Teaching Multivariational Reasoning through AI-Guided Inquiry in Interactive Simulations

Ekaterina Shved[1], Engin Bumbacher[2], and Seyed Parsa Neshaei[1] and Tanja Käser[1]

[1] EPFL, Lausanne, Switzerland
{kate.kutsenok, seyed.neshaei, tanja.kaeser}@epfl.ch
[2] University of Teaching Education Vaud, Lausanne, Switzerland
engin.bumbacher@hepl.ch

Abstract. Understanding how students learn and enact multivariate quantitative inquiry strategies during science inquiry remains challenging, particularly in scalable instructional settings. Although strategies such as the extended Control of Variables Strategy (CVS++) and the General Principle Strategy (GPS) are known to support multivariate reasoning, less is known about how instruction can foster their uptake and sustained use during inquiry. We investigate this issue using a rule-based AI chatbot that scaffolds strategy-specific reasoning during simulation-based inquiry, providing context-sensitive prompts to guide students in enacting either CVS++ or GPS. In a between-groups experiment, 104 chemistry apprentices engage in a three-phase inquiry sequence: inquiry with AI-supported guidance (*Learning Phase*), independent inquiry in the same simulation with a new variable (*Near Transfer*), and inquiry in a novel simulation (*Far Transfer*). Students' inquiry actions and strategy enactment were captured through fine-grained interaction log data. Results show that CVS++ leads to higher post-instruction performance than GPS, driven primarily by single-variable reasoning items, while no group differences emerge in *Near* or *Far Transfer*. Behavioral analyses indicate that students largely enact the instructed strategies in phases close to instruction, but explicit strategy use declines in the novel simulation, particularly for GPS. Across phases, higher learning is consistently associated with more systematic inquiry behaviors, including coordinated exploration, recording, and analysis actions, rather than strategy use alone. Supplementary Materials are released on our GitHub repository[3].

Keywords: Quantitative multivariate reasoning · AI-chatbot guidance · Inquiry simulation

1 Introduction

Inquiry-based learning is one of the central approaches in science education, engaging students in exploring phenomena, designing experiments, and constructing evidence-based explanations. When implemented effectively, it supports conceptual understanding, scientific reasoning, and knowledge transfer [1, 20, 22].

[3] https://github.com/TeachMultivarReason2026/supplementary-materials-2026

However, inquiry tasks are often complex and ill structured, requiring coordination of domain knowledge and strategic skills, which many learners find challenging [13, 28]. These difficulties are particularly pronounced in multivariate quantitative reasoning. Although many scientific phenomena involve interactive relationships among multiple variables [17], students often struggle to integrate variable effects or derive formal relationships, even when individual variable effects are understood [14, 15]. Supporting this form of reasoning requires coordinating several foundational skills [11], including proportional reasoning [24] and covariational reasoning across multiple dimensions [25].

Consequently, prior work has focused on supporting students' multivariate reasoning within inquiry contexts by helping them recognize that scientific phenomena arise from the joint influence of multiple variables. Across domains, guided inquiry, scaffolding, and structured reflection have been shown to foster qualitative understanding of multivariable causality, such as recognizing interactions among factors, reconciling multiple causal claims, and moving beyond single-variable explanations [4, 12, 17]. Simulation-based approaches further suggest that engaging students in exploratory tasks, either before or alongside instruction, can improve awareness of multivariable relationships and support transfer across contexts [26]. However, these approaches primarily emphasize qualitative sense-making and provide limited guidance for systematically designing quantitative experiments, coordinating proportional changes across variables, or constructing formal multivariable relationships. As a result, students may recognize that multiple variables matter without knowing how to investigate their joint effects productively.

Only a small body of work addresses quantitative multivariate reasoning during inquiry [6,8,21]. For example, invention-based activities that prompt learners to compare multiple contrasting cases have been shown to support the construction of deeper structural explanations and quantitative relationships among variables when using simulations [6]. Similarly, [21] introduced an extended Control of Variables Strategy (CVS++) that guides learners to isolate individual variables and then examine joint effects through proportional variation, supporting reasoning beyond single-variable effects in expert-guided settings. The General Principle Strategy (GPS) takes a complementary approach by encouraging learners to identify compensatory relationships through outcome-preserving comparisons [8]. However, across these approaches, strategy use is typically examined through post-test responses or expert-guided interactions, and instruction relies largely on lectures or interviews. Together, prior work leaves open how students independently enact quantitative multivariate reasoning strategies during inquiry and how these strategies unfold in real time, in part because existing approaches rely on lectures or expert-guided interviews that limit both scalability and fine-grained observation of students' inquiry processes.

