



ARTICLE

Exploring behavioural patterns during complex problem-solving

Beate Eichmann¹ | Samuel Greiff² | Johannes Naumann³ | Liene Brandhuber⁴ | Frank Goldhammer¹

¹Centre for International Student Assessment (ZIB), Educational Quality and Evaluation, DIPF Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany

²Cognitive Science and Assessment, Department of Behavioral and Cognitive Sciences, University of Luxembourg, Esch-sur-Alzette, Luxembourg

³School of Education, Institute of Educational Research, University of Wuppertal, Wuppertal, Germany

⁴Institut für Allgemeine Erziehungswissenschaften, Goethe University Frankfurt, Frankfurt am Main, Germany

Correspondence

Beate Eichmann, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt am Main, Germany.
Email: beate.eichmann@dipf.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Numbers: 01LSA1504A, 01LSA1504B; Fonds National de la Recherche Luxembourg, Grant/Award Number: The Training of Complex Problem Solving; "TRIOPS"

Peer Review

The peer review history for this article is available at <https://publons.com/publon/10.1111/jcal.12451>.

Abstract

In this explorative study, we investigate how sequences of behaviour are related to success or failure in complex problem-solving (CPS). To this end, we analysed log data from two different tasks of the problem-solving assessment of the Programme for International Student Assessment 2012 study ($n = 30,098$ students). We first coded every interaction of students as (initial or repeated) exploration, (initial or repeated) goal-directed behaviour, or resetting the task. We then split the data according to task successes and failures. We used full-path sequence analysis to identify groups of students with similar behavioural patterns in the respective tasks. Double-checking and minimalistic behaviour was associated with success in CPS, while guessing and exploring task-irrelevant content was associated with failure. Our findings held for both tasks investigated, from two different CPS measurement frameworks. We thus gained detailed insight into the behavioural processes that are related to success and failure in CPS.

KEYWORDS

complex problem-solving, exploration, log data, PISA, sequence analysis

1 | INTRODUCTION

Our world and our society are becoming increasingly complex. In particular the fast development of technology confronts people more and more frequently with challenges in dealing with unknown situations (e.g., installing a smart TV, using driving assistance technology,

using a new app on the smartphone) (Autor, Levy, & Murnane, 2003). Also in non-technical contexts complexity increases. Globalization connects people all around the world, leading to global markets and organizations in which interests of more interdependent parties have to be managed than in small, local structures (Wilpert, 2009). The ability to cope with novel situations of these kinds is addressed in

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *Journal of Computer Assisted Learning* published by John Wiley & Sons Ltd.

research on complex problem-solving (CPS). One common definition of CPS that we use in this article is put forward by Frensch and Funke (1995, p. 36):

CPS occurs to overcome barriers between a given state and a desired goal state by means of behavioural and/or cognitive, multistep activities. The given state, goal state, and barriers between given state and goal state are complex, change dynamically during problem-solving, and are intransparent. The exact properties of the given state, goal state, and barriers are unknown to the solver at the outset. CPS implies the efficient interaction between a solver and the situational requirements of the task, and involves a solver's cognitive, emotional, personal, and social abilities and knowledge.

The skill to solve complex problems can be regarded as a so-called 21st century skill—a skill becoming increasingly relevant for both work and private life in the 21st century (Binkley et al., 2012). Therefore, the question arises which “multistep activities” (Frensch & Funke, 1995, p. 36) and behavioural patterns, respectively, lead to success or failure in CPS. Gaining knowledge about crucial processes in CPS is fundamental to making students better problem solvers and prepare them for the challenges of the future. Previous studies investigated the effects of different behaviours on success in CPS (Eichmann, Goldhammer, Greiff, Pucite, & Naumann, 2019; Greiff, Niepel, Scherer, & Martin, 2016; Schult, Stadler, Becker, Greiff, & Sparfeldt, 2017; Sonnleitner, Brunner, Keller, & Martin, 2014). However, these studies mostly used single unit measures of CPS behaviour, for example, they investigated how the occurrence or frequency of certain behaviours or the time spent on (parts of) a task is related to success. Extending these lines of research, in the present study we want to get more comprehensive insights into the relation between behavioural characteristics and success in CPS by considering complete sequences of certain behaviours. Thereby we want to identify effects of patterns or compositions of behaviours. To achieve this, we will use full-path sequence analysis to identify and group together similar behavioural sequences. According to the definition by Frensch and Funke (1995) mentioned above, multistep interactions (i.e., sequences of behaviour) are key to CPS.

1.1 | Assessment of CPS behaviour

CPS research regularly employs computer simulations of real world problems (e.g., handling an MP3 player, OECD, 2013) to assess CPS skills. Since in computer-based assessment the interactions of participants with the assessment system are recorded, this behavioural data can then be analysed, and inferences about relations between behaviour and success in CPS can be made. There are two widely used CPS assessment frameworks that are explained in detail in the method section of this article (Funke, 2001; Greiff, Fischer, Stadler, & Wüstenberg, 2001). The advantage of these formal frameworks is that

they make results from different studies comparable and allow for a systematic description of the underlying task structures (Greiff, Wüstenberg, & Funke, 2012). CPS tasks built under both frameworks are usually highly interactive and lead to rich log data. In log data the behaviour of the problem solver while working on a task is stored. Therefore, this data enables the investigation of behaviour while solving complex problems. In the following sections, we will describe and discuss the different approaches and results of previous research on CPS log data.

1.2 | Top-down approaches to investigate CPS log data

As mentioned before, previous research often used single unit measures to investigate the relation between behaviour and success in CPS (Kroehne & Goldhammer, 2018; Naumann, Goldhammer, Rölke, & Stelter, 2014; Richter, Naumann, & Noller, 2003). In this top-down approach, theory-driven hypotheses about the relations between certain behaviours and success in CPS are formulated. These behavioural states are then identified by events that are included in the log data (Kroehne & Goldhammer, 2018). Single unit measures rely on single events and do not take into account sequential information (Richter et al., 2003). They are derived, for instance, by determining the (cumulated) time in a certain state or the frequency of a certain event or type of event. Naumann et al. (2014) investigated the relation between the number of interactions and success in technology-based problem-solving. They assumed that in everyday technology-based problems most people not solving the problems behave too passive (rather than too active). Their results revealed that low achieving students indeed often show too little interaction with the problem at hand. Moreover, several studies showed a positive relation between the amount of exploration and success in CPS (Dormann & Frese, 1994; Eichmann, Goldhammer, Greiff, Brandhuber, & Naumann, 2018). Exploration can serve several purposes in CPS. First, exploration can be required to gather necessary information to solve a problem. Second, exploration can be non-targeted; that is exploration of task-irrelevant information. For example, if the problem requires the problem solver to buy a subway ticket on a ticket machine, interacting with buttons for bus tickets would be regarded as non-targeted exploration, since these interactions are not directly goal-related. However, non-targeted exploration can serve the purpose of getting to know the problem space and can therefore support the problem solver to build a mental model of the problem (Dormann & Frese, 1994). Bell and Kozlowski (2008) found a positive relation between exploration and metacognitive activity. They argue that metacognitive activity also facilitates CPS. However, Bell and Kozlowski (2008) did not differentiate between exploration of necessary information task-irrelevant information. Greiff et al. (2016) found a low intervention frequency to be advantageous in CPS. Hence, they argue that CPS benefits from planned behaviour. Accordingly, Eichmann et al. (2019) showed that, especially in the beginning of a CPS process, taking time to plan ahead has a positive impact on

success. They also argued that in the course of CPS, time allocation into phases of higher and lower activity plays an important role, especially for difficult problems. A quite well investigated CPS strategy applicable to certain CPS tasks is the vary-one-thing-at-a-time strategy (VOTAT). The VOTAT strategy implies manipulating single variables, while all other possible input variables are kept constant. Thus, the influence of the manipulated variable on other variables can be investigated and knowledge about the problem can be generated, which has a positive effect on success in CPS (Greiff, Wüstenberg, & Avvisati, 2015; Tóth, Rölke, Greiff, & Wüstenberg, 2014; Wüstenberg, Greiff, Molnár, & Funke, 2014). The VOTAT strategy has also received much attention in research on science inquiry (Apedoe & Schunn, 2013; Jirout & Zimmerman, 2015). Problem solvers' use of the VOTAT strategy is usually also operationalized by count indicators that reflect whether and to what extent the strategy was used.

The top-down approach of using single unit measures to analyse log data has the advantage that theory-based assumptions about effects of behaviour can be investigated. Relations between single unit measures (frequencies or durations of single behaviours) and success in CPS can easily be investigated using the approach adopted in many of the studies mentioned before. However, this approach reduces behavioural sequences to single numbers. Through this reduction of data effects of combinations or sequences of behaviours (i.e., the exact order of actions) might be overlooked and important information might get lost. Of course, information reduction can also be useful to reduce noise in the data. However, Richter et al. (2003) argue that single unit measures might lead to similar measures for in fact very different behaviours since they do not take into account sequential information. Therefore, they recommend the (additional) use of sequential measures. The differences between the aforementioned top-down approaches and bottom-up approaches used with sequential measures are depicted in Figure 1. In the top-down approach, theory-based indicators are formed (e.g., single unit measures), in the bottom-up approach regularities in the data (e.g., clusters of behavioural sequences) are interpreted on the basis of theoretical assumptions. Bottom-up approaches used to analyse log data will be explained in more detail in the following section.

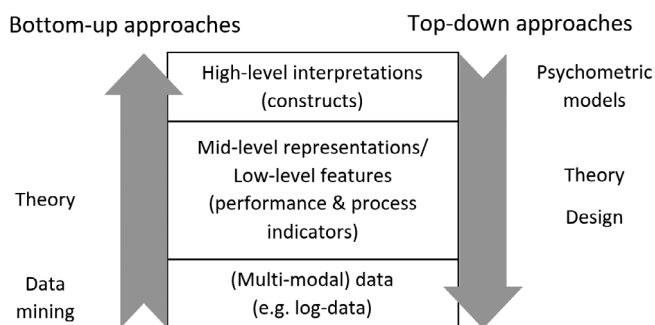


FIGURE 1 Comparison of top-down and bottom-up approaches to identify evidence for drawing inferences about a construct. *Source:* Adapted from Mislevy (2019, p. 35)

1.3 | Bottom-up approaches to investigate CPS log data

To overcome the limitations of single unit measures, in the recent literature methods to investigate (sub)sequences of behaviour were applied (He & von Davier, 2015; Stadler, Fischer, & Greiff, 2019). He and von Davier (2015) used the *n*-gram approach to identify patterns of behaviour related to success in CPS. The *n*-gram approach decomposes CPS behaviour into small subsequences and analyses the relation between the frequency of these subsequences and success in CPS. In this data-driven approach log data is analysed without prior hypotheses about specific behaviours. The substantive interpretation of the results will then take place a posteriori. He and von Davier (2015) found actions that were *not* part of the shortest path to success to be associated with *not* solving a complex problem, while more goal-directed actions were associated with solving it. They investigated all the possible actions in the investigated task as distinct behaviours obtaining 27 different behaviours, which were combined into 144 bigrams and 257 trigrams. Using the same approach Stadler et al. (2019) showed behaviour sequences were more likely to be related to success if they could be assumed to generate less cognitive load. They argue that cognitive load is lowest if the problem solver follows the direct path to the correct solution. In contrast to the study of He and von Davier (2015), Stadler et al. (2019) pre-coded the actions of their test takers according to two distinct behaviour categories (manipulating variables and annotating the observed results in a CPS task). They argue that the direct path to success (generating the least cognitive load) is characterized by annotating results immediately after every variable manipulation instead of performing several manipulations subsequently. Thus, still using a bottom-up approach they integrated the top-down element of using a (rather task-driven) coding for their data. Both studies showed goal-directed behaviour to be beneficial in CPS, while non-targeted exploration was found to be detrimental. However, this finding runs contrary to the results of Dormann and Frese (1994) and Eichmann et al. (2018) who reported a positive relation between exploration (both goal-directed and non-targeted) and success in CPS. Therefore, the question remains under which circumstances the respective effects arose.

The apparent ambiguity in these findings concerning the usefulness of exploration may be a result of the varying definitions of exploration in contrast to goal-directed behaviour. He and von Davier (2015) and Stadler et al. (2019) emphasize the shortest path to the successful solution to be beneficial. They did not consider repeated actions in this regard and therefore concluded parsimonious behaviour to be beneficial. Eichmann et al. (2018) also considered the shortest path to success as goal-directed. However, all interactions that went beyond this shortest path including repeated recapitulation (of goal-directed parts of the task) were regarded as exploration. Thus, the effects of goal-directed exploration (i.e., repeated goal-directed interactions) and non-targeted exploration could not be disentangled. However, investigating repeated goal-directed interactions and non-targeted exploration separately might be a way to clarify the usefulness of these behaviours in CPS. As Greiff, Molnár, Martin,

Zimmermann, and Csapó (2018) argue, not only the quantity but also the quality of exploration might be related to CPS performance. Moreover, using the full path of behaviour might help to clarify the usefulness of not only single instances of these behaviours but also of behaviour patterns that might be more complex than three or four subsequent actions, which is the maximum length of n -grams used by He and von Davier (2015) and Stadler et al. (2019).

The ambiguous results on the effects of exploration and goal-directed behaviour between studies using the n -gram approach and studies using single unit measures demand a method to clarify the role of exploration and goal-directed behaviour in CPS. By combining the advantages of theory-driven top-down approaches and data-driven bottom-up approaches we aim at taking an even closer look at behavioural processes in CPS.

1.4 | A top-down bottom-up mixed approach to investigate CPS log data

To take this closer look, in this study we apply full-path sequence analysis (Gabadinho, Ritschard, Müller, & Studer, 2011), an exploratory approach that does not only take into account sub-sequences of CPS but the whole behavioural sequence of every problem solver. In full-path sequence analysis, complete sequences (of behaviour) are clustered according to their similarity leading to clusters of similar CPS behaviours. With these clusters we hope to identify possibly heterogeneous behaviours, which alike might lead to either correct or false solutions in CPS. The aforementioned sequence analysis methods were originally used for comparing DNA sequences and use string matching algorithms to determine the similarity of sequences (Abbott & Forrest, 1986). There are different string matching algorithms available that take into account different attributes of the sequences to be compared. Comparable to the n -gram approach, sequence analysis methods can be applied to rather raw (see He & von Davier, 2015) or pre-coded (see Stadler et al., 2019) log data.

In this study, we use behavioural categories that have been shown to be relevant to success in CPS in previous research for coding our data. Therefore, we integrate the top-down approach of theory-driven single unit measures (through pre-coding) with the bottom-up approach of exploratory full-path sequence analysis. We distinguish non-targeted exploration behaviour, which has been shown to be positively related to success in CPS by Dormann and Frese (1994) and Eichmann et al. (2018), from goal-directed behaviour (including goal-directed exploration), which was found to be most positively related to success in the studies of He and von Davier (2015) and Stadler et al. (2019). According to Dormann and Frese (1994) exploration denotes metacognitive activities and helps building a mental model of the problem at hand. In contrast, goal-directed behaviour (as we defined it) reflects an efficient processing of the tasks content. Since from previous research it is unclear whether parsimony is beneficial for CPS performance (Eichmann et al., 2018; He & von Davier, 2015; Stadler et al., 2019), we also distinguish between initial and repeated actions. As Wirth (2004) argues, repeating actions could be an

attempt to integrate information that has been identified before. Repeating goal-directed actions could therefore reflect thoroughness, while repeating non-targeted exploration could reflect an overestimation of the relevance of the inspected information. Students' behaviour in a specific CPS task could therefore indicate general student characteristics such as perseverance or motivational states. Therefore, students' behaviour might not only predict success in this very task but might also be an expression of this student's overall CPS performance. Therefore, we want to investigate the relation between students' behaviour and both their performance in the very item, in which the behaviour was shown, and their overall CPS performance. Sequence analysis methods have the advantage that they take into account both the frequency of behaviours as well as the order of behaviours throughout the whole behavioural path (Gabadinho et al., 2011; Studer, Ritschard, Gabadinho, & Müller, 2011). Through this we hope to clarify the circumstances under which parsimony, non-targeted exploration, and goal-directed behaviour are beneficial for successful CPS.

1.5 | Hypotheses and research questions

Although Dormann and Frese (1994) and Eichmann et al. (2018) found positive effects of non-targeted exploration, we expect these effects to be possibly confounded with repeated goal-directed behaviour. Since He and von Davier (2015) argue that non-goal-directed behaviour should be detrimental, we expect the positive effect found by Dormann and Frese (1994) and Eichmann et al. (2018) to be due to repeated goal-directed actions being part of their measure of exploration. Therefore, our hypotheses are as follows:

Hypothesis 1 Non-targeted exploration (both initial and repeated) is more frequent among false CPS solutions than correct CPS solutions.

Hypothesis 2 Goal-directed behaviour (both initial and repeated) is more frequent among correct CPS solutions than false CPS solutions.

In addition, we want to explore the complex behaviour patterns that result in success or failure in CPS. Thus, we formulated the following research questions:

Research question 1 Which clusters of behavioural patterns (in terms of complete CPS behaviour sequences) are related to success or failure in CPS? To the best of our knowledge, previous research did not address the relation between complete CPS processing paths and success in the respective CPS tasks.

Research question 2 Can clusters of behavioural patterns and their relation to success in CPS be generalized across different task frameworks? The question of generalizability of these relations between behaviour sequences and success in CPS across different CPS task frameworks has not been addressed by previous research.

2 | METHOD

2.1 | Sample

We used data from the computer-based assessment of the Programme for International Student Assessment (PISA) 2012. In the PISA study, the competencies of 15-year-olds are assessed in several countries. We used the data of those students who worked on at least one of the two tasks for our analysis. The sample consisted of $N = 30,098$ students from 42 countries; 50.37% were female.

2.2 | Instruments

There are two widely used frameworks to measure CPS skills: Finite State Automata (FSA) and Linear Structural Equations (LSE) (Funke, 2001; Greiff, Fischer, Stadler, & Wüstenberg, 2014). FSA are characterized by a finite number of distinct states the system can attain. The problem solver can use a defined set of operators to switch between

these states. An example for such a system is a ticket machine on which the problem solver can use different buttons (=operators) to navigate through several options (e.g., daily ticket or individual trips) for buying tickets (=states).

Tasks from the LSE framework are based on a number of interrelated input (exogenous) and output (endogenous) variables. The relations between the variables are unknown to the problem solver. By manipulating the exogenous and observing the endogenous variables the relations between them can be investigated. An example for such a system is the control of room temperature and humidity (=endogenous variables) using the sliders of a climate control (=exogenous variables). These two frameworks can be used to design CPS tasks covering a wide range of real world problems within different fields of knowledge. Also, these frameworks allow for intentional, theory-driven manipulation of item difficulty (Stadler, Niepel, & Greiff, 2016).

To cover both the LSE and the FSA framework we chose one CPS task from either framework for analysis. Both tasks were released by the OECD. We used two tasks that provide prototypical instances for LSE and FSA type problems, respectively, and that had a comparable

FIGURE 2 The climate control task from PISA 2012 after the arrows in the diagram have been drawn. PISA, Programme for International Student Assessment [Colour figure can be viewed at wileyonlinelibrary.com]

number of minimum actions required to solve the task. In both tasks it was straightforward to distinguish between goal-directed behaviour and non-targeted exploration. From the LSE framework we chose the climate control task (see Figure 2). In this task, students were required to investigate the relations between exogenous and endogenous variables and then visualize these relations by drawing lines in a diagram. The exogenous variables were control sliders regulating the endogenous variables, which were temperature and humidity. The goal of this task was to obtain a diagram that correctly represents all existing relations between exogenous and endogenous variables. To obtain a correct solution in this task a minimum of six goal-directed actions was required. From the FSA framework, we chose the tickets task (see Figure 3). In this task, students were required to navigate through the states of a ticket machine to reach a desired goal state. The goal was to buy the cheapest ticket available considering particular requirements. Students had to compare and revisit several states to decide which ticket was the cheapest. To compare all relevant and find the correct ticket a minimum of seven goal-directed actions and one reset was required. We used the students' responses on all other 25 CPS tasks from the PISA 2012

assessment to determine their overall CPS skills. We did not use the plausible values for problem-solving from the PISA 2012 database for two reasons: First, the plausible values include both complex and analytical problem-solving performance. However, we only want to investigate complex problem-solving performance. Second, the plausible values also include the raw scores of the two items analysed in our study. Therefore, investigating the relation between performance in our two items and PISA's plausible values would lead to an overestimation of the relation between behaviour and performance. For details about the PISA problem-solving assessment see OECD (2013).

2.3 | Procedure

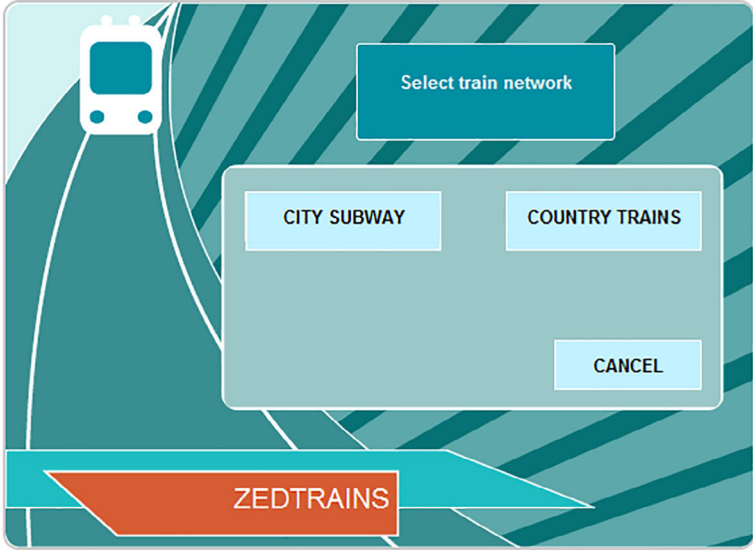
Problem-solving was part of the optional computer-based assessment in PISA 2012. In the participating countries, the computer-based assessment was carried out after the paper-based assessment. Students first received a tutorial to practice the required actions with the computer-based assessment environment to eliminate any effects of

TICKETS

A train station has an automated ticketing machine. You use the touch screen on the right to buy a ticket. You must make three choices.

- Choose the train network you want (subway or country).
- Choose the type of fare (full or concession).
- Choose a daily ticket or a ticket for a specified number of trips. Daily tickets give you unlimited travel on the day of purchase. If you buy a ticket with a specified number of trips, you can use the trips on different days.

The BUY button appears when you have made these three choices. There is a CANCEL button that can be used at any time BEFORE you press the BUY button.



Question 13: TICKETS CP038Q01

You plan to take four trips around the city on the subway today. You are a student, so you can use concession fares. Use the ticketing machine to find the cheapest ticket and press BUY. Once you have pressed BUY, you cannot return to the question.

?
➔

FIGURE 3 The tickets task from PISA 2012. PISA, Programme for International Student Assessment [Colour figure can be viewed at wileyonlinelibrary.com]

students' ICT skills. According to the booklet design of PISA, students received either two problem-solving clusters or one problem-solving cluster and one task cluster from a different assessment domain. Students were then given 20 min time to complete each computer-based cluster. Each problem-solving cluster contained four problem-solving units. The problem-solving units consisted of two to three tasks each (OECD, 2014a). The two tasks we used for our analyses were located on position 2 (tickets task) and position 4 (climate control task) of their cluster. Depending on whether this cluster was administered as first or second cluster, the tasks were either in an early position of the test or in a middle position.

2.4 | Data preparation

Before coding the log data, we deleted log events that were not caused by student action (e.g., log events that mark the loading or unloading of a task). As mentioned above, we coded all remaining log events in our data separately as categories of behaviour. As discussed earlier, previous research points out the importance of exploration for success in CPS (Bell & Kozlowski, 2008; Dormann & Frese, 1994; Eichmann et al., 2018) but also the positive effects of parsimonious, goal-directed behaviour (He & von Davier, 2015; Stadler et al., 2019). Therefore, we chose to use the categories non-targeted exploration and goal-directed behaviour for our sequence analysis. We defined goal-directed behaviour as every interaction *necessary* for the students to solve the respective task correctly (i.e., every interaction that is part of the shortest path to task success given the knowledge the problem solver has at the beginning of each task). Therefore, goal-directed behaviour also includes exploration that is required to solve the task. In contrast to this, non-targeted exploration was operationalized as every interaction *not* necessary to solve the task. The distinction between non-targeted exploration and goal-directed behaviour was implemented differently for the tasks from the LSE and FSA framework. In FSA tasks it is possible to simply distinguish between states of the problem that are required to be visited for a correct solution and states that were not. In LSE tasks this is not the case. Therefore, we decided to define the use of the VOTAT strategy, the drawing of a correct line, and the deletion of a wrong line in the diagram as goal-directed behaviour in the climate control task (Wüstenberg, Greiff, & Funke, 2012). Other interactions in the climate control task (e.g., manipulating multiple variables at a time, drawing a wrong or deleting a correct line in the diagram) were coded as non-targeted exploration. We further refined the two categories goal-directed behaviour and non-targeted exploration by also distinguishing whether an interaction was performed for the first time or repeatedly. Both tasks contained a reset button, which would restore the initial state of the task. Pressing the reset button could not be categorized as goal-directed or non-targeted exploration, since resetting could be both part of non-targeted exploration or part of goal-directed interaction. Therefore, we defined resetting as a unique category. Thus, we ended up with five categories of CPS behaviour: initial goal-directed behaviour, repeated goal-directed behaviour, initial non-targeted exploration, repeated non-targeted exploration and resetting.

2.5 | Data analysis

We divided the dataset by task (climate control vs. tickets) and by the correctness of the given response (correct vs. false). By dividing the data into correct and false trials, we wanted to obtain behaviours that can be clearly assigned to correct or false responses, rather than obtaining more or less successful behaviours. In the climate control task 51.08% of the students gave a correct response. In the tickets task, 43.04% of the students did so. The following analyses were performed for the four resulting subsets of the data separately.

For Hypotheses 1 and 2 we conducted chi-squared tests to compare the relative frequencies of (initial and repeated) non-targeted exploration and (initial and repeated) goal-directed behaviour between correct and false responses in the two tasks, respectively. For the sequence analysis, we determined the differences between the sequences of behaviour categories of the students using optimal matching and the R package TraMineR (R Core Team, 2016; Studer & Ritschard, 2016). The optimal matching algorithm determines the dissimilarity between sequences by calculating the costs of transferring one sequence into the other. There are two types of costs to be specified. The costs for inserting or deleting an element of the sequence (indels), which reflect differences in sequence length, and the costs for substituting one sequence element with another element. We chose indels = 2 and the substitution cost matrix shown in Table 1 for our analysis. The substitution costs reflect the theoretical similarity between the behaviour categories (e.g., initial and repeated non-targeted exploration are more similar to each other than initial non-targeted exploration and initial goal-directed behaviour). Our indels equal the maximum of the substitution costs, so a difference in length between two sequences would result in the same difference value as a difference between one behaviour category and a very dissimilar one. Therefore, both sequence lengths as well as qualitative differences between sequences are taken into account to determine the dissimilarity between sequences. We also normalized the dissimilarity between the sequences dividing it by the length of the longer of each two sequences to account for potentially larger (non-normalized) dissimilarity between longer sequences (Gabadinho et al., 2011). Note that the comparison between sequences refers to the order of interactions and not to timing (i.e., two sequences are regarded as being identical if they contain the same interactions in the same order, no matter if the interactions were performed with different speed).

Based on the differences of students' sequences, we conducted a hierarchical cluster analysis using the Ward algorithm (Studer, 2013). We used the PISA 2012 final student weights in the analysis to account for oversampling. We used the normalized point-biserial correlation (PBC), average silhouette width (ASW) and Hubert's C index (HC) as quality criteria to determine the optimal number of clusters. The PBC measures the capacity of a clustering solution to reproduce the differences between sequences obtained through string matching. The ASW compares the average weighted distance of a cluster member from other members of the same cluster with its average weighted distance from the closest other cluster. The HC reflects the difference between the obtained cluster solution and the best cluster

TABLE 1 Substitution cost matrix for optimal matching

	Initial exploration	Repeated exploration	Initial goal-directed behaviour	Repeated goal-directed behaviour	Resetting
Initial exploration	0	1	1.5	2	1.5
Repeated exploration	1	0	2	1.5	1.5
Initial goal-directed behaviour	1.5	2	0	1	1.5
Repeated goal-directed behaviour	2	1.5	1	0	1.5
Resetting	1.5	1.5	1.5	1.5	0

solution that could have been obtained with the given dataset and number of clusters. While PBC and ASW should be maximized to obtain the optimal clustering solution, HC should be minimized to do so (Studer, 2013). PBC, ASW and HC take into account different properties of the cluster solutions. Considering these different indices we aim at a well-balanced evaluation of the different cluster solutions. For our four different data subsets we tested hierarchical clustering with two to eight clusters, respectively and chose the optimal solutions according to our quality criteria.

To compare the obtained clusters with respect to students' overall CPS skills, we used the responses of the students on the other 25 CPS items of the PISA 2012 assessment. The responses were coded as no credit, partial credit, full credit or not reached. We recoded not reached items as no credit and fitted a one-parameter logistic (1PL) partial credit item response theory (IRT) model to the response data to obtain weighted likelihood estimators (WLEs) of students' overall CPS skills using marginal maximum likelihood estimation of the TAM package (Robitzsch, Kiefer, & Wu, 2019). Maximum likelihood estimation allows to compute unbiased means of ability estimates (even though the variance might be overestimated) (Mislevy, Beaton, Kaplan, & Sheehan, 1992). Since we used WLEs based on maximum likelihood estimation to compare group means only, we refrained from the more complex analysis approach using plausible values. Due to PISA's rotated block design, there were missing responses by design in all CPS items. The WLE scale is centred so its mean is zero. Therefore, negative WLEs represent CPS skills below average. We used the PISA final student weights to account for the stratified sampling (OECD, 2014b). Subsequently, we applied analysis of variance to compare the mean CPS skills across clusters. To obtain group-wise comparisons, we used Tukey Honest Significance Difference test, which controls for Type I error inflation (Field, Miles, & Field, 2013). We also compared the clusters regarding their occurrence depending on tasks' positions in the test (early vs. middle position). Therefore, we used chi-squared tests to compare if clusters occur significantly more often at an early or a middle position in the test. For all chi-squared tests we calculated the effect size φ using the DescTools package in R (Signorell, Andri et al., 2019). According to the conventions of Cohen (1988), a φ value of 0.1 is considered a small effect, 0.2 a medium effect and 0.3 a large effect. For all Tukey tests we calculated Cohen's d using the psych package in R (Revelle, 2018). According to the conventions of Cohen (1988), a d value of 0.2 is considered a small effect, 0.5 a medium effect and 0.8 a large effect.

3 | RESULTS

The results of the chi-squared tests comparing the relative frequencies of goal-directed behaviour and non-targeted exploration are shown in Tables 2 and 3. In line with Hypothesis 1, both initial and repeated non-targeted exploration was more frequent among false responses. The results for goal-directed behaviour were only significant for repeated goal-directed behaviour in the tickets task (Table 3). This indicates that in the tickets task repeated goal-directed behaviour was more frequent among correct responses. Therefore, Hypothesis 2 is only supported in this particular case.

The quality criteria PBC, ASW and HC according to the different numbers of clusters are shown in Figure 4. Since not all the quality criteria favoured the same solution in all cases, we decided on those solutions that were favoured by at least one quality criterion while also showing good results for the other two criteria.

Following this rule, we decided on the 5-cluster solution for false solutions in the climate control task, which reflects the maximum of the PBC, a local minimum of HC and a medium value for ASW. A high value of PBC indicates that the partition reflects the patterns of dissimilarities between sequences quite well. A low value of HC reflects a favourable ratio of within- and between-cluster dissimilarities (Studer, 2013). We chose a 3-cluster solution for correct solutions in the climate control task, again maximizing PBC, choosing a local minimum for HC and a medium value for ASW. We chose a 7-cluster solution for false solutions in the tickets task, this time optimizing both PBC and HC while ASW had a value close to its maximum. A high value of ASW reflects a good ratio of sequences' similarities to their cluster members and dissimilarities to members of other clusters. We chose a 6-cluster solution for correct solutions in the tickets task, maximizing PBC and ASW while HC had a value close to its minimum. The resulting clusters are displayed in the following paragraphs.

3.1 | Climate control task

3.1.1 | False solutions

We chose a solution with five clusters of sequences for false solutions in the climate control task. The clusters are depicted in Figure 5. The

TABLE 2 Relative frequencies of different behaviours in the climate control task

	% of behaviour in		χ^2	df	p	φ
	False responses	Correct responses				
Initial exploration	29.63	14.03	5.57	1	.018	0.36
Repeated exploration	30.78	14.63	5.75	1	.017	0.36
Initial goal-directed behaviour	18.21	31.41	3.50	1	.061	0.27
Repeated goal-directed behaviour	14.17	26.10	3.53	1	.060	0.30

TABLE 3 Relative frequencies of different behaviours in the tickets task

	% of behaviour in		χ^2	df	p	φ
	False responses	Correct responses				
Initial exploration	24.77	6.78	10.26	1	.001	0.57
Repeated exploration	9.60	0.85	7.32	1	.007	0.84
Initial goal-directed behaviour	40.30	42.36	0.05	1	.821	0.02
Repeated goal-directed behaviour	17.95	37.37	6.82	1	.009	0.35

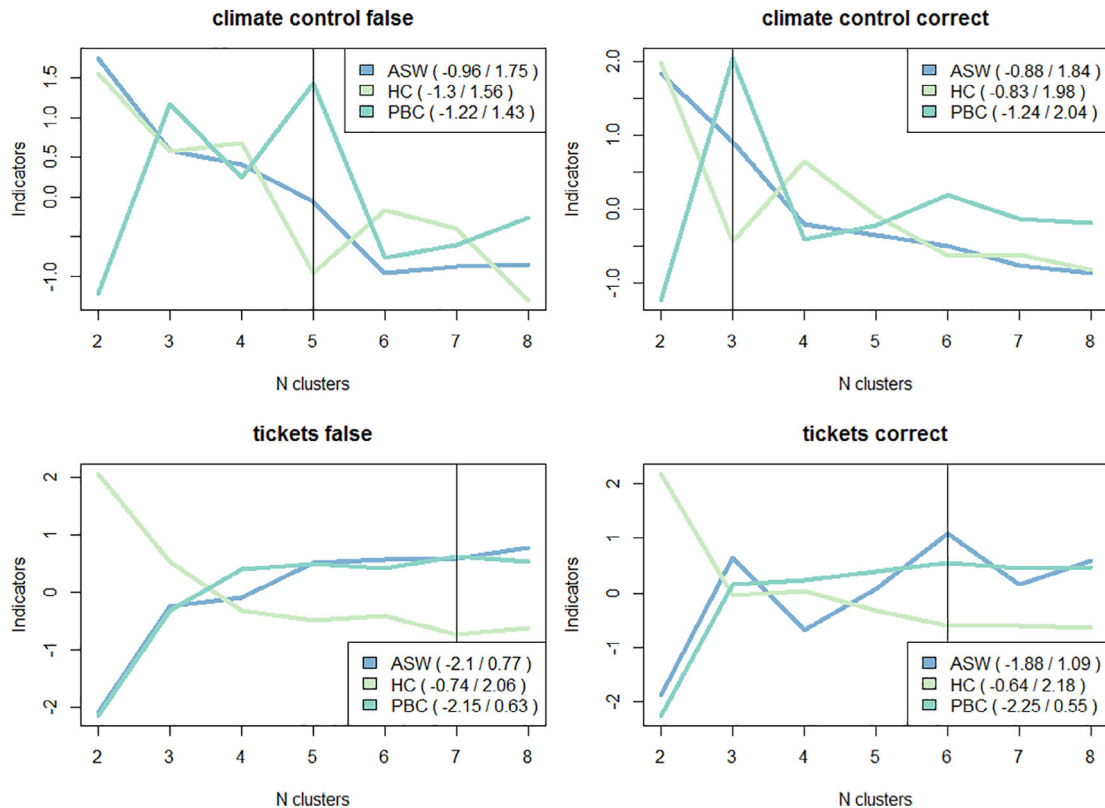
**FIGURE 4** Z-standardized quality criteria (ASW, PBC and HC) according to number of clusters in the different data sets. Minimum and maximum values are displayed in brackets. The vertical lines mark the chosen solution. ASW, average silhouette width; HC, Hubert's C index; PBC, point-biserial correlation [Colour figure can be viewed at wileyonlinelibrary.com]

figure displays state distribution plots for each cluster that show the relative distribution of behaviour categories at each interaction. Values on the x-axis represent the numbered interactions from the behaviour sequences. Values on the y-axis represent the relative frequencies of behaviour categories displayed by the students in the

respective cluster at the respective interaction. For example, in Figure 5 in the first cluster (top left) roughly 50% of the students showed non-targeted exploration behaviour in their first interaction, about 40% showed goal-directed behaviour in their first interaction, and about 10% reset the task in their first interaction (which has no

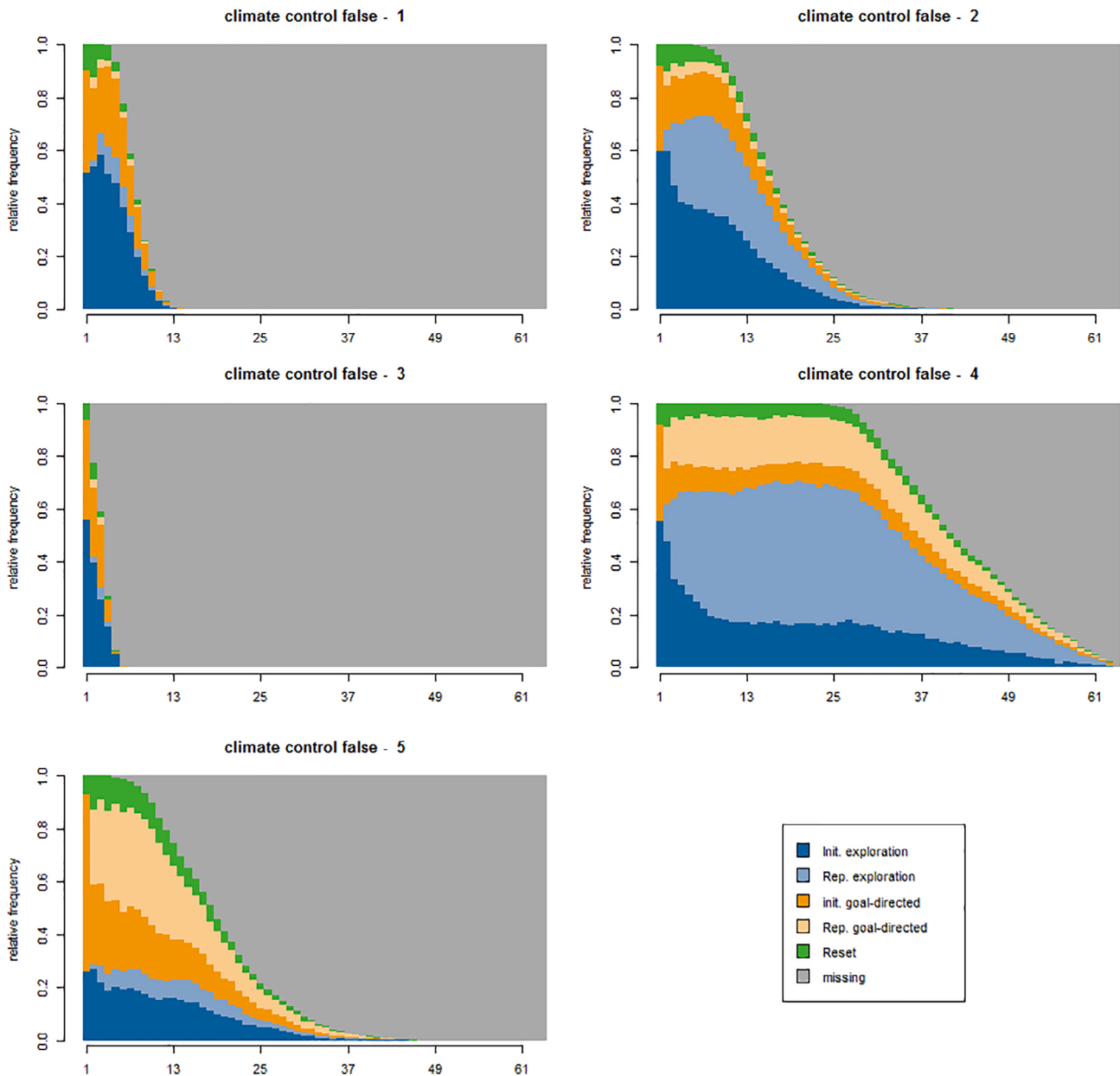


FIGURE 5 State distribution plots for each cluster for false solutions in the climate control task [Colour figure can be viewed at wileyonlinelibrary.com]

effect in the tasks initial state). Missing values in the figure are caused by sequences shorter than the value range of the x-axis.

Cluster 1 contains the sequences of 2,672 students (19.31% of all false solutions in the climate control task). They show slightly more non-targeted exploration than goal-directed behaviour. The vast majority of interactions are initial (and not repeated) and the reset button is rarely used. Overall the sequences are quite short with the longest sequence comprising of 15 interactions and an average sequence length of 7.24 interactions. These students seem to unsystematically try out VOTAT and non-VOTAT actions.

Cluster 2 contains the sequences of 4,484 students (32.40% of all false solutions in the climate control task). They show approximately

70% non-targeted exploration behaviour. Especially the non-targeted exploration interactions are often repeated; the reset button is again rarely used. Overall, the sequences are relatively long with the longest sequence containing 49 interactions and an average sequence length of 16.98 interactions. These students also seem to unsystematically try out different actions and engage increasingly in reinvestigation of non-targeted exploratory actions.

Cluster 3 contains the sequences of 1846 students (13.34% of all false solutions in the climate control task). They show approximately 60% non-targeted exploration behaviour. The vast majority of interactions are initial; the reset button is again rarely used. The sequences in this cluster are especially short with the longest sequence consisting of seven interactions and an average sequence length of 2.71

interactions. These students seem to abandon the task quite early resulting in not answering or guessing a false solution.

Cluster 4 contains the sequences of 2,158 students (15.59% of all false solutions in the climate control task). Like cluster 2, they also show approximately 70% non-targeted exploration behaviour. However, in this cluster most interactions (even in the beginning of the sequences) are repeated. The reset button is again rarely used. The sequences in this cluster are the longest sequences among false solutions in the climate control task with an average sequence length of 42.26 interactions. These students reinvestigate the same content again and again. However, most of the investigated content is irrelevant.

Cluster 5 contains the sequences of 2,680 students (19.36% of all false solutions in the climate control task). They show about 60% goal-directed behaviour and about 30% non-targeted exploration. About 10% of the interactions are with the reset button. In this cluster most goal-directed interactions are repeated while most non-targeted exploration is initial. The sequences in this cluster are of similar length as those in cluster 2 with an average sequence length of 18.67 interactions. From the false solutions in the climate control tasks, students in this cluster show the highest proportion of goal-directed behaviour. However, the proportion of initial goal-directed behaviour is quite small. This could indicate an incomplete application of the VOTAT strategy not investigating the effect of every input variable in isolation.

The comparison of the clusters regarding overall CPS skills using WLEs shows that all clusters have a negative average estimated skill (see Figure 6). Students guessing or not answering the task (cluster 3) show the lowest average CPS skills, whereas students applying the incomplete approach (cluster 5) show the highest CPS skills. CPS skills increase for clusters with longer sequences and with higher frequencies of goal-directed behaviour. The results of the Tukey Honest Significance Difference test show that significant differences in mean CPS skills exist between all clusters except for the two groups of medium to high sequence length showing mainly non-targeted exploration (clusters 2 and 4) (see Table 4).

The comparison of the clusters regarding their occurrence in early or middle positions in the test reveals that the clusters with very short sequences (clusters 1 and 3) occurred more often in the middle item position (see Table 5). The longer clusters containing mainly exploration (clusters 2 and 4) occur more frequently at the early item position. For cluster 5, no significant difference in occurrence at either position was found.

3.1.2 | Correct solutions

We chose a solution with three clusters of sequences for correct solutions in the climate control task. The clusters are depicted in Figure 7. Cluster 1 contains the sequences of 6,569 students (45.46% of all correct solutions in the climate control task). They show between 10 and 30% non-targeted exploration with a decreasing trend in the course of the problem-solving process and about 70% goal-directed

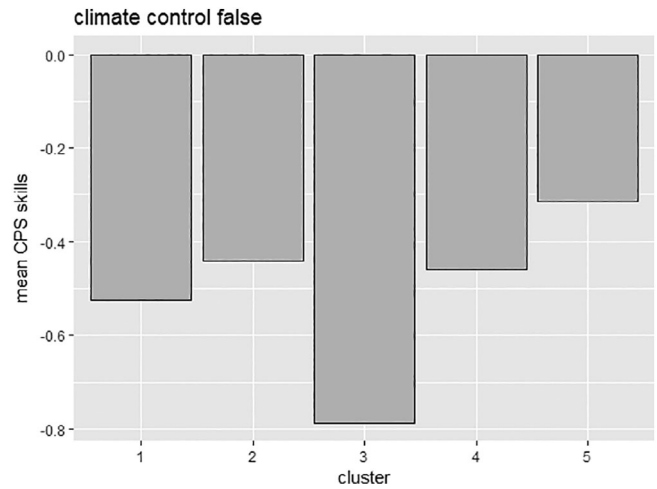


FIGURE 6 Mean CPS skills per cluster for false responses in the climate control task. CPS, complex problem-solving

behaviour. About 20% of the interactions are with the reset button. The majority of interactions are initial. Overall the sequences are of short or medium length with the longest sequence containing 30 interactions and an average sequence length of 13.11 interactions. Since the sequences are quite short and little non-targeted exploration takes place, these students show a quite efficient behaviour.

Cluster 2 contains the sequences of 2,984 students (20.65% of all correct solutions in the climate control task). They show about 30% non-targeted exploration and about 65% goal-directed behaviour. About 5% of the interactions are with the reset button. The majority of interactions is repeated. This cluster contains the longest sequences in the climate control task with sequences up to 80 interactions length and an average sequence length of 48.20 interactions. The large number of repeated goal-directed actions indicates an approach of double-checking relevant information.

Cluster 3 contains the sequences of 4,898 students (33.89% of all correct solutions in the climate control task). They show about 35% non-targeted exploration and about 50% goal-directed behaviour. About 15% of the interactions are with the reset button. The narrow majority of the goal-directed interactions is initial and the narrow majority of the non-targeted exploration is repeated. The sequences in this cluster are rather long with sequences up to 75 interactions length and an average sequence length of 25.69 interactions. These students seem to apply a rather mixed approach with some non-targeted exploration and double-checking but mainly initial goal-directed behaviour.

The comparison of clusters regarding overall CPS skills using WLEs shows that all clusters have a positive average estimated skill (see Figure 8). Students applying the double-checking approach (cluster 2) show the highest average CPS skills. The results of the Tukey Honest Significance Difference test show that the CPS skills of students applying the double-checking approach (cluster 2) are significantly higher than those of students using more efficient (cluster 1) or a mixed approach (cluster 3) (see Table 6). There is no significant difference in overall CPS skill between clusters 1 and 3.

Compared clusters	Difference	Confidence interval		p value	Cohen's d
		Lower bound	Upper bound		
2-1	0.08	0.04	0.13	<.001	0.16
3-1	-0.26	-0.32	-0.20	<.001	-0.39
4-1	0.07	0.01	0.12	<.001	0.11
5-1	0.21	0.16	0.26	<.001	0.37
3-2	-0.35	-0.40	-0.29	<.001	-0.53
4-2	-0.02	-0.07	0.03	.869	-0.04
5-2	0.13	0.08	0.17	<.001	0.22
4-3	0.33	0.27	0.39	<.001	0.49
5-3	0.47	0.41	0.53	<.001	0.72
5-4	0.14	0.09	0.20	<.001	0.25

TABLE 4 Comparison of mean CPS ability across clusters for false responses in the climate control task

Abbreviation: CPS, complex problem-solving.

TABLE 5 Frequencies of clusters depending on item position for false responses in climate control task

Cluster	Item position in test		χ^2	df	p	φ
	Early	Middle				
1	1,152	1,349	15.52	1	<.001	0.08
2	2,196	1,859	28.01	1	<.001	0.08
3	500	722	40.33	1	<.001	0.18
4	1,135	855	39.40	1	<.001	0.14
5	1,151	1,246	3.77	1	.052	0.04

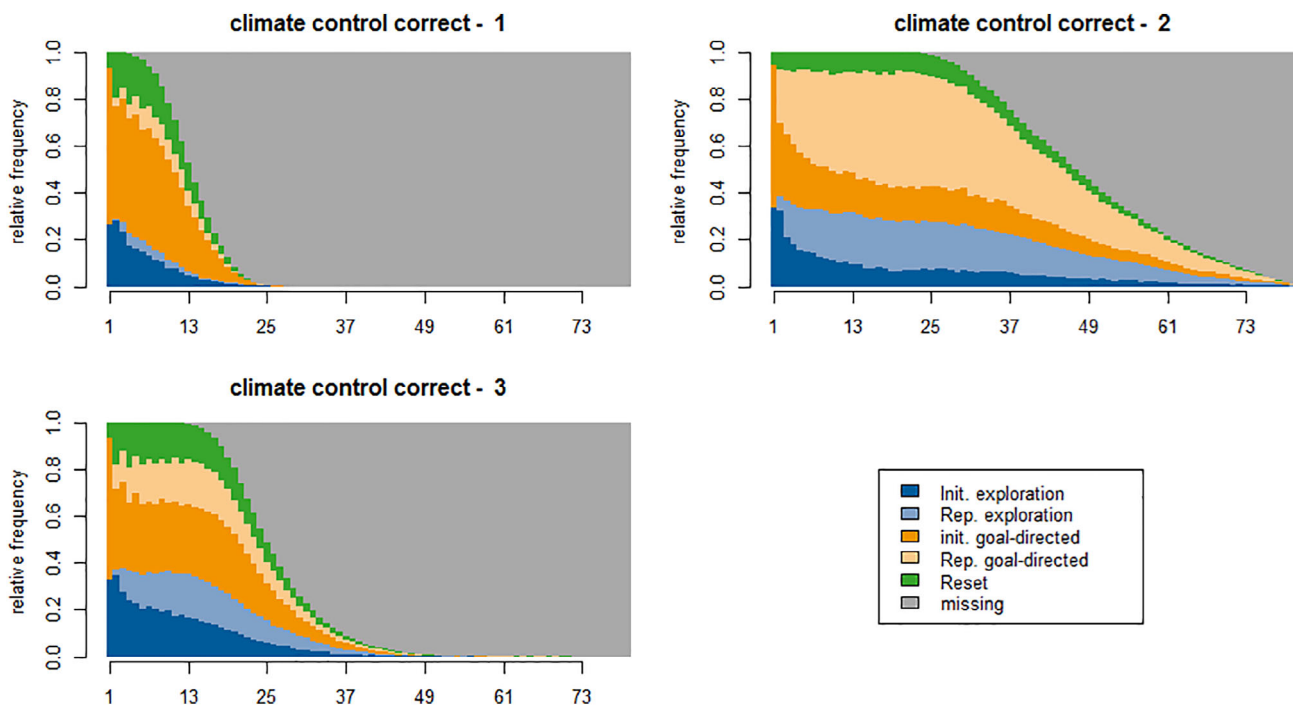


FIGURE 7 State distribution plots for each cluster for correct solutions in the climate control task [Colour figure can be viewed at wileyonlinelibrary.com]

The occurrence of all three clusters differed significantly between the two item positions (see Table 7). While clusters 1 and 2 appeared more frequently in the middle position, cluster 3 was more frequently observed in the early position.

3.2 | Tickets task

3.2.1 | False solutions

We chose a solution with seven clusters of sequences for false solutions in the tickets task. The clusters are depicted in Figure 9. Cluster 1 contains the sequences of 2,334 students (14.61% of all false solutions in the tickets task). They show almost exclusively non-targeted exploration. The reset button was not used by this group. In the beginning all interactions were initial; towards the end almost all interactions were repeated. The sequences are extremely short with the longest sequence containing five interactions and an average sequence length of 4.36 interactions. Since the sequences contain hardly any goal-directed interactions and are also too short to reveal the correct solution, the students seem to apply unsystematic guessing behaviour.

Cluster 2 contains the sequences of 1,668 students (10.44% of all false solutions in the tickets task). They show mainly goal-directed behaviour in their first interaction, afterwards it was almost exclusively non-targeted exploration. This group made little use of the reset button. In the beginning, all interactions were initial; towards the end there were more repeated interactions and resets. The sequences were as short as the sequences in cluster 1 with the longest sequence containing five interactions and an average sequence length of 3.59 interactions. Similar to cluster 1 these students show unsystematic guessing behaviour or do not respond at all.

Cluster 3 contains the sequences of 6,937 students (43.41% of all false solutions in the tickets task). They showed almost exclusively

goal-directed behaviour. This group did not use the reset button. During the first three interactions, all interactions were initial; only in the last interaction there were repeated interactions. The sequences are similarly short as the sequences in cluster 1 and 2 with the longest sequence containing four interactions and an average sequence length of 3.98 interactions. Again the short sequences indicate guessing a solution. However, the students seemed to intentionally choose goal-directed actions. Therefore, their behaviour could be called "goal-directed guessing".

Cluster 4 contains the sequences of 1,120 students (7.01% of all false solutions in the tickets task). They showed almost exclusively non-targeted exploration behaviour. This group used the reset button a few times. In the course of the task, the frequency of repeated interactions increases. The sequences are of medium length with the longest sequence containing 18 interactions and an average sequence length of 8.95 interactions. These students seem to be quite persevering in engaging with irrelevant content.

Cluster 5 contains the sequences of 1,078 students (6.75% of all false solutions in the tickets task). They showed almost exclusively goal-directed behaviour. In this group, there is a peak of uses of the reset button at interaction 4. This peak indicates that students navigated to the first ticket option and reset the task to consider further options. Prior to this peak, there were mostly initial interactions. After the peak there were mostly repeated interactions. The sequences are of medium length with the longest sequence containing 12 interactions and an average sequence length of 7.94 interactions. These students show a quite systematic approach. However, instead of comparing different ticket options most students investigated the first ticket twice, which becomes evident in the high proportion of repeated interactions after the peak of resets. Therefore, their approach is rather incomplete.

Cluster 6 contains the sequences of 2,191 students (13.71% of all false solutions in the tickets task). They showed almost exclusively goal-directed behaviour during their first three interactions followed by a peak of reset at interaction 4, similar to cluster 5. However, after the peak there were mostly repeated goal-directed interactions and non-targeted exploration. The sequences are quite long with the longest sequence containing 18 interactions and an average sequence length of 11.66 interactions. These students either compare a relevant ticket with a non-relevant ticket or reinvestigate the first ticket multiple times.

Cluster 7 contains the sequences of 652 students (4.08% of all false solutions in the tickets task). Remarkably, all the sequences in this cluster are identical. They showed three initial goal-directed interactions followed by one initial and one repeated non-targeted exploration. The sequences are similarly short as those in clusters 1 and 2 containing five interactions. These students show guessing behaviour that is partly goal-directed.

The comparison of the clusters regarding overall CPS skills using WLEs shows that all clusters have a negative average estimated skill (see Figure 10). Students engaging in unsystematic guessing or not answering (cluster 2) show the lowest average CPS skills, while students applying an incomplete approach (cluster 5) show the highest

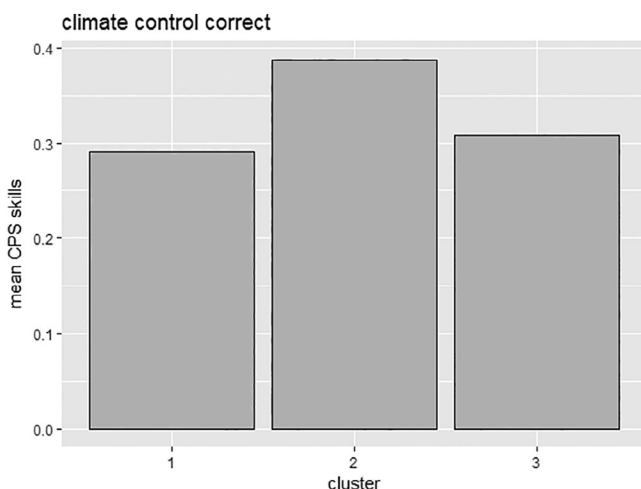


FIGURE 8 Mean CPS skills per cluster for correct responses in the climate control task. CPS, complex problem-solving

Compared clusters	Difference	Confidence interval		<i>p</i> value	Cohen's <i>d</i>
		Lower bound	Upper bound		
2-1	0.10	0.06	0.13	<.001	0.14
3-1	0.02	-0.01	0.05	.395	0.03
3-2	-0.08	-0.12	-0.04	<.001	-0.11

Abbreviation: CPS, complex problem-solving.

TABLE 7 Frequencies of clusters depending on item position for correct responses in climate control task

Cluster	Item position in test		χ^2	<i>df</i>	<i>p</i>	φ
	Early	Middle				
1	3,141	3,341	6.17	1	.013	0.03
2	1,400	1,540	6.67	1	.010	0.05
3	2,585	2,236	25.27	1	<.001	0.07

CPS skills in this group. CPS skills increase for clusters with longer sequences and with higher frequencies of goal-directed behaviour. The results of the Tukey Honest Significance Difference test show that significant differences in mean CPS skills exist between most clusters (see Table 8). Exceptions are clusters 4 and 1; and clusters 3, 5 and 6 showing similar CPS skills.

The comparison of the frequency of clusters between the early and the middle position in the test revealed that all guessing clusters (clusters 1, 2, 3 and 7) appeared more frequently at the early position (see Table 9). Also cluster 4 appeared more frequently at the early position. Cluster 6 is the only cluster that was observed more frequently at the middle position. For cluster 5, no significant difference was found.

3.2.2 | Correct solutions

We chose a solution with six clusters of sequences for correct solutions in the tickets task. The clusters are depicted in Figure 11. Cluster 1 contains the sequences of 2,629 students (21.77% of all correct solutions in the tickets task). The vast majority of the sequences showed initial goal-directed behaviour during the first three interactions followed by a peak of resets at interaction 4. This peak indicates that students navigated to the first ticket option and reset the task to consider further options. Starting from interaction 5, a lot of repeated goal-directed behaviour followed again by initial goal-directed behaviour took place. The sequences are of medium length with the longest sequence containing 21 interactions and an average sequence length of 9.29 interactions. The students apply a quite efficient (rather minimalistic) approach, since most of them show the minimum behaviour that is needed to solve the task.

Cluster 2 contains the sequences of 1978 students (16.38% of all correct solutions in the tickets task). The sequences show again almost exclusively initial goal-directed behaviour in the beginning followed by much repeated goal-directed behaviour. Only little non-targeted exploration took place. There are several peaks of resets at

TABLE 6 Comparison of mean CPS ability across clusters for correct responses in the climate control task

interaction 4, 9 and 13. The sequences are quite long with the longest sequence containing 27 interactions and an average sequence length of 20.99 interactions. These students seem to double-check the relevant tickets again and again.

Cluster 3 contains the sequences of 769 students (6.37% of all correct solutions in the tickets task). The sequences show again almost exclusively initial goal-directed behaviour in the beginning followed by repeated goal-directed behaviour. There is a peak of non-targeted exploration at interaction 5 and peaks of resets at interaction 4 and 6. Apart from the peak not much non-targeted exploration took place. The sequences are of medium length with the longest sequence containing 17 interactions and an average sequence length of 10.99 interactions. These students (as those in cluster 1) also show quite efficient behaviour, only they also show some non-targeted exploration.

Cluster 4 contains the sequences of 2,157 students (17.86% of all correct solutions in the tickets task). The sequences show again almost exclusively initial goal-directed behaviour in the beginning followed by much repeated goal-directed behaviour. Only little non-targeted exploration took place. There are peaks of resets at interaction 5 and 9. The sequences are rather long with the longest sequence containing 27 interactions and an average sequence length of 14.84 interactions. These students also show rather minimalistic behaviour (as those students in cluster 1). However, they investigated the tickets in a different order, forcing them to reset once more and navigate back to the ticket they inspected first.

Cluster 5 contains the sequences of 672 students (5.57% of all correct solutions in the tickets task). The sequences show exclusively initial goal-directed behaviour in the first three interactions followed by almost half non-targeted exploration and goal-directed behaviour. In the course of the task students exhibited more and more repeated interactions. In this group the reset button was not used at all. The sequences are quite short with the longest sequence containing 17 interactions and an average sequence length of 7.68 interactions. These students show guessing behaviour that is partly goal-directed and partly non-targeted exploration.

Cluster 6 contains the sequences of 3,869 students (32.04% of all correct solutions in the tickets task). The sequences show exclusively initial goal-directed behaviour in the beginning followed by repeated goal-directed behaviour. In this group neither the use of the reset button nor non-targeted exploration was displayed. All sequences are identical and contain 5.00 interactions. Therefore, these students show "goal-directed guessing".

The comparison of the clusters regarding CPS skills using WLEs shows that students in all clusters except one have a positive average

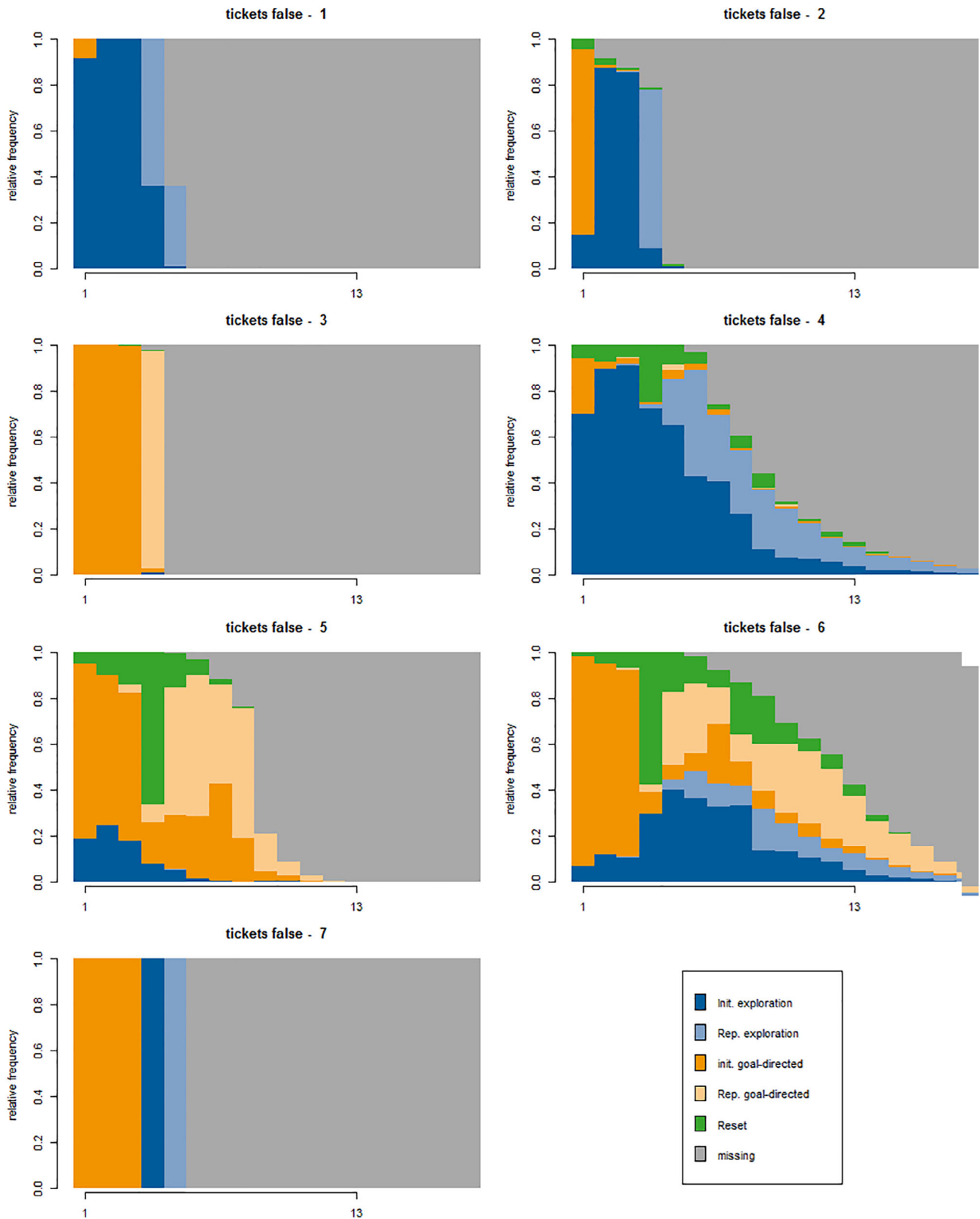


FIGURE 9 State distribution plots for each cluster for false solutions in the tickets task [Colour figure can be viewed at wileyonlinelibrary.com]

estimated skill (see Figure 12). Students, who double-checked their solution (cluster 2), show the highest average CPS skills, while students unsystematically guessing (cluster 5) show the lowest CPS skills. CPS skills increase for clusters with longer sequences and

with higher frequencies of goal-directed behaviour. The results of the Tukey Honest Significance Difference test show that significant differences in mean CPS skills exist between all clusters (see Table 10).

The comparison of clusters with regard to their occurrence at different positions in the test revealed that the goal-directed guessing cluster (cluster 6) was observed more frequently at the early position (see Table 11). The minimalistic and the double-checking clusters (clusters 1, 2, 3 and 4) were observed more frequently at the middle position. For cluster 5, no significant difference was found.

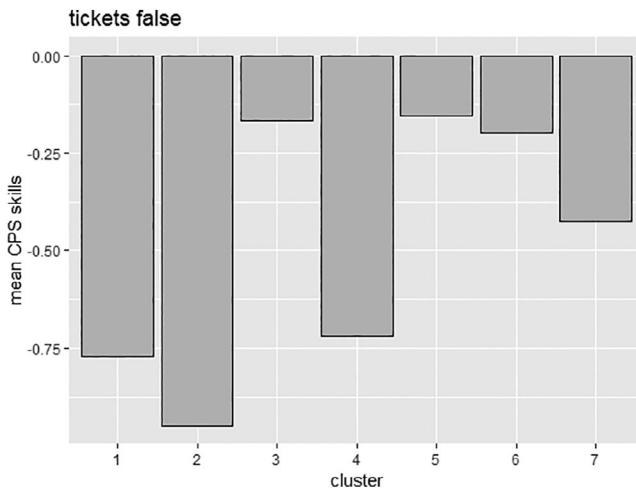


FIGURE 10 Mean CPS skills per cluster for false responses in the tickets task. CPS, complex problem-solving

4 | DISCUSSION

The aim of this study was to investigate how behavioural sequences are related to success or failure in CPS. Moreover, we wanted to clarify inconsistent findings of previous research regarding the usefulness of exploration in CPS. To this end, we used log data from the PISA 2012 CPS assessment and conducted full-path sequence analysis. We were able to clarify the inconsistent previous findings regarding exploration. Moreover, we identified several behavioural patterns associated with success or failure in CPS. Most patterns were found in both investigated tasks and thus across the two CPS frameworks of FSA and LSE. In the following paragraphs we will first discuss our results concerning non-targeted exploration and goal-directed behaviour (Hypotheses 1 and 2) before we discuss the behavioural patterns we found and their relation to students' overall CPS performance (Research questions 1 and 2).

4.1 | Non-targeted exploration and goal-directed behaviour

The results of the chi-squared test revealed that both initial and repeated non-targeted exploration were found more frequently among false responses. This finding supports our Hypothesis 1. Therefore, the results of Dormann and Frese (1994) and Eichmann et al. (2018), who reported a positive relation between exploration

Compared clusters	Difference	Confidence interval		<i>p</i> value	Cohen's <i>d</i>
		Lower bound	Upper bound		
2-1	-0.18	-0.24	-0.11	<.001	-0.17
3-1	0.60	0.55	0.65	<.001	0.89
4-1	0.05	-0.02	0.12	.428	0.11
5-1	0.61	0.54	0.69	<.001	0.98
6-1	0.57	0.51	0.63	<.001	0.95
7-1	0.35	0.26	0.44	<.001	0.56
3-2	0.78	0.72	0.84	<.001	1.00
4-2	0.23	0.15	0.31	<.001	0.27
5-2	0.79	0.71	0.87	<.001	1.10
6-2	0.75	0.68	0.82	<.001	1.10
7-2	0.52	0.43	0.62	<.001	0.70
4-3	-0.55	-0.62	-0.49	<.001	-0.77
5-3	0.01	-0.06	0.08	.999	-0.00
6-3	-0.03	-0.08	0.02	.533	0.01
7-3	-0.26	-0.34	-0.17	<.001	-0.38
5-4	0.56	0.48	0.65	<.001	0.84
6-4	0.52	0.45	0.60	<.001	0.81
7-4	0.30	0.20	0.40	<.001	0.43
6-5	-0.04	-0.12	0.03	0.648	0.01
7-5	-0.27	-0.37	-0.17	<0.001	-0.41
7-6	-0.23	-0.32	-0.13	<0.001	-0.40

TABLE 8 Comparison of mean CPS ability across clusters for false responses in the tickets task

TABLE 9 Frequencies of clusters depending on item position for false responses in tickets task

Cluster	Item position in test		χ^2	df	p	φ
	Early	Middle				
1	1,463	814	184.98	1	<.001	0.29
2	753	383	120.51	1	<.001	0.33
3	3,485	3,013	34.29	1	<.001	0.07
4	590	415	30.47	1	<.001	0.17
5	462	437	0.70	1	.404	0.03
6	851	973	8.16	1	.004	0.07
7	349	293	4.88	1	.027	0.09

and success in CPS, seem to be the result of not differentiating between goal-directed and non-targeted exploration. While goal-directed exploration might indeed be related to success, non-targeted exploration is in our data clearly related to failure. Therefore, non-targeted exploration could rather be a sign of confusion or distraction. These results were consistent in both investigated tasks.

We found that goal-directed behaviour (opposed to non-targeted exploration) was only if it was repeated and only in the tickets task more frequently found among correct responses. In the climate control task, this difference was not significant. We did not find significant differences for initial goal-directed behaviour in either task. Therefore, Hypothesis 2 was only supported for repeated goal-directed behaviour in the tickets task. A possible reason for that could be the higher opacity of the tickets task compared to the climate control task. While in the climate control task gathered information stays visible until the reset button is used, in the tickets task this is not the case. Therefore, in the tickets task repeated goal-directed behaviour might have been used to recall information, which was not required in the climate control task.

4.2 | Climate control task

4.2.1 | False solutions

We identified five clusters of behaviour sequences that did not result in correct solutions. Students in the cluster with the highest overall CPS skills among those who did not solve the climate control task showed an incomplete approach. They seem to have correctly applied the VOTAT strategy, but not to every input variable leading to an incomplete solution (cluster 5). Students that failed to apply the VOTAT strategy and instead investigated irrelevant content a lot showed medium overall CPS skills (clusters 2 and 4). Students in these two clusters seem to have applied a similar approach and differed mainly by sequence length. They engaged mainly in repeated non-targeted exploration behaviour, which could be a sign of over-estimating the relevance of in fact irrelevant information. Students mostly exploring irrelevant information and stopping their attempts early on showed even lower overall CPS skills (cluster 1). Since they show mainly initial non-targeted exploration, they do not seem to find

any satisfying solution. However, the lowest overall CPS skills were shown by the group guessing an answer or leaving the task unanswered (cluster 3). These students' sequences were shorter than minimally required by the task. Therefore, they were assumed to be guessing.

4.2.2 | Correct solutions

We found three clusters of behaviour sequences resulting in correct solutions. Those students that seem to double-check their solutions showed the highest overall CPS skills among students working on the climate control task (cluster 2). Notably, only about 21% of the correct responses were in the double-checking cluster. There was no significant performance difference in overall CPS between students who applied a quite efficient approach (cluster 1) and students who applied a mixed approach of goal-directed behaviour and non-targeted exploration (cluster 3).

4.3 | Tickets task

4.3.1 | False solutions

We identified seven clusters of behaviour resulting in false solutions in the tickets task. Among those students who showed the highest overall CPS skills in the false solutions tickets group was one group that applied the incomplete approach of showing mostly goal-directed behaviour but failed to investigate all relevant tickets (cluster 5). But also those who showed a mixed approach of goal-directed behaviour and non-targeted exploration (cluster 6) and those who showed goal-directed guessing (cluster 3) were found among the highest performing group (of students who got the tickets task wrong). In the tickets task, all students who did not use the reset button, were assumed to be guessing, since they did not inspect all relevant tickets. Notably, the group of goal-directed guessers was by far the largest among the false responses in the tickets task (43.41%). Lower overall CPS skills were found among students who were also guessing but whose guesses were only partly goal-directed (cluster 7). They also seemed to apply goal-directed guessing, but in the end got

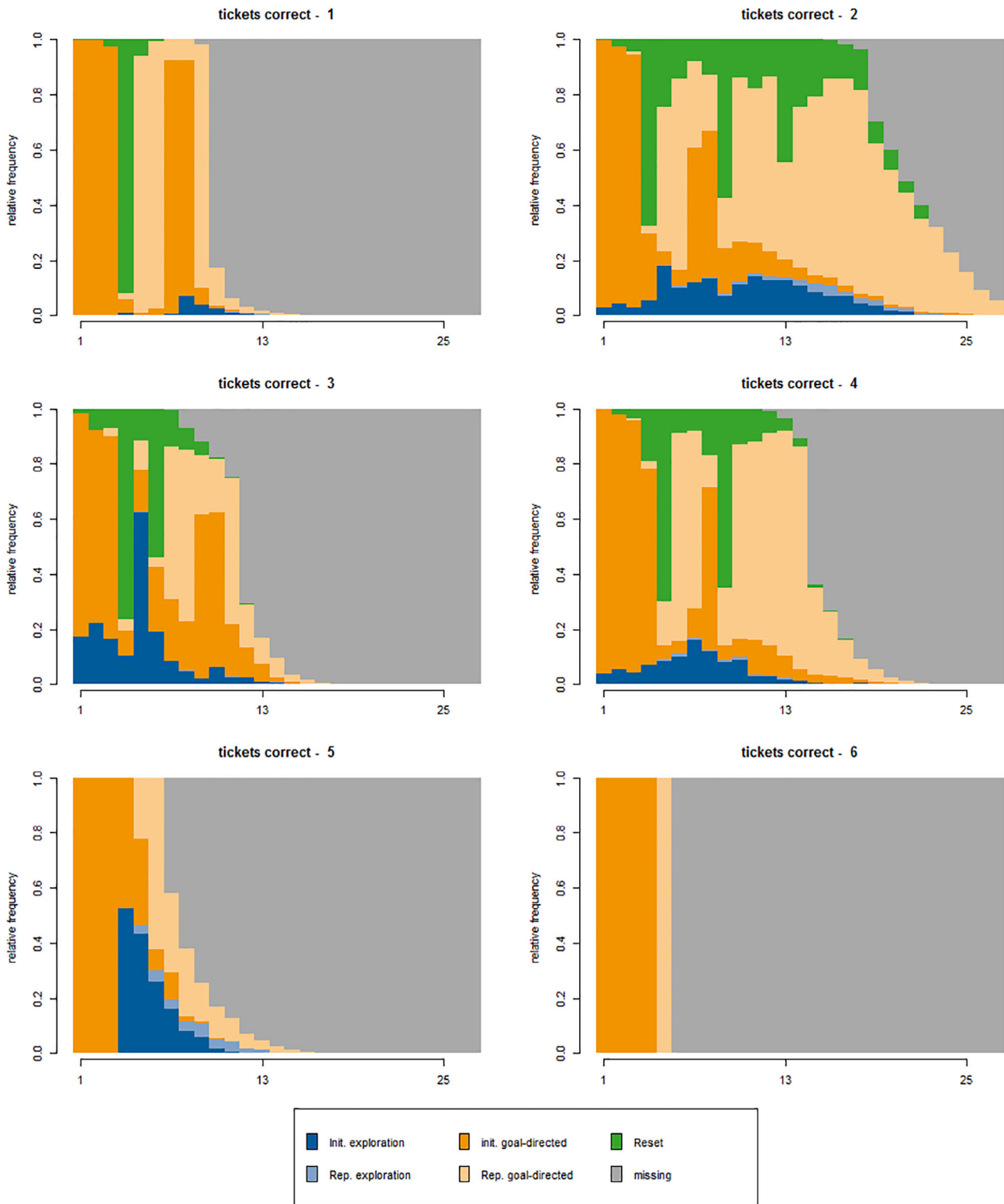


FIGURE 11 State distribution plots for each cluster for correct solutions in the tickets task [Colour figure can be viewed at wileyonlinelibrary.com]

sidetracked. An even lower overall CPS performance was found among those students who hardly showed any goal-directed behaviour. Within this group, it made no difference with regard to overall CPS performance whether students were guessing (cluster 1) or

showed longer non-targeted exploration (cluster 4). However, the lowest overall CPS performance among those students who did not solve the tickets task was found with students who started goal-directed but then guessed an implausible solution or left the task

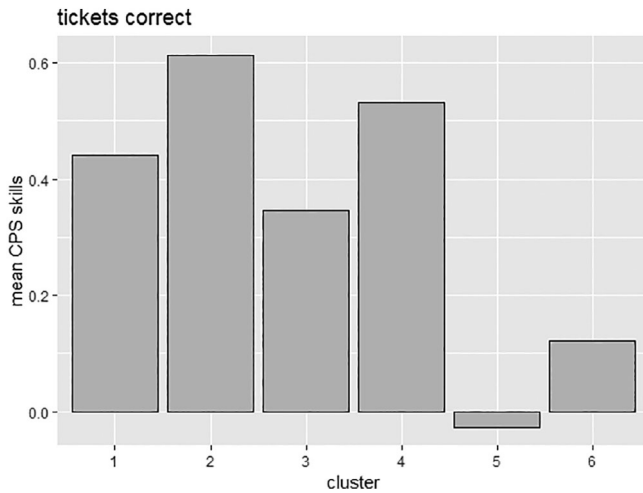


FIGURE 12 Mean CPS skills per cluster for correct responses in the tickets task. CPS, complex problem-solving

unanswered (cluster 2). In sum, 72.54% of the false solutions in the tickets task were the result of guessing (clusters 1, 2, 3 and 7). Therefore, guessing was the most frequent behaviour leading to false solutions in the tickets task.

4.3.2 | Correct solutions

We found six clusters of behaviour resulting in correct solutions in the tickets task. The highest overall CPS skills were again found among students who double-checked their solution (cluster 2). The second highest overall CPS skills were found among students who showed quite efficient behaviour in the tickets task (clusters 4 and 1). Those students, who were also efficient, but got distracted by non-relevant content at some point showed lower overall CPS skills (cluster 3). Even lower CPS skills were observed among students showing goal-directed guessing behaviour (cluster 6). The lowest skills of students correctly solving the

TABLE 10 Comparison of mean CPS ability across clusters for correct responses in the tickets task

Compared clusters	Difference	Confidence interval		<i>p</i> value	Cohen's <i>d</i>
		Lower bound	Upper bound		
2-1	0.17	0.12	0.23	<.001	0.26
3-1	-0.10	-0.17	-0.02	.005	-0.14
4-1	0.09	0.04	0.15	<.001	0.14
5-1	-0.47	-0.55	-0.39	<.001	-0.71
6-1	-0.32	-0.37	-0.27	<.001	-0.49
3-2	-0.27	-0.35	-0.19	<.001	-0.40
4-2	-0.08	-0.14	-0.02	.001	-0.12
5-2	-0.64	-0.72	-0.56	<.001	-0.96
6-2	-0.49	-0.54	-0.44	<.001	-0.75
4-3	0.19	0.11	0.27	<.001	0.27
5-3	-0.37	-0.47	-0.27	<.001	-0.53
6-3	-0.22	-0.30	-0.15	<.001	-0.34
5-4	-0.56	-0.64	-0.48	<.001	-0.82
6-4	-0.41	-0.46	-0.36	<.001	-0.62
6-5	0.15	0.07	0.23	<.001	0.23

Abbreviation: CPS, complex problem-solving.

TABLE 11 Frequencies of clusters depending on item position for correct responses in tickets task

Cluster	Item position in test		χ^2	<i>df</i>	<i>p</i>	φ
	Early	Middle				
1	1,152	1,432	30.34	1	<.001	0.11
2	777	1,158	75.02	1	<.001	0.20
3	336	417	8.71	1	.003	0.11
4	985	1,136	10.75	1	.001	0.07
5	338	319	0.55	1	.459	0.03
6	2,016	1,779	14.80	1	<.001	0.06

tickets task were shown by students who showed partly goal-directed and partly non-goal-directed guessing (cluster 5). In sum, the most frequent behaviour among the correct responses in the tickets task were rather minimalistic approaches (clusters 1, 3 and 4) with about 46%; the second most frequent were guessing approaches (clusters 5 and 6) with about 37.61%; and the least frequent was the double-checking approach (cluster 2) with 16.38%.

4.4 | General discussion

4.4.1 | Exploration behaviour

There are some characteristic behaviours that appear to be related to success in both tasks analysed in this study. In both tasks, the chi-squared tests revealed that non-targeted exploration is more frequent in false responses. This becomes also evident in the observed clusters, supporting the view of He and von Davier (2015) and Stadler et al. (2019), who found minimalistic behaviour to be most successful in CPS. However, only *repeated* goal-directed behaviour was found more frequently among correct responses and only in the tickets task. Moreover, since in both tasks the longest sequences were found among the correct responses and the shortest sequences were found among the false responses, sequence length seems to be positively related to success, in line with findings reported by Eichmann et al. (2018) and Naumann et al. (2014). Therefore, we assume long sequences of goal-directed behaviour, or in other words revisiting solution-relevant information, to be positively related to success while engagement with non-goal-directed information appears counterproductive. The distinction between goal-directed behaviour and non-targeted exploration we applied in the present study, therefore, reveals which specific types of behaviour are actually beneficial in CPS. What we defined as non-targeted exploration in our study seems to reflect confusion or distraction. In general, exploration is a rather broad category of behaviour that can include different things. The distinction of behaviours we applied, clarified the different roles exploration behaviour can fulfil in CPS. This finding has implications for the interpretation of past research outcomes as well as for the design of future research.

Our results help to clarify the ambiguous previous findings concerning the usefulness of exploration in CPS. Moreover, in line with our Research question we were able to identify patterns of behaviour that were associated with correct or false solutions in the CPS tasks analysed in this study and with overall CPS performance. We were also able to find commonalities and differences regarding these patterns between the tasks from two different CPS frameworks as intended with our Research question 2. We will further discuss the observed patterns in the following paragraphs.

4.4.2 | Perseverant approaches

In both tasks, we found students, who showed long sequences of goal-directed behaviour (i.e., who were presumably double-checking

their solutions), to show the highest overall CPS skills. However, this was the least frequently used approach in both tasks. This double-checking approach could indicate these students' tendency to act perseverant or conscientious. This behaviour seems to be positively related to CPS performance, which is in line with the results of Naumann (2015), who found a positive effect of goal-directed actions in digital reading tasks, which likewise can be classified as complex. Therefore, perseverant goal-directed behaviour might be an adaptive strategy of dealing with complexity in general. However, at least in the climate control task we also found quite long sequences of mostly non-targeted exploration behaviour among the false responses. This, on the other hand, indicates that perseverance or conscientiousness does not necessarily lead to high CPS performance. Nevertheless, these "perseverant non-targeted explorers" still showed higher overall CPS performance than students, who exhibited shorter sequences of mostly non-targeted exploration behaviour, which again confirms the results of Naumann et al. (2014). Among those perseverant non-targeted explorers, there were groups of students showing more repeated non-targeted exploration while others showed more initial non-targeted exploration. Showing mostly initial non-targeted exploration might indicate that students identify much information but struggle to identify the relevant one. However, showing mostly repeated non-targeted exploration might indicate that students are convinced of an incorrect way to solve the problem. These behaviours, characterized by much non-targeted exploration, were more frequent in the climate control task. One possible reason is that acquiring knowledge through variable manipulation (as it is required in LSE tasks) might be quite uncommon to students and lead to behaviour of trial and error. Beckmann and Guthke (1995) argue that this kind of behaviour can be associated with high motivation and poor performance.

4.4.3 | Minimalistic approaches

However goal-directed perseverant approaches were quite successful, shorter sequences were more frequent in our data. Students' apparent preference of shorter sequences of behaviour might be due to the set time limit or students' limited motivation to engage in a low-stakes assessment such as PISA. However, it should be kept in mind that short sequences refer only to a small number of interactions and not to timing, which was not looked at in the present study. Efficient goal-directed behaviour (or minimalism) was found mostly among the correct responses. This finding is in line with the results of Stadler et al. (2019) who found minimalistic behaviour to be related to high CPS performance. However, the overall CPS skills of these students were lower than those applying double-checking behaviour. A reason for the lower overall CPS performance of the minimalists compared to the double-checking students could be that minimalists have a higher chance to oversee mistakes they made and therefore have a higher probability of giving a false response than students, who double-check their responses. Of course this interpretation implies that students exhibit similar behaviour across all CPS tasks.

4.4.4 | Guessing approaches

Even shorter sequences than the minimalistic ones were shown by students guessing a solution. In the tickets task, there was a remarkably large group of students guessing the correct solution. However, students in this group showed only medium overall CPS performance. Nearly one third of the correct solutions in the tickets task were the result of a goal-directed guessing approach. One reason why students applied this approach (quite successfully) mostly to the tickets and not to the climate control task might be that in the tickets task the solution required only one guess (i.e., buying one ticket) while in the climate control task guessing a solution would require independently guessing several relations between the variables decreasing the chance of guessing the correct solution. Therefore, in LSA tasks guessing might not be regarded as an adaptive strategy by students while it might be regarded as adaptive in FSA tasks, if one is not capable of solving the task properly. Especially, in scenarios like the FSA task we investigated, which was about buying a subway ticket, students might apply goal-directed guessing that does not guarantee an optimal but a sufficiently good solution: Instead of investing time and effort to find the cheapest ticket, students might choose to use a heuristic by buying any ticket that would satisfy their requirements (Evans, 2008; Gigerenzer & Goldstein, 1996). This behaviour could be an expression of either not being motivated to invest much time and effort or not being able to invest time and effort due to perceived time pressure or due to not having understood that there might be a better option available. In the context of a low-stakes assessment such as PISA a rather low motivation of students seems not surprising. In contrast, when buying real subway tickets, the motivation to save money might be higher. Although goal-directed guessing led to a high number of correct solutions in the tickets task, most goal-directed guessing led to incorrect (yet plausible) solutions. In a real situation these plausible solutions translate to not buying the cheapest but a valid ticket.

Opposed to goal-directed guessing, guessing randomly led mostly to false solutions in both tasks. Moreover, guessing or not answering seems to be the most frequent behaviour leading to false solutions in the tickets task. The difference between the goal-directed and the random guessers might be that goal-directed guessers read and understood the task (otherwise they could not identify goal-directed actions), whereas random guessers do not seem to have read and understood the task at hand. Random guessers also show a lower overall CPS performance than goal-directed guessers. Therefore, students applying random guessing seem to have more fundamental difficulties in CPS than students applying goal-directed guessing. The source for these difficulties could be (apart from lacking motivation), for example, low reading abilities, which prevent the students from properly understanding and processing the task, or struggling with different components of problem-solving (Carlson, Khoo, Yaure, & Schneider, 1990). Similarly to our result, Naumann et al. (2014) reported low achieving students in technology-based problem-solving to exhibit particularly little interactions with the tasks. Therefore, not engaging enough with problem-solving tasks might be one of the most frequent maladaptive behaviours.

4.4.5 | Incomplete approaches

Among the false solutions in both tasks there are also groups showing an incomplete goal-directed approach. These students started their process quite promising but failed to find the correct solution in the end. However, these students showed a relatively high overall CPS performance compared to other groups who did not solve the respective task. The overall CPS performance of this group in the tickets task was comparable to that of the goal-directed guessers. Since the incomplete goal-directed group's sequences are not especially short, these students do not seem to be particularly unmotivated. They seem to focus on plausible but wrong solutions. Repeating their goal-directed actions a lot, they do not seem perfectly convinced of their solution. However, they struggle with considering other options.

4.4.6 | Resetting

In both tasks, there seems to be more frequent use of the reset button in correct responses than in the false responses. This implies students' use of the reset button in both frameworks is indicative of a systematic approach rather than mere trial and error. Especially in the tickets task, which has a fixed sequence of actions following each other, an unsystematic approach becomes evident in the rare and unsystematic use of the reset button in false solutions, since this task requires the use of the reset button at certain points. Resetting might also reduce cognitive load. Especially in the climate control task, resetting clears the visible information of past actions, and therefore also reduces unnecessary information as potential sources of distraction (Sweller, 1988). According to Stadler et al. (2019), a reduction of cognitive load should be related to higher CPS performance. However, since generally little resetting was done, no final conclusions should be drawn with respect to resetting. Therefore, these results should be verified by future research.

4.4.7 | Effects of task position

We found most of the clusters either more frequently at the early or in the middle item position. The results were quite different for the two tasks. In the climate control task, shorter sequences were more often observed at the middle position, that is the item was presented at the beginning of the second half of the test. Longer sequences were observed more frequently at the early position. This finding could indicate a loss of students' motivation during the first half of the test. Greiff et al. (2018) reported a similar result in a latent class analysis investigating students' exploration behaviour in the course of six CPS tasks from the LSA framework. They argued that students who exhibited declining exploration probably experienced a decrease in motivation.

In the tickets task, however, the results were quite different. In this task, most of the guessing clusters were more frequently observed in the early item position, while the more successful,

minimalistic and perseverant approaches were observed more frequently in the middle position. This suggests a different process than the results regarding the climate control task. Since the tickets task was more difficult than the climate control task (fewer students were able to solve it correctly), the tickets task might have been too difficult at the beginning of the test. However, during the test students might have learned how to approach tasks from the FSM framework, so the tickets task was easier when presented at the middle position. Greiff et al. (2018) found that some students improve their exploration strategies during assessment leading to more elaborated task processing at later task positions. Since this pattern was only observed in the tickets task, students might learn adaptive strategies for processing FSM tasks more easily than strategies for LSA tasks. Another reason why tasks from both frameworks might differ with respect to strategy learning and motivation might be that tasks from the FSM framework can look quite heterogeneous whereas LSA tasks usually appear very similar. In the PISA 2012 test, all LSA tasks share surface features that make the tasks look quite similar, even if they concern different topics. Therefore, students who struggled with an LSA task before might be unmotivated when a second LSA task is administered to them. In contrast, in the PISA 2012 assessment FSA tasks varied for example with respect to response type and interface design. Therefore, students got the impression of rather heterogeneous tasks that possibly did not demotivate those students, who experienced difficulties before. On the contrary, it seems students acquired more adaptive strategies for processing FSM tasks during the assessment. Future research might address this issue by comparing students' behaviour in a larger number of tasks.

4.5 | Limitations

The present study used an exploratory and correlational approach to investigate behaviour in CPS. Therefore, the interpretation of our results needs to be validated by future research. Moreover, despite the fact that we used two tasks that may be seen as prototypical examples within their respective frameworks, the generalizability of our results will have to be established by analysing behaviour in a wider range of tasks. In addition, it should be investigated whether the results can also be replicated in more complex problem scenarios (e.g., systems such as those used by Stemmann & Lang, 2018). Further, our sample did only include 15-year-old students. Therefore, we cannot make assumptions about the behaviour in CPS in other age groups. Additionally, our large sample size limits the meaningfulness of the statistical significances we found to some extent. Another limitation is that we did not use a model-based approach such as latent class or profile analysis to identify groups of students. Therefore, the choice of our clustering solution relies on the comparison of relative quality criteria and not on model selection comparing goodness of fit measures (Oberski, 2016). Also the assumption that students exhibit similar behaviour across different CPS tasks needs to be further investigated. Moreover, it should be kept in mind that the analysis of log data does not allow to identify the intention that students had when

exhibiting certain behaviours. Therefore, assumptions about causes of actions have to be validated in future research including experimental setups as well as think-aloud studies. Since PISA is a low-stakes assessment students' intentions might also be affected by a rather low motivation. Moreover, future research should also take timing information into account, since timing could be part of students' strategies in CPS. Further research is needed to overcome these limitations.

5 | CONCLUSION

We identified several behaviours associated with success or failure in CPS tasks. We observed a high proportion of goal-directed behaviour mostly among correct responses and a high proportion of non-targeted exploration mostly among false responses. Note that non-targeted exploration was defined as interactions *not necessary* to solve the task in this study, while required exploration was categorized as goal-directed behaviour. However, students applying double-checking approaches showed even higher CPS skills than students applying efficient, minimalistic approaches. Among the false solutions, extremely short behaviour sequences ultimately resulting in guessing a response are frequently observed especially in the tickets task. Therefore, the most frequent obstacles in CPS we found are abandoning a problem early and being sidetracked by goal-irrelevant content. Our findings hold true for both our investigated CPS tasks, however, the different behaviour patterns were found differently often in the two tasks. Thus, it seems our results are applicable to tasks from the LSE as well as from the FSA framework.

Overall, our results contribute to a better understanding of the processes that are related to success and failure in CPS. Moreover, they are a promising basis to make students more competent problem solvers. Encouraging students not to abandon problems early and teach them to identify and stick to the relevant aspects of problems might help them to become better problem solvers and prepare them for complex tasks they will most certainly encounter in their future. Moreover, knowledge about behaviour sequences related to success or failure in CPS makes it possible to identify the particular difficulties individual students are facing while solving complex problems. This information could be used to give students feedback about the aspects of their behaviour that are considered to be related to low CPS performance (Shute, 2008).

The detailed analysis of students' behaviour while solving complex problems allowed us to gain deeper insights into the processes in CPS. Most importantly, our results may help clarifying the role of exploration behaviour, specifically concerning the question of whether this kind of behaviour is beneficial in CPS. This knowledge may help to strengthen students' CPS skills and prepare them for the challenges of the 21st century.

ACKNOWLEDGEMENTS

This research was funded by the German Federal Ministry of Education and Research (Grant Numbers: 01LSA1504A and 01LSA1504B)

and by a project funded by the Fonds National de la Recherche Luxembourg (The Training of Complex Problem Solving; "TRIOPS").

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created in this study.

ORCID

Beate Eichmann  <https://orcid.org/0000-0001-7135-7945>

REFERENCES

- Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 14, 471–494.
- Apedoe, X. S., & Schunn, C. D. (2013). Strategies for success: Uncovering what makes students successful in design and learning. *Instructional Science*, 41, 773–791. <https://doi.org/10.1007/s11251-012-9251-4>
- Autor, D. H., Levy, F., & Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, 118, 1279–1333. <https://doi.org/10.1162/003355303322552801>
- Beckmann, J., & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 177–200). Hillsdale, NJ: L. Erlbaum Associates.
- Bell, B. S., & Kozlowski, S. W. J. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology*, 93, 296–316. <https://doi.org/10.1037/0021-9010.93.2.296>
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht, the Netherlands: Springer. https://doi.org/10.1007/978-94-007-2324-5_2
- Carlson, R. A., Khoo, B. H., Yaure, R. G., & Schneider, W. (1990). Acquisition of a problem-solving skill: Levels of organization and use of working memory. *Journal of Experimental Psychology: General*, 119, 193–214.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.
- Dormann, T., & Frese, M. (1994). Error training: Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction*, 6, 365–372. <https://doi.org/10.1080/10447319409526101>
- Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2018, April). *Using process data to explain group differences in complex problem solving*. Annual Conference of the the National Council on Measurement in Education (NCME), New York.
- Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, 128, 1–12. <https://doi.org/10.1016/j.compedu.2018.08.004>
- Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–278. <https://doi.org/10.1146/annurev.psych.59.103006.093629>
- Field, A., Miles, J., & Field, Z. (2013). *Discovering statistics using R (Reprint)*. Los Angeles, Calif.: Sage.
- Frensch, P. A., & Funke, J. (1995). Definitions, traditions, and a general framework for understanding complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 3–22). Hillsdale, NJ: L. Erlbaum Associates.
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, 7, 69–89. <https://doi.org/10.1080/13546780042000046>
- Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40. <https://doi.org/10.18637/jss.v040.i04>
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650–669.
- Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2014). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning*, 21, 356–382. <https://doi.org/10.1080/13546783.2014.989263>
- Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, 126, 248–263. <https://doi.org/10.1016/j.compedu.2018.07.013>
- Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36–46. <https://doi.org/10.1016/j.chb.2016.02.095>
- Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, 91, 92–105.
- Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving. *Applied Psychological Measurement*, 36, 189–213. <https://doi.org/10.1177/0146621612439620>
- He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with N-grams. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Springer proceedings in Mathematics & Statistics. Quantitative Psychology Research* (Vol. 140, pp. 173–190). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-19977-1_13
- Jirout, J., & Zimmerman, C. (2015). Development of science process skills in the early childhood years. In K. C. Trundle & M. Saçkes (Eds.), *Research in early childhood science education* (pp. 143–165). Dordrecht, the Netherlands: Springer.
- Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log-data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, 45, 527–563.
- Mislevy, R. J. (2019). On integrating psychometrics and learning analytics in complex assessments. In H. Jiao, R. W. Lissitz, & A. van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 1–52). Charlotte, NC: Information Age.
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, 133–161.
- Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior*, 53, 263–277. <https://doi.org/10.1016/j.chb.2015.06.051>
- Naumann, J., Goldhammer, F., Rölke, H., & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen [Successful problem solving in technology rich environments: Interactions between number of actions and task demands]. *Zeitschrift für Pädagogische Psychologie*, 28, 193–203. <https://doi.org/10.1024/1010-0652/a000134>
- Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson & M. Kaptein (Eds.), *Human-computer interaction series. Modern statistical methods for HCI* (1st ed., pp. 275–287). Cham, Switzerland: Springer.

- OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. PISA. Paris, France: OECD.
- OECD. (2014a). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. V). Paris, France: OECD. <https://doi.org/10.1787/9789264208070-en>
- OECD. (2014b). *PISA 2012: Technical report*. Paris, France: OECD. Retrieved from <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>
- R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Revelle, W. (2018). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from <https://CRAN.R-project.org/package=psych>
- Richter, T., Naumann, J., & Noller, S. (2003). LOGPAT: A semi-automatic way to analyze hypertext navigation behavior. *Swiss Journal of Psychology*, 62, 113–120. <https://doi.org/10.1024//1421-0185.62.2.113>
- Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test Analysis Modules*. Retrieved from <https://CRAN.R-project.org/package=TAM>
- Schult, J., Stadler, M., Becker, N., Greiff, S., & Sparfeldt, J. R. (2017). Home alone: Complex problem solving performance benefits from individual online assessment. *Computers in Human Behavior*, 68, 513–519. <https://doi.org/10.1016/j.chb.2016.11.054>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153–189. <https://doi.org/10.3102/0034654307313795>
- Signorell, A. et al. (2019). DescTools: Tools for descriptive statistics. Retrieved from <https://cran.r-project.org/package=DescTools>
- Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology*, 106, 681–695. <https://doi.org/10.1037/a0035506>
- Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*, 10, 248. <https://doi.org/10.3389/fpsyg.2019.00777>
- Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior*, 65, 100–106. <https://doi.org/10.1016/j.chb.2016.08.025>
- Stemmann, J., & Lang, M. (2018). Eignet sich die logfilegenerierte Explorationsvollständigkeit als Prozessindikator für den Wissenserwerb im problemlösenden Umgang mit technischen Alltagsgeräten? [Is logfile-generated exploration completeness suitable as a process indicator for knowledge acquisition in handling of everyday technical devices?]. *Journal of Technical Education*, 6, 185–199.
- Studer, M. (2013). *WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R*. LIVES Working Papers, 24. doi: <https://doi.org/10.12682/lives.2296-1658.2013.24>.
- Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179, 481–511. <https://doi.org/10.1111/rssa.12125>
- Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40, 471–510. <https://doi.org/10.1177/0049124111415372>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Tóth, K., Rölke, H., Greiff, S., & Wüstenberg, S. (2014). Discovering Students' Complex Problem Solving Strategies in Educational Assessment. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Chairs), Proceedings of the 7th International Conference on Educational Data Mining London : CEUR Workshop Proceedings. Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/short%20papers/225_EDM-2014-Short.pdf
- Wilpert, B. (2009). Impact of globalization on human work. *Safety Science*, 47, 727–732. <https://doi.org/10.1016/j.ssci.2008.01.014>
- Wirth, J. (2004). *Selbstregulation von Lernprozessen* [Self-regulation of learning processes]. *Pädagogische Psychologie und Entwicklungspsychologie* (Vol. 39). Münster, Germany: Waxmann.
- Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, 29, 18–29. <https://doi.org/10.1016/j.lindif.2013.10.006>

How to cite this article: Eichmann B, Greiff S, Naumann J, Brandhuber L, Goldhammer F. Exploring behavioural patterns during complex problem-solving. *J Comput Assist Learn*. 2020; 1–24. <https://doi.org/10.1111/jcal.12451>