior pattern-finding algorithm, we identify different problem-solving profiles in both

countries and compare students' behavioral features based on their class profiles and final scores (Research Question 4).

2. The present study

CPS has been extensively assessed on international large-scale assessments (see PISA, 2012). However, as an international option, not all countries that participated in the 2012 PISA cycle also participated in the assessment of problem-solving, and only a few countries from the Middle East did so. For instance, Jordan did not. Thus, the present study is likely to be the first to report on CPS among Jordanian students. Of note, despite the widespread use of CPS in international samples, less attention has been paid to its measurement invariance across different nations and cultures. On PISA, according to the general procedures, "items were singled out whenever they showed differential item functioning in the Field Trial" (OECD, 2014b, p. 98). According to the PISA technical report, no measurement invariance was tested in the structural equation modeling framework. Wüstenberg et al. (2014) investigated and showed measurement invariance of CPS between Hungarian and German 8^{th}–11^{th}-grade students. Wu and Molnár (2021) analyzed measurement invariance of CPS among Hungarian and Chinese 12-year-old students. Their results indicated that the measurement of CPS across these two cultures was not measurement-invariant. This indicated that cultural and educational differences can indeed influence the measurement of CPS. That is, before looking at substantial differences, for instance, in Hungarian and Jordanian students, it is important to examine measurement equivalence across cultures. Thus, the first research question in this study asks whether we can measure CPS equally in both countries.

Research Question 1 (RQ1): Do Jordanian and Hungarian students interpret CPS problems the same way? Is CPS measurement-invariant across Jordanian and Hungarian university students?

Several studies have shown that students with different educational and cultural backgrounds perform differently in CPS environments (Greiff, Wüstenberg, & Avvisati, 2015; OECD, 2014a; Wu & Molnár, 2021; Wüstenberg et al., 2014). However, this picture is incomplete and limited to certain geographical areas and countries as of now. There is little research about levels of CPS achievement, even within a traditional performance-oriented approach, among students from the Middle East. Further, based on the international research results on

developmental changes in students' exploration strategies and test-taking behavior in a CPS environment, which will be the focus of the third research question, most of the information is based on international comparison studies, where, beyond students from different European or Asian countries, students from Hungary form the bases of comparison (see Csapó & Molnár, 2017; Greiff et al., 2013, 2018; Molnár & Csapó, 2017; Wu & Molnár, 2021; Wüstenberg et al., 2014). That is, research results based on data from Hungarian students (as a common aspect of earlier analyses) could act as a benchmark in comparison studies involving students' developmental differences (from a more traditional performance-oriented approach, RQ2) and behavioral differences (from a more innovative behavioral pattern-oriented approach, RQ3) in a CPS environment at an international level. Built on the knowledge acquired from answering RQ1, that is, assuming that CPS can be measured equivalently in the two countries, the following research question has been formed on the second issue from a more traditional perspective:

Research Question 2 (RQ2): Can developmental differences be identified in CPS skills between Jordanian and Hungarian university students? If so, what is the nature of these developmental differences?

Having established that we can measure CPS equivalently across the two countries (in RQ1) and that students in our two subsamples from Hungary and Jordan differ both in their overall level and their development (RQ2), we want to better understand these differences by actually looking at underlying behaviors. To this end, technology-based assessment offers an opportunity to collect contextual information gathered beyond the final response data and analyze different behavioral indicators, such as strategy effectiveness, number of trials, and time-on-task. Logfile analysis has the potential to look at developmental differences from different perspectives (Nicolay et al., 2021) and to provide more sophisticated feedback instead of using single indicators, such as an overall test score.

In addition, according to earlier research results (see Greiff et al., 2018; Moln'ar & Csapo', 2018), final scores may conceal true developmental and behavioral differences as regards CPS. For example, students with average achievement can engage in three completely different behavior patterns: (a) they can be average achievers on all of the problems; (b) they can be high achievers on the easiest problems but low achievers on the hardest ones; and (c) they can be

grouped as rapid learners, that is, learners with low achievement at the beginning of the test but, as a result of a rapid learning effect, high achievers on the most difficult problems at the end. The interactivity of the CPS problems offers opportunities to analyze, describe, and cluster the behavior of the students during the test and thus to understand the patterns that lead to final CPS performance scores.

Molnár and Csapó (2018) investigated the relation between (1) theoretical strategy effectiveness, which was linked to the amount of information extracted from the problem environment and empirical effectiveness, and (2) ultimate CPS achievement in 3rd–12th-grade students. Results showed that the use of a theoretically effective strategy does not necessary result in high performance.

Goldhammer et al. (2014) studied the link between number of interactions and problem-solving achievement in technology-rich environments, which "assumes two concepts, accessing information and making use of it, that seem similar to knowledge acquisition and application" (Goldhammer et al., 2014, p. 7). Results showed that low-achieving students typically engage in fewer interactions with problems that require controlled processing. Other studies have confirmed the positive correlation between CPS achievement with number of clicks and amount of exploration (see Eichmann, Goldhammer, Greiff, Brandhuber, & Naumann, 2020).

Research findings referring to time-on-task as regards CPS are more heterogeneous. According to Greiff et al. (2016), spending too much time on CPS problems was associated with poor performance. Authors claimed that there was an optimal time frame for working on CPS tasks. In contrast, Alzoubi et al. (2013) argued that spending more time on CPS problems resulted in significantly higher achievement; that is, more time allows for longer planning and better planned solutions. This finding was, by and large, confirmed by Eichmann et al. (2019). They argued that, especially at the beginning of the CPS process, more planning has a positive impact on final achievement. According to Goldhammer et al. (2014), time-on-task correlated positively with item difficulty and more time was helpful for compensating for the lack of problem-solving ability.

That is, to understand the reasons behind overall achievement differences in CPS between groups (here Hungarian and Jordanian students; see RQ2), we analyze students' test-taking behavior in solving CPS problems with the aim of answering the following research questions:

Research Question 3 (RQ3): What kind of test-taking behavior do Jordanian and Hungarian university students exhibit when solving complex problems? Are there differences between them in the theoretical effectiveness of their strategy use, their time-on-task, and the number of trials they use?

In RQ2 and RQ3, we explored quantitative differences between the two samples in general; that is, we looked for differences in the Hungarian and Jordanian students' test-taking behavior: final score (empirical effectiveness), amount of information extracted (theoretical effectiveness), number of trials, and time-on-task. In RQ4, we expand on this perspective by highlighting different types of problem-solvers with a person-centered approach. We thus explore whether there are also qualitatively different problem-solvers in the two groups under examination. That is, we want to investigate whether there are different profiles and whether there are other CPS-related differences between the two countries. We will thus focus on the exploration of student-level problem-solving behaviors and investigate whether there are different types of problem-solvers and how these compare in the two groups (Tóth, Rölke, Goldhammer, & Barkow, 2017).

As input variables for this person-centered approach, the vary-one-thing-at-a-time (VOTAT) strategy was chosen because it has received the most attention in CPS research as a process indicator and has been shown to be one of the most relevant indicators of high CPS achievement. Its effectiveness in connection with high CPS achievement has frequently been discussed (e.g., Eichman et al., 2020; Greiff et al., 2018; Greiff, Wüstenberg, & Avvisati, 2015; Moln'ar and Csapó, 2018; Mustafic et al., 2019; Stadler et al., 2020; Wu & Molnár, 2021). According to the definition of the VOTAT strategy, students systematically vary only one input variable while keeping the others unchanged. This is akin to the principle of isolated variation. We used the extent to which this special strategy was employed in the exploration phase and conducted a latent class analysis in a person-centered approach to see whether there are qualitative differences in the classes across the two samples.

There are a few studies that have examined different classes of Hungarian students, but only one so far has compared different countries. Greiff et al. (2018) analyzed 6^{th}–8^{th}-grade Hungarian students' exploration strategy class profiles in CPS environments. Molnár and Csapó (2018) examined 3rd–12th-grade (aged 9–18) Hungarian students' problem-solving behavior to distinguish qualitatively different exploration strategies. At the university level, Molnár (2021) identified four latent class profiles in Hungarian students: (1) proficient explorers; (2) almost high performers on the easiest problems but low performers on the

complex ones with a slow learning effect; (3) rapid learners; and (4) low to intermediate performers on the easiest problems but non-performers on the complex ones with a slow learning effect. With regard to groups from different countries, Wu and Molnár (2021) compared Hungarian and Chinese 6th-graders' (twelve-year-old students) exploration profiles in a CPS environment. They identified three qualitatively different class profiles with remarkable differences in both the Chinese and Hungarian samples: for example, the class of "low performers" did not exist in the Chinese sample, and the proportion of proficient explorers was significantly higher in the Chinese sample than in the Hungarian one.

To sum up, students' behavior on the CPS tasks separately not only predicts their problem-level achievement but might also be an indicator of their general test-taking behavior and a predictor of their overall CPS performance. Therefore, as a validation strategy for the qualitatively different classes identified, we investigate the relation between students' class membership on the basis of the VOTAT strategy in connection with their behavior and their overall CPS performance. We will thus answer the following research question:

Research Question 4(RQ4): Based on the exploration strategy (i.e., VOTAT), which profiles can be extracted from the Jordanian and Hungarian students? Are there differences in the types of profiles that emerge from the two groups?

3. Methods

3.1. Participants

The participants in the Jordanian sample were studying in different years at two large Jordanian universities. One of the universities has 15 schools (students from five of the schools took part in the assessment: Arts, Economics and Administrative Sciences, Shari'a and Islamic Studies, Education, and Information Technology and Computer Science). The other one has 13 schools (students from four of the schools participated in the study: the School of Information Technology, School of Arts and Humanities, School of Science, and School of Educational Sciences). After cleaning the data, that is, deleting all the students (less than 5%) from the sample who had completed less than half of the test, the sample consisted of 431 students (mean age=20.6; SD=3.11), with 53.4% of them being female. Students' participation was voluntary; as an incentive, they earned credit for successful completion of the test.

Participants in the Hungarian sample were commencing their studies at one of the largest and highest-ranked Hungarian universities. The university has twelve schools (e.g., Humanities and Social Sciences, Science, Medicine, Law, and Economics), all of which were involved in the assessment. A total of 1844 students, that is, 44.8% of the target population, participated in the study (mean age=19.9; SD=1.82), with 59.8% of them being female. After data cleaning, that is, deleting all the students from the sample who had completed less than half of the test, 1828 students remained in the sample (less than 1% omitted). Students' participation was voluntary; as an incentive, they earned one credit for successful completion of the tests.

Some differences were noted in the background data for the two samples. In Hungary, only first-semester students took part in the assessment, whereas there were also students in higher semesters in Jordan. Nontheless, there was no significant difference in the mean age of the participating students. In Jordan, in terms of proportions, 6% more male students were part of the study than in Hungary. The study goal of the students was the same in both countries. Parents' educational level, number of books, and available ICT infrastructure at home proved to be higher in Hungary than in Jordan (see Table 1). These differences need to be taken into account when interpreting the results.

Table 1.

Comparison of the Jordanian and Hungarian samples along the same variables

Demographic data	Jordan			Hungary		t test/ Welch test
	Mean	SD		Mean	SD	
Age	20.6	3.11	=	19.99	1.82	n.s.
Gender (1: male; 2: female)	1.53	.50	<	1.59	.49	t=-2.5 p<.05
Year of Matura examination	2015	5	<	2019	1.7	t =-+6.9 p<.01
Average result of Matura examination – compulsory parts*	85.81	9.04		76.22	15.03	not comparable
Study goal**	1.67	1.17	=	1.66	1.03	n.s.
Parental education***	4.26	2.47	<	5.44	1.26	t=-8.5 p<.01
Number of books****	2.94	1.98	<	4.41	1.68	t=-12.8 p<.01
ICT infrastructure*****	3.16	1.3	<	4.19	0.967	t=17.11 p<.001

Note. *The compulsory subjects are different in the two countries. In Jordan, they are Arabic, English, History of Jordan, and Islamic Education. In Hungary, they are Hungarian, Mathematics, History, and a foreign language.

**Study goal is measured on a 3-point scale: 1: BA; 2: MA; 3: PhD – the level of education he or she ultimately wishes to complete.

***Parental education was measured on a 7-point scale: 1: below primary…7: university degree equivalent to MA or MSc.

****Number of books: 7-point scale: 1: less than 1 bookshelf…7: more than 1000 books.

***** ICT infrastructure at home: 1: none at all….5: a great deal

3.2. Instrument

A complex problem-solving test based on the MicroDYN approach was administered in both countries. The tests consisted of the same complex problems (ten problems) with increasing item complexity (number of input and output variables and number of relations) and fictitious cover stories. At the beginning of the test, participants were provided with the same instructions on engaging with the user interface, including the same warm-up task. MicroDYN is designed to allow students to acquire the general exploring skill through problem-solving with a limited number of variables and relations, while in most cases nothing changes in the problem scenario if the participant has not changed any variables. Thus, the test is designed so that students can

learn during the test-taking process as previous problems and previous problem-solving processes can influence subsequent problem-solving in the MicroDYN task. Because of these special features of MicroDYN problems and tests, the learning process can be explored and quantified, thus providing the possibility to measure the learning potential of the students occurring in the problems and during the test-taking procedure.

From the perspective of the traditional psychometric approach, each problem consisted of two phases: knowledge acquisition (first phase) and knowledge application (second phase), which were scored separately. Consequently, each problem consisted of two scoreable items.

In the first phase of the problem-solving process, the free exploration phase, the relations between the input and output variables needed to be explored by interacting with the problem environment. During this interaction process, students were expected to manipulate the values of the input variables (Greiff & Funke, 2010) as many times as they liked within 180 seconds and to identify the resultant changes in the output variables (direct effects) to acquire new knowledge (Fischer et al., 2012). The test contained tasks where output variables could have changed not only as an effect of manipulation of the input variables but also spontaneously, with internal dynamics (eigendynamic; Greiff et al., 2013). Independent of the type of effects and relations, it was possible to detect the structure of the problems with an adequate problem-solving strategy (Greiff et al., 2012) and with an adequate, systematic manipulation strategy. To do this, test-takers were expected to click on a button with a + or – sign or by using a slider linked to the respective input variable (See Figure 1) and press the Application button, which made it possible to test the effect of the set values of the input variables on the output variables, which was defined as a trial. The effect in terms of the changes in the values of the output variables was presented on a graph next to each output variable, similarly to the history of the earlier settings of the input variables within the same scenario, which was also presented on a graph next to each input variable. According to the user interface settings, within the same phase of each problem, the input values remained at the same level until the Reset button was pressed or they were changed manually. The Reset button set the system back to its original status, that is, the values of the input and output variables were reset to zero, and the history of the earlier settings and effects disappeared from the graphs. In the present paper, we have labeled and analyzed these strategies using the log data collected during the exploration phase of the problem-solving process. During this 180 seconds in the first phase of the problem-solving process, they were expected to draw the relations they noticed in the form of arrows between the variables presented on the concept map under the MicroDYN scenario on screen.

This first phase of the problem-solving process, including the free exploration and the model building process, is often called the knowledge acquisition phase (see Greiff et al., 2013).

Figure 1. Screenshot of the MicroDYN task "Game Night." See the original version of the task in Greif et al. (2011). The controllers of the input variables range from "- -" (value=-2) to "++" (value=+2). They are presented on the left side of the problem environment in the Hungarian-language version and on the right side in the Arabic one. The model is shown at the bottom of the figure. (The English-language version is presented in Figure 2.)

In the second part of each of the problems, in what is called the knowledge application phase (Greiff et al., 2013), students were expected to reach the given target values of the output variables within a given time frame (90 seconds), at most in four clicks of the Application button. In this phase the right concept map was presented to the students on screen to make the different parts of the problem-solving process as independent as possible. Finally, students were able to navigate between the different phases within the same MicroDYN scenario and between the different MicroDYN scenarios using the Next button (there was no Back button available on the test).

The language of the problems differed in the two samples. In Hungary, the language of the instructions was Hungarian, whose writing proceeds from left to right, while in Jordan it was Arabic, whose writing goes from right to left. Figure .1 shows a sample item from the Hungarian and Arabic versions of the complex problem-solving test. The translation was conducted in the following way. The German and Hungarian versions of the instructions were independently translated into English. The two English versions were compared and discussed.

The final English version was translated into Arabic by two independent translators. The two Arabic versions were then compared and discussed. Sentences which were subject to different interpretations were further discussed among the Arabic and Hungarian researchers and translators.

The CPS approach and the CPS tasks have been employed extensively at both the national and international levels (see Csapó & Funke, 2017; Eichmann et al., 2020; Greiff et al., 2013; Greiff, Fischer, Stadler, & Wüstenberg, 2015; Mustafic et al., 2019; Nicolay et al., 2021; OECD, 2014a). The psychometric indices of the test proved to be good, independent of the cultures and nations (see Wüstenberg et al., 2014; Wu & Molnár, 2021).

3.3. Procedures

3.3.1. Data collection in Jordan.

Because of the COVID-19 situation, the Jordanian assessment could not be administered in a monitored environment in the university buildings. Students received the password and link to the complex problem-solving test and were asked to complete the test at home. Consistent with Schultz et al.'s (2017) results, individual online testing of complex problem-solving favored the Jordanian sample over the Hungarian one. Like the Hungarian version, the Jordanian testing time was limited. The tests and questionnaire were administered using the eDia online platform (Csapó & Molnár, 2019). After entering the eDia system, students had 60 minutes to solve the problems and complete the related questionnaire. They received immediate feedback on their average achievement after test completion.

3.3.2. Data collection in Hungary.

The Hungarian assessment was carried out in a large computer room at the university learning and information center using several security protocols due to COVID-19 (e.g., every other computer was switched off, use of face masks and hand sanitizer was compulsory, and all the keyboards and mice were disinfected during the breaks). The assessment was carried out in the first four weeks of the semester, when the university was engaged in hybrid education. The tests and questionnaire were administered using the eDia online platform as was the case in the Jordanian data collection. The testing time was limited; students had 60 minutes to complete the test and the related questionnaire. They received immediate feedback on their average achievement after test completion and detailed feedback with normative comparative data on their performance a week later.

To sum up, there are similarities and differences between the two data collections. The later one may cause some limitations in the research result as the samples cannot be directly compared. The Hungarian sample is likely to be more positively selected than the Jordanian one, and we would therefore already expect that the Hungarian students outperformed their Jordanian peers. Similarities are the following: introduction, problems, test, test platform, immediate feedback, credit for completion, university students, age, same period of the year, and large universities. Differences are language, gender distribution, and supervision or no supervision during data collection.

3.3.3. Scoring.

Achievement was scored the same way in both countries. Performance on each problem in both the first and second phases of the problem-solving process (i.e., the knowledge acquisition and knowledge application phases) was scored dichotomously, that is, as either right or wrong. For the knowledge acquisition phase, a set of fully correct arrows on the concept map, that is, the completely matching problem structure, was assigned a score of 1; otherwise, the response was incorrect, and students received 0 points. For the knowledge application phase, the answers were marked as correct ("1") if students managed to reach the given target values of the output variables in no more than four clicks of the Application button and within the given time frame; otherwise, the answer was marked as incorrect ("0"). That is, each student received two scores on each of the ten MicroDYN tasks, one for knowledge acquisition and one for knowledge application.

3.3.4. Labeling and scoring the log data.

We scored the manipulation behavior of the students in the first phase of the problem-solving process (i.e., the knowledge acquisition phase) based on the collected logfiles. In order to map and describe the students' manipulation strategy, we used a labeling procedure developed by Molnár and Csapó (2018), which is applicable to problems based on minimal complex systems, such as MicroDYN problems. The unit of this labeling process was a setting of the input variables (a trial), which was executed by clicking on the Application button. For example, Figure 2. demonstrates four trials. In the first trial, the value of a variable called blue gambling chips was set to 1, and the two remaining variables were kept at a neutral level, zero. In the second trial, the first input variable was reset to zero, and the second (green gambling chips) was set to one. The third one was kept at zero. In the third trial, the effect of the third input variable was tested by setting the values of the third variable to one and keeping the first two

in their earlier status (zero and one). In the last trial, once again, only the value of a single variable (the first one) was changed, and the other two retained their earlier status (one), resulting in a trial where all of the values for the input variable were set to one.

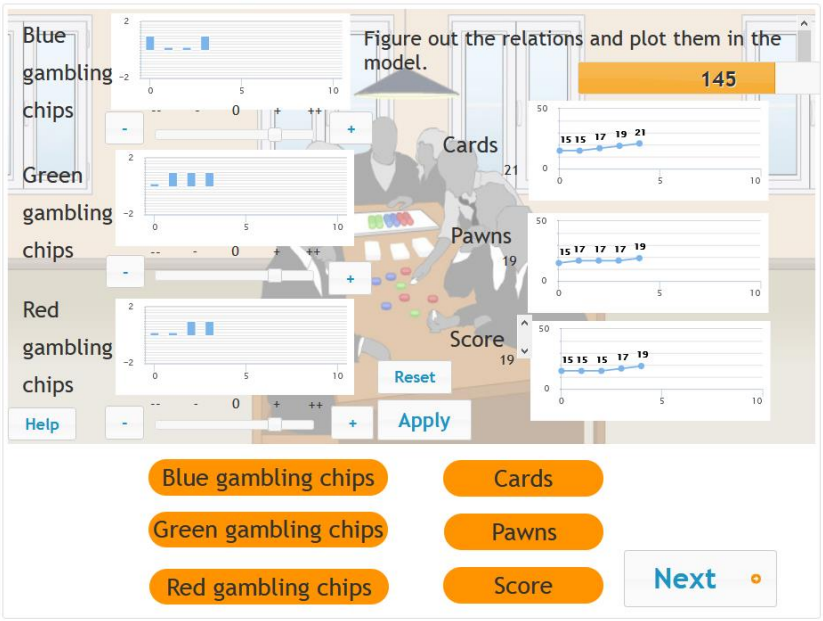

Figure 2. Demonstrating the meaning of a trial within the "Game Night" problem (English-language version of the task presented in Figure 1). The instruction for the task: Your friends invite you to a game night. They show you an interesting game you do not know the rules to. Find out how the blue, green, and red gambling chips affect the number of cards, the number of pawns, and your score.

The sum of these trials within the same problem environment (i.e., within each MicroDYN problem) describes students' manipulation behavior in its entirety. In the present paper, we analyzed and scored the log data from two different perspectives: (1) theoretical effectiveness or strategic effectiveness: if the manipulations of the problem-solver provided all the relevant information on the relations that can be identified, the manipulation was called a theoretically effective strategy and was assigned a code of 1 (that is, the information generated across trials within the knowledge acquisition phase of one problem was complete in the sense that all the relevant information was generated); otherwise, the manipulation was ineffective and students received 0 points. (2) As regards the type of manipulation strategy, we used an extra three-category scoring procedure based on the level of optimal exploration strategy use for each of

the CPS tasks (i.e., use of the VOTAT strategy). According to Fischer et al. (2012), the VOTAT strategy is one of the most effective strategies for identifying causal relations between variables. In applying the VOTAT strategy, the problem-solver systematically varies only one input variable, while the others remain unchanged. One of the most obvious and systematic VOTAT strategies is when only one input variable is different from the neutral level in all the trials and all the other input variables are systematically maintained at the neutral level (the isolated variation strategy; Müller et al., 2013). The following three categories have been defined: (a) no isolated variation at all: when no isolated variation was employed for the input variables – scoring 0 points; (b) partially isolated variation: when isolated variation was employed for some but not all of the input variables – scoring 1 point; and (c) fully isolated variation: when isolated variation was employed for all of the input variables – scoring 2 points.

In the example presented in Figure 2., the manipulation strategy was theoretically successful in that students generated all the information on the relations of the input and output. In the first two trials, the effect of the first and second input variables was tested separately, keeping the values of all the remaining input variables at zero. In the third trial, the test-taker was expected to keep the result of the second trial in mind – the second input variable has an effect on the first output variable – because the value of the second input variable was not set to zero but kept at the earlier level; however, the value of the third input variable was changed. That is, the resulting change in the output variables was not only caused by the third input variable but also by the effect of the second input variable. If the students took care of this, during the third trial they were able to learn about the effect of the third input variable on the output variables. As the problem did not involve internal dynamics, it was appropriate to test the manipulation strategy described here to ascertain the effect of the input variables on the output variables separately; that is, students generate all the relevant information needed to solve the problem properly. As regards the type of exploration strategy used and presented in Figure 6.2, all of the manipulations are part of the VOTAT strategy; however, only the first two trials are part of the isolated variation strategy, while the third and the fourth trials are partially isolated variations.

Beyond scoring performance in the two phases and the two strategy scores (i.e., strategic effectiveness and level of isolated variation), additional log data were analyzed, including time-on-task and number of trials. That is, CPS knowledge acquisition (traditional scoring), CPS knowledge application (traditional scoring), effective strategy use (logfile-based), isolated variation strategy use (logfile-based), time-on-task (logfile-based), and number of trials

(logfile-based), i.e., six variables in total, were used for each CPS problem in the analyses. Given that ten problems were presented, each student was scored on 60 variables overall.

3.3. Analyses.

Multi-group confirmatory factor analysis was used to test measurement invariance between the two countries (RQ1). Weighted least squares, mean- and variance-adjusted (WLSMV) estimation, and THETA parameterization were employed in the analyses (Muthén & Muthén, 2012). χ^2 values, an absolute fit index (the root mean square error of approximation, RMSEA), and two incremental fit indices (the Tucker–Lewis Index, TLI, and the comparative fit index, CFI) were computed to evaluate model fit. According to Byrne and Stewart (2006), a series of hierarchical models with increasing restrictions on model parameters were estimated. According to them, measurement invariance is met if model restrictions do not generate a substantially worse model fit in comparison to the unrestricted model or with a stricter traditional approach and the special χ^2 difference test does not indicate significant differences in model fit. In this paper, because of the large difference in sample size (see Chen, 2007; Kaplan & George, 1995; Yoon & Lai, 2017), we evaluated measurement invariance from a practical perspective (see Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008; Putnick & Bornstein, 2016; Rutkowski & Svetina, 2014). We used the following criteria based on Chen (2007) and Cheung and Rensvold (2002): measurement invariance obtains if the difference for ΔCFI is smaller than -.01 and that for ΔRMSEA is smaller than .01.

To find developmental differences between our two samples of Jordanian and Hungarian students (RQ 2, assuming that measurement invariance holds; cf. RQ1), we used standard statistical procedures, such as the independent t-test and effect size (Cohen's d), to compare traditional mean CPS performance scores between the two groups of students. Measurement invariance obtained between the groups; that is, latent mean differences could be interpreted as true differences in the measured construct and were not due to psychometric issues (while keeping in mind that there are limitations in the comparability of the two samples). A latent mean comparison was conducted by constraining thresholds and factor loadings so that they were equal in both groups. The factor intercepts for the Hungarian group were set to zero so it could serve as a reference group during the analyses, and the latent means of the Jordanian group were freely estimated (Ingles et al., 2011).

In the analyses for RQ1 and RQ2, we only used data collected on the overall CPS performance scores in knowledge acquisition and knowledge application. After examining these overall CPS

performance differences in both countries, in RQ3 we looked more deeply into the behavior patterns and continued the comparative analyses at the logfile level, focusing not only on students' final scores but also on their test-taking behavior. That is, in answering RQ3, we involved process data in the analyses to find what was happening "behind the scenes," that is, which behavioral procedures could have led to the overall CPS performance differences between the Jordanian and Hungarian students. More specifically, standard statistical procedures (similar to RQ2) were used to find the mean differences in theoretical strategy effectiveness, time-on-task, and number of trials between the Hungarian and Jordanian students.

In RQ4, we employed a person-centered approach in terms of a latent profile analysis (Collins & Lanza, 2010; Tein, Coxe, & Cham, 2013). We searched for patterns on how the VOTAT strategy, more particularly, fully isolated or partially isolated variation, developed across tasks among the Hungarian and Jordanian students separately, especially learning patterns across one testing session composed of different tasks. The Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted Bayesian information criterion (aBIC), entropy, and the Lo–Mendell–Rubin adjusted likelihood ratio were used to approximate and determine an adequate number of classes in the LCA models. In addition, the average latent class probabilities (ALCP) indicated the most likely latent class membership for every student. Once the most likely class membership for a student was decided, we looked at mean differences in theoretical strategy effectiveness (as regards the amount of extracted information), in CPS knowledge acquisition (traditional scoring), in CPS knowledge application (traditional scoring), in time-on-task, and in number of trials between students in different latent classes in both Hungary and Jordan. Standard statistical procedures such as ANOVA were used for these comparisons.

4. Results

4.1. Reliabilities

The reliability of the CPS problems as a measure of knowledge acquisition and knowledge application, the traditional CPS indicators for phases 1 and 2, was good in both countries (Jordan: $\alpha_{\_ph1}$=.842, $\alpha_{\_ph2}$=.719; Hungary: $\alpha_{\_ph1}$=.858, $\alpha_{\_ph2}$=.750; see Table 2). After we labeled the students' behavior in the exploration phase of the problem-solving process at the beginning of the problem-solving process and used the new dichotomous variables as indicators to describe the effectiveness of strategy for each task and person, the overall

reliability of the test scores improved in both cases (α=.921 and 944, respectively; see Table 2). The reliability of the test improved further by using the categorically scored variables to describe the level of isolated variation strategy use (α=.950 and 946, respectively; see Table 2). That is, the data proved to be reliable at both the test and phase levels, independently of the educational context. The results of the analyses can thus be generalized. Please note that at this point the different scores do not all measure the same phenomenon. In fact, at a conceptual level, all five scores measure something different, which is the reason we do not compare them directly.

Table 2.

Reliabilities of the CPS test in the Arabic and Hungarian-language contexts with and without the use of log data

Type of data	Arabic	Hungarian
Reliabilities of the test with traditional scoring (knowledge acquisition phase)	.842	.858
Reliabilities of the test with traditional scoring (knowledge application phase)	.719	.750
Reliabilities of the test with traditional scoring (phases 1 and 2)	.872	.882
Reliabilities of the test (knowledge acquisition phase) consisting of the new dichotomously scored variables in terms of the effectiveness of strategy usage at the beginning of the problem-solving process (ten items)	.921	.944
Reliabilities of the test (knowledge acquisition phase) consisting of the new categorically scored variables describing the level of isolated variation strategy usage (ten items)	.950	.946

4.2. Results for Research Question 1 (RQ1): Do Jordanian and Hungarian students interpret CPS problems the same way? Is CPS measurement invariant across Jordanian and Hungarian university students?

To tackle RQ1 we investigated measurement invariance across the Jordanian and Hungarian students. The baseline model with the two latent CPS factors (knowledge acquisition and knowledge application) fitted the data well in both countries (Jordan: χ^2=369.02, df=186, CFI=.975, TLI=.972, RMSEA=.044; Hungary: χ^2=865.53, df=186, CFI=.980, TLI=.978, RMSEA=.045). As can be seen in the results below, CPS can be measured invariantly across nationalities in Jordan and Hungary in our sample (see Table 3). Because of the large differences in sample size, we evaluated measurement invariance by looking at CFI and

RMSEA differences (instead of the stricter χ^2 differences). We accepted less than .01 for ΔCFI and no more than .01 for ΔRMSEA, that is, a less than .01 drop in fit indices between the nested models that meet stricter and more stringent conditions of equivalence. In other words, students with identical scores on the latent level can be expected to have the same chance of scoring on the observed measure regardless of the nation to which they belong (Millsap, 2012); that is, the measure is not biased against either of the groups.

Table 3.

Goodness of fit indices for testing invariance of CPS across nationalities

Model	χ^2	df	CFI	TLI	RMSEA	ΔCFI	ΔRMSEA
Configural invariance	944.33	334	.979	.982	.040	-	-
Strong factorial invariance	1062.87	350	.978	.977	.042	.001	.002
Strict factorial invariance	1205.66	370	.975	.974	.045	.003	.003

4.3. Results for Research Question 2 (RQ2): Can we find developmental differences in CPS skills between Jordanian and Hungarian university students? If so, what is the nature of these developmental differences?

Table 4 summarizes the mean and standard deviation of the CPS performance scores in both phases for problems with different levels of complexity (Greiff et al., 2013) and for the respective sum scores. The level of complexity was defined by the number of input and output variables and the number and type of connections (Molnár & Csapó, 2017). We distinguished three levels of complexity: (1) less complex task (2 input variables, 2 output variables, and 2 connections), (2) more complex task with only direct effects (3 input variables, 3 output variables, and 3 or 4 connections), and (3) more complex tasks with internal dynamics (3 input variables, 3 output variables, and 2 or 3 direct effects beyond the internal dynamics). The students in the Hungarian sample achieved significantly higher scores at all complexity levels and in both of the CPS phases (the knowledge acquisition and knowledge application phases). Please note that this might be due to different selections in our samples (cf. limitations).

The differences between the two countries grew as the complexity of the items increased within both groups of problems with only a direct effect or with internal dynamics. This phenomenon

was found in both CPS phases, the knowledge acquisition and knowledge application phases (all of the t-values are significant at p<.001, see Table 4).

As measurement invariance was sufficiently met between the Jordanian and Hungarian students in RQ1, latent mean differences were not due to psychometric issues but could be interpreted as true differences in the measured construct between the two samples. As regards latent mean differences across the nations, the results showed that the Hungarian students performed significantly better in knowledge acquisition ($M_{HU}=0$; $M_{J}=-.79$, $p<.001$) and knowledge application ($M_{HU}=0$; $M_{J}=-1.01$, $p<.001$) than their Jordanian peers, confirming research results obtained at a manifest level.

Table 4.

Cross-national achievement differences in CPS: Problem complexity and problem phase-level differences

Complexity of problem (Number of input and output variables and number of connections)	Jordanian		Hungarian		t	p	d
	Mean	SD	Mean	SD			
Knowledge acquisition							
2-2 (2)	0.59	0.492	0.77	0.422	7.48	<.001	-0.39
3-3 (3 or 4)	0.46	0.493	0.76	0.424	12.96	<.001	-0.66
3-3 (2+1 or 3+1)	0.13	0.319	0.28	0.447	6.72	<.001	-0.40
Sum	0.36	0.422	0.57	0.44	13.21	<.001	-0.49
Knowledge application							
2-2 (2)	0.56	0.498	0.72	0.450	6.62	<.001	-0,33
3-3 (3 or 4)	0.05	0.233	0.37	0.472	13.14	<.001	-0,82
3-3 (2+1 or 3+1)	0.02	0.126	0.17	0.348	7.93	<.001	-0,51
Sum	0.15	0.258	0.35	0.416	16.88	<.001	-0.58

Note. The '+' sign by the number of connections denotes the presence of internal dynamics (associated with a higher level of complexity) in the problem environment.

4.4. Results for Research Question 3 (RQ 3): What kind of test-taking behaviors do Jordanian and Hungarian university students exhibit in solving complex problems? Are there differences between them in the theoretical effectiveness of their strategy use, their time-on-task, and the number of trials they use?

To answer RQ3, we looked at three different behavioral indicators that students exhibited in working on the CPS environments: theoretical strategy effectiveness, time-on-task, and number of trials.

4.4.1 Theoretical strategy effectiveness based on the amount of extracted information.

In the Jordanian sample, 44% of the students used a theoretically effective strategy; that is, they were able to extract all the information from the problem environment necessary to solve the problem properly, while this rate was 93% in the Hungarian sample. As the CPS performance differences based on the traditional scoring were not consistent with these results (see Table 4), to be able to understand the behavioral differences between the students from the two countries more deeply, we went further and compared the rate of theoretically effective strategy use and final problem-solving achievement.

In the Hungarian university sample, the percentage of theoretically effective strategy use and high CPS performance based on the traditional scoring changed from 28% to 76%, depending on the complexity of the CPS tasks (see Table 5). On average, 56.4% of the students used a theoretically effective strategy, were able to interpret the extracted information, and succeeded in drawing the right concept map; that is, they solved the first part of the problem properly. 36.4% of the students used a theoretically effective strategy but were unable to solve the first part of the problem correctly based on the extracted information. This rate was significantly higher on problems with only direct effects (on average, 75% of the students were successful). The students achieved significantly lower and were less successful on problems with internal dynamics. In the case of the most complex problems and problems with internal dynamics independent of their complexity, the rate of students who applied a theoretically non-effective strategy and still solved the problems correctly (by guessing) was very low (0.8% of the sample). This confirms earlier research results that have found that tasks with internal dynamics are generally considered more difficult to complete than those without them. Those tasks require additional exploration, where everything is maintained in a neutral (zero) position, that

is, a higher number of variable manipulations, which can significantly contribute to an increased chance of performance success (Beckmann & Goode, 2017; Lotz et al., 2017).

Table 5.

Percentage of theoretically effective and non-effective strategy use and traditional CPS scoring

Complexity of problem (Number of input and output variables and connections)	Frequency (%)									
	Theoretically effective strategy use					Theoretically non-effective strategy use				
	Low achievement (%; in proportion to whole sample)		High achievement (%; in proportion to whole sample)		Independent of final score, in proportion to whole sample	Low achievement (%; in proportion to whole sample)		High achievement (%; in proportion to whole sample)		Independent of final score, in proportion to whole sample
Jordanian sample										
2-2 (2)	30.2	(13.5)	69.8	(31.4)	44.9	38.5	(21.2)	61.4	(33.8)	55.1
3-3 (3 or 4)	40.3	(18.3)	59.6	(27.1)	45.5	61.8	(33.6)	38.1	(20.8)	54.5
3-3 (2+1 or 3+1)	83.3	(34.7)	16.6	(6.9)	41.6	87.8	(51.1)	12.1	(7.1)	58.3
Test	55.5	(23.9)	44.4	(19.9)	43.9	67.5	(38.1)	32.4	(17.9)	56.1
Hungarian sample										
2-2 (2)	21.8	(20.8)	78.2	(74.4)	95.25	75.6	(4.9)	24.4	(1.3)	6.2
3-3 (3 or 4)	18.1	(16.7)	81.9	(75.9)	92.6	94.4	(6.9)	8.5	(0.7)	7.4
3-3 (2+1 or 3+1)	69.4	(63.7)	30.6	(28.1)	91.8	76.7	(6.6)	1.1	(0.1)	8.4
Test	39.4	(36.4)	60.6	(56.4)	92.8	81.9	(6.3)	11.4	(0.8)	7.6

Note. Students' achievement was considered high if they managed to achieve a score of 1 based on the traditional scoring method.

This pattern was different in the Jordanian sample (see Table 5). Only half of the students (44%) managed to use a theoretically effective exploration strategy on the CPS problems compared to the Hungarian sample (93%). The percentage of theoretically effective strategy use and high CPS performance changed from 7% to 31%, depending on the complexity of the CPS tasks. On average, 20% of the students used a theoretically effective strategy, were able to interpret the extracted information, and managed to draw the right concept map. Almost one fourth of the students used a theoretically effective strategy but were unable to interpret the extracted information and solve the first part of the problem correctly. Like the Hungarian sample, this rate was significantly higher on problems with only direct effects (on average, 29% of the students were successful). The guessing factor, that is, ad hoc optimization, when students used a theoretically non-effective strategy and still solved the problem correctly, was

significantly higher in the Jordanian sample than in the Hungarian one. On the test level – independent of the complexity and structure of the problem – it was less than 1% for the Hungarian students and nearly 18% in the Jordanian sample.

4.4.2. Time-on-task.

There were large differences found in students' test-taking behavior as regards time-on-task (see Table 6). On average, the Jordanian students spent 36 seconds exploring the problem, while the Hungarian students spent more time on exploration (56 sec). On the one hand, the differences become smaller parallel to the increasing complexity of the tasks; on the other hand, they become large again when problems with internal dynamics appeared on the test. This phenomenon was caused mainly by the Hungarian students, who spent ever less time on problem exploration.

The Jordanian students' test-taking behavior was more stable over time and across different levels of problem complexity. However, there was a backward but weaker tendency identified compared to the Hungarian sample. The Jordanian students spent increasingly more time with more trials – but significantly less than their Hungarian peers – in the exploration phase of the problem-solving process as the problems became ever more complex.

Table 6.

Cross-national differences in students' test-taking behavior: time-on-task and number of trials

Complexity of problem	Jordanian			Hungarian			t	p	d
	Low achievement	High achievement	Mean	Low achievement	High achievement	Mean			
Time-on-task									
2-2 (2)	49.5	26.2	33.8	74.9	59.0	63.1	14.0	<.001	-0.67
3-3 (3 or 4)	38.5	37.0	37.6	55.9	47.2	49.2	6.0	<.001	-0.31
3-3 (2 or 3+1)	35.2	39.8	35.5	56.0	70.9	60.1	13.6	<.001	-0.76
Sum	39.4	35.9	36.0	59.7	59.1	56.4	18.4	<.001	-.57
Number of trials									
2-2 (2)	1.9	1.6	1.7	5.8	6.3	6.1	21.2	<.001	-1.3
3-3 (3 or 4)	1.8	2.3	2.0	3.8	4.4	4.2	15.9	<.001	-0.9
3-3 (2 or 3+1)	1.9	3.4	2.0	4.9	7.7	5.6	19.1	<.001	-1.19
Sum	1.9	2.6	1.9	4.8	6.2	5.3	26.2	<.001	-1.15

4.4.3. Number of trials.

There were also large differences found in the students' test-taking behavior in number of trials (see Table 6). On average, the Jordanian students attempted two trials, while the number of trials among the Hungarian students was more than five. The Hungarian students' time-on-task data and number of trials data were consistent with each other, while this was not the case in the Jordanian sample. Based on the tendencies in time-on-task and number of trials by high and low CPS achiever, we can conclude that the Hungarian students became increasingly aware of their exploration behavior and they engaged in ever fewer trial-and-error moves and attempted ever fewer trials in less and less time. However, both of the behavioral factors grew immensely when problems with internal dynamics appeared on the test.

4.5. Results for Research Question 4 (RQ 4): Based on the exploration strategy (i.e., VOTAT), which profiles can be extracted from the Jordanian and Hungarian students? Are there differences in the types of profiles that emerge from the two groups?

To tackle RQ 4, we investigated latent class analyses in both samples among the behavior patterns in the log data. They were scored according to the level of optimal exploration strategy use: 2: fully isolated variation strategy; 1: partially isolated variation strategy; 0: no isolated variation at all. The Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted Bayesian information criterion (aBIC), entropy, and the Lo–Mendell–Rubin adjusted likelihood ratio were used to approximate and determine the correct number of classes in the LCA models. In addition, the average latent class probabilities (ALCP) indicated the most likely latent class membership for every student.

After running the LCA in both samples, the information theory criteria used (AIC, BIC, and aBIC) indicated an almost continuous decrease with a growing number of latent classes up to the 4-class model. The likelihood ratio statistical test (the Lo–Mendell–Rubin adjusted likelihood ratio test) showed the best model fit – in both countries – for the 4-class model and was no longer significant with the 5-class model. The entropy-based criterion reached the maximum values for the 2-class solutions, but it was also high for the 4-class models based on the information theory and likelihood ratio criteria. Thus, the entropy index for the 4-class

model demonstrated that 95% of the Jordanian students and 96% of the Hungarian students were accurately categorized based on their class membership (Table 7).

As noted above, four latent classes were distinguished in the Jordanian sample (as well as in the Hungarian sample). The classes were interpreted as follows based on their profiles: (1) non-performing explorers, (2) non-persistent explorers, (3) restarting explorers with a learning effect, and (4) almost proficient explorers.

Table 7.

Information theory, likelihood ratio, and entropy-based fit indices for latent class analyses

Number of latent classes	AIC	BIC	aBIC	Entropy	L–M–R test	P
Jordanian						
2	5266	5433	5303	.979	2797	.000
3	5008	5260	5063	.949	298	.000
4	*4948*	*5286*	*5022*	*.948*	*100*	*.006*
5	4935	5358	5028	.934	54	.838
Hungarian						
2	10376	10602	10471	.990	6089	.000
3	9683	10025	9828	.958	729	.000
4	*9513*	*9970*	*9707*	*.959*	*210*	*.001*
5	9479	10052	9721	.949	75	.169

Note. 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 italics.

Non-performing explorers (40% of the Jordanian students) employed no fully or partially isolated strategy at all. Non-persistent explorers proved to be intermediate explorers on the easiest problems but low explorers on the complex ones (6.6% of the Jordanian students), having employed the partially isolated variation strategy less and less parallel to the increasing level of complexity of the CPS problems. Restarting explorers with a learning effect (15.3% of the Jordanian students) were able to learn between problems of similar complexity (similar number of input and output variables and number and type of connections), but the probability of applying a partially or fully isolated strategy dropped again as the complexity of the

problems grew. Almost proficient explorers (38.4% of the Jordanian students) used the isolated variation strategy with 80% probability on problems with only direct effects. Then, after a rapid learning process, they managed to continue this exploration behavior even with the CPS problems with internal dynamics (see Figure 3 and Table 8).

The following four latent classes were distinguished in the Hungarian sample, albeit somewhat different ones as compared to the Jordanian sample: (1) non-performing explorers, (2) restarting slow learners, (3) rapid learners, and (4) proficient explorers (see Figure 4 and Table 8).

Non-performing explorers (7.4% of the Hungarian students) did not use any isolated or partially isolated variation at all throughout the tasks. Restarting slow learners (3.2% of the Hungarian students) were among the intermediate-performing explorers who only rarely employed a fully or partially isolated variation strategy with a very slow learning effect. Rapid learners (7% of the Hungarian students) were basically low performers with regard to the efficacy of the exploration strategy they used on the easiest problems, but they become proficient explorers as a result of rapid learning, with achievement on the complex ones that equaled that of the top performers. Proficient explorers (82.4% of the Hungarian students) used the isolated variation strategy with high probability on all the proposed CPS problems.

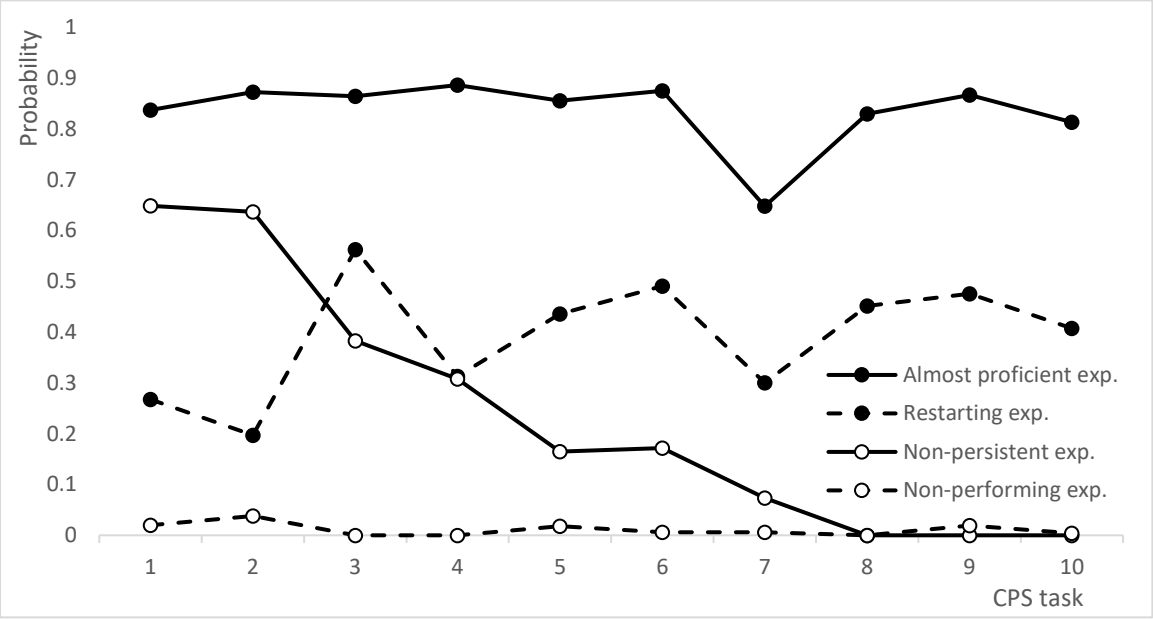

Figure 3. Four qualitatively different class profiles in the Jordanian sample

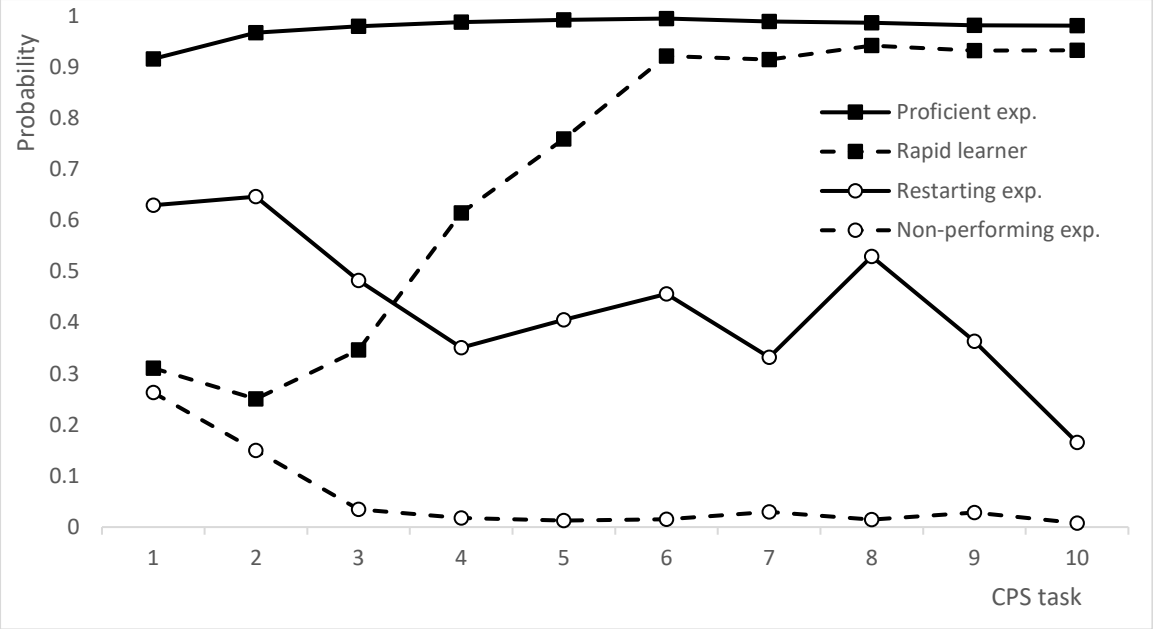

Figure 4. Four qualitatively different class profiles in the Hungarian sample

Table 8.

Relative frequencies and average latent class probabilities in the Arabic and Hungarian-language samples

Profiles	Arabic		Hungarian	
	Frequency	Average Latent Class Probabilities	Frequency	Average Latent Class Probabilities
Non-performing explorers	39.7	0.987	7.4	.985
Non-persistent explorers	6.6	0.937	-	-
Restarting slow learners	15.3	0.958	3.2	.934
Rapid learners	-	-	7.0	.906
Almost proficient explorers	38.4	0.970	-	-
Proficient explorers	-	-	82.4	.989

Note. Latent classes are ordered along their levels of isolated variation strategy.

We analyzed students' test-taking behavior (time-on-task and number of clicks) and their overall CPS performance based on their latent class membership (Figure 5). Similar to our earlier findings, the two samples showed slightly different patterns.

In the Hungarian sample, there was a quadratic relation (see Figure 5) between latent class membership and students' overall performance scores in CPS and between students' achievement and number of trials but not time-on-task. That is, proficient explorers achieved significantly higher in both knowledge acquisition and knowledge application based on the traditional scoring method and attempted more trials than rapid learners. Rapid learners achieved significantly higher than restarting slow learners, and restarting slow learners achieved significantly higher but only in knowledge acquisition, than non-performing explorers, who applied the fewest trials and spent the least time on the problem-solving process. Rapid learners and restarting slow learners spent the most time in the problem environments on average.

In the Jordanian sample, the pattern was different, and there was no clear parallel identified between latent class membership and the students' overall performance scores in CPS, a finding which runs counter to our previous expectations but in line with the findings in RQ 3 about Jordanian students' high (18%) guessing factor. As a result, the Jordanian non-performing explorers achieved significantly higher than the students who fall in the non-persistent explorers' group with a very low number of trials (almost no trials) and time spent on the problem-solving process. This indicates that it was mostly the students from the non-performing explorers group that used the guessing strategy in the problem-solving process, which resulted in higher final achievement than the manipulation strategy suggests. The students with the lowest and highest CPS achievement (non-persistent explorers and almost proficient explorers, see Figure 5) spent the same amount of time solving the problems. This amount of time was exactly the same as that of the Hungarian non-explorers.

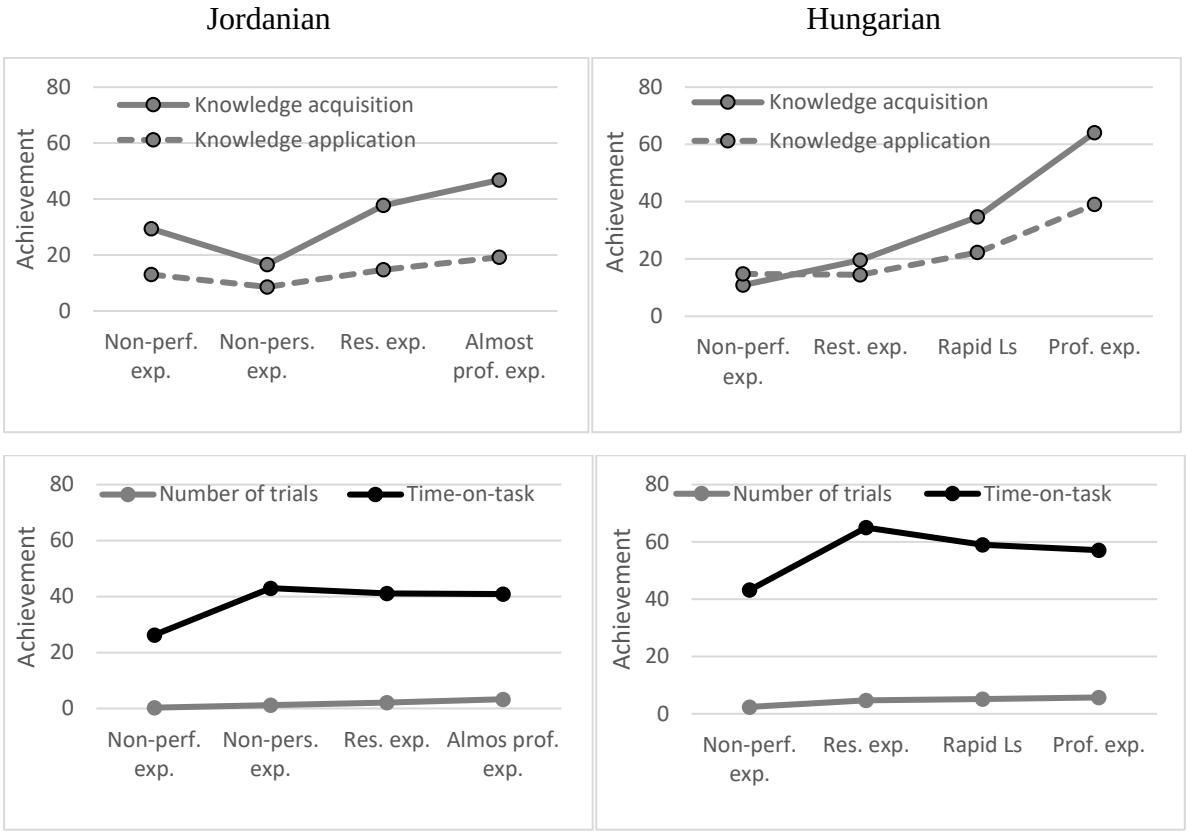

Figure 5. Performance and test-taking behavior among students with different latent class profiles. (We have connected the data points to visualize the tendencies.)

5. Discussion

This study shows that complex problem-solving can be measured validly, reliably, and equivalently in the Hungarian and Jordanian educational contexts. It provides important insights into the international validity of CPS measurements and sheds light on the different behavior patterns of Hungarian and Jordanian university students, thus expanding our understanding beyond what we can learn from traditional performance indicators for CPS. We used state-of-the art analyses on logged process data to quantify qualitative behavioral differences in students' problem-solving behavior. Hungarian data on CPS were used as benchmark indicators in this study. This research is a good reminder that results obtained in one culture or one country are not necessarily generalizable to other countries or cultures even if these results cover general skills, such as problem-solving, which is less developed explicitly

in school context. Students socialized in one school context can think differently and can reach the same results with the same aims on different routes.

RQ1: Do Jordanian and Hungarian students interpret CPS problems the same way? Is CPS measurement-invariant across Jordanian and Hungarian university students?

We found invariance in CPS measurement across the Jordanian and Hungarian university students; that is, both groups interpret CPS problems the same way, so the language-based conceptual representational differences did not influence the way the students interpreted the problems. Despite the large cultural and educational differences, which can influence measurement invariance, it is possible for CPS to be measurement-invariant across nationalities in the Jordanian and Hungarian contexts. That is, measurement invariance was influenced neither by the substantial language differences nor by the expansion of technology-based assessment. Earlier studies indicated (see Wüstenberg et al., 2014) measurement invariance of CPS between Hungarian and German students. We have expanded and broadened the usability of CPS instruments to the Middle East region. Earlier studies also pointed to measurement non-invariance of CPS across Hungarian and Chinese students (Wu & Molnár, 2021). The inconsistency of these research findings and the non-invariance between the Hungarian and Chinese results may lie in students' different cognitive styles (Wu & Molnár, 2021) connected to the different encoding and conceptual representations in the languages and in the different behavior during testing, which can be rooted in educational and cultural differences. Limitations on the generalization of these research results may be that all the research was conducted with students of different ages and used different sampling procedures. To sum up, we can hypothesize that measurement invariance holds across Western and Eastern cultures, at least to the extent of the countries that have been involved in such studies.

RQ2: Can we identify developmental differences in CPS skills between Jordanian and Hungarian university students? What is the nature of these developmental differences?

We identified developmental differences between the Jordanian and Hungarian university students' CPS skills in favor of the Hungarian group, which is consistent with earlier research results, indicating that students with different educational and cultural backgrounds can perform differently in a CPS environment (see Greiff, Wüstenberg, & Avvisati, 2015; OECD, 2014a; Wu & Molnár, 2021; Wüstenberg et al., 2014); that is, the development of CPS skills is not universal. We used Hungarian CPS data as a benchmark indicator in the present comparison study. Additional research is needed to validate the results using representative samples in both countries.

The score-based achievement differences were smaller at the beginning of the test when the students were expected to solve less complex problems and grew as the complexity of the problems increased. This phenomenon was noticeable in both CPS phases (knowledge acquisition and knowledge application). The trend was broken by problems with internal dynamics, which proved to be too difficult for the students. There were only a few Jordanian students who managed to cope with this kind of problem, resulting in very low group-level achievement. The Hungarian students' mean achievement also dropped immensely but was still significantly higher than the performance of their Jordanian peers.

The traditional scoring-based achievement differences between the Hungarian and Jordanian students were independent of the problem-solving phase; they were more a function of the complexity of the problems. This means that if the Jordanian students' achievement dropped immensely, the Hungarian students' mean performance also dropped at the same level; it was only the starting value that differed significantly, resulting in significant differences in achievement in both phases among all complexity levels. That is, despite the fact that most of the Hungarian students in the study sample started out as expert problem-solvers, their achievement was influenced just as much by the level of problem complexity as it was in the case of the Jordanian sample (on these hypotheses, see RQ3 and RQ4).

Reasons for these differences in achievement may lie in major cultural and educational differences as well as in the experience of computer use in educational context. The educational use of computers has long been addressed in Western countries, and one important area is supporting learning of scientific knowledge and skills (e.g., testing hypotheses while interacting with software that simulates scientific phenomena). Differences in experience with such computer use might also cause differences in exploring behaviors (on these hypotheses, see also RQ2 and RQ3).

RQ3: What kind of test-taking behaviors do Jordanian and Hungarian university students exhibit in solving complex problems? Are there differences between Jordanian and Hungarian students in the theoretical effectiveness of their strategy use, their time-on-task, and the number of trials they use?

Having learned that we can measure CPS equivalently (in RQ1) and that the Hungarian and Jordanian students (in this particular sample) differ in their level of CPS skills (in RQ2), we wanted to better understand these differences and take a closer look at their test-taking behavior. Based on the logfile analyses, there were large differences noted in the use of a

theoretically effective exploration strategy in both samples. A total of 93% of the Hungarian university students used a theoretically appropriate strategy compared to 44% in the Jordanian sample. This confirms our earlier explanation that most of the Hungarian students started out as expert problem-solvers. The percentages of theoretically effective strategy use and high CPS performance were also different. It was 60.6% on average in the Hungarian sample and 44.4% among the Jordanian students. The Hungarian findings are consistent with earlier large-scale research results (Molnár & Csapó, 2018) on changes in theoretically effective strategy use among 3rd–12th-grade Hungarian students. Molnár and Csapó found an increasing tendency by age: 40% of 3rd–5th-grade children, 55% of 6th–8th-grade students, and 65% of 9th–12th-graders managed to use a theoretically effective CPS strategy. In the present study, this grew to 93% in the university sample. They found a similar tendency in students' interpretation of extracted information; that is, 20% of young people in Grades 3–5, 30% of students in Grades 6–8, and 40% of those in Grades 9–12 were able to interpret the extracted information correctly and solve the problem properly. This rate increased to 56% in the present case, confirming that, based on the effectiveness of the exploration strategy they used and the level of interpretation of extracted information, the Jordanian university students in the study are in an earlier phase of CPS development than their Hungarian peers. That is, there were not only large differences in the appropriateness of the exploration strategy they used but also in the effectiveness of their interpretation of extracted information between the two samples, resulting in large differences in final CPS achievement.

Beyond the effectiveness of the exploration strategies used in the CPS environment, there were large differences identified in the students' test-taking behavior as regards time-on-task and number of trials at the international level. At the sample level, we confirmed Eichmann et al. (2019) and Goldhammer et al.'s (2014) research findings that low-achieving students typically engage in less interaction with the problem than high achievers (cf. the Jordanian and Hungarian results); that is, there is a positive correlation between CPS achievement and number of clicks, i.e., amount of exploration. If students spent more time on a CPS task, their performance improved significantly (Alzoubi et al., 2013; Goldhammer et al., 2014). Taking a closer look at the results, we identified two more important behavioral differences.

The differences identified grow smaller compared to the increasing complexity of the tasks. This tendency was caused by the Hungarian students, who spent generally less and less time attempting fewer and fewer trials despite the increasing complexity of the tasks in comparison to the Jordanian students, who spent almost the same time and used almost the same number

of trials throughout the test. In our view, one which was tested in RQ12, this tendency indicated that the Hungarian students grew increasingly aware of their effective exploration behavior and required ever fewer trial-and-error moves and ever fewer trials. This may also explain the different research results for time-on-task and high CPS achievement (cf. Alzoubi et al., 2013; Greiff et al., 2016; Scherer, Greiff, & Hautamäki, 2015), which Goldhammer et al. (2014) concluded was due to the lack of a common definition of time-on-task and achievement.

RQ4: Based on the exploration strategy (i.e., VOTAT), what profiles can be extracted from among the Jordanian and Hungarian students? Are there differences in the types of profiles that emerge from the two groups?

In RQ2 and RQ3, we found several sample-level behavioral differences. In RQ3, we used a more person-centered approach to see further CPS-related differences between the two samples and search for more detailed explanations for the tendentious differences between high and low CPS achievers found previously in the two cultures and beyond.

Based on the level of the optimal exploration strategy, we employed latent class analyses to describe students' exploration strategies in a CPS environment. We identified four latent classes in both samples. The classes of non-performing explorers and restarting slow learners proved to be almost identical in the two samples, indicating existing differences between the behaviors of Jordanian and Hungarian students. Our study confirmed Moln'ar's (2021) result on the presence of rapid learners in the Hungarian university sample, which was not found in the Jordanian sample. Rapid learners showed a remarkable learning curve while working on the problems and reached the same level as the proficient explorers in terms of their exploration behavior by the sixth problem on the test. They have the ability to adapt quickly and flexibly to the expectations of a specific situation (see Greiff et al., 2018). Instead of rapid learners, a class of non-persistent explorers was identified in the Jordanian sample (cf. Greiff et al., 2018). These students applied the partial variation strategy on the easiest problems but were unable to transfer this knowledge to the more complex problems. Finally, we identified behavioral differences in the top explorer groups – Hungarian vs. Jordanian. The proficient explorers in the Hungarian sample seemed to have more explicitly specific schemata (see Greiff et al., 2018); they were thus able to use the optimal exploration strategy throughout the CPS tasks, independently of their complexity, while the Jordanian students' schemata proved to be less well-founded and transferable, independently of the complexity of the CPS environment. However, the students in this group were able to learn rapidly and adapt to a given situation

flexibly and quickly, like the rapid learners in the Hungarian sample. The proportion of students in the different class profiles in Jordan and Hungary varied strongly.

Confirming earlier research results (Greiff et al., 2018) on time-on-task, both the rapid learners and restarting slow learners might have varying amounts of general cognitive schemata that they can adapt quickly and flexibly or slowly and less flexibly to the demands of a specific situation, CPS problems in the present case. This adaptation requires time to take effect. Non-performing explorers, who were not motivated in the test-taking process, and proficient explorers, who were aware of their strategy use, spent less time on the problem exploration process. The number of trials showed different patterns and was not strongly correlated to time-on-task, contrary to our hypotheses. Time-on-task increased with the amount of optimal strategy use in both samples; that is, students' exploration profiles proved to be a better predictor of the expected number of trials than time-on-task or final achievement.

6. Limitations

The study used a widely used model, the MicroDYN approach, for measuring students' problem-solving skills. These problems are artificial, with a limited number of variables and relations, but appropriate and reliable for measurement purposes. Problems in the MicroDYN approach do not cover all kinds of problems and complex systems found in life, which are dynamic in nature in most cases (i.e., they change regardless of attempts to address them); thus, problem-solving behavior observed in problem scenarios developed through the MicroDYN approach cannot be generalized to all kinds of complex problems we face in life. However, their special features make it possible to monitor students' learning processes and learning potential during the problem-solving process.

Similarly, there is an optimal exploration strategy for problems with a limited number of variables and relations, such as MicroDYN problems. Nonetheless, optimal exploration strategies do not apply to everyday complex problems, as observed by Funke (2021) with regard to problems of "minimal complexity" (i.e., the subject of most research on CPS and a focus of PISA) and real-world complex (wicked) problems, which represent an urgent priority but cannot be experienced in laboratory environments, in which variables can be selectively controlled for educational purposes. In fact, real-world complex problems are characterized precisely by non-fully knowable or controllable variables which interact over time in changing ways, independent of any attempt to address the problem situation. Relatively large differences in sample size are among the limitations of the present cross-national comparison study as well

as differences in gender distribution, differences in time elapsed since the Matura examination (in Hungary, only first-year students took part in the assessment, while students in higher years also participated in Jordan), differences in parental education and socio-economic background (e.g., number of books in the home), differences in the subjects studied by the students (in Hungary, students from all twelve schools within an entire university took part in the assessment, while students from two universities, mostly focused on economics, education, the humanities, IT, and science subjects, participated in the study in Jordan – thus not covering such areas of study as medicine and engineering), and differences in data collection (supervised and not supervised). Compared to the Hungarian sample, the relatively small Jordanian one may lead to limitations in the validity of the findings, especially for RQ10, and restrict the generalizability of the results at a population level. Based on the current findings, some initial trends can be identified, which can form a solid foundation for further large-scale empirical studies on Jordanian students' exploration behavior in a CPS environment with a focus on comparing problem-solving behavior among students with different cultural backgrounds.

7. Conclusions

The results of the current study provide important insights into the international validity of CPS measurements and shed light on the different hidden behavior patterns and test-taking behaviors of Jordanian and Hungarian university students as they solve complex problems, thus expanding our understanding beyond what we can learn from traditional performance indicators. As for educational implications, we are confident that a more thorough grasp of the differences and similarities in students' problem-solving behavior will not only help educators to recognize relevant individual differences more effectively and become more sensitive towards these differences in learning but also provide valuable input for the design of appropriate training tasks and the training of students to become better problem-solvers.

References

Al Suwaidi, M. (2008). When an Arab executive says "Yes": Identifying different collectivistic values that influence the Arabian decision-making process. Master of Science in Organizational Dynamics Theses. 19. https://repository.upenn.edu/od_theses_msod/19

Alzoubi, O., Fossati, D., Di Eugenio, B., Green, N., & Chen, L. (2013). Predicting students' performance and problem solving behavior from iList log data. In ICCE 2013, 21st International Conference on Computers in Education.

Arieli, S., & Sagiv, L. (2018). Culture and problem-solving: Congruency between the cultural mindset of individualism versus collectivism and problem type. *Journal of Experimental Psychology: General, 147*(6), 789–814. https://doi.org/10.1037/xge0000444

Beckmann, J. F., Birney, D. P., & Goode, N. (2017). Beyond psychometrics: the difference between difficult problem solving and complex problem solving. *Frontiers in psychology*, 8, 1739. https://doi.org/10.3389/fpsyg.2017.01739

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch, & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale, NJ: Erlbaum.

Byrne, B.M., Stewart, S.M., 2006. The MACS approach to testing for multigroupinvariance of a second-order structure: A walk through the process. Struct. Equ.Model. 13 (2), 287–321.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Collins, L. M., & Lanza, S. T. (2010). Latent class and latent transition analysis. In *With applications in the social, behavioral, and health sciences*. New York: Wiley

Csapó, B., & Funke, J. (2017). Epilogue. In B. Csapó and J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 265–268). Paris: OECD Publishing.

Csapó, B., & Molnár, G. (2017). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Frontiers in Psychology, 8*, 2022. 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*, 1522. https://doi.org/10.3389/fpsyg.2019.01522

Dörner, D. (1986). Diagnostik der operativen Intelligenz [Assessment of operative intelligence]. *Diagnostica, 32*(4), 290–308.

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8*:1153. doi: 10.3389/fpsyg.2017.01153

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, 128, 1–12.

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.

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving, 4*, 19–42. doi: 0.7771/1932-6246.1118

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning*, 7(1), 69–89.

Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Frontiers in Psychology*, 5. 739. doi: 10.3389/fpsyg.2014.00739

Funke, J (2021). It requires more than intelligence to solve consequential world problems. Journal of Intelligence, 9(3), 38.

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective – 10 years after. In D. H. Jonassen (Ed.), Learning to solve complex scientific problems (pp. 25–47). New York: Erlbaum.

Gleitman, L., & Papafragou, A. (2012). New perspectives on language and thought. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (2nd edition) (pp. 543–568). New York, NY: Oxford University Press.

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.

Greiff, S., & Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme. *Zeitschrift für Pädagogik, 56*, 216–227.

Greiff, S. (2012). Assessment and theory in complex problem solving: A continuing contradiction. *Journal of Educatioanal and Developmental Psychology*, *2*, 49–56.

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* (pp. 93–105). Paris: OECD Publishing.

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning*, *21*(3), 356–382. 10.1080/13546783.2014.989263

Greiff, S., Holt, D. V., & Funke, J. (2013). Perspectives on problem solving in cognitive research and educational assessment: analytical, interactive, and collaborative problem solving. *Journal of Problem Solving, 5*, 71–91. https://doi.org/10.7771/1932-6246.1153

Greiff, S., Krkovic, K., & Hautamäki, J. (2015). The prediction of problem-solving assessed via microworlds. *European Journal of Psychological Assessment, 32*, 298–306.

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 & Education*, 126, 248–263.

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 files. *Computers in Human Behavior*, *61*, 36–46.

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.

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. doi: 10.1037/a0031856

Hofstede, G., Hofstede, G.J., (2005). Cultures and Organizations: Software of the Mind.McGraw-Hill, New York.

Holicza, P., (2016). Understanding magyar: an analysis of Hungarian identity within the framework of cultural dimensions theory and additional metrics. In: 4th International Scientific Correspondence Conference. Slovak University of Agriculture in Nitra,pp. 118–124.

Holyoak, K.J., (1985). The pragmatics of analogical transfer. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation. Academic Press, New York, NJ, pp. 59–87.

Ingles, C.J., Marzo, J.C., Castejon, J.L., Nunez, J.C., Valle, A., Garcia-Fernandez, J.M., Delgado, B., (2011). Factorial invariance and latent mean differences of scores on the achievement goal tendencies questionnaire across gender and age in a sample of Spanish students. Learn. Indiv Differ 21, 138–143.

Kaplan, D., George, R., (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. Struct. Equ. Model. 2, 101–118.

Klahr, D., Triona, L.M., Williams, C., 2007. Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. J. Res. Sci. Teach. 44, 183–203.

Landau, B., Dessalegn, B., Goldberg, A.M., (2010). Language and space: momentary interactions. In: Chilton, P., Evans, V. (Eds.), Language, Cognition, and Space: the State of the Art and New Directions. Advances in Cognitive Linguistics Series. Equinox Publishing, London, United Kingdom, pp. 51–78.

Lotz, C., Scherer, R., Greiff, S., Sparfeldt, J.R., (2017). Intelligence in action – effective strategic behaviors while solving complex problems. Intelligence 64, 98–112.

Meade, A.W., Johnson, E.C., Braddy, P.W., (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. J. Appl. Psychol. 93, 568–592.

Moln'ar, G., (2021). How to make different thinking profiles visible through technology: the potential for log file analysis and learning analytics. In: Virvou, M., Tsihrintzis, G.A., Tsoukalas, L.H., Jain, L.C. (Eds.), Advances in Artificial Intelligence-Based Technologies. Springer, Cham, pp. 125–146.

Moln'ar, G., Greiff, S., Csapo', B., (2013). Inductive reasoning, domain specific and complex problem solving: relations and development. Think. Skills Creativ. 9 (8), 35–45.

Moln'ar, G., Csapo', B., (2017). Exploration and learning strategies in an interactive problem-solving environment at the beginning of higher education studies. In: Spender, J.C., Gavrilova, T., Schiuma, G. (Eds.), Knowledge Management in the 21st century: Resilience, Creativity and Co-creation. Proceedings IFKAD2017. St Petersburg University, St. Petersburg, pp. 283–292.

Moln'ar, G., Csapo', B., (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: logfile analyses. Front. Psychol. 9, 302. Millsap, R.E., 2012. Statistical Approaches to Measurement Invariance. Routledge.

Mustafi'c, M., Yu, J., Stadler, M., Vainikainen, M.-P., Bornstein, M.H., Putnick, D.L.,Greiff, S., (2019). Complex problem solving: profiles and developmental paths revealed via latent transition analysis. Dev. Psychol. 55 (10), 2090–2101.

Muth'en, L.K., Muth'en, B.O., (2012). Mplus User's Guide, seventh ed. Muth'en and Muth'en, Los Angeles, CA.

Müller, J.C., Kretzschmar, A., Greiff, S., (2013). Exploring exploration: inquiries into exploration behavior in complex problem solving assessment. In: D'Mello, S.K., Calvo, R.A., Olney, A. (Eds.), Proceedings of the 6th International Conference on Educational Data Mining, pp. 336–337.

Nicolay, B., Krieger, F., Stadler, M., Gobert, J., Greiff, S., (2021). Lost in transition – learning analytics on the transfer from knowledge acquisition to knowledge application in complex problem solving. *Comput. Hum. Behav.* 115.

OECD, (2014a). Creative Problem Solving: Students' Skills in Tackling Real-Life Problems – Volume V. OECD, Paris.

OECD, (2014b). PISA 2012 Technical Report. OECD, Paris.

Ourfali, E., (2015). Comparison between western and middle eastern cultures: research on why american expatriates struggle in the Middle East. Otago Manag. *Grad. Rev.* 13, 33–43.

Putnick, D.L., Bornstein, M.H., (2016) Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. Developmental review 41, 71–90.

Rom'an, A., Flumini, A., Lizano, P., Escobar, M., Santiago, J., (2015). Reading directioncauses spatial biases in mental model construction in language understanding. Sci. Rep. 5 (1), 1–8.

Rutkowski, L., Svetina, D., (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. Educ. Psychol. Meas. 74 (1), 31–57.

Schoppek, W., Kluge, A., Osman, M., Funke, J., (2018). Editorial: complex problem solving beyond the psychometric approach. Front. Psychol. 9, 1224.

Schult, J., Stadler, M., Becker, N., Greiff, S., Sparfeldt, J.R.,(2017). Home alone: complex problem solving performance benefits from individual online assessment. Comput. Hum. Behav. 68 (March), 513–519.

Scherer, R., Greiff, S., Hautamaki, J., (2015). Exploring the relation between time on taskand ability in complex problem solving. I*ntelligence* 48, 37–50.

Schwartz, S.H., Bilsky, W., (1990). Toward a theory of the universal content and structure of values: extensions and cross-cultural replications. *J. Pers. Soc. Psychol.* 58 (5), 878–891.

Schweizer, F., Wüstenberg, S., Greiff, S., (2013). Validity of the MicroDYN approach: complex problem solving predicts school grades beyond working memory capacity. *Learn. Indiv Differ* 24, 42–52.

Stadler, M., Hofer, S., Greiff, S., (2020). First among equals: log data indicates ability differences despite equal scores. Comput. Hum. Behav. 111.

Tein, J.Y., Coxe, S., Cham, H., (2013). Statistical power to detect the correct number of classes in latent profile analysis. Struct. Equ. Model.: A Multidiscip. J. 20 (4), 640–657.

To'th, K., Rolke, H., Goldhammer, F., Barkow, I., (2017). Educational process mining: new possibilities for understanding students' problem-solving skills. In: Csapo', B., Funke, J. (Eds.), The Nature of Problem Solving: Using Research to Inspire 21st Century Learning. OECD Publishing, Paris.

Triandis, H.C., (1994). McGraw-Hill Series in Social Psychology. Culture and Social Behavior. Mcgraw-Hill Book Company.

Ünal, E., Papafragou, A., (2018). The Relation between Language and Mental State Reasoning. Metacognitive Diversity: an Interdisciplinary Approach, pp. 153–169.

Wu, H., Moln'ar, G., (2021). Logfile analyses of successful and unsuccessful strategy use in complex problem-solving: a cross-national comparison study. *Eur. J. Psychol. Educ.* 1–24.

Wüstenberg, S., Greiff, S., Funke, J., (2012). Complex problem solving, more than reasoning? Intelligence 40 (1), 1–14.

Wüstenberg, S., Greiff, S., Molnar, G., Funke, J., (2014). Determinants of cross-national gender differences in complex problem solving competency. Learn. Indiv Differ 29, 18–29.

Yoon, M., Lai, M.H.C., (2018). Testing factorial invariance with unbalanced samples. Struct.Equ. Model.: A Multidiscip. J. 25 (2), 201–213.

.

6

CONCLUSION AND LIMITATIONS

Technology has greatly improved the effectiveness of testing procedures: it has sped up data collection, enabled real-time automatic scoring, accelerated data processing, enabled immediate feedback, and revolutionized the entire assessment process, including creative task presentation (Csapó et al., 2012). It also opens up new possibilities for both item and test development. In addition to these alternatives, technology allows for the storage and analysis of contextual data. Educational data mining, logfile analysis, and learning analytics are all terms used to describe this new approach, each describing a somewhat different type of study. Because of the numerous benefits, the most important assessments will most likely be administered in a technological environment in the near future; however, more research and development on the application of CBA among Jordanian students for all levels are needed.

Theoretical studies confirm that offering extra indicators during the educational process, particularly in assessment, attracts researchers' attention to this new topic, therefore supporting the educational process with a multitude of indicators. Indeed, as a research field, using technology in assessments has matured. In recent decades, attention has been focused on analyzing contextual data in educational contexts using educational data mining, which employs data mining techniques to transform initial data collected through educational systems into meaningful information. Logfile analysis entails analyzing behavioural processes, time-on-task, and the sequence of actions captured in logfiles and thus introduces novel methods to analyze the instruction and learning process and educational assessment.

Recent work has focused on logfile analysis, educational data mining, and learning analytics. Developments in information technologies have made it possible to design different assessments, thus boosting the number of ways students can demonstrate their skills and abilities. Parallel to these advances, the focus of technology-based assessment has shifted from an individual and summative approach to cooperative, diagnostic and more learning-centered to implement efficient testing for personalized learning.

According to results from a comparison analysis based on the Scopus database, (1) research interest in this field has grown immensely in the last few years; that is, EDM is an emerging discipline. In addition, (2) EDM and logfile analysis examine earlier hidden information to provide explanations of students' learning and testing behaviour from a new perspective. Thus, broadening our understanding of students' behaviour, interests, learning processes, motivational aspects, and test results and the reason for their learning outcomes.

We have tested the feasibility of computer-based assessment, especially, the applicability of a third generation, innovative test measuring a 21st century skills, such as complex problem solving in Jordanian higher education context. We have tested the behaviour and the psychometric indices of the CPS test, adapted within the confines of the present project. With this project we filled an important niche as computer-based assessment of thinking skills was not commonly implemented in Jordan and in Jordanian higher education context. First, a pilot study was conducted to prove the feasibility of computer-based assessment and the reliability of the CPS test in Jordanian cultural and higher education environments. Second, a large-scale assessment was organized to confirm research results of the pilot study and get more knowledge about Jordanian students' test-taking and problem solving behaviour. For example, larger dataset was required to run pattern discovery algorithms on the collected logfiles (Greiff et al., 2018; Molnár & Csapó, 2018; Wu & Molnár, 2021) in order to map and classify the students' exploration tactics while solving interactive challenges. Finally, the results of an international comparison study highlight the differences and similarities between Arabic (Jordanian) and European (Hungarian) students' test-taking and problem solving behaviour in interactive problem solving environments.

According to the results the CPS assessments had a high level of internal consistency – similarly to the European results, but the interactive problems proved to be generally hard for the Jordanian participants. We could conclude from the descriptive results that computer-based assessment and the use of innovative online tests are feasible and valid in Jordan in the higher education environment. The results regarding the third-generation CPS test are generalizable in the Jordanian higher educational environment. Since CPS skills are necessary component of educational achievement and address important gaps in modern education: the gap between students' ability to acquire and apply knowledge in uncertain conditions, these results are important in the twenty-first century.

Logfile-based analyses extended the scope of previous research results connected to CPS, particularly in the Arabic context. We monitored and identified the way students understand interactive problems, especially minimal complex systems and causal relationships within the problems. Despite the fact that most of the Jordanian university students showed systematic strategies were unable to solve the problem and on the contrary several students managed to solve the problem without applying an effective problem solving strategy. Thus, solving an interactive problem does not necessarily require the application of a strategy that gives the problem solver sufficient information about the problem environment to achieve the correct

solution and the application of a right problem solving strategy does not always result in high problem solving achievement. This confirms de Jong and van Joolingen (1998) research results, who claim that learners often have trouble understanding data. Generally, these results are in line with previous research results (e.g., Greiff et al., 2015; Molnár & Csapó, 2018; Vollmeyer et al., 1996).

There was a significant correlation between KAC and KAP on both the manifest and latent levels. The KAC and KAP processes were empirically distinguished and confirmed by the international research results (e.g. Funke, 2001; Wüstenberg et al., 2012). More specifically, previous studies have found that KAC and KAP correlations range from weak to strong relationships between the two phases. The wide range of correlation indices associated with the use of multiple CPS assessments with varied approaches to measuring KAC and KAP.

To interpret and understand Jordanian research results more deeply and detect the cross-national aspects of CPS, we organized an international assessment in Hungary and Jordan. Based on the results of the cross-national large scale study, we can conclude that complex problem-solving can be measured in the Hungarian and Jordanian educational contexts validly, reliably, and equivalently. The results revealed the different behaviour patterns of Hungarian and Jordanian undergrads, providing valuable insights into the international validity of CPS assessments and increasing our understanding beyond what we can learn from traditional CPS performance indicators. Even students who are socialized in the same school environment can think differently and achieve the same results with the same aims via different routes.

We investigated the way Jordanian and Hungarian students interpret CPS problems. The results showed measurement invariance of CPS across Jordanian and Hungarian undergraduates, that is, students independent of their culture interpreted CPS problems the same way. Not even the language-based conceptual representational differences and the differences in frequency usage of computer-based assessments impacted measurement invariance of CPS.

Developmental differences have been found in CPS skills between Jordanian and Hungarian university students in favor of the Hungarian students. This is consistent with previous research findings indicating that students from different educational and cultural backgrounds can perform differently in CPS environment (see Greiff et al., 2015; OECD, 2014; Wu & Molnár, 2021; Wüstenberg et al., 2014). The development of CPS skills is not universal. In the dissertation presented comparison analysis, we used Hungarian CPS data as a benchmark

indicator. Additional studies using representative samples from both nations are required to validate these findings.

When students were expected to explore and solve easier problems, having less complexity, the score-based achievement differences were less between Hungarian and Jordanian students, but they grew as the complexity of the problems increased. This phenomenon was observed in both CPS phases (knowledge acquisition and knowledge application). Problems having internal dynamics (the more complex problems) enlarged this trend as only a few Jordanian students were able to deal with such problems, resulting in an extremely low group-level average achievement. By these type of problems the mean achievement of the Hungarian students dropped significantly as well, but it was still much higher than that of their Jordanian peers.

Beyond differences, there were also similarities detectable in Jordanian and Hungarian students' CPS behaviour. If Jordanian students' achievement dropped, Hungarian students' mean performance dropped too; but there was a major difference in their starting values, resulting in significant achievement differences in both phases across all complexity levels. The reasons for these differences in achievement could be due cultural and educational differences and probably their prior computer experience in academic environment. In Western countries, the use of computers in education has long been addressed, and belongs to key areas supporting learning.

After we have confirmed that CPS can be measured cross-nationally in the same way and the level of CPS skills of Hungarian and Jordanian students (in this sample) differ, we wanted to expand our understanding of these differences and analyse their test-taking behaviour. There were significant differences in using a theoretically effective exploration strategy in both samples based on the results of the logfile analyses. In total, 93% of Hungarian university students used a theoretically effective strategy, compared to 44% of Jordanian university students. This supports our earlier discussion regarding the less complex problems in the test that the majority of Hungarian students belong to the expert problem solvers. There were also differences in the percentages of using a theoretically effective strategy and having high CPS performance. It was 60.6% on average in the Hungarian sample and 44.4% among the Jordanian students. That is, there were large differences between the two samples not only in the efficacy of their interpretation of the extracted information, but also in the suitability of the

exploration strategy they employed, resulting in large significant differences in their final CPS performance.

Beyond the effectiveness of the exploration strategies applied in the CPS environments, there were significant differences identified in students' test-taking behaviour in relation to the time spent on the task and the number of trials internationally. Our findings confirmed Eichmann et al. (2019) and Goldhammer et al. (2014) research findings that low-achieving students generally interact with the problem less than high achievers; that is, the amount of exploration (number of clicks) are positively correlated with the CPS achievement. Students' performance improved significantly when they spent more time on a CPS task (Alzoubi et al., 2013; Goldhammer et al., 2014).

We identified other significant behavioural differences between Hungarian and Jordanian students, which differences were becoming smaller as the task complexity increased. During the test-taking process, Hungarian students spent less and less time attempting fewer and fewer trials despite the fact that the tasks were becoming more complex in comparison to Jordanian students, who spent nearly the same time and used nearly the same number of trials throughout the test. This tendency indicated that Hungarian students became increasingly aware of their effective exploration behaviour and required fewer trial-and-error moves and trials.

We used a more person-centered approach to see any other CPS-related differences between the two samples and look for more detailed explanations for the previously observed tendentious differences between high and low CPS achievers in the two cultures and beyond. To characterize students' exploration strategies in a CPS context, we employed latent class analyses based on the level of the optimal exploration strategy. Four latent classes have been identified in both samples. In both samples, the classes of non-performing explorers and restarting slow learners were nearly identical, indicating differences in the behaviours of Jordanian and Hungarian students. There was a rapid learner in the Hungarian sample, which did not exist in the Jordanian sample. Rapid learners revealed a noticeable learning curve when working on the CPS tasks. By the sixth problem having no eigendynamic on the test, they had reached the same level as proficient explorers in terms of exploration behaviour. They have the ability to quickly and flexibly adapt to the expectations of a given situation (see Greiff et al., 2018). Instead of rapid learners, in the Jordanian sample we could have identified a class of non-persistent explorers. These students were able to apply the partial variation strategy on the easiest problems, but they couldn't apply it to the more difficult ones.

Finally, behavioural differences between the top explorer groups — Hungarian vs. Jordanian – were identified. The proficient explorers in the Hungarian sample appeared to have more explicit schemata (see Greiff et al., 2018). As a result, they were able to use the optimal exploration strategy throughout the CPS tasks, regardless of their complexity, whereas the Jordanian students' schemata proved to be less well-founded and transferable, regardless of the complexity of the CPS scenario. However, students in this group, like the rapid learners in the Hungarian sample, were able to learn quickly and adapt to a given situation flexibly and rapidly. In Jordan and Hungary, the proportion of students in different class profiles varied significantly. Confirming previous research findings on time-on-task (Greiff et al., 2018), both the rapid learners and restarting slow learners in CPS problems may have varying degrees of general cognitive schemata. They may adjust rapidly and flexibly to the needs of a given situation, or they may adapt slowly and less flexibly. This adaptation needs time to take effect. Non-performing explorers who were unmotivated in the test-taking process spent less time solving the problem than proficient explorers who were conscious of their strategy use.

The number of trials revealed different patterns and was not strongly correlated to the time on task in the CPS test. In both samples, time on task increased with the amount of an optimal strategy usage; nevertheless, students' exploration profiles were a stronger predictor of the expected number of trials than time on task or final achievement.

Limitations

This study used the MicroDYN approach to assess students' problem-solving skills. Problems with the MicroDYN approach scenarios do not cover all types of problems and complex systems faced in everyday life. Thus, problem-solving behaviour detected in problem scenarios developed with the MicroDYN approach cannot be generalized to all types of complex problems we face in daily life. On the other hand, their special features make it possible to track students' learning processes and potential during the problem-solving process. Similarly, there is an efficient exploration strategy for problems with a small number of variables and relations, such as MicroDYN problems. However, optimal exploration strategies do not apply to complex everyday problems, as Funke (2021) observed. One of the limitations of the present cross-national comparison study is the relatively considerable differences in sample size and differences in gender distribution and the study year for students since only first-year students in Hungary took part in the assessment. As a part of the Matura examination, students in various study years participated in Jordanian sample. Other factors to consider as differences are parental education, and financial status (e.g., the number of books in the home), as well as

differences in the subjects studied by undergraduates (in Jordan, the sample not cover all areas of study, while in the Hungarian sample, students from an entire university participated in the assessment. Compared to the Hungarian sample, the Jordanian sample is relatively small, which may be considered a limitation in the validity of the test and the generalizability of the results at the population level.

Some initial trends can be identified based on the current findings, which can serve as a strong foundation for further large-scale empirical studies on Jordanian students' exploration behaviour in a CPS environment, focusing on comparing problem-solving behaviour among students from different cultural backgrounds.

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 ICCE 2013, 21st International Conference on Computers in Education.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century* skills (pp. 143–230). New York: Springer. https://doi.org/10.1007/978-94-007-2324-5_4

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research, 68*(2), 179-201

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, 128, 1–12.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning*, 7(1), 69–89.

Funke, J. (2021). It requires more than intelligence to solve consequential world problems. *Journal of Intelligence*, 9(3), 38.

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.

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*(3), 356–382. 10.1080/13546783.2014.989263

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 & Education*, 126, 248–263.

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.

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Log-file analyses. *Frontiers in Psychology, 9*, 302.

Organisation for Economic Co-operation and Development. (OECD) (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Volume V). Paris: OECD. https://doi.org/10.1787/9789264208070-5-en

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science*, 20(1), 75–100.

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*, 1–24. https://doi.org/10.1007/s10212-020-00516-y

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

Wüstenberg, S., Greiff, S., Molnar, G., & Funke, J. (2014): Determinants of cross-national gender differences in complex problem solving competency. *Learning and Individual Differences, 29*, 18–29.

DECLARATION

I hereby certify that the content of this dissertation is my original work of production. This dissertation has not been submitted for any other degree previously or at any other educational institution.