Educational AI technologies offer a potential way to address these limitations. Prior work demonstrates that chatbots and conversational agents can scaffold complex scientific practices by guiding learners' exploration, modeling, and reasoning through interactive, context-sensitive prompts. For example, rule-based systems in chemistry and physics tasks provide structured guidance and maintain pedagogical consistency, supporting engagement, conceptual understanding, and motivation [16,18]. In more open-ended inquiry, systems such as Inquirybot guide students through experimental design using scripted conversational prompts [5], showing the potential of chatbots to support exploratory problem-solving. Other AI-driven tutors and collaborative agents dynamically assess learners' actions, provide individualized scaffolds, and facilitate reasoning dialogues [2,3,10]. Across these approaches, most studies focus on task progression, content scaffolding, or dialogue quality rather than explicitly supporting students in applying multivariate reasoning strategies. Consequently, research on AI-supported enactment of learning strategies remains limited.

In this study, we investigate whether and how multivariate inquiry strategies can be integrated into an instructional approach that supports students' quantitative reasoning. Using a between-groups experimental design, we examine the impact of strategy-based instruction on students' multivariate reasoning. We equip an inquiry-based simulation with a rule-based AI chatbot that prompts students to use either CVS++ or GPS, providing pedagogically consistent and context-sensitive guidance during exploration. 104 chemistry apprentices first completed inquiry with AI support, then applied the same strategy independently in the same simulation with a new variable, and finally transferred their reasoning to a novel simulation, allowing us to assess both guided and independent strategy use as well as transfer to new contexts. In this study we address the following research questions: (**RQ1**) Which instructional strategy, CVS++ or GPS, leads to greater improvement in students' conceptual understanding? (**RQ2**) How do students in each instructional group differ in their demonstration and transfer of the taught strategies across tasks? (**RQ3**) Which inquiry behaviors are associated with improved conceptual learning in multivariate reasoning?

Our results indicate that differences between CVS++ and GPS groups appear only immediately after following AI-supported strategy instruction, while performance is similar when transferring to a new task. Behavioral analyses show that students enact the taught strategies in contexts close to instruction, but strategy use declines in a new simulation. High-performing students nevertheless engage in more structured and systematic inquiry across tasks, including the use of the taught strategies, highlighting the importance of systematic inquiry in addition to strategy instruction. These findings suggest that AI-supported strategy guidance can effectively shape initial multivariate reasoning, but sustained application and transfer require fostering structured inquiry behaviors.

2 Methodology

We conducted a three-phase study to teach and measure strategies for multivariate quantitative reasoning (Fig. 1). Instruction focused on interactively teaching a specific multivariate reasoning strategy (CVS++ or GPS) using an AI-chatbot. Across all phases, students interacted with simulations to collect and interpret data. To measure strategy use, we derived features reflecting the instructed strategies and applied sequence mining to distinguish strategy enactment between high- and low-performing students.

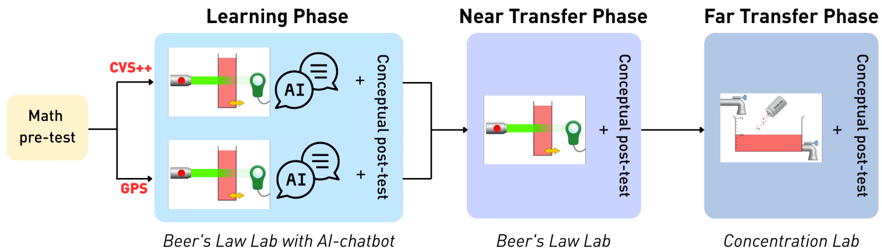


Fig. 1: Study design across *Learning*, *Near Transfer*, and *Far Transfer Phases*.

2.1 Study Context

We conducted an experimental study with two instructional conditions: instruction on CVS++ and instruction on GPS. After completing a mathematics pretest, students were assigned to one condition and entered the *Learning Phase*, where a rule-based AI chatbot guided them in learning and practicing a multivariate reasoning strategy. This phase concluded with a conceptual post-test. Students then applied the learned strategy in the same simulation with an additional variable during the *Near Transfer Phase*, followed by another conceptual post-test. Finally, in the *Far Transfer Phase*, students worked with a new simulation and completed the last conceptual post-test.

Participants. The study included 104 first-year chemistry apprentices (49 female) from Switzerland, aged 15 to 19 years. The study took place during regular classroom instruction. All participants consented to data collection, and a parental opt-out procedure was applied for underage students. The study protocol was approved by the university ethics review board (Approval Nr. 010-2023).

Learning environments. The study employed two interactive PhET simulations[4]: Beer's Law Lab and Concentration Lab (Fig. 2). These simulations support inquiry learning by allowing students to explore scientific phenomena and experiment in a controlled environment. In Beer's Law Lab, students investigated absorbance as the outcome variable by manipulating flask width and solute concentration, and later wavelength, which determines the extinction coefficient. In Concentration Lab, students explored factors affecting molar concentration by varying solute mass, water volume, and solute type, linked to molecular mass. Both simulations included tools for recording variable values and visualizing data through plots.

Procedure. The study lasted 90 minutes and was framed as an industry laboratory task guided by a fictional senior researcher. Participants were randomly assigned within classroom to the CVS++ or GPS condition. At the beginning, students completed a brief mathematics pre-test (5–10 minutes) assessing concepts relevant to quantitative reasoning. The study then proceeded through three phases. In the *Learning Phase*, students explored the Beer's Law Lab simulation with guidance from a rule-based AI chatbot portraying the senior researcher (Fig. 2 (left)), positioned alongside the simulation. Students were tasked to investigate how width, concentration and their combination influenced absorbance

[4] https://phet.colorado.edu/en/simulations

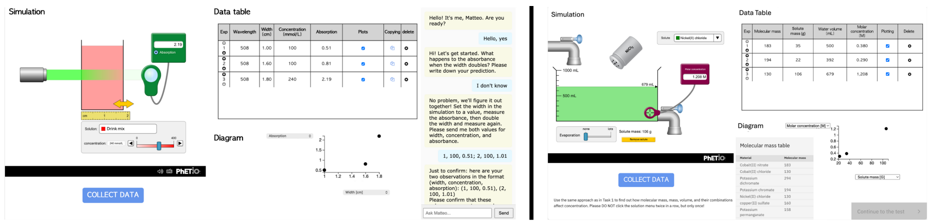


Fig. 2: Beer's Law Lab (left) and Concentration Lab (right) simulations used in the study; containing exploration field (variables to manipulate), table with the data recorded, and a diagram to analyse and plot the data. Beer's Law Lab simulation in the *Learning Phase* was integrated with an AI-chatbot to teach students strategies for quantitative multivariate reasoning.

by systematically applying the assigned strategy. After completing the inquiry, they took a conceptual post-test. In the *Near Transfer Phase*, students continued in the Beer's Law Lab, but investigated an additional independent variable, the extinction coefficient. Without AI support, they applied the learned strategy to reason about how this variable, together with width and concentration, affected absorbance, followed by a conceptual post-test. In the *Far Transfer Phase*, students moved to the Concentration Lab simulation and applied the learned strategy in a novel context to figure out how molecular mass, solute mass and water volume, and their interactions, influenced concentration. This phase also concluded with a conceptual post-test. All conceptual post-tests included questions targeting both single-variable effects and multiple-variable effects to probe different levels of conceptual understanding of the underlying phenomenon. Including both types of questions allowed us to examine how different instructional approaches influenced distinct levels or components of conceptual understanding. Pre- and post-tests are available in Supplementary Materials.

2.2 AI-guided instruction of multivariate reasoning strategies

Instruction in multivariate reasoning was delivered through a rule-based AI chatbot embedded alongside the Beer's Law simulation in the *Learning Phase* (Fig. 2 (left)). The chatbot used GPT-4.1 with a temperature of 1.00 for language generation, while all instructional flow, evaluation criteria, and progression rules were predefined and identical within each instructional condition, ensuring conversational flexibility with high instructional consistency. As illustrated in Fig. 3 (full prompts in Supplementary Materials), each chatbot prompt consisted of two components: (1) general information defining the chatbot's role, context, and global instructional constraints, and (2) strategy-specific instructional logic that differed between the CVS++ and GPS conditions. Students could not access the post-test until they completed all required instructional checkpoints defined by this logic. The chatbot followed a deterministic interaction flow. At each step, it posed a single question, evaluated the student's response against strategy-specific criteria embedded in the prompt, and determined whether to advance, repeat, or redirect the interaction. These criteria included expected proportional dependencies or mathematical relationships and were never disclosed to students.

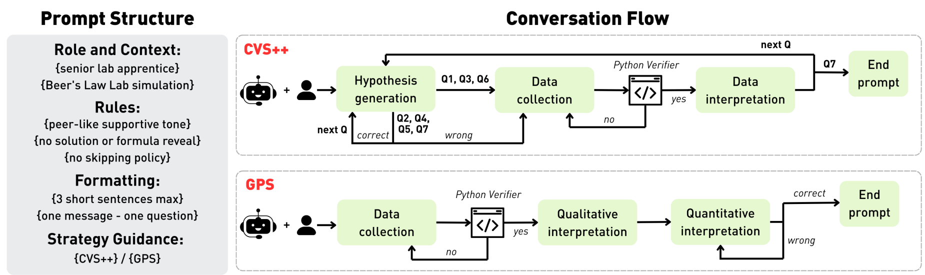


Fig. 3: Structure of the AI-chatbot prompts for teaching CVS++ and GPS multivariate reasoning strategies and their conversation flows.

When responses were incorrect, the chatbot provided targeted nudges encouraging additional data collection or reflection within the simulation. Repeated errors triggered further guidance but did not alter the overall instructional trajectory. Limited adaptivity was implemented through rule-based shortcuts. When a student provided a correct hypothesis in some questions or relationship immediately, the chatbot skipped redundant data collection or validation steps and moved on to the next instructional unit. Post-hoc inspection of chatbot–student interactions confirmed that all students completed the instructional checkpoints, covered the full content for their assigned strategy, and spent approximately 20 minutes on the activity, independent of instructional condition.

CVS++ strategy. The CVS++ strategy supports multivariate reasoning by first examining individual variable effects and then varying multiple variables in proportional steps to observe their combined influence on the outcome [21]. For instance, when exploring how variables A and B influence a result, students might first change A while keeping B constant, then change B while keeping A constant, and finally try combinations such as doubling A and doubling or halving B to observe how the outcome responds. Thus, our instructional goal was to support articulation of proportional and multiplicative relationships without providing formal formulas. The CVS++ chatbot followed a fixed interaction structure comprising hypothesis generation, data collection, validation, and interpretation (Fig. 3). Students completed seven interaction chains: two focused on proportional changes in flask width (**Q1**, **Q2**), two on solute concentration (**Q3**, **Q4**), and three on combined variable changes (**Q5**, **Q6**, **Q7**). For each chain, students predicted how absorbance would change, collected data, and reported observed values. A Python-based verification routine checked whether reported values were consistent with plausible simulation outputs based on Beer's Law equation. Depending on this check, the chatbot prompted either interpretation the results in terms of proportional change or data recollection. For **Q2**, **Q4**, **Q5**, **Q7**, correct hypotheses triggered skipping of data collection and validation, allowing instruction to remain efficient while preserving the strategy structure.

GPS strategy. The GPS strategy encourages learners to vary multiple variables simultaneously to identify different combinations that yield the same outcome [8]. By comparing these equivalent outcomes, students can discern compensatory relationships among variables. GPS is particularly suited to contexts where isolating variables is impracticable or where relationships must be derived from patterns of coordinated variation. Our instructional goal was to support recognition of outcome-preserving combinations and inference the underlying mathematical relationship. In this condition, students first collected five variable pairs producing the same outcome. A Python-based routine verified equivalence by computing expected outcome values, after which the chatbot guided students to describe qualitatively how changes in one variable compensated for changes in another. Students were then prompted to express the quantitative relationship that remained constant across cases. Rule-based adaptivity allowed the chatbot to skip remaining steps when students immediately identified the correct relationship. Throughout, the chatbot did not reveal formulas but supported students in articulating them independently.

2.3 Measuring multivariate reasoning strategies

To measure students' multivariate reasoning strategies, we extracted temporally ordered action sequences from students' interaction logs and encoded them to capture exploration, recording, and analysis behaviors [7]. We then identified strategy-specific action patterns by labeling sequences segments corresponding to CVS++ and GPS reasoning, quantified strategy use, compared behavioral patterns across instructional groups, and applied sequence mining to examine the differences between high- and low-performing students.

Action Extraction. For each student and study phase, we categorized actions as *explore*, *record*, or *analysis*. Explore actions were labeled by the manipulated variable (e.g., *explore_width*). Following [7], consecutive explore actions on the same variable separated by less than 3 seconds were merged. Record actions were labeled by the variables that changed relative to the previously recorded state. For example, *record_width_concentration* label in the *Near Transfer Phase* indicates that width and concentration changed while wavelength remained constant.

Strategy Labeling. Action sequences were then labeled with respect to GPS and CVS++. For GPS, we identified record actions belonging to sets of at least three distinct data points with equivalent outcomes within a tolerance of 0.01 due to the simulation precision, excluding exact repetitions of collected points. For each such record action, we traced backward in the sequence and labeled the closest preceding explore actions involving the changed for the record variables as *GPS_explore_variable*. For example, a *record_width_concentration* action led to labeling the nearest preceding *explore_width* and *explore_concentration* actions as GPS-related. All remaining actions were considered for CVS++ labeling. Because CVS++ involves examining single-variable effects and then combining variables, labeling reflected these steps. An explore action was labeled as *explore_proportional* if the change in the variable corresponded to a proportional factor of 1.5, 2, 2.5, 3, 4, 5, 6, 7, 8, 9, or 10, within a tolerance of 0.05. These factors captured feasible proportional changes given the simulation's variable ranges and precision. When more than two consecutive *explore_proportional* actions involved different variables, these actions were further labeled as *CVS++_explore_variable*, indicating systematic investigation of combined variable effects.

Comparison of strategies usage. To compare students' behaviors, we computed, for each phase and instructional group, the proportion of GPS-labeled actions, CVS++-labeled actions, and *explore_proportional* actions relative to all explore actions. Because CVS++ can be applied at multiple levels, we aggregated *explore_proportional* and *CVS++*-labeled actions into a single measure capturing both foundational and advanced strategy use. Group differences were analyzed using Kruskal–Wallis tests, followed by Dunn's post hoc tests with Benjamini–Hochberg correction (BH).

Performance-related inquiry patterns. To discover action sequences associated with learning outcomes, we divided students into high-performing (HP) and low-performing (LP) groups for each phase using a median split of conceptual post-test scores (HP: score $\geq$ median; LP: score $<$ median). We then preprocessed labeled action sequences to improve the interpretability of the patterns. Specifically, we collapsed consecutive identical actions and removed variable-specific information, so that actions targeting different variables were treated as the same type (e.g., *GPS_explore_wavelength* and *GPS_explore_width* were both encoded as *GPS_explore*). This preprocessing step allowed us to focus on strategic structure rather than surface-level differences. From each processed action stream, we extracted all unique action sequences of length two to ten and retained those occurring in at least 35% of either HP or LP students, ensuring focus on group-level behaviors rather than rare individual actions [23].

To characterize overall behavioral differences, we computed Jensen-Shannon divergence between HP and LP sequence usage profiles. Follow-up analyses focused on the five sequences with the highest divergence. For these sequences, we applied Pearson's chi-square tests to examine differences in sequence presence, and Mann–Whitney U tests to compare usage frequency. All p-values were corrected using the BH procedure.

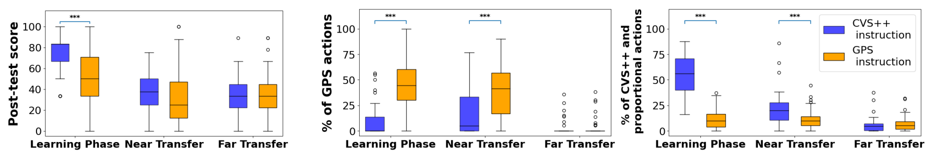
3 Results

We applied our measurement framework to analyze differences in conceptual learning between the CVS++ and GPS groups (**RQ1**), differences in strategic behaviors across instructional groups (**RQ2**), and associations between performance and specific inquiry patterns (**RQ3**).

3.1 Differences in conceptual learning of instructional groups (RQ1)

We first tested for group differences in mathematical pre-test performance and found no significant difference between CVS++ and GPS ($U = 1515.5$, $p = .21$). Overall performance was high ($M = 76\%$, $SD = 21\%$), indicating substantial prior knowledge.

Figure 4a shows post-test performance across phases. To assess instructional effects on conceptual learning while controlling for prior mathematical proficiency, we fitted linear regression models for each phase ($\text{Post}_{\text{concept}}(\text{phase}) \sim \text{Group} + \text{Pre}_{\text{math}}$). In the *Learning Phase*, instructional group and mathematical pre-test score significantly predicted post-test performance. CVS++ students outperformed GPS students ($M = 74\%$, $SD = 21\%$ vs. $M = 52\%$, $SD = 29\%$).



(a) Post-test scores

(b) Usage of GPS and CVS++ strategies

Fig. 4: Conceptual post-test scores (a) and usage of GPS (b, left) and CVS++ (b, right) strategies across all phases. *** p $\leq$ .001

The overall model was significant ($F(2, 97) = 15.41$, $p < .001$, $R^2 = .24$). Regression results confirmed lower scores for GPS relative to CVS++ ($B = