Learning and Individual Differences 128 (2026) 102893

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Double-checks or better no repeated steps? The role of effective and efficient exploration behavior for successful complex problem solving

Julia Ruby [a,*], Julius J. Weise [a,b], Samuel Greiff [c], Jörn R. Sparfeldt [a]

[a] Saarland University, Department of Educational Science, D-66123, Saarbrücken, Germany
[b] Saarland University, Institute for Medical Biometry, Epidemiology and Medical Informatics, D-66421, Homburg, Germany
[c] Technical University of Munich, TUM School of Social Sciences and Technology, D-80333, München, Germany

ARTICLE INFO

Keywords:
Complex problem solving
Intelligence
Latent class analysis
Log-file analyses
VOTAT

ABSTRACT

Complex problem solving (CPS) is a prominent competence, significantly related to educational achievements. Regarding CPS, the domain-general exploration strategy VOTAT (vary-one-thing-at-a-time; Tschirgi, 1980) is strongly correlated with CPS success. However, the number of additional VOTAT exploration steps remained unclear. Going beyond previous studies, we combined these approaches and analyzed logfile based process data. The effectiveness (i.e., using VOTAT) and the efficiency (i.e., non-necessary, additional VOTAT-steps) of the exploration behavior were investigated across a sequence of CPS items in $N = 469$ high-school students. Across the item sequence, the exploration behavior became more effective and efficient. In latent class analyses, four classes were identified. Students in classes that demonstrated more effective exploration behavior reached higher intelligence and CPS performance scores than other students. Regarding the remaining classes, the relation between efficient exploration behavior with intelligence and CPS performance was less consistent. Implications for fostering students' explorations and CPS success are discussed.

Educational relevance and implications statement: The results of our study emphasize the role of strategic exploration behaviors while exploring complex problem solving (CPS) tasks for successfully solving complex problems. Based on more or less effective (i.e., using VOTAT) and efficient (i.e., the number of additional, non-necessary exploration steps) strategy use, we found four different groups (i.e., latent classes) of students: (1) ineffective explorers, (2) inefficient explorers, (3) emerging explorers, and (4) proficient explorers. Students of the four classes differed in their mean intelligence and CPS scores. These results form a basis for adaptive interventions aimed at fostering beneficial exploration strategies to improve problem solving skills as an important educational goal.

1. Introduction

Complex problem solving is an intensively studied 21st century skill (e.g., Greiff et al., 2013; OECD, 2014) that is closely related to and significantly predicts educational achievements (e.g., Kretzschmar et al., 2016; Lotz et al., 2016). Notably, more differentiated scientific knowledge regarding beneficial strategic behaviors while solving complex problems is still needed (e.g., Lotz et al., 2022; Nicolay et al., 2023). Furthermore, groups of students might differ in their strategic behaviors while working on single complex problem items and regarding the strategic behavior across a sequence of complex problems. Elucidating the mechanism by which effective and efficient strategic behaviors promote successful solutions of complex problems seems to be a promising approach and basis for designing future training programs to foster students' problem solving skills.

According to Buchner's widely accepted definition (in Frensch & Funke, 1995, p. 14), complex problem solving (CPS[1]) is considered as the "successful interaction with item environments that are dynamic (i.e., change as a function of user's intervention and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process". Typically, CPS starts with the

* Corresponding author at: Saarland University, Campus A5 4, D-66123, Saarbrücken, Germany.
E-mail address: julia.ruby@uni-saarland.de (J. Ruby).

[1] Please note that the abbreviation CPS is also used in other fields of research, for example Creative Problem Solving (e.g., Shin et al., 2025) or Collaborative Problem Solving (e.g., Rojas et al., 2025).

https://doi.org/10.1016/j.lindif.2026.102893
Received 21 May 2024; Received in revised form 28 January 2026; Accepted 1 February 2026
Available online 16 March 2026
1041-6080/

exploration of the unknown problem space followed by attempts to solve the problem (e.g., Funke, 2001). To successfully explore CPS items with direct effects linking one input to one output variable, the so-called VOTAT (vary-one-thing-at-a-time) strategy (Tschirgi, 1980) is an important and optimal, or in other words effective, strategic exploration behavior. Related to the number of exploration steps, individuals might show an effective VOTAT behavior once or more than once for an input variable in a single item (i.e., efficiency). Such repeated VOTAT behaviors do not increase the amount of information about the problem space, but such double-checking behaviors might be useful by increasing confidence or consolidating and integrating knowledge about these effects. However, the effectiveness and the efficiency of strategic behaviors, their progression over an item sequence, and the relations of these strategic behaviors to successful complex problem solving remain to be clarified. Furthermore, complex problem solving and intelligence are closely connected (e.g., Kröner et al., 2005; Stadler et al., 2015). To date, however, knowledge is still limited about the connection between beneficial strategic behaviors while solving complex problems and intelligence (e.g., Lotz et al., 2017; Weise et al., 2022).

Following this, the aims of the present study were fourfold. First, we explored how many students showed an effective and efficient exploration behavior for five CPS items with only direct effects (i.e., effective VOTAT, efficient VOTAT). Second, we searched for (latent) classes of students' strategic exploration behavior observed over the item sequence. Third, we investigated whether this class membership was related to the intelligence scores. Fourth, we examined whether the class membership was related to the CPS performance scores. Thereby, our analyses went beyond earlier studies by analyzing the effectiveness and the efficiency of students' exploration behavior. Consequently, we inspected log-file data that provided information about how effectively and efficiently the students explored the system. Hence, we investigated the strategic exploration behavior, CPS performance, and intelligence in 10th and 11th graders to enrich our knowledge about strategic behavior, its development across an item sequence, and the relationship to CPS performance and intelligence.

1.1. Complex problem solving and its assessment

Problem solving is described as a non-routine action or sequence of actions, in which a given state is transformed into a goal state (Mayer & Wittrock, 2006). Complex problem solving as a subtype is assessed with complex problems, characterized by five features (Dörner et al., 1983): complexity (several variables), connectivity (various connections between the input and output variables), dynamics (problem changes over time), non-transparency (not all information to solve the problem is initially available, requiring exploration of the problem), and polytely (several criteria need to be achieved). These characteristics require an active search for information and corresponding manipulations to identify and understand the underlying structure of the problem space. In contrast to successful problem solving (where "only" cognitive processing is sufficient), successful CPS requires active interactions of the problem solver with the problem. Accordingly, the OECD framework employed in the PISA 2012 cycle outlined four central processes for CPS: exploring and understanding, representing and formulating, planning and executing, as well as monitoring and reflecting (OECD, 2014).

Concerning the assessment of CPS, simulated real world problems were implemented in corresponding computer models of so called microworlds dealing with, for example, acting as the mayor of a small town (Lohhausen: Dörner et al., 1983) or manager of a factory (Tailorshop: Putz-Osterloh, 1981). Although these aforementioned features resulted in high face validity (Greiff, Stadler et al., 2015), these historically early assessments often exhibited significant psychometric deficiencies (e.g., limited evidence regarding reliability), as repeatedly mentioned (e.g., Greiff, Stadler, et al., 2015; Kröner et al., 2005; Süß, 1996). Notably, those single task simulations like Lohhausen (Dörner et al., 1983) are time consuming and task difficulties do not differ within one assessment. Furthermore, incorrect or random steps at the beginning of working on such a CPS program significantly impede overall performance. To overcome the deficiencies of those single task assessments and to improve the psychometric characteristics, the minimal complex system approach with various, rather short, and less complex scenarios was established, where Finite State Automata (FSA) or Linear Structural Equations (LSE) were used as frameworks. FSA (e.g., HeiFi; Funke et al., 1998) consist of a limited number of distinct states that can be transformed into one another, for example, through user interaction; this is found in everyday life in many technical devices such as ticket machines and vending machines. In contrast, tasks from the LSE framework (e.g., MicroDYN; Greiff et al., 2012) consist of a limited number of input and output variables, which are connected through linear equations. In both frameworks, CPS tasks have been constructed covering a variety of theory-based item difficulties (Stadler et al., 2016). In particular the MicroDYN approach is frequently used, as also in our study, based on theoretical and promising psychometric characteristics. These items cover a wide range of item difficulties (e.g., Greiff et al., 2012; Lotz et al., 2022). The reliability estimates of the corresponding scales are at least adequate (e.g., Greiff et al., 2012: $.85 \leq \alpha \leq .95$; Schweizer et al., 2013: $.71 \leq \alpha \leq .86$). With respect to validity estimates, repeated findings indicated at least moderate to high correlations with intelligence (see below) and school grades (e.g., Kretzschmar et al., 2016; Lotz et al., 2016).

In each MicroDYN item, a cover story like training a handball team or providing medical aid is implemented, connecting a set of input and output variables. The relations between the variables can be classified into two basic types: direct effects and eigendynamic effects. Direct effects represent a relationship in which an input variable is linked directly with an output variable (see Fig. 1: e.g., *blue chips → Grande*). Eigendynamic effects are characterized by an output variable changing over time without any intervention of the problem solver indicating growth processes. Each item consists of two phases, based on theoretical and empirical considerations (e.g., Funke, 2001; Wüstenberg et al., 2012): the knowledge acquisition phase and the knowledge application phase. In the knowledge acquisition phase, test takers are expected to explore the system and its underlying structure to construct a causal diagram that represents the correct relations between the input and output variables (see Fig. 1). To explore the system, test takers manipulate the input variables by using the plus and minus buttons on the left; after clicking the Apply button, the effects on the output variables are shown in the diagrams on the right. The input variables remain the same from round to round unless they are changed by the test taker (or the Reset button is clicked). Finally, test takers are supposed to identify the relations between the variables and mark them with arrows in the diagram at the bottom of the user interface. In the subsequent knowledge application phase, test takers are provided with full information about the variables and their relations (i.e., diagram with correct arrows). Then, test takers are expected to use this information to reach the target values of the output variables within four manipulation steps of the input variables (see Fig. 2). Empirically, the performances in the knowledge acquisition phase and the knowledge application phase correlated substantially, reaching at least medium size (e.g., Greiff et al., 2012; Wüstenberg et al., 2012).

In the knowledge acquisition phase, test takers might strategically explore a yet unknown complex problem. An effective strategic behavior to identify direct effects in a MicroDYN item or another CPS item based on the LSE framework is the so-called VOTAT strategy (vary-one-thing-at-a-time, Tschirgi, 1980; see also "control of variables strategy", Chen & Klahr, 1999), meaning that one input variable is varied and the others are set to 0. In contrast, eigendynamic effects can be reliably detected with the strategic behavior NOTAT (vary-no-thing-at-a-time; all input variables are set to 0; Lotz et al., 2017). Although not yet implemented in MicroDYN items such as those used in our study, additional effects could be investigated. For example, interaction effects could be detected successfully by applying HOTAT (hold-one-thing-at-a-time; one variable is

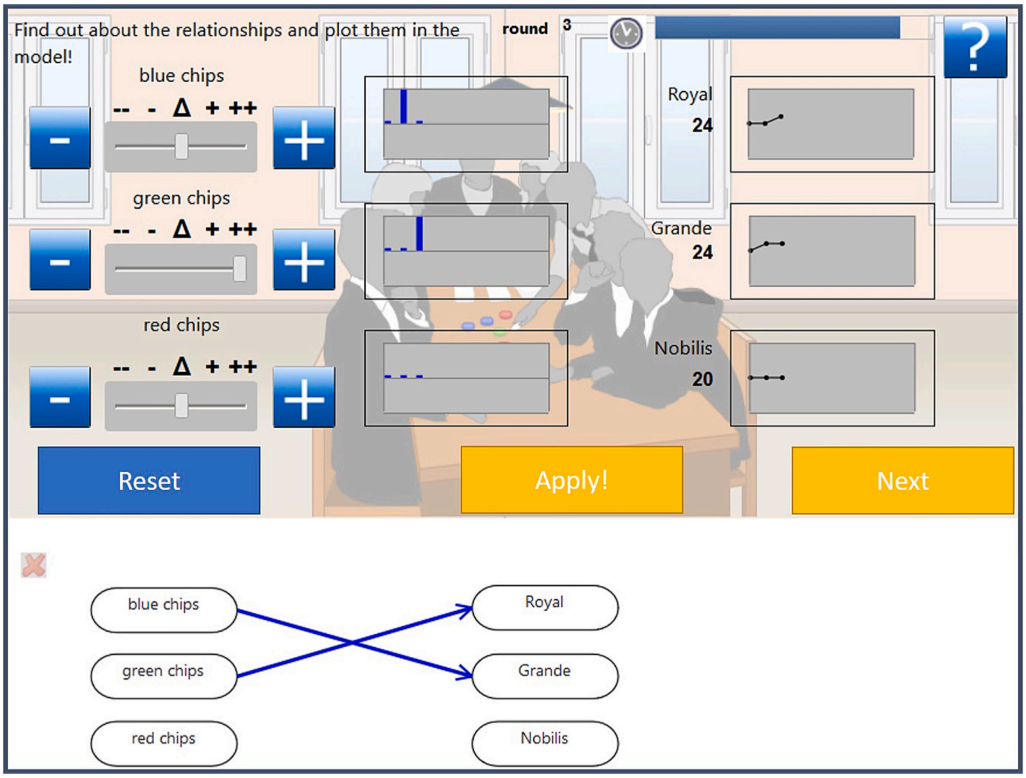


Fig. 1. Screenshot of the MicroDYN item *Game* during the knowledge acquisition phase.
Note. At the point in time shown in the figure, the test taker has executed two exploration steps: VOTAT was executed once for *blue chips* (in the first exploration step) and once for *green chips* (in the second exploration step). Concerning the diagram in the lower part of the figure, the test taker has drawn two arrows regarding the information obtained about the system structure from these exploration steps.

kept constant, while the others are varied; Tschirgi, 1980). Similarly, non-linear relations could be detected by setting one input variable multiple times to different non-zero values and keeping the remaining input variables at zero. As only direct effects were implemented in the items used in our study, we focused mainly on VOTAT (see for further strategies, e.g., Nicolay et al., 2023; Rollett, 2008).

1.2. The relevance of an effective and efficient exploration behavior

The well-researched strategy VOTAT is described as optimal strategic exploration behavior to detect direct effects, thereby providing all information about direct effects (e.g., Csányi & Molnár, 2025; Lotz et al., 2022). Accordingly, medium to high correlations were reported between VOTAT application and CPS performance in the knowledge acquisition/knowledge application phase (e.g., Kröner et al., 2005: r = .47/.40; Greiff, Wüstenberg, & Avvisati, 2015: r = .67/.61; Vollmeyer et al., 1996: r = .76/.32; see also e.g., Greiff et al., 2016; Lotz et al., 2022; Pejić & Molcer, 2021; Wüstenberg et al., 2014). Furthermore, test takers using VOTAT solved complex problems more successfully than test takers using other strategies, although some of those other strategies also provided all the necessary information (Molnár & Csapó, 2018). Due to the outstanding importance of VOTAT, we focused on VOTAT in our study. In accordance with the typically used dichotomous VOTAT scoring (Greiff et al., 2012), we labeled the exploration behavior as effective when VOTAT was applied at least once for each input variable, and as ineffective when VOTAT was not applied at least once for each input variable.

Notably, 85.2% of the undergraduate students in the study by Csányi and Molnár (2025) applied VOTAT (almost) consistently in all MicroDYN items. These so-called proficient explorers showed, however, a substantial amount of variability in the number of exploration steps (i.e., clicks). VOTAT can be applied once or repeatedly (besides applying non-VOTAT steps). Importantly and due to the underlying system structure of the CPS program used in our study (i.e., MicroDYN), using VOTAT once for an input variable provides all information about the direct effects from this input variable; repeating VOTAT provides no additional information about the system. Notably, test takers are typically not instructed about which types of effects can (not) occur, and repeated VOTAT application might help to clarify this. Furthermore, such repetitions might be helpful to consolidate students' knowledge about the system or indicate cognitive integrations. In sum, it might be useful to further explore the large group of proficient explorers, possibly consisting of relevant subgroups differing in the number of exploration steps.

Concerning the number of interactions, students with equal CPS performance scores showed substantial variations (Stadler et al., 2020). Furthermore, lower numbers of exploration steps were significantly related to better CPS performance in the knowledge acquisition and the knowledge application phase (Greiff et al., 2016). In another study, the relations differed depending on item complexity: For easier and less complex items, a lower number of interaction steps was associated with more successful transitions between exploring, drawing the model, and controlling the system (and, thus, also higher CPS performance scores), whereas for more difficult and more complex items, a higher number of exploration steps was beneficial (Molnár & Greiff, 2023). Additionally, the effectiveness and the number of exploration steps differed between the students from two countries. Specifically, Hungarian students explored more effectively and carried out more interactions than Jordanians (Molnár et al., 2022). For students from both nationalities, the number of clicks correlated positively with CPS performance. For Hungarian students, the number of interactions decreased across the item sequence, even though the complexity and the difficulty of the items increased. Inspecting a German sample and another problem solving task (i.e., not MicroDYN, not CPS), Naumann et al. (2014) reported, for

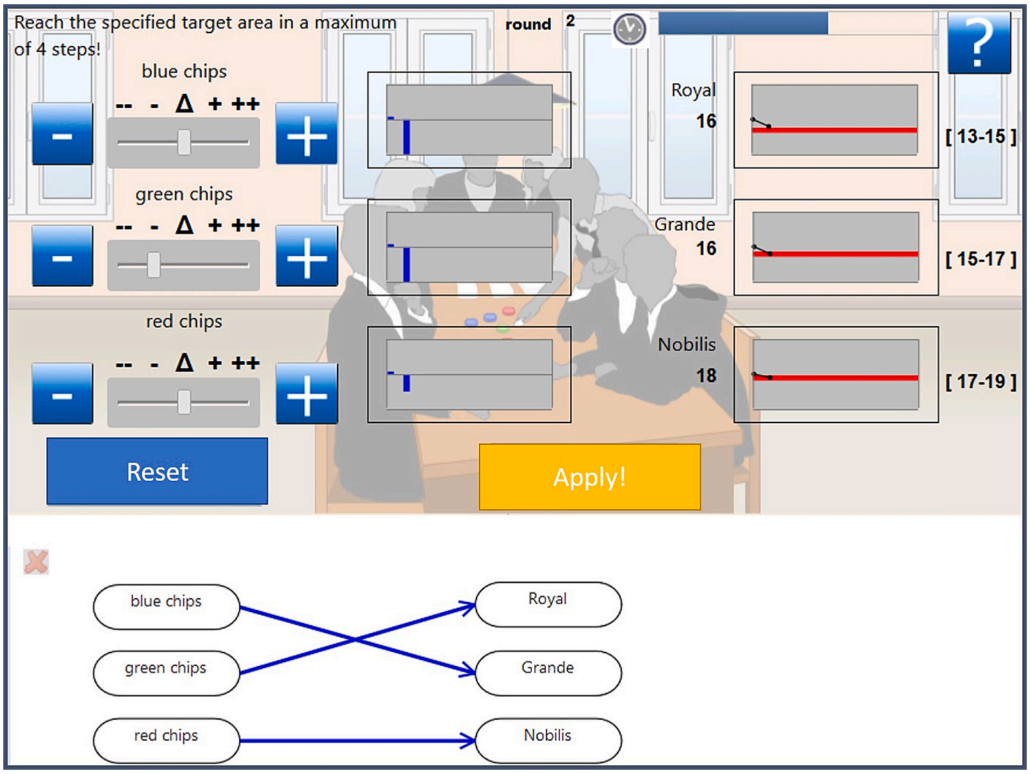


Fig. 2. Screenshot of the MicroDYN item *Game* during the knowledge application phase.
Note. The test taker was provided with the correct diagram showing all relations between input and output variables (lower part of the figure). At the point in time shown in the figure, the test taker has executed one control step (sliders for *blue chips* and *green chips* moved to "- -", slider for red chips moved to "-"; goal states were reached for *Grande* and *Nobilis*) and is about to execute the second control step (i.e., sliders were moved, Apply button not yet pressed).

items requiring higher numbers of interactions, a positive relation between the number of interactions and performance as well as a quadratic relation indicating that performance increased with more interactions up to an optimum and then decreased again. For items requiring lower numbers of interactions, no significant correlation between the number of interactions and CPS performance was found. Relying on a person-centered approach of the exploration behavior, Eichmann, Greiff, et al. (2020) identified three clusters of students who successfully solved one MicroDYN item: efficient exploration, double-checking, and mixed approach. The double-checking students reached the best overall-CPS performance, whereas students from the remaining clusters did not differ significantly in their CPS performance. Importantly, these authors distinguished between goal-directed actions (i.e., every interaction necessary to solve the task correctly, irrespective of whether the interaction was initial or repeated) and non-goal-directed actions (i.e., every action not necessary to solve the task). Double-checking students showed a high amount of goal-directed actions that were seemingly beneficial for successful CPS performance. Inspecting the exploration behavior in a finite state automata task, Zhang et al. (2024) found seven clusters of students who solved this task successfully, but who differed, among other things, in the number of exploration steps and the performance score in a set of CPS tasks. Despite the distinction between goal-directed behavior and non-goal-directed behavior, the authors have unfortunately not considered in their analyses whether all of the necessary exploration steps for an effective exploration behavior had been executed. In conclusion, not only the quantity but also the quality of the exploration steps was important for CPS performance in the aforementioned studies.

