Heliyon 8 (2022) e08775

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

How we explore, interpret, and solve complex problems: A cross-national study of problem-solving processes

Gyöngyvér Molnár [a,*], Saleh Ahmad Alrababah [b], Samuel Greiff [c]

[a] Institute of Education, University of Szeged, MTA-SZTE Digital Learning Technologies Research Group, Hungary
[b] Doctoral School of Education, University of Szeged, Hungary
[c] Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg

HIGHLIGHTS

- CPS was measured equivalently in the sample of Jordan and Hungary university students.
- The development of CPS skills was not universal in the two samples.
- Students in the two samples showed different test-taking behaviors.
- Different types of exploration profiles were identified in the two samples.
- Process indicators in the two samples were non-invariant across the different latent profiles.

ARTICLE INFO

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

ABSTRACT

Complex problem solving (CPS) is considered an important educational outcome in the 21st century. Despite its importance, we have only little only knowledge of its measurability, development, or comparability in some countries, in particular in those with a short history of computer-based assessment. The results of the current study provide insights into the validity of CPS measurements and shed light on the different behavioral patterns and test-taking behavior in two convenience samples with different sample characteristics 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 identified differences in students' test-taking behavior in terms of the effectiveness of their exploration strategy, time-on-task, and number of trials. 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 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 profiles. This study contributes to our understanding of how students from different educational contexts behave while solving complex problems.

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, there 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.

* Corresponding author.
E-mail address: gymolnar@edpsy.u-szeged.hu (G. Molnár).

https://doi.org/10.1016/j.heliyon.2022.e08775
Received 2 July 2021; Received in revised form 15 September 2021; Accepted 12 January 2022

Complex problem solving (CPS)[1] is among the most extensively studied areas among the problem-solving skills in educational context over the past few decades (Csapó and Molnár, 2017; Greiff et al., 2013a, b; Greiff et al., 2015a,b; Dörner and 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 et al., 2012). CPS tasks focus on domain-general processes and disregard the role of content knowledge and rote learning (see Funke, 2001; Funke and Frensch, 2007; Greiff et al., 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 et al., 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 et al., 2013b; Molnár 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 et al., 2013), it has become essential to understand the fine mechanisms of CPS, especially to understand the reasons students' behaviors may lag behind in CPS performance and 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., 2015b; Molnár and Csapó, 2018; Mustafic et al., 2019; OECD, 2014a; Wu and 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 students with different cultural backgrounds and learning experiences may 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 little 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 and students don't have as much access to computers and associated learning experiences as in some other countries. 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 and Papafragou, 2012; Landau et al., 2010); it is a type of tool that influences human representational resources (Ünal and Papafragou, 2018). Thus, language-related factors can cause invariance in measuring rroblem solving. In addition, cultural mindset can also influence problem-solving (Arieli and 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 and 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 and Bilsky, 1990), where group performance is more important than individual task performance (Hofstede and 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). Further, the problems that are used here are likely to be more focused much on the task and on individual performance. Countries also differ to the extent in which they teach, for instance, problem solving at school, which in turn might lead to distinct differences between countries. All these cultural, national, linguistic, and educational factors may have a powerful effect on students' skills, resulting in potential 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 (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 in 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. In 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

[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 and Funke, 2017). In the present paper, we use the most widely used modifier, complex.

modeling framework. Wüstenberg et al. (2014) investigated and showed measurement invariance of CPS between Hungarian and German 8th–11th-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 mean 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 samples.

Research Question 1 (RQ1): Do Jordanian and Hungarian students in the samples interpret CPS problems the same way? Thus, is CPS measurement-invariant across our samples of Jordanian and Hungarian university students?

Several studies have shown that students with different educational and cultural backgrounds perform differently in CPS environments (Greiff et al., 2015b; OECD, 2014a; Wu and 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 and Molnár, 2017; Greiff et al., 2013b, 2018; Molnár and Csapó, 2017; Wu and 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 samples, 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 in our samples of 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 samples (in RQ1) and that students in our two subsamples from Hungary and Jordan may differ both in their CPS test scores (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; Molnár and Csapó, 2018), final scores may conceal true developmental and behavioral differences as regards CPS. For example, students with average achievement can engage in different behavior patterns. For instance, (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 et al., 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 two convenience samples of 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 in our samples 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 samples under examination. That is, we want to investigate whether there are different profiles and whether there are other CPS-related differences. 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 samples (Tóth et al., 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., Eichmann et al., 2020; Greiff et al., 2018; Greiff et al., 2015b; Molnár and Csapó, 2018; Mustafic et al., 2019; Stadler et al., 2020; Wu and 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 6th–8th-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" was not found 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 student samples? Are there differences in the types of profiles that emerge from the two samples?

Please note that unlike large-scale assessments that often use representative samples, the samples in this study are not representative and also not directly comparable (see participant description). To this end, this study (and the four research questions therein) does not allow to compare Jordanian and Hungarian students on a general level and interpretation is restricted to the two samples on the background of their different characteristics (see below for a description of these differences). However, given the scarcity of data in this area, we believe that this study has knowledge to add to the current state of the literature.

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 important 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. Nonetheless, 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 did not differ in both samples. Parents' educational level, number of books, and available ICT infrastructure at home proved to be higher in the Hungarian than in the Jordanian sample (see Table 1). These differences need to be taken into account when interpreting the results because the two samples differ on important background variables. These differences could be due to different learning experiences as well as due to differences in sample composition. This needs to be kept in mind when interpreting the findings of this study.

3.2. Instrument

A complex problem-solving test based on the MicroDYN approach was administered in both samples. 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

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.

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 and Funke, 2010) as many times as they liked within 180 s 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., 2013b). 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 s 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., 2013b).

In the second part of each of the problems, in what is called the knowledge application phase (Greiff et al., 2013b), students were expected to reach the given target values of the output variables within a given time frame (90 s), 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 researchers and translators.

The CPS approach and the CPS tasks have been employed extensively at both the national and international levels (see Csapó and Funke, 2017; Eichmann et al., 2020; Greiff et al., 2013b; Greiff et al., 2015b; 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 and 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 Schult et al. (2017) results (even though collected in German samples that might not be directly comparable to Jordan), individual online testing of complex problem-solving might have (to a small extent) 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 and Molnár, 2019). After entering the eDia system, students had 60 min to solve the problems and complete the related questionnaire. They received immediate feedback on their average achievement after test completion.

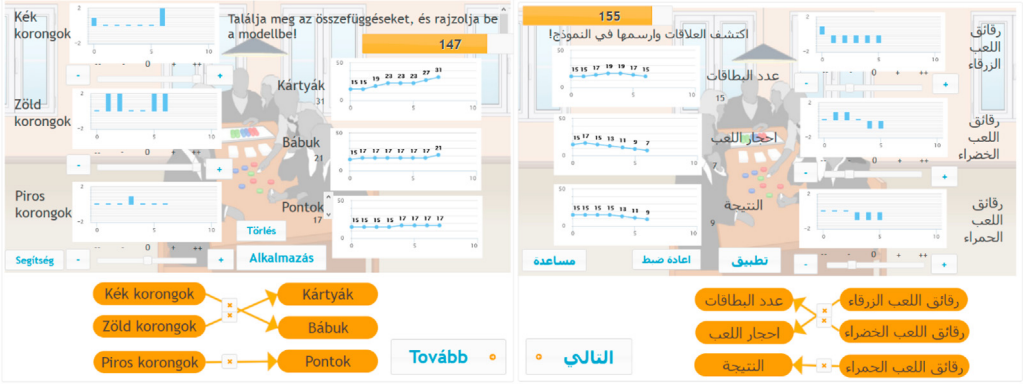

Figure 1. Screenshot of the MicroDYN task "Game Night." See the original version of the task in Greiff et al., (2013b). The controllers of the input variables range from "- -" (value = -2) to "++" (value = +2). They are presented on the left side of the problem environment in the Hungarian-language version 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).

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 min 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 as outlined above. 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 on the basis of the sample characteristics. 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 samples. 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.

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 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.

3.3.5. Analyses

Multi-group confirmatory factor analysis was used to test measurement invariance between the two samples (RQ1). Weighted least squares, mean- and variance-adjusted (WLSMV) estimation, and THETA parameterization were employed in the analyses (Muthén and 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

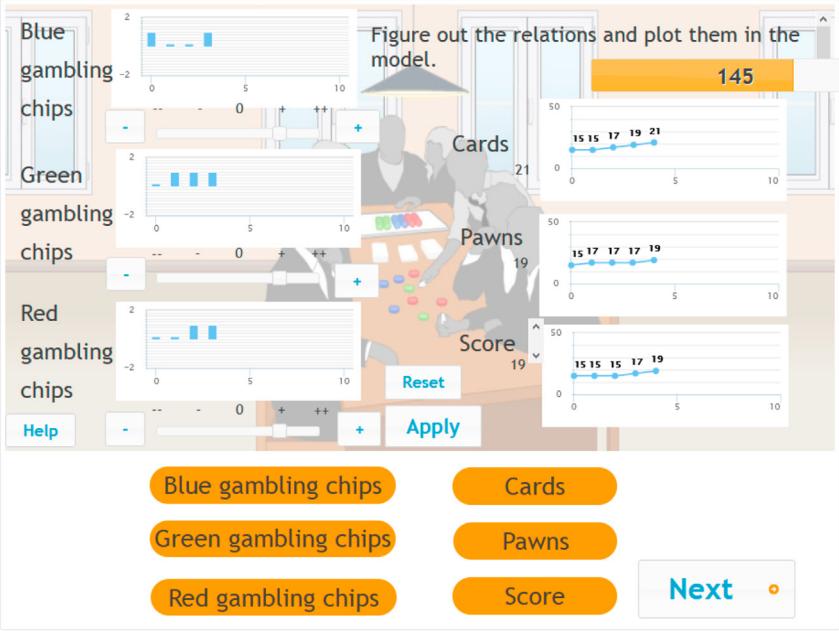

Figure 2. Demonstrating the meaning of a trial within the "Game Night" problem (English-language version of the task presented in Figure 1). The instruction for the task: Your friends invite you to a game night. They show you an interesting game you do not know the rules to. Find out how the blue, green,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

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 and George, 1995; Yoon and Lai, 2018), we evaluated measurement invariance from a practical perspective (see Chen, 2007; Cheung and Rensvold, 2002; Meade et al., 2008; Putnick and Bornstein, 2016; Rutkowski and 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 (RQ2, 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 sample were set to zero so it could serve as a reference group during the analyses, and the latent means of the Jordanian sample 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 samples, 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 and Lanza, 2009; Tein et al., 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 in our samples, 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 the Hungarian and Jordanian samples. 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 samples (Jordanian sample: $\alpha_{\_ph1} = .842$, $\alpha_{\_ph2} = .719$; Hungarian sample: $\alpha_{\_ph1} = .858$, $\alpha_{\_ph2} = .750$; 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 ($\alpha = .921$ 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 ($\alpha = .950$ and 946, respectively; see Table 2). That

Table 2. Reliabilities of the CPS test in the Jordanian and Hungarian-language contexts with and without the use of log data.

Type of data	Jordanian	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

is, the data proved to be reliable at both the test and phase levels. Please note that at this point the different scores do not 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.

4.2. Results for research question 1 (RQ1): Do Jordanian and Hungarian students in the samples interpret CPS problems the same way? Thus, is CPS measurement invariant across our samples of Jordanian and Hungarian university students?

To tackle RQ1, we investigated measurement invariance across the Jordanian and Hungarian samples. The baseline model with the two latent CPS factors (knowledge acquisition and knowledge application) fitted the data well in both samples (Jordan: $\chi^2 = 369.02$, df = 186, CFI = .975, TLI = .972, RMSEA = .044; Hungary: $\chi^2 = 865.53$, df = 186, CFI = .980, TLI = .978, RMSEA = .045). As can be seen in the results below, CPS can be measured invariantly across 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 sample (i.e., Hungarian or Jordan) to which they belong (Millsap, 2012); that is, the measure is not biased against either of the groups.

4.3. Results for research question 2 (RQ2): Can we find developmental differences in CPS skills in our samples of 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., 2013b) 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 and 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 samples 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 (of note, not between the populations; the interpretation is sample-based). As regards latent mean differences across the samples, 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 3. Goodness of fit indices for testing invariance of CPS across the two samples.

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

Table 4. Cross-sample 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 (RQ3): What kind of test-taking behaviors do Jordanian and Hungarian university students in our samples 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,

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 (0.8).4	7.6

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

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 samples 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 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 et al., 2017; Lotz et al., 2017).

This pattern was somewhat different in the Jordanian sample (see Table 5). Only half of the students (44%) used 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 s exploring the problem, while the Hungarian students spent more time on exploration (56 s). 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.

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 per task, while the number of trials among the Hungarian students was more than five on average. The Hungarian students' time-on-task data and number of trials data were consistent with each other, while this was not necessarily the case in the Jordanian sample.

4.5. Results for research question 4 (RQ4): Based on the exploration strategy (i.e., VOTAT), which profiles can be extracted from the two samples of Jordanian and Hungarian students? Are there differences in the types of profiles that emerge from the two samples?

To tackle RQ4, 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.

Table 6. Cross-sample differences in students' test-taking behavior: time-on-task and number of trials.

Complexity of problem	Jordanian sample			Hungarian sample					
	Low achievement	High achievement	Mean	Low achievement	High achievement	Mean	t	p	d
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

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 samples – 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.

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 increasingly 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).

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 sample						
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 sample						
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.

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). Please note that despite the same labels, the classes between the two samples (as displayed in Table 8) are not directly comparable.

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.

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 (here rougly considered as ordinal variable) 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 RQ3 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 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

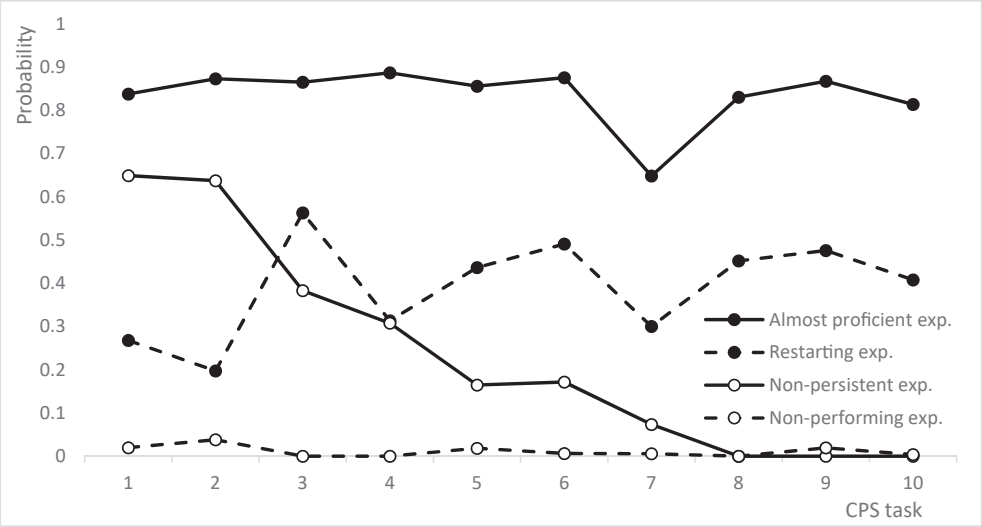

Figure 3. Four qualitatively different class profiles in the Jordanian sample.

Table 8. Relative frequencies and average latent class probabilities in the Jordanian and Hungarian-language samples.

Profiles	Jordanian		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.

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.

5. Discussion

This study shows that complex problem-solving can be measured validly, reliably, and equivalently in the Hungarian and Jordanian university contexts and samples derived therein. It provides important insights into the international validity of CPS measurements and sheds light on the different behavior patterns of two samples 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,

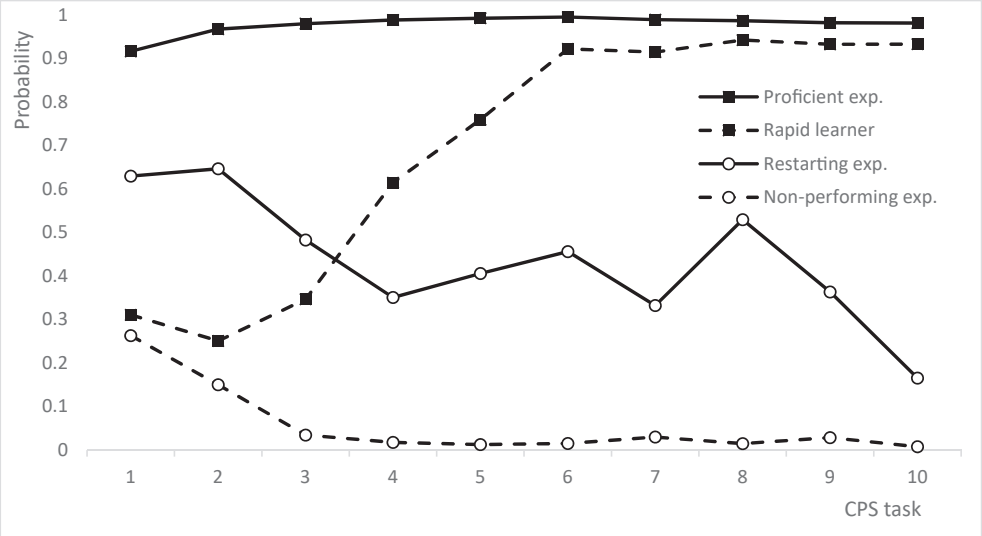

Figure 4. Four qualitatively different class profiles in the Hungarian sample.

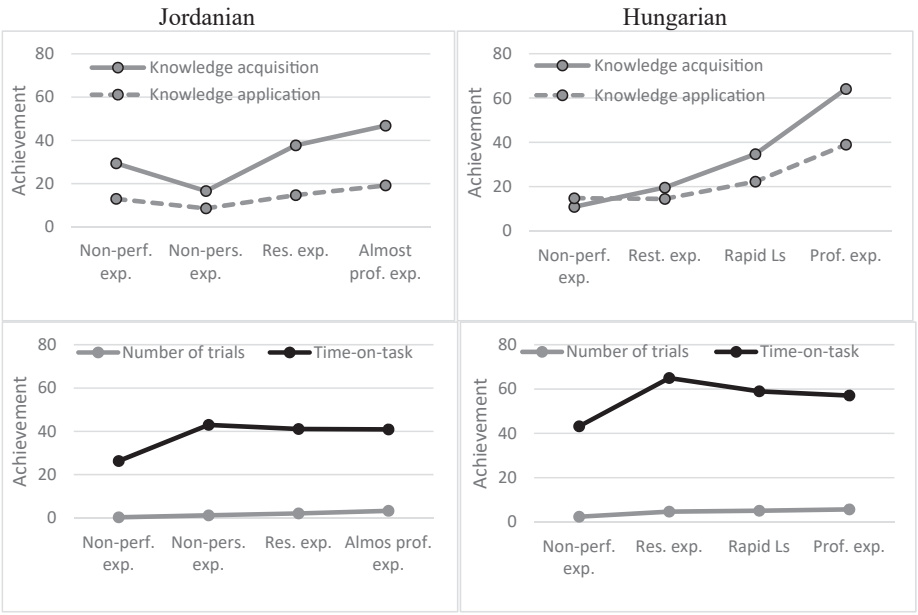

Figure 5. Performance and test-taking behavior among students with different latent class profiles in the two samples. (We have connected the data points to visualize the tendencies.)

country, or sample are not necessarily generalizable to other countries, cultures, or samples 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.

Research Question 1 (RQ1): Do Jordanian and Hungarian students in the samples interpret CPS problems the same way? Thus, is CPS measurement-invariant across our samples of Jordanian and Hungarian university students?

We found invariance in CPS measurement across the Jordanian and Hungarian university student samples; that is, both samples 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 cultural and educational differences, which can influence measurement invariance, it is possible for CPS to be measurement-invariant across 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 and 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 and 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 (as was partly already confirmed by the PISA study), at least to the extent of the countries that have been involved in such studies.

Research Question 2 (RQ2): Can developmental differences be identified in CPS skills in our samples of Jordanian and Hungarian university students? If so, 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 sample, which is consistent with our expectations given different background characteristics of the two samples. Please also note that earlier research results have indicated that students with different educational and cultural backgrounds can perform differently in a CPS environment (see Greiff et al., 2015b; OECD, 2014a; Wu and 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 in light of the differing sample characteristics in this study.

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.

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, the Hungarian students' mean performance also dropped at the same level; it was only the starting level 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).

Substantial reasons for achievement differences (assuming they hold and can be replicated also in representative samples) may lie in educational differences as well as in the experience of computer use in educational context. The educational use of computers has long been addressed in the majority of 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 RQ3 and RQ4).

Research Question 3 (RQ3): What kind of test-taking behavior do Jordanian and Hungarian university students in our samples 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?

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 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 in the sample 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 and 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 differences in final CPS achievement. Please again note differences in sample composition that do not allow to draw generaliable conclusions between student populations.

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 sample level, we confirmed Eichmann et al. (2019) and Goldhammer et al. (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 grew smaller compared to the increasing complexity of the tasks. This tendency was mainly driven by the Hungarian students, who spent generally increasingly less time attempting increasingly 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. 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 et al., 2015), which Goldhammer et al. (2014) concluded was due to the lack of a common definition of time-on-task and achievement.

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

In RQ2 and RQ3, we found several sample-level behavioral differences. In RQ4, we used a more person-centered approach to investigate 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 different cultures.

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 (through separate analyses). The classes of non-performing explorers and restarting slow learners proved to be almost identical in the two samples on a descriptive level, indicating existing differences between the behaviors of Jordanian and Hungarian students. Our study confirmed Molnár (2021) result on the presence of rapid learners in the Hungarian university sample, which was not found in this particular 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). 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 or unwilling (as motivation could also be a good explanation for achievement differences) 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. 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. However, this type of problem is 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. In particular in samples with a somewhat lower experience with technology and lower access to it, this may pose distinct disadvantages. 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 are difficult to experience 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 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, which may be particularly prone to learn problem solving at university), 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 RQ2, and restrict the generalizability of the results beyond the specific samples (i.e., on a population level). That is, we have two convenience and non-representative samples that differ in several features and, thus, cannot be directly compared. With this in mind, we consider our findings a first starting point for generating new hypotheses and for identifying initial patterns, which can form a starting point for further large-scale empirical studies on CPS in different groups, for instance, students with different cultural backgrounds, different access to technology, and different learning experiences.

7. Conclusions

The results of the current study provide important insights into the validity of CPS measurements and shed initial light on the different behavior patterns and test-taking behaviors in two samples 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.

Declarations

Author contribution statement

Gyöngyvér Molnár: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Saleh Ahmad Alrababah: Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Samuel Greiff: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

Gyöngyvér Molnár was supported by nemzeti kutatási, fejlesztési és innovációs alap (OTKA K135727).

Samuel Greiff was supported by fonds national de la recherche luxembourg (CORE 'TRIOPS').

Data availability statement

Data will be made available on request.

Declaration of interests statement

SG is one of two authors of the commercially available COMPRO-test that is based on the multiple complex systems approach and that employs the same assessment principle as MicroDYN. However, for any research and educational purpose, a free version of MicroDYN is available. SG receives royalty fees for COMPRO.

Additional information

No additional information is available for this paper.

References

Al Suwaidi, Mohammed, 2008. When an Arab executive says "Yes": identifying different collectivistic values that influence the Arabian decision-making process. Mast. Sci. Organiz. Dynam. 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. J. Exp. Psychol. Gen. 147 (6), 789–814.

Beckmann, J.F., Birney, D.P., Goode, N., 2017. Beyond psychometrics: the difference between difficult problem solving and complex problem solving. Front. Psychol. 8, 1739.

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

Byrne, B.M., Stewart, S.M., 2006. The MACS approach to testing for multigroup invariance 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. Struct. Equ. Model. 14, 464–504.

Cheung, G.W., Rensvold, R.B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct. Equ. Model. 9, 233–255.

Collins, L.M., Lanza, S.T., 2009. Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences, 718. John Wiley & Sons.

Csapó, B., Funke, J., 2017. The Nature of Problem Solving. OECD, Paris.

Csapó, B., Molnár, G., 2017. Potential for assessing dynamic problem-solving at the beginning of higher education studies. Front. Psychol. 8, 2022.

Csapó, B., Molnár, G., 2019. Online diagnostic assessment in support of personalized teaching and learning: the eDia system. Front. Psychol.

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. Front. Psychol. 8, 1153.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., Naumann, J., 2019. The role of planning in complex problem solving. Comput. Educ. 128, 1–12.

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., Goldhammer, F., 2020. Exploring behavioural patterns during complex problem-solving. J. Comput. Assist. Learn. 36 (6), 933–956.

Fischer, A., Greiff, S., Funke, J., 2012. The process of solving complex problems. J. Prob. Solv. 4, 19–42.

Funke, J., 2001. Dynamic systems as tools for analysing human judgement. Think. Reas. 7 (1), 69–89.

Funke, J., 2014. Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. Front. Psychol. 5, 739.

Funke, J., Frensch, P.A., 2007. Complex problem solving: the European perspective – 10 years after. In: Jonassen, D.H. (Ed.), Learning to Solve Complex Scientific Problems. Erlbaum, New York, pp. 25–47.

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

Gleitman, L., Papafragou, A., 2012. New perspectives on language and thought. In: Holyoak, K., Morrison, R. (Eds.), The Oxford Handbook of Thinking and Reasoning, second ed. Oxford University Press, New York, NY, pp. 543–568.

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. J. Educ. Psychol. 106 (3), 608–626.

Greiff, S., Funke, J., 2010. Systematische Erforschung komplexer Problemlösefähigkeit an hand minimal komplexer Systeme. Z. für Padagogik 56, 216–227.

Greiff, S., Fischer, A., Stadler, M., Wüstenberg, S., 2015a. Assessing complex problem-solving skills with multiple complex systems. Think. Reas. 21 (3), 356–382.

Greiff, S., Holt, D.V., Funke, J., 2013a. Perspectives on problem solving in cognitive research and educational assessment: analytical, interactive, and collaborative problem solving. J. Prob. Solv. 5, 71–91.

Greiff, S., Wüstenberg, S., Avvisati, F., 2015b. Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012. Comput. Educ. 91, 92–105.

Greiff, S., Wüstenberg, S., Funke, J., 2012. Dynamic problem solving: a new assessment perspective. Appl. Psychol. Meas. 36 (3), 189–213.

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., Csapó, B., 2013b. Complex problem solving in educational contexts – something beyond g: concept, assessment, measurement invariance, and construct validity. J. Educ. Psychol. 105 (2), 364–379.

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., Csapó, B., 2018. Students' exploration strategies in computer-simulated complex problem environments: a latent class approach. Comput. Educ. 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. Comput. Hum. Behav. 61, 36–46.
Güss, C.D., Tuason, M.T., Gerhard, Ch., 2010. Cross-national comparisons of complex problem-solving strategies in two microworlds. Cognit. Sci. 34.
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., Nuñez, 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.
Molnár, 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.
Molnár, G., Greiff, S., Csapó, B., 2013. Inductive reasoning, domain specific and complex problem solving: relations and development. Think. Skills Creativ. 9 (8), 35–45.
Molnár, G., Csapó, 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.
Molnár, G., Csapó, 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.
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. Dev. Psychol. 55 (10), 2090–2101.
Muthén, L.K., Muthén, B.O., 2012. Mplus User's Guide, seventh ed. Muthén and Muthén, 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.
Román, A., Flumini, A., Lizano, P., Escobar, M., Santiago, J., 2015. Reading direction causes 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., Hautamäki, J., 2015. Exploring the relation between time on task and ability in complex problem solving. Intelligence 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.
Tóth, K., Rölke, H., Goldhammer, F., Barkow, I., 2017. Educational process mining: new possibilities for understanding students' problem-solving skills. In: Csapó, 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., Molnár, 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.