Correspondingly, investigating the combination of effective exploration behavior (i.e., VOTAT) with the number of exploration steps is fruitful to further clarify students' interactions with an unknown system. Csányi and Molnár (2025) already combined those two approaches in the aforementioned study, whereby the application of VOTAT was used to classify the students and the number of interactions was then examined as dependent variable to characterize the different classes. Their largest class comprised students who applied VOTAT (almost) consistently and included 85.2% of the participants. In further research, a closer look at students applying VOTAT consistently might be fruitful; for example, corresponding subgroups of students might differ in the number of exploration steps. Based on these considerations, we combined consistent VOTAT application with the number of exploration steps to form effective and efficient exploration behavior. Specifically, we labeled applying VOTAT at least once per input variable as effective exploration behavior, as aforementioned. Furthermore, we defined a maximum efficient exploration behavior through applying VOTAT exactly once per input variable and operationalized the degree of efficiency as the number of exploration steps exceeding the minimum number of necessary VOTAT steps. Therefore, effectiveness constitutes a prerequisite for efficiency. For example, the exploration behavior of a student showing VOTAT exactly once for each input variable is defined as effective and maximum efficient; the exploration behavior of another student showing VOTAT for all input variables, but with two steps more than necessary, is also defined as effective, but less efficient (efficiency score: 2). Whereas effective exploration behavior is closely related to CPS performance, it is yet unclear whether more or less efficient exploration behavior is related to better performance. Building on previous studies (e.g., Csányi & Molnár, 2025; Molnár et al., 2022), further research should clarify how the effectiveness and the efficiency of the exploration behavior develop across an item sequence and how they relate to CPS performance.

1.3. Development of the exploration behavior across an item sequence

When students are working on more than one CPS item, inspecting the development of the exploration behavior over the item sequence might be fruitful. When problem solvers have figured out VOTAT as an effective strategy to gain relevant information about the effects of the item, one can assume that they proceed to explore the next item by applying VOTAT for each input variable. Therefore, the exploration behavior of test takers, who realize the underlying system structure of CPS items, should become more effective across the item sequence. Regarding efficiency, assumptions about the relations with CPS success are less clear, because either repeated or no unnecessary exploration steps might be related to high performance, as mentioned before. Assumingly, realizing the underlying item structures and the involved effects should go hand in hand with a better understanding about necessary and useful exploration steps; this might be related to an increase in efficiency of the exploration behavior across the item sequence. Correspondingly, better understanding of the system structure might go hand in hand with better (i.e., effective and possibly more efficient) exploration behavior, which in turn enhances problem solving performance scores.

Investigating the problem solving behavior across an item sequence in students from grade three to twelve, Molnár and Csapó (2018) ran latent class analysis (LCA; see also Greiff et al., 2018). Three classes comprised students without relevant changes of their exploration behavior across the item sequence. Specifically, students of class 1 did not use and students of class 2 rarely used an effective exploration behavior, whereas students of class 3 consistently used effective exploration strategies. In the other three classes, the exploration behavior changed across the item sequence. Specifically, students of class 4 partially used effective exploration strategies in the first and easy items, but did not explore effectively in the later and more complex ones. The students of class 5 mostly used effective exploration strategies in the first and easy items, but not in the later and more complex items. The students of class 6 did not show an effective exploration behavior in the beginning, but in the end. In conclusion, these authors found six meaningful classes characterizing differences in the effectiveness of the exploration behavior. However, there is a need for further research to examine (beside the effectiveness) the efficiency of the exploration behavior.

Using latent transition analysis as another person-centered approach, Mustafić et al. (2019) inspected the VOTAT behavior of students across a sequence of six MicroDYN-items and identified four groups: Low-performing explorers failed to use VOTAT most likely in all items. Rapid learners did not apply VOTAT in the first three items, but were more likely to use VOTAT in the last three ones. In contrast, emerging explorers used VOTAT in the first three items but only partially in the last items. Proficient performers applied VOTAT consistently in all items. In conclusion, the students of these four groups showed important differences in their VOTAT based exploration behavior. Although named slightly different, four very similar classes resulted from an LCA (Wu & Molnár, 2022): proficient exploration strategy users, rapid learners, non-persistent explorers, and non-performing explorers. Concerning the knowledge acquisition and the knowledge application phase, proficient exploration strategy users achieved the highest performance scores, followed by rapid learners, non-persistent explorers, and finally non-performing explorers. By inspecting the exploration behavior with an LCA in Hungarian and Jordanian students in the aforementioned study, Molnár et al. (2022) identified four classes in both nationalities, two classes were identical in both samples (non-performing explorers, restarting slow learners). Furthermore, rapid learners and proficient explorers were found in the Hungarian sample, and non-persistent explorers and almost proficient explorers in the Jordanian sample. Additionally, Csányi and Molnár (2025) found four classes in a Latent Profile Analysis based on the VOTAT application: proficient explorers, rapid learners, ineffective learners, and non-performers. Proficient explorers showed the highest number of interactions and the best CPS performance, rapid learners a medium number of interactions and medium performance, ineffective learners a medium number of interactions and low performance, and non-performers took the fewest exploration steps and reached low CPS scores. In the first six items, the number of interactions decreased for all classes, with non-performers carrying out the least interactions in contrast to the remaining, rather similar classes. In the last four, more difficult items, the number of interactions remained rather similar for non-performers and increased for the other three classes. Proficient explorers showed a substantial increase, rapid learners and ineffective learners a smaller increase of the number of interactions.

In conclusion, the number of classes, their profiles, and the interpretations of the profiles differed in prior studies, calling for further research. Furthermore, class formation was based on the effectiveness of the exploration behavior in previous studies. In contrast, efficiency has not yet been considered as a basis for class formation in person-centered analyses. Students, who have realized that applying VOTAT once per input variable is sufficient to obtain all the information about the system and who have thus understood the basic structure of the tasks, might show efficiency increases and higher CPS performance scores. In contrast, double-checking might also be related to better CPS performance.

1.4. The relevance of intelligence

CPS and intelligence overlap conceptually. For example, Gottfredson (1997, p. 13) conceptualized problem solving as a constituting part of intelligence in her well-known definition. In intelligence test items such as figural matrices, test takers are provided with all information to solve the item from the start, and the problem does not change over time (e.g., Jonassen, 2000). In contrast, complex problems change over time and the solving process requires the interaction of the test taker with the problem (e.g., Stadler, Niepel, & Greiff, 2019). Nevertheless, there are many similarities between intelligence and CPS items concerning the required solving process such as rule induction and reasoning. Correspondingly, Stadler et al. (2015) reported in a meta-analysis an unadjusted mean relationship of $M(g) = .43$ between intelligence and CPS performance. Moreover, in studies with psychometrically strong CPS assessments and operationalizations of intelligence in the sense of g (e.g., Jensen & Weng, 1994), even higher correlation coefficients were found (e.g., Lotz et al., 2017: $r = .69/.76$; Kretzschmar et al., 2016: $r = .85$; Kröner et al., 2005: $r = .75$). Additionally, CPS correlated significantly with inductive reasoning ($r = .44$, Molnár et al., 2013).

Regarding the relations between beneficial strategic behaviors while solving CPS items and intelligence, positive correlations have also been repeatedly reported (for a review see Lotz et al., 2017: $.17 \leq r \leq .64$). For example, significant correlations were shown between fluid intelligence and VOTAT use ($r = .64$, Wüstenberg et al., 2014) and between a broader assessment of intelligence in the sense of g and the frequency of VOTAT applications ($.40 \leq r \leq .48$; Lotz et al., 2017). Furthermore, more intelligent individuals applied more effective strategic behaviors (i.e., VOTAT) while exploring an unknown problem space of complex problems, and those students who had selected and applied more effective strategic behaviors while exploring the complex problems solved the problems more successfully afterwards. Correspondingly, effective strategic behaviors meditated the intelligence-CPS relation (Lotz et al., 2022). In the aforementioned study by Wu and Molnár (2022), the inductive reasoning scores of proficient exploration strategy users were higher than those of rapid learners, followed by non-persistent explorers and, finally, non-performing explorers. Although noteworthy results regarding the relations between intelligence (and inductive reasoning as important intelligence factor closely related to g) with CPS performance and strategy use had been reported in former studies, their relations with the effectiveness and especially the efficiency of the strategy use should be further clarified.

To the best of our knowledge, the relation between intelligence and the number of exploration steps in CPS tasks has, unfortunately, not yet been researched. Theoretically, repeated VOTAT applications do not provide any new information, but may contribute to a better understanding of the underlying structure of the item. Possibly, more intelligent explorers might realize earlier and faster than less intelligent explorers that the underlying structure of MicroDYN items (based on linear equation systems) remains consistent across an item sequence although the number of variables and the number of effects may vary. Correspondingly, more intelligent explorers should not only realize that VOTAT is an effective strategic exploration behavior, but might also detect that non-necessary, repeated exploration steps do not provide any additional information about the system and that a single execution of VOTAT per input variable is sufficient. Thus, they might gradually explore more efficiently across the item sequence.

1.5. *The present study*

As previously mentioned, it is needed to take a closer look at the effectiveness and especially the efficiency of the exploration behavior, the development of these two aspects across an item sequence, as well as their relations with CPS performance and intelligence. Therefore, we inspected the effectiveness and the efficiency of the exploration behavior in five low to medium complex (derived from the results of Molnár & Greiff, 2023) MicroDYN items with direct and without eigendynamic effects. Subsequently we investigated the development of the exploration behavior over the item sequence. Furthermore, we explored qualitatively differing classes of students' exploration behavior across the item sequence. Based on those results, we analyzed mean differences between the classes' intelligence and CPS performance. Specifically, we investigated the following research questions (RQ):

RQ1. How many students showed an effective exploration behavior and a more or less efficient exploration behavior in each item and across the item sequence? Across the item sequence, we expected an increasing number of students who showed an effective exploration behavior as well as an effective and efficient exploration behavior. Accordingly, we expected a decreasing number of students showing an ineffective or an effective but less efficient exploration behavior.

RQ2. Can different (latent) classes of students, based on the development of their exploration behavior, be found? In the aforementioned studies inspecting effectiveness (e.g., Csányi & Molnár, 2025; Molnár et al., 2022; Molnár & Csapó, 2018; Mustafić et al., 2019; Wu & Molnár, 2022), a number of latent classes was detected. Specifically, we inspected the development of effective and efficient exploration behavior across the item sequence forming the basis for class formation. Although we ran an explorative analysis, we also expected a number of (qualitatively) different classes of students. We expected one class of students who showed consistently an ineffective exploration behavior; such a class was also found in the aforementioned studies. We expected another class of students whose exploration behavior was consistently effective and became more efficient across the item sequence; a class of students exhibiting consistently VOTAT use was found in the aforementioned studies as well. In addition, we expected at least one further class with rather moderate, but increasingly effective and efficient exploration behavior (in accordance with the aforementioned former studies reporting effectiveness improvements across the item sequence). Thus, we expected at least three different classes.

RQ3. Do the students of the different classes from RQ2 differ in their mean intelligence? Relying on former studies consistently reporting at least small to moderate positive correlations between intelligence and VOTAT use (e.g., Lotz et al., 2017; Wu & Molnár, 2022; Wüstenberg et al., 2014), we expected higher mean intelligence scores for classes comprising students with an effective and more efficient exploration behavior than for classes comprising students with an ineffective or an effective but less efficient exploration behavior.

RQ4. Do the students of the different classes from RQ2 differ in their mean CPS performance scores? Based on consistently high positive correlations between VOTAT use and CPS performance reported in former studies (e.g., Greiff et al., 2016; Lotz et al., 2022; Pejić & Molcer, 2021; Wüstenberg et al., 2014) regarding effectiveness, and inconclusive evidence regarding the number of steps (e.g., Csányi & Molnár, 2025; Eichmann, Greiff, et al., 2020; Greiff et al., 2016; Molnár & Greiff, 2023; Zhang et al., 2024) regarding efficiency, we expected higher CPS achievement scores for classes comprising students with an effective and more efficient exploration behavior than for classes comprising students with an ineffective or an effective but less efficient exploration behavior.

2. Methods

2.1. *Sample and procedure*

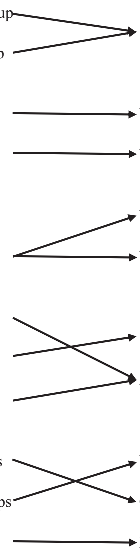
The sample consisted of $N = 495$ German high-school students ($n = 264$ females, $n = 228$ males, $n = 3$ without gender specification; age $M = 16.40$, $SD = 0.94$ years).[2] The students attended two academic-tracked school types from either 10th grade (Gymnasium, graduation after 12th grade; 15 classes out of 6 schools) or 11th grade (Gesamtschule, graduation after 13th grade; 11 classes out of 5 schools). Ethical approval was provided by the corresponding governing board. Students' and their parents' informed consent was obtained prior to testing. The participation rate was 87%; the parents of 11% of the students did not allow their children to participate; 2% of the students were absent due to reasons unrelated to the study (e.g., illness). Students were not graded or rewarded for their participation. Data collection was administered by trained experimenters within regular school lessons lasting 45 min each. During one lesson students worked on the intelligence test, during another lesson on the CPS items.

2.2. *Variables*

Complex problem solving was assessed with the entirely computer-based micro-world program MicroDYN (Greiff et al., 2012), consisting of nine independent items. Cover stories like training a handball team or providing medical aid were implemented for each item. At the beginning, the students were instructed by the program on how to explore and control the interface of the program and what they were expected to do in the sequential items. Each item consisted of two phases: In the knowledge acquisition phase of each item, the students explored an unknown system with a set of related input and output variables and drew their conclusions in a model (180 s.). The variables were labeled without semantic meaning or with fictitious meaning to minimize the influence of prior knowledge. For example, in the item *Game*, the three input variables were *blue chips*, *green chips*, and *red chips*, and the three output variables were *Royal*, *Grande*, and *Nobilis*; see Fig. 1). At the start of the knowledge application phase of each item, the correct model was presented to the students. Subsequently, they had to control the system by reaching given target values of the output variables within four steps (90 s.). In this study, we focused on the first five items without eigendynamic effects. Details about these items regarding their topic, input and output variables, and effects are presented in Table 1. These items had two or three input variables and one, two, or three output variables.

[2] Please note that earlier studies (Lotz et al., 2016; Lotz et al., 2017; Lotz et al., 2022; Weise et al., 2020; Weise et al., 2022), that investigated different research questions, have been conducted with parts of these data. Please note also the unfortunately transposed numbers regarding the numbers of schools and classes of the two school types in these former publications.

Table 1
Details about the implemented MicroDYN items.

Name	Topic	Input variables	Effects	Output variables
(1) Lemonade	Effects of two types of syrup on the sweetness of a drink	Green syrup		Sweetness
		Blue syrup		
(2) Drawing	Effects of two powders on the properties of a color	Topax		Brightness
		Floba		Fluidness
(3) Cat	Effects of two feed types on the behavior of a cat	Brekon		Movement
		Mikas		Purring
(4) Moped	Effects of three types of fuel on the driving performance of a moped	Carenol		Velocity
		Noresal		Exhaust gases
		Farunin		
(5) Game	Effects of three types of game chips on the game	Blue chips		Royal
		Green chips		Grande
		Red chips		Nobilis

Note. All effects had a path coefficient equal to each other.

Intelligence was assessed with the Berlin Intelligence Structure test – Form 4 (BIS; Jäger et al., 1997). Based on the results from earlier studies (Brunner & Süß, 2005; Valerius & Sparfeldt, 2014), we selected the following subtests, covering a wide range of the BIS' content-operation-combinations: Figural analogies, Crossing out letters, Charkow, City map, Number sequences, X greater, Estimation, Story, Fact-opinion, and Verbal analogies.

2.3. *CPS scoring*

2.3.1. *Effectiveness and efficiency of the exploration behavior*

Concerning these five structurally similar items with only direct effects, the strategic behavior VOTAT is the most effective and efficient way for exploring the effects. The log-file data of the knowledge acquisition phase was used to analyze for every single exploration step whether VOTAT was used or not. If VOTAT was shown at least once for each input variable, the corresponding strategic behavior was classified as “effective”. Based on the number of input variables, we determined the number of necessary exploration steps (i.e., one VOTAT step for each input variable) and the number of non-necessary exploration steps (i.e., the difference between the number of actually performed steps and the number of necessary steps) indicating the efficiency of the exploration behavior.

For each CPS item, we adopted the following coding scheme: If students did not apply VOTAT for every input variable of an item, the behavior was considered non-effective and coded with 8. If students showed VOTAT for every input variable, their exploration behavior was coded according to the efficiency, which is the number of non-necessary steps: For zero steps more than the number of necessary steps also referred to as maximum efficient behavior, the exploration behavior was coded with 0; for one step more than the number of necessary steps, the exploration behavior was coded with 1; for two steps more than the number of necessary steps, the exploration behavior was coded with 2; …; for six steps more than the number of necessary steps, the exploration behavior was coded with 6; and for seven or more steps more than the number of necessary steps, the exploration behavior was coded with 7. If students did not explore the respective CPS item, exploration behavior was coded as missing. By applying this coding scheme, we considered differences in the number of input variables between the five CPS items.

2.3.2. *CPS performance*

For both CPS phases, a dichotomous scoring was applied to obtain performance scores in accordance with prior studies (e.g., Greiff et al., 2012; Kröner et al., 2005; Wüstenberg et al., 2012). Concerning the knowledge acquisition phase, credit was given, if the drawn model was completely correct (coded as 1; otherwise 0). Concerning the knowledge application phase, credit was given if the target values of all variables were reached within four steps (coded as 1; otherwise 0). Based on the performance scores for each item, mean performance scores of the five items were calculated separately for both phases.

2.4. *Analysis*

Analyses were conducted with the statistical packages IBM SPSS Statistics 26 and *Mplus* 7.11 (Muthén & Muthén, 1998–2015). Concerning RQ1, the exploratory analysis of students' effective and efficient exploration behavior was conducted with SPSS. As described before, we coded the effectiveness and the efficiency of students' exploration behaviors separately for the five CPS items. Subsequently, we inspected the frequencies of students for each numerical code.

To address RQ2, we conducted an exploratory Latent Class Analysis (LCA) in *Mplus*. LCA is a latent variable modeling approach, in which a discrete latent variable (i.e., latent class variable) is assumed to clarify the relations between observed variables (Geiser, 2012). Therefore, the probability of a particular item response is assumed to depend on the probability of being a member of a particular class (i.e., class proportions) and the probability of giving a particular item response by members of this particular class (i.e., conditional response probabilities). Students who share a similar pattern of responses in a set of observed variables are grouped in the same latent class; each latent class is characterized through its proportion of students and a pattern of conditional response probabilities for the observed variables. Concerning RQ2, we estimated models with one to six latent classes using robust maximum likelihood estimation (MLR). The observed variables were the categorically ordered indicators of exploration efficiency (0 to 7: consistent VOTAT use, number of non-necessary exploration steps; 8: no consistent VOTAT application) in the five CPS items. For this analysis and all following analyses, only students who worked on all five items were included, which resulted in $n = 469$ students with no missing data. The optimal number of classes was determined on the basis of the following analysis strategy (Collins & Lanza, 2010; Geiser, 2012): To determine absolute fit, we inspected the Pearson χ^2 statistic, the Likelihood-Ratio χ^2 statistic, and χ^2/df. In large samples, a χ^2/df ratio $<$ 2 is usually considered an indicator of adequate absolute fit. To determine relative fit, we inspected the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the sample-size adjusted BIC (adjBIC). Lower values of the respective relative fit indices indicate a better model fit. Simulation studies showed that BIC and adjusted BIC are more informative indices than AIC when determining the number of classes (Nylund et al., 2007). Furthermore, we inspected the average latent class attribution probabilities of the LCA models with different numbers of latent classes. The average assignment probabilities indicate the certainty of students being assigned to classes and should exceed 0.80 for all classes (Rost, 2006). In accordance with prior studies (e.g., Greiff et al., 2018), we additionally inspected the parametric bootstrapped likelihood ratio test (BLRT), which compares a model with k classes and a model with k − 1 classes. The null hypothesis specifies that an additional class does not improve absolute model fit; therefore, a statistically significant *p*-value indicates that the model with k classes provides a better fit to the data than the model with k − 1 classes. Although the BLRT is somehow similar to the Lo-Mendell-Rubin (LMR) test that *Mplus* calculates by default, the BLRT is superior to the LMR test in determining the number of classes for LCA models (e.g., Nylund et al., 2007); accordingly, we chose this test.

Regarding RQ3 related to potential intelligence differences between the students of the different classes from RQ2, analyses were performed in *Mplus*. We modeled the intelligence *g*-factor in accordance with the common higher order structure: The manifest scores of the ten BIS intelligence subtests loaded onto the corresponding content-specific first-order factors of verbal (V_{BIS}), numerical (N_{BIS}), and figural (F_{BIS}) intelligence, which, in turn, loaded on the second-order *g*-factor (g_{BIS}). The nonstandardized loadings of the first indicator of a corresponding latent factor were fixed to 1. Beside the three content facets referring to verbal, numerical, and figural intelligence, we also considered the operation facets of the BIS-model (Jäger et al., 1997). Specifically, we estimated residual correlations among those manifest subtest indicators that belonged to the same operation-facet of the BIS-model. To account for students' clustering within classes, we used the "type = complex" option in *Mplus* which adjusts for potential effects of clustering on standard errors and χ^2-statistics. To compare the students of the latent classes regarding their intelligence, we inspected linear regression models using a maximum likelihood estimation with robust standard errors (referred to as MLR in *Mplus*). Class membership served as predictor, the latent *g*-factor as criterion. To compare the students of the different latent classes, we conducted pairwise Scheffé comparisons and inspected regression coefficients, standard errors, *p*-values, and R^2-values for the outcome variables. The significance level was set to $p < .05$; in order to control for the accumulation of alpha errors, we adjusted according to the Bonferroni-Holm procedure (Holm, 1979).

Regarding RQ4 related to potential differences in CPS performance between the students of the different classes from RQ2, we ran analyses in the same way as in RQ3 (i.e., linear regression analyses with MLR estimator). We modeled CPS performance scores of the knowledge acquisition phase and the knowledge application phase as latent factors. Specifically, the manifest scores of the five MicroDYN items of the respective CPS phase loaded on the corresponding factor and the nonstandardized loadings of the first indicator of a corresponding latent factor were fixed to 1.

3. Results

Means and standard deviations of the five CPS items, the performance scores for the knowledge acquisition and the knowledge application phase, and of the BIS subtests (supplemented by ICCs and design effects) are shown in Table 2.

3.1. Exploratory analysis of student's exploration behavior

The numbers and proportions of students who showed more or less effective and efficient exploration behavior addressed in RQ1 are depicted in Table 3. In item one, 39.8% of the students did not explore effectively (i.e., not consistently using VOTAT for each input variable).[3] In contrast, about 60% explored effectively. These effective explorers differed in efficiency, reaching from students who explored effectively, but in a very inefficient way (31.5%, coded as 7) to students who explored effectively and maximally efficiently (1%, coded as 0). Considering in contrast the last item (item 5), the proportion of ineffective exploration behavior was lower (29.7%, coded as 8) and the proportion of effective exploration behavior was higher. Concerning efficiency, the proportion of students showing a highly efficient exploration behavior was higher and the proportion of students showing less efficient exploration behavior was lower (Table 3). Regarding the interjacent items 2, 3, and 4, the overall pattern showed an increase in the effectiveness and the efficiency of the exploration behavior.

Table 2
Means (*M*), standard deviations (*SD*), intra-class correlations (*ICC*), and design effects for the five MicroDYN items, the sum scores of the two CPS phases, and for the 10 intelligence subtests.

Name (and abbreviation, if present) of MicroDYN items and BIS subtests	*M*	*SD*	*ICC*	Design effect
Knowledge acquisition				
Lemonade	0.74	0.44	.00	1.00
Drawing	0.65	0.48	.08	2.38
Cat	0.70	0.46	.05	1.89
Moped	0.72	0.45	.06	2.07
Game	0.70	0.46	.09	2.55
Sum score	0.70	0.33	.08	2.33
Knowledge application				
Lemonade	0.89	0.31	.00	1.00
Drawing	0.70	0.46	.10	2.72
Cat	0.75	0.43	.12	2.96
Moped	0.33	0.47	.12	2.96
Game	0.31	0.46	.07	2.14
Sum score	0.59	0.29	.11	2.93
Intelligence subtests				
Charkow (CH)	2.32	1.54	.07	2.21
City map (OG)	14.81	4.39	.04	1.68
Crossing out letters (BD)	56.60	12.95	.10	2.69
Estimation (SC)	3.56	1.76	.09	2.50
Fact-opinion (TM)	9.93	2.55	.06	2.07
Figural analogies (AN)	3.36	1.58	.06	1.94
Number sequences (ZN)	4.23	2.27	.13	3.20
Story (ST)	8.77	3.43	.06	2.04
Verbal analogies (WA)	2.37	1.52	.08	2.35
X greater (XG)	19.47	8.53	.15	3.59

Note. $N = 469$.

3.2. Latent class analysis

Regarding RQ2 related to the number of latent classes, the indicators of model fit for models with one to six classes are presented in Table 4. Concerning the optimal number of classes, relative fit indices indicated that either the three- or the four-class solution were most appropriate. The three-class solution was supported by the BIC, whereas the four-class solution was supported by the AIC. Notably, the adjusted BIC was comparably low for both solutions. The BLRT showed a significantly better fit for the four-class solution compared with the three-class solution ($\Delta\chi^2 = 121.60$; $df = 41$; $p < .001$). A closer inspection of students' latent class profiles revealed that both solutions showed similar profile patterns for two classes of the three-class solution; however, the remaining third group of the three-class solution which was comprised of the students who explored most efficiently was split up in the four-class solution meaningfully into two groups. Taking the relative fit indices, the results of the BLRT, and the interpretation of the different solutions into account, the four-class solution appeared to be most appropriate. Therefore, we selected the four-class solution as the basis for all further analyses.

The profiles of the four-class solution are visualized in Fig. 3 displaying the conditional probabilities of students' exploration behavior in the five CPS items. As expected, the four classes differed in regards to the effectiveness and efficiency of their exploration behavior. Students of the first class (29.4%) showed for items 1/2/3/4/5 a high probability of not applying VOTAT consistently (coded as 8) of 0.76/0.93/0.92/0.96/0.89. Correspondingly, the probabilities for more or less efficient VOTAT behavior (coded as 0 to 7) were very low and mostly between

[3] The group coded as "8" comprised students not consistently using VOTAT for each input variable. For items 1 to 3 with two input variables, about 55% to 60% of these students applied VOTAT for one input variable, the remaining students did not use VOTAT at all. For items 4 and 5 with three input variables, 55% to 60% of these students did not use VOTAT at all, about 35% applied VOTAT for one input variable, and the remaining students (10% to 15%) applied VOTAT for two input variables. To inspect whether inconsistent VOTAT application differed from the complete absence of VOTAT use, we ran post-hoc MANOVAS with VOTAT use as independent variable and CPS performance as dependent variable. Although the findings were not totally consistent across the five items and the two CPS phases regarding statistical significance, statistically significant results indicated that students with more consistent VOTAT use achieved higher CPS scores, also confirming the overall importance of (consistent) VOTAT use for successful problem solving.

Table 3
Distribution of the exploration behaviors for the five MicroDYN items.

Item	Consistent VOTAT use, efficiency coded according to the number of non-necessary exploration steps while applying VOTAT for each input variable								Non-consistent VOTAT use, coded as	Missing
	0	1	2	3	4	5	6	7+	8	
Item 1	5 (1.0%)	6 (1.2%)	18 (3.6%)	12 (2.4%)	29 (5.9%)	23 (4.6%)	23 (4.6%)	156 (31.5%)	197 (39.8%)	4 (0.8%)
Item 2	25 (5.0%)	23 (4.6%)	32 (6.5%)	34 (6.9%)	34 (6.9%)	21 (4.2%)	23 (4.6%)	84 (17.0%)	193 (39.0%)	3 (0.6%)
Item 3	12 (2.4%)	22 (4.4%)	46 (9.3%)	39 (7.9%)	33 (6.7%)	20 (4.0%)	25 (5.0%)	106 (21.4%)	166 (40.2%)	7 (1.4%)
Item 4	72 (14.5%)	42 (8.5%)	21 (4.2%)	39 (7.9%)	31 (6.3%)	21 (4.2%)	18 (3.6%)	49 (9.9%)	176 (35.6%)	12 (2.4%)
Item 5	110 (22.2%)	37 (7.5%)	33 (6.7%)	43 (8.7%)	22 (4.4%)	21 (4.2%)	13 (2.6%)	43 (8.7%)	147 (29.7%)	15 (3.0%)

Note. $N = 495$.

Table 4
Absolute fit indices, relative fit indices, and attribution probabilities of the LCA for solutions with one to six classes.

Number of classes	1	2	3	4	5	6
Absolute fit						
Likelihood ratio χ^2 (df)	1677.64 (58,856)	1493.95 (58,854)	1392.26 (58,843)	1279.88 (58,801)	1379.0 (58,779)	1420.45 (58,747)
Pearson χ^2 (df)	8072.30 (58,856)	7962.59 (58,854)	6447.79 (58,843)	5721.53 (58,801)	6262.16 (58,779)	6295.91 (58,747)
χ^2/df < 2 for both χ^2	Yes	Yes	Yes	Yes	Yes	Yes
Boundary estimates	0	24	48	64	80	122
Relative Fit						
AIC	8532.25	7834.25	7499.20	7459.68	7426.14	7424.53
BIC	8698.27	8170.45	8005.57	8136.22	8272.86	8441.43
aBIC	8571.32	7913.38	7618.37	7618.90	7625.41	7663.85
Attribution probabilities						
Mean *p*	n.a.	.965	.954	.947	.941	.937
$p < 0.80$	n.a.	0	0	0	0	0
Bootstrapped likelihood ratio test						
2* LD (difference in number of parameters)	n.a.	779.99 (33)	417.06 (41)	121.60 (41)	115.54 (41)	83.604 (41)
p	n.a.	<.001	<.001	<.001	<.001	.008
Entropy	n.a.	0.880	0.893	0.910	0.907	0.919

0.00 and 0.05. Therefore, the first class was labeled *ineffective explorers* (1). Concerning the second class (29.7%), the main characteristic of students' behavior was a consistent VOTAT, but inefficient exploration behavior (i.e., seven or more additional steps; coded as 7). Importantly, the probability of the maximum efficient exploration behavior (no additional step; coded as 0) remained close to zero. Hence, we labeled the second class as *inefficient explorers* (2). The students of the third and fourth class showed high probabilities of the most efficient VOTAT-related exploration behavior (i.e., zero additional steps) at the end of the item set. In the first two items, the students of class three explored ineffectively, whereas students of class four explored effectively. Specifically, students of the third class (11.5%) showed a rather high probability not to use VOTAT consistently (i.e., coded as 8) for the first and second item (0.78 and 0.76), that decreased for the third item (0.37) and was very low for the last two items (items 4/5: 0.05/0.00). Simultaneously, the probability for the maximum efficient exploration behavior increased over the set of items (items 1/2/3/4/5: 0.00/0.03/0.05/0.23/0.54). In turn, the third class was labeled *emerging explorers* (3). The students of the fourth class (31.1%) explored in an effective way, whereby the less efficient behavior (coded 1 to 7) decreased and the most efficient behavior (coded 0) increased over the set of items (items 1/2/3/4/5: 0.04/0.16/0.07/0.40/0.50). Accordingly, we labeled the fourth class as *proficient explorers* (4). Descriptively, the efficiency increased substantially from item three to item four in students of these two classes.

In summary, the profiles of the four latent classes (ineffective, inefficient, emerging, and proficient explorers) exhibited distinct patterns of exploration behaviors. The main characteristics differentiating the four groups of students were (i) whether students applied VOTAT consistently (distinguishing class one from two and respectively three from four) and (ii) whether students' exploration efficiency increased over the course of CPS items (distinguishing class two from classes three and four).

3.3. Class membership and intelligence

Concerning RQ3, we expected that students in classes with an effective and more efficient exploration behavior would achieve higher intelligence scores than participants in classes with an ineffective or an effective, but less efficient exploration behavior. In accordance with our expectations, the results of the Scheffé tests revealed that students of classes with a mostly effective exploration behavior (classes 2, 3, and 4) achieved significantly higher intelligence scores than students of the class with an ineffective exploration behavior (class 1; all $p < .001$, see Table 5). Contrary to our expectations the inefficient explorers reached significantly higher intelligence scores than the emerging explorers (class 3; $p < .001$) and the proficient explorers (class 4; $p < .004$). Furthermore, the proficient explorers (4) did not differ significantly regarding the intelligence score from the emerging explorers (class 3; $p = .254$).

3.4. Class membership and CPS performance

Concerning RQ4, we expected that students in classes with an effective and more efficient exploration behavior achieved higher CPS performance scores than students in classes with an ineffective or an

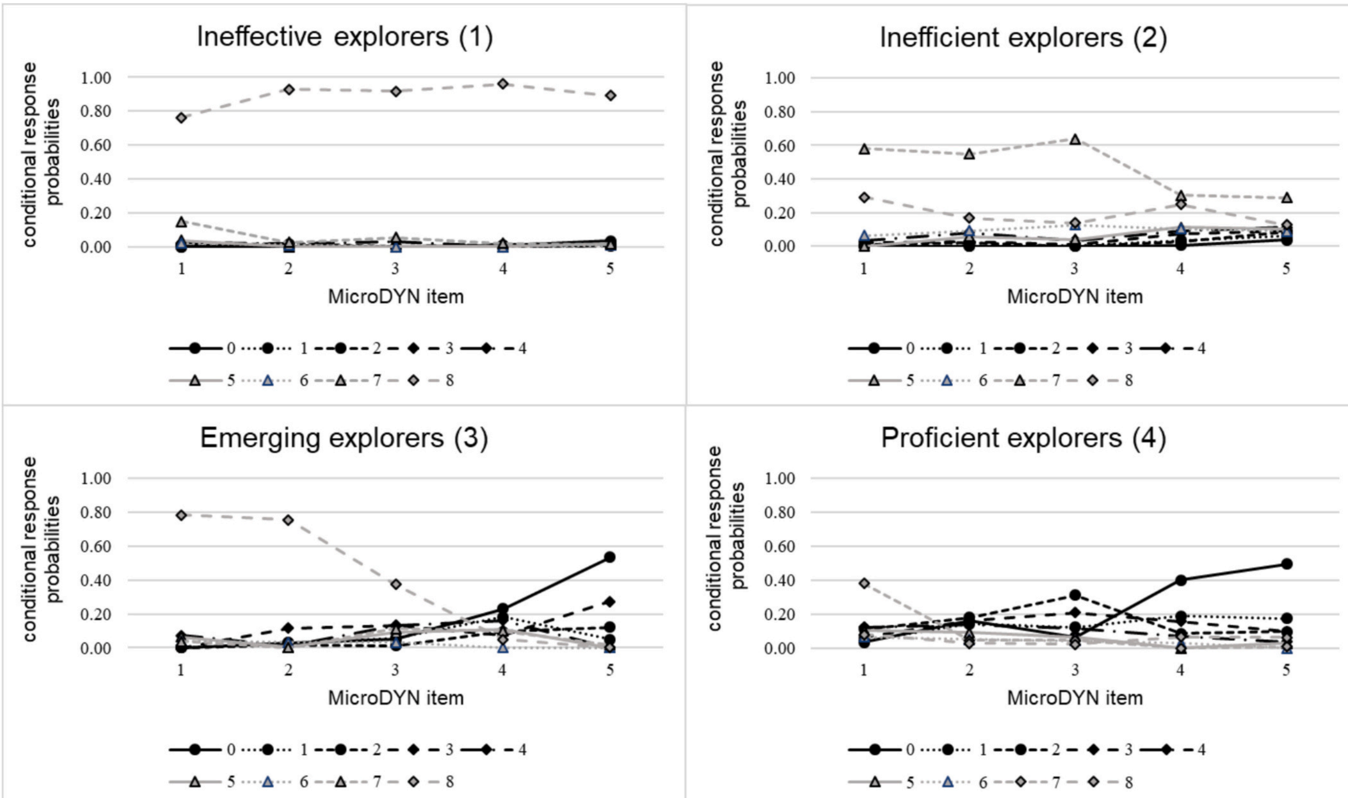


Fig. 3. Profiles of the four classes of the LCA.
Note. Displayed are the conditional probabilities for the nine ordered categories (from "0" to "6", "7", and "8") across the five CPS items.

Table 5
Results of the linear regression models.

Comparisons	Knowledge acquisition (mean score item 1–5)			Knowledge application (mean score item 1–5)			Intelligence		
	Est.	*SE*	*p*	*Est.*	*SE*	*p*	*Est.*	*SE*	*p*
1 vs. 2	0.78	0.03	<.001*	0.50	0.05	<.001*	0.44	0.05	<.001*
1 vs. 3	0.67	0.04	<.001*	0.41	0.05	<.001*	0.28	0.06	<.001*
1 vs. 4	0.73	0.05	<.001*	0.49	0.05	<.001*	0.27.	0.06	<.001*
2 vs. 3	−0.29	0.06	<.001*	−0.24	0.05	<.001*	−0.27	0.05	<.001*
2 vs. 4	0.01	0.08	.91	−0.02	0.05	.67	−0.16	0.06	.004*
3 vs. 4	0.29	0.07	<.001*	0.20	0.06	<.001*	0.08	0.07	.254

Note. $N = 469$.
* Statistically significant according to the Bonferroni-Holm procedure ($p < .05$).

effective, but less efficient exploration behavior. The interpretation of the results of the Scheffé tests were the same for both phases (Table 5). In accordance with our expectations, students of classes with a mostly effective exploration behavior (classes 2, 3, and 4) performed statistically significantly better than students of the class with an ineffective exploration behavior (class 1; all $p < .001$). Only partially in line with our expectations, inefficient explorers (2) reached significantly higher CPS performance scores than emerging explorers (class 3; $p < .001$). Furthermore, inefficient explorers did not differ significantly regarding their CPS performance from proficient explorers (class 4; $p = .91$). Consistent with our hypothesis, proficient explorers (4) achieved significantly higher CPS score in both CPS phases than emerging explorers (class 3; $p < .001$).

4. Discussion

Based on its importance for successful complex problem solving, we aimed to further clarify the role of beneficial strategy use while exploring CPS scenarios. Specifically, we investigated effective and efficient VOTAT applications across a sequence of five items. Concerning our main results, the exploration behavior became more effective and more efficient across the item sequence. Furthermore, we identified four meaningfully distinct classes of students: (1) ineffective, (2) inefficient, (3) emerging, and (4) proficient explorers. Ineffective explorers showed lower intelligence test and CPS performance scores than students from the other classes. Inefficient explorers achieved higher intelligence test scores than students from the two remaining classes (3, 4), who did not differ among each other. Regarding CPS performance, emerging explorers reached lower scores than inefficient explorers and proficient explorers, who did not differ.

4.1. *Descriptive analysis of the exploration behavior*

Regarding the exploration behavior in single items addressed in our first research question, important differences appeared: in item one, about 40% of the students did not explore effectively; this proportion decreased to about 30% in item five. In contrast, the proportion of students exploring maximally efficient increased from 1% in item one to

22% in item five. Thus, the exploration behavior changed across the item sequence as hypothesized. This increase of a consistent VOTAT use corresponds well with previous results (e.g., Vollmeyer et al., 1996). Going beyond prior studies, we did not only inspect the effectiveness of the exploration behavior and the number of exploration steps, but also its efficiency. Notably, previous studies demonstrated the importance of the quality of exploration steps, whereas findings regarding the number of exploration steps remained inconclusive. Correspondingly, we introduced the idea of the efficiency of the exploration behavior to gain deeper insights of CPS. These descriptive results already revealed that the efficiency of the exploration behavior changed across items, indicating increasingly efficient explorations. This trend serves as important starting point for our subsequent research questions. In RQ2, we had hypothesized one class consisting of students with an increasingly efficient exploration behavior. To conclude, this result pattern indicates that some sort of learning had occurred across the item sequence, with increasingly effective and efficient exploration behavior.

4.2. Latent class analysis of exploration behaviors

Building on these descriptive results, four classes of students were identified in a latent class analysis differing in effective and efficient strategy use. Comparing our findings with earlier person-centered results (e.g., Csányi & Molnár, 2025; Molnár et al., 2022; Molnár & Csapó, 2018; Mustafić et al., 2019; Wu & Molnár, 2022), remarkable similarities appeared. As in the previous studies, we identified a class of students not (or only rarely) exploring effectively (i.e., ineffective explorers). Likewise, we found a class of students not exploring effectively in the first, but in the last items (i.e., emerging explorers). Furthermore, students showing consistently an effective exploration behavior constituted one class in earlier studies (e.g., Csányi & Molnár, 2025; Molnár & Csapó, 2018; Mustafić et al., 2019) and two classes in our study (inefficient explorers, proficient explorers). Presumably, we found two classes, because we did not investigate only the effectiveness, but also the efficiency of the exploration behavior. Whereas the students of both classes explored effectively, proficient explorers became more efficient across the item sequence and inefficient explorers remained rather inefficient.

Referring to the increasingly effective and efficient exploration behavior described in RQ1, the results of the LCA revealed a more fine-grained pattern. While the ineffective explorers and the inefficient explorers showed a widely consistent exploration behavior across the item sequence, particularly the emerging explorers and the proficient explorers contributed to the descriptive changes in exploration behavior identified in RQ1. Concerning the increased effectiveness, mainly the emerging explorers showed corresponding changes as reflected by the numbers of students. Specifically, the number of students exploring ineffectively decreased by 50 students from item 1 to item 5, while the class of emerging explorers comprises 54 students. Regarding the efficiency increase, both the emerging and the proficient explorers shifted towards a more efficient exploration behavior.

In contrast to earlier studies, we did not find a class of students exploring effectively only in the first, but not in the last items. In previous studies, it was suggested that these participants failed to transfer their successful exploration behavior from simpler to more complex and later placed items (e.g., Molnár & Csapó, 2018). We might not have identified such a class because item difficulties increased only moderately. Notably, we focused exclusively on items without eigendynamic effects, which had often been used in later items with higher item difficulty in former studies (e.g., Stadler et al., 2016). Therefore, we suggest to examine the efficiency of exploration behaviors in tasks involving eigendynamic effects in future studies to reveal further insights into the exploration behavior.

4.3. Class membership and intelligence

As expected, students of the class consistently exploring ineffectively (class 1) reached lower mean intelligence test scores than students of the remaining classes, who mostly explored effectively. This finding corresponds with previous studies that showed positive correlations between VOTAT application and intelligence (e.g., Kröner et al., 2005; Lotz et al., 2017; Wüstenberg et al., 2014; see also for the relations with inductive-reasoning: Wu & Molnár, 2022). Furthermore, VOTAT strategy application mediated the intelligence/CPS performance relation (e.g., Lotz et al., 2022). Comparing the intelligence scores of our three remaining classes, inefficient explorers (2) reached higher intelligence test scores than emerging explorers (3) and proficient explorers (4), who did not differ among each other. This partially unexpected result pattern should be clarified in further research. Possibly, inefficient explorers (2) tended to check their hypotheses repeatedly (characterizing less efficient exploration behavior) to become more confident about their conclusions. Furthermore, the more intelligent students (class 2) might have repeated certain exploration steps in order to inspect whether the path coefficients differed. Less intelligent students (classes 3, 4) may not have considered possibly differing path coefficients. Although not explicitly mentioned in the instructions, these paths coefficients were always equal to each other in our study as also in many prior studies (e.g., Greiff et al., 2013; Krieger et al., 2021). Moreover, more intelligent students might have employed additional strategies beyond VOTAT to ensure that no other than direct effects occurred. Additionally, repeated actions might indicate knowledge consolidation processes and the deeper integration of the information, making this information possibly more available and applicable in the knowledge application phase (see e.g., Wirth, 2004). In reference to cognitive and meta-cognitive learning strategies (Weinstein & Mayer, 1986), one can characterize VOTAT as cognitive exploration strategy and the systematic, effective, and efficient application and orchestration of corresponding exploration strategies as meta-cognitive. Future studies further elaborating meta-cognitive or motivational explanations for our partially unexpected results and integrating corresponding assessments in future studies possibly even with, for example, experimentally manipulated corresponding instructions might further clarify these aspects.

4.4. Class membership and CPS performance

Regarding CPS performance, students of the class consistently exploring ineffectively (class 1) achieved lower mean scores in both CPS phases than students of the remaining classes (classes 2, 3, 4), who applied VOTAT more consistently and explored more effectively. This result is in line with our expectations and previous results showing positive correlations between VOTAT strategy application and CPS performance (e.g., Greiff et al., 2016; Kröner et al., 2005; Molnár & Csapó, 2018; Pejić & Molcer, 2021; Vollmeyer et al., 1996; Wüstenberg et al., 2014). Comparing our three remaining classes, emerging explorers (3) reached lower CPS scores for both phases than inefficient explorers (2) and proficient explorers (4). This result pattern further underlines the importance of an effective exploration behavior (i.e., using VOTAT at least once for each input variable). In contrast to inefficient explorers (2) and proficient explorers (4), emerging explorers (3) did not explore effectively in the first two items (and partially in the third item). Thus, the mean effectiveness across the item set was lower in emerging explorers (3) than in the other two classes, resulting in lower CPS performance scores. However, and in contrast to our expectations, the inefficient explorers (2) and the proficient explorers (4) did not differ concerning CPS performance scores. The reasoning underlying our expectations was that using items with rather low complexity in our study might result in the consequence that more efficient exploration behavior is beneficial for CPS performance (following Molnár & Greiff, 2023). However, students of more efficient exploring classes did not achieve superior CPS scores. Comparably unequivocal were the previously reported relations between the number of exploration steps and CPS performance as mentioned in the introduction, where higher CPS performance scores sometimes went hand in hand with higher (e.g., Csányi & Molnár, 2025; Eichmann, Goldhammer, et al., 2020) and sometimes with lower (e.g., Greiff et al., 2016; He & von Davier, 2015; Stadler, Fischer, & Greiff, 2019) numbers of exploration steps. Notably, repeated actions in the sense of double-checking behaviors (Eichmann, Greiff, et al., 2020) might be useful. Those authors focused on goal-directed exploration behaviors (i.e., every action that is necessary to solve the task successfully) and not specifically on the number of necessary and unnecessary VOTAT actions as analyzed in our study. Therefore, the theoretical status of the number of exploration steps and the efficiency of the exploration behavior for successful CPS performance as well as their relations call for further research.

Interestingly, inefficient explorers (2) and proficient explorers (4) differed in the efficiency of the exploration behavior, but not in their CPS performance in both phases. Furthermore, the students of both classes applied VOTAT consistently. Thereby, they explored effectively, and the complete information about the underlying relations (i.e., direct effects) was provided by the system. Possibly, the students of both classes differed in their repeated actions (in the sense of less efficient explorations) because they differed also in the amount of information required to come to confident judgements about direct effects. Correspondingly, the interindividual difference variable confidence judgement (e.g., Rudolph et al., 2017) might explain these class differences in the efficiency of the exploration behavior. Specifically, students with lower confidence judgements might require more evidence and, correspondingly, double-checking to come to conclusions about the effects. To further elucidate these ideas in future studies, one should assess confidence judgements and characteristics of the problem solving style (see e.g., Gao et al., 2022).

Interestingly, our findings offer an empirical basis for adaptive and individualized approaches to foster students' explorations: To ineffective explorers, one should teach the VOTAT strategy and its application to enhance more effective exploration behavior and, ultimately, more successful complex problem solving. For inefficient or less efficient explorers, however, instructions might be useful to highlight that applying unnecessary exploration steps does not provide any further information, which is particularly important for situations with limited exploration time. This should be examined in future studies.

4.5. Limitations, strengths, and outlook

As every study, our study has some limitations. First and concerning the operationalization, we defined effective exploration behavior as using VOTAT at least once for each input variable, which is characterizing a very structured, systematic, and expedient exploration behavior, providing all information about the underlying relations between the input and output variables of the item. Nevertheless, not using VOTAT consistently for each input variable, but using other strategic behaviors might sometimes be intertwined with successful problem solving as well (e.g., Rollett, 2008). However, numerous authors pointed to the importance of VOTAT as indicated by, for example, substantial relations between VOTAT use and intelligence (e.g., Lotz et al., 2017; Wüstenberg et al., 2014) as well as VOTAT use and CPS performance (e.g., Greiff et al., 2016; Kröner et al., 2005; Lotz et al., 2022; Molnár & Csapó, 2018; Nicolay et al., 2023; Pejić & Molcer, 2021; Vollmeyer et al., 1996; Wüstenberg et al., 2014). Furthermore, and regarding the VOTAT application across the item sequence, we widened the scope to effectiveness and efficiency, thereby replicating earlier VOTAT related results to a large extent (e.g., Molnár & Csapó, 2018). Second, we did not further differentiate the group of students who did not use VOTAT consistently in the analyses; this group (coded as "8" in Table 3) included students who did not use VOTAT at all and students who used VOTAT only partially (i.e., for some, but not all input variables). As noted in the footnote in the results section, we examined this group in more detail in a post-hoc analysis pointing to the relevance of partial VOTAT use for CPS performance scores (although the pattern of results was not completely consistent regarding statistical significance). Correspondingly, it might be fruitful in future research to subdivide this group into certain subgroups. Third, we did not further inspect the unnecessary exploration steps beyond their number. More in-depth analyses could investigate the patterns of repeated actions (e.g., distribution of double-checking behavior across input variables) and their relations to, for example, reaching the target values of the related output variables. Thereby, the processes of knowledge consolidation and integration should be further clarified. Fourth, we identified four classes based on students' exploration behavior across a sequence of CPS items and showed significant mean differences between the classes in their intelligence test scores and CPS achievement scores. Although the authors of some earlier studies have considered intelligence as a predictor of the exploration behavior, and exploration behavior as a predictor of CPS outcomes (Lotz et al., 2022; Nicolay et al., 2023), we focused on the class-based classification of students' exploration behavior and the corresponding mean differences among the classes in a cross-sectional design. Fifth, concerning the generalizability of our results, using more items or items from other CPS scenarios (e.g., *MultiFlux*; Kröner et al., 2005) in other samples (e.g., age group, country) represents an important endeavor for future research. Sixth, an open question is related to the relevance of our results concerning transfer to other facets of problem solving. The MicroDYN items used in our study are in line with the definition of CPS mentioned in the introduction. However, it may be worthwhile to additionally investigate the relevance of our results regarding the effectiveness and efficiency of beneficial strategic behaviors for other forms of problem solving (e.g., creative problem solving, collaborative problem solving) and other assessment tools (e.g., stealth assessment; see e.g., Shute et al., 2016). Considering various forms of problem solving and assessment tools simultaneously in one study might allow to come to a more comprehensive evaluation of problem solving skills. Furthermore, such results may provide a solid basis for adaptive training programs. Seventh, the order of the CPS items in our study was fixed. Therefore, one cannot distinguish effects resulting from the structure of the specific item, the position of the specific item in the item sequence, nor item sequence effects regarding the exploration behavior. Possibly, students might have learned how to explore effectively and efficiently in specific preceding items. Correspondingly, permuting the items might provide useful insights.

Despite these limitations, our study has some important strengths. First, we combined log-file based exploration behaviors (i.e., process data) across a sequence of CPS items with CPS achievement scores, thereby further elucidating the relations between process and product data. Second, our person-centered approach identifying four classes of students with qualitatively differing exploration behaviors across the item sequence provided first insights into the corresponding short time developments of the exploration behavior. Third, we investigated VOTAT as a well-researched and domain general exploration strategy in CPS and widened the research focus to its effectiveness and efficiency. Concerning effectiveness, we replicated former results pointing to the extraordinary relevance of VOTAT for successful problem solving. Furthermore, and extending previous studies on the exploration behavior, we combined both, qualitative (i.e., VOTAT, effectiveness) and quantitative (i.e., number of unnecessary VOTAT steps, efficiency) aspects of the exploration behavior. In contrast to former studies relying on effective VOTAT use as basis for class formation, we combined effective and efficient VOTAT applications, providing novel insights into beneficial exploration behavior and successful problem solving as well as the changes of the exploration behavior across the item sequence in different classes. Importantly, these results might serve as foundation for the aforementioned adaptive and individualized approaches to foster beneficial explorations and successful problem solving. Further investigations of the efficiency should further clarify the role of the number of repeated exploration steps. Fourth, to the best of our knowledge our study is the first to examine the relationship between intelligence and the efficiency of the exploration behavior. Fifth and in contrast to many earlier studies which assessed intelligence with only one subtest or very few subtests, we operationalized intelligence broader with ten differing subtests in the sense of a "good *g*" (Jensen & Weng, 1994).

Concerning educational practice, fostering effective exploration behavior (i.e., consistent VOTAT use) should be regarded as a central instructional goal when students are confronted with new systems, tasks, and problems. Regarding scientific phenomena, the instruction of the control of variables strategy is established, for example through carefully designed experiments in physics education (Schwichow et al., 2015; see also Schwichow et al., 2016). When designing a specific curriculum, teachers or instructors should consider the more general and content overarching strategy development besides teaching certain topics. Accordingly, the educational standards for science education in, for example, Germany define students' competencies for knowledge acquisition through experimentation (including the control of variables strategy) as a central goal of physics instruction (KMK, 2005). The importance of an efficient exploration behavior has not yet been conclusively established despite its relevance in designing experiments when, for instance, dealing with expensive consumables. Regarding the relationship between class membership, students' intelligence, effective and efficient exploration behavior, and successful problem solving as well as their relations with successful learner-tailored instruction, further research is needed concerning corresponding aptitude-treatment interactions (Cronbach & Snow, 1977), especially regarding the question whether higher intelligence interacts as enhancer or compensator with educational instructions (see also Kühl et al., 2022).

Taking the design and the results of our study into account, several directions for future research are recommended. First, participants should be explicitly instructed to explore efficiently (at least in one condition) to further clarify the relevance of the efficiency of the exploration behavior. Second, providing more details about possibly occurring effects in the MicroDYN introduction (such as potential effect with non-linear relations that are easily identified through double-checking) and prior to item engagement may promote the application of corresponding strategies. Third, an experimental design with intervention groups, in which targeted exploration strategies are instructed (e.g., effective exploration vs. effective and efficient exploration) should be contrasted with a control group characterized by not specifically influenced exploration behavior. Fourth, including personality traits (e.g., conscientiousness) and problem solving styles (e.g., acting, reflecting, shirking; Gao et al., 2022) should provide further insights about their interplay with exploration effectiveness and efficiency, possibly suggesting even more details about tailored support—within and, hopefully even, beyond CPS contexts. Fifth, an experimental design with different orders of the items could help to detect effects of the item structure or item order. Sixth, future studies might benefit from subdividing the group we coded as "8" in our study and examining how students who apply VOTAT only partially (i.e., not for all input variables) fit into the existing findings. In this context, it might also be worthwhile to consider an alternative operationalization of effectiveness—one that recognizes not only VOTAT but also other theoretically effective combinations of steps as effective. This approach could further be used to compare different operationalizations in terms of their impact on the results.

4.6. Conclusions

By analyzing the log-file data of the exploration behavior, we identified four distinct classes of students differing in their VOTAT behavior across the sequence of five items. Specifically, the students of these classes differed in whether they used VOTAT, how consistent and efficient they used VOTAT, and how their exploration behavior changed across the item sequence. The person-centered approach is well suited to examine the development of the exploration behavior over an item sequence and to detect and classify patterns of strategic behaviors. The four classes of students differed in their mean intelligence test scores as well as their CPS performance scores in both phases. The fundamental distinction of students applying VOTAT (mostly) consistently or not, and corresponding intelligence test and CPS performance differences underline the exceptional relevance of effective exploration behavior. The efficiency of the VOTAT behavior across the item sequence was, however, less consistently related to the intelligence test scores and CPS performance scores. In conclusion, the insights gained in our study widened our understanding of the exploration behavior forming a basis for adaptive interventions aimed at fostering beneficial strategies to ultimately become better problem solvers and successful students in an increasingly complex world.

CRediT authorship contribution statement

Julia Ruby: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Julius J. Weise:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Samuel Greiff:** Writing – review & editing, Project administration, Investigation. **Jörn R. Sparfeldt:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Conceptualization.

Declaration of competing interest

Samuel Greiff is one of two authors of the commercially available COMPRO-test that is based on the multiple complex systems approach and that employs the same assessment principle as MicroDYN. For any research and educational purpose, a free version of MicroDYN is available.

Acknowledgements

We are particularly grateful to Christin Lotz for her valuable conceptual and empirical work regarding effective strategic behavior while solving complex problems and coordinating the data collection. We thank Anna Auth, Elena Groh, and Kerstin Mayer, who participated in an excellent manner on the data collection. Furthermore, we are also grateful to the TBA group at DIPF (http://tba.dipf.de) for providing the authoring tool CBA Item Builder and technical support. The contribution of the first author was supported within the German project SaLUt. The project (SaLUt, 01JA1906A) is part of the "Qualitätsoffensive Lehrerbildung", a joint initiative of the Federal Government and the Länder, which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research. The authors are responsible for the content of this publication. The authors developed the topic and content of this manuscript independent of this funding.

References

Brunner, M., & Süß, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement, 65*, 227–240. https://doi.org/10.1177/0013164404268669

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

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

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

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

Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (1983). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität [Lohhausen: About dealing with Uncertainty and Complexity]*. Huber.

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology, 112*(8), 1546–1562. https://doi.org/10.1037/edu0000446

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

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. 3–25). Erlbaum.

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

Funke, J., Töpfer, S., & Wagener, S. (1998). *Finite Automaten als Instrument zur Erfassung von Problemlösefähigkeiten bei Schülern [Finite automata for assessing problem solving competence of students]*. Psychologisches Institut der Universität Heidelberg.

Gao, Y., Zhai, X., Bulut, O., Cui, Y., & Sun, X. (2022). Examining humans' problem-solving styles in technology-rich environments using log file data. *Journal of Intelligence, 10*, 38. https://doi.org/10.3390/jintelligence10030038

Geiser, C. (2012). *Datenanalyse mit Mplus [Data analysis with Mplus]*. Guilford Press. https://doi.org/10.1007/978-3-531-93192-0

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*, 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8

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 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., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence, 53*, 92–101. https://doi.org/10.1016/j.intell.2015.02.007

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. https://doi.org/10.1016/j.compedu.2015.10.018

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

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*(2), 364–379. https://doi.org/10.1037/a0031856

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.), *Vol. 140. Springer proceedings in mathematics & statistics. Quantitative psychology research* (pp. 173–190). Springer. https://doi.org/10.1007/978-3-319-19977-1_13.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. BIS-Test, Form 4 [Berlin Intelligence-Structure Test. Version 4]*. Hogrefe.

Jensen, A. R., & Weng, L.-J. (1994). What is a good *g*? *Intelligence, 18*(3), 231–258. https://doi.org/10.1016/0160-2896(94)90029-9

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*, 63–85. https://doi.org/10.1007/BF02300500

KMK. (2005). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss: Beschluss vom 16.12.2004* [Educational Standards in Physics for the Intermediate School-Leaving Certificate]. Wolters Kluwer Deutschland.

Kretzschmar, A., Neubert, J. C., Wüstenberg, S., & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence, 54*, 55–69. https://doi.org/10.1016/j.intell.2015.11.004

Krieger, F., Stadler, M., Bühner, M., Fischer, F., & Greiff, S. (2021). Assessing Complex Problem-Solving Skills in Under 20 Minutes. *Psychological Test Adaptation and Development, 2*, 80–92. https://doi.org/10.1027/2698-1866/a000009

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*(4), 347–368. https://doi.org/10.1016/j.intell.2005.03.002

Kühl, T., Fehringer, B. C. O., & Münzer, S. (2022). Unifying the ability-as-compensator and ability-as-enhancer hypotheses. *Educational Psychology Review, 34*(2), 1063–1095. https://doi.org/10.1007/s10648-021-09650-5

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action — Effective strategic behaviors while solving complex problems. *Intelligence, 64*, 98–112. https://doi.org/10.1016/j.intell.2017.08.002

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2022). g's little helpers–VOTAT and NOTAT mediate the relation between intelligence and complex problem solving. *Intelligence, 95*, 1–13. https://doi.org/10.1016/j.intell.2022.101685

Lotz, C., Sparfeldt, J. R., & Greiff, S. (2016). Complex problem solving in educational contexts — Still something beyond a "good *g*"? *Intelligence, 59*, 127–138. https://doi.org/10.1016/j.intell.2016.09.001

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander, & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 287–303). Erlbaum.

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

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Logfile analyses. *Frontiers in Psychology, 9*, Article 302. https://doi.org/10.3389/fpsyg.2018.00302

Molnár, G., & Greiff, S. (2023). Understanding transitions in complex problem-solving: Why we succeed and where we fail. *Thinking Skills and Creativity, 50*, Article 101408. https://doi.org/10.1016/j.tsc.2023.101408

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

Mustafić, M., Yu, J., Stadler, M., Vainikainen, M. P., Bornstein, M. H., Putnick, D. L., & Greiff, S. (2019). Complex problem solving: Profiles and developmental paths revealed via latent transition analysis. *Developmental Psychology, 55*(10), 2090–2101. https://doi.org/10.1037/dev0000764

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Naumann, J., Goldhammer, F., Rölke, H., & Stelter, A. (2014). Erfolgreiches Problemlösen in technologiebasierten Umgebungen: Wechselwirkungen zwischen Interaktionsschritten und Aufgabenanforderungen [Successful problem solving in technology-based environments: Interactions between interaction steps and item requirements]. *Zeitschrift für Pädagogische Psychologie, 28*, 193–203. https://doi.org/10.1024/1010-0652/a000134

Nicolay, B., Krieger, F., Kuhn, J.-T., Graesser, A. C., Ifenthaler, D., Baker, R., & Greiff, S. (2023). Unsuccessful and successful complex problem solvers — A log file analysis of complex problem solving strategies across multiple items. *Intelligence, 101*, 1–14. https://doi.org/10.1016/j.intell.2023.101793

Nylund, K., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535–569.

OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling reallife problems* (Vol. V). OECD Publishing.

Pejić, A., & Molcer, P. S. (2021). Predictive machine learning approach for complex problem solving process data mining. *Acta Polytechnica Hungarica, 18*(1), 48–63. https://doi.org/10.12700/aph.18.1.2021.1.4

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [The relation between test intelligence and problem solving success]. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie, 189*(1), 79–100.

Rojas, M., Nussbaum, M., & Moreno, C. (2025). Validating a collaborative problem-solving assessment tool across educational stages. *Computers & Education, 227*, Article 105228. https://doi.org/10.1016/j.compedu.2024.105228

Rollett, W. (2008). *Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme [Strategy usage, generated information and used information in the exploration and control of complex dynamic systems]*. LIT-Verlag.

Rost, J. (2006). Latent-Class-Analyse. In F. Petermann, & M. Eid (Eds.), *Handbuch der Psychologischen Diagnostik* (pp. 275–287). Hogrefe.

Rudolph, J., Niepel, C., Greiff, S., Goldhammer, F., & Kröner, S. (2017). Metacognitive confidence judgments and their link to complex problem solving. *Intelligence, 63*, 1–8. https://doi.org/10.1016/j.intell.2017.04.005

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences, 24*, 42–52. https://doi.org/10.1016/j.lindif.2012.12.011

Schwichow, M., Christoph, S., & Härtig, H. (2015). Förderung der Variablen-Kontroll-Strategie im Physikunterricht [Promotion of the control-of-variables strategy in physics education]. *Der Mathematische und Naturwissenschaftliche Unterricht, 68*(6), 346–350.

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review, 39*, 37–63. https://doi.org/10.1016/j.dr.2015.12.001

Shin, Y., Lee, H., & Kim, W. (2025). Exploring creative problem-solving in computer-supported collaborative learning: Focusing on group cohesiveness and socially shared metacognitive regulation. *Thinking Skills and Creativity, 56*, Article 101771. https://doi.org/10.1016/j.tsc.2025.101771

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53*, 92–101. https://doi.org/10.1016/j.intell.2015.09.005

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*, Article 777. https://doi.org/10.3389/fpsyg.2019.00777

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior, 111*, Article 106442. https://doi.org/10.1016/j.chb.2020.106442

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

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence, 72*, 1–12. https://doi.org/10.1016/j.intell.2018.11.003

Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen. Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen* [Intelligence, knowledge and problem solving. Cognitive prerequisites for success in problem solving with computer-simulated problems]. Hogrefe.

Tschirgi, J. E. (1980). Sensible Reasoning: A Hypothesis about Hypotheses. *Child Development, 51*(1), 1–10.

Valerius, S., & Sparfeldt, J. R. (2014). Consistent *g*- as well as consistent verbal-, numerical- and figural-factors in nested factor models? Confirmatory factor analyses using three test batteries. *Intelligence, 44*, 120–133. https://doi.org/10.1016/j.intell.2014.04.003

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75–100. https://doi.org/10.1016/S0364-0213(99)80003-2

Weinstein, C., & Mayer, R. (1986). The teaching of learning strategies. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 315–327). Macmillan.

Weise, J. J., Greiff, S., & Sparfeldt, J. R. (2020). The moderating effect of prior knowledge on the relationship between intelligence and complex problem solving – testing the Elshout-Raaheim hypothesis. *Intelligence, 83*, Article 101502. https://doi.org/10.1016/j.intell.2020.101502

Weise, J. J., Greiff, S., & Sparfeldt, J. R. (2022). Focusing on eigendynamic effects promotes students' performance in complex problem solving: A log-file analysis of strategic behavior. *Computers & Education, 189*, Article 104579. https://doi.org/10.1016/j.compedu.2022.104579

Wirth, J. (2004). *Selbstregulation von Lernprozessen [Self-regulation of learning processes]*. Waxmann.

Wu, H., & Molnár, G. (2022). Analysing complex problem-solving strategies from a cognitive perspective: The role of thinking skills. *Journal of Intelligence, 10*, 46. https://doi.org/10.3390/jintelligence10030046

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

Wüstenberg, S., 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

Zhang, X.-Q., Gao, Q.-H., Tian, W., & Xin, T. (2024). Dynamic and typological explanations of planning in complex problem-solving. *Learning and Individual Differences, 110*, Article 102417. https://doi.org/10.1016/j.lindif.2024.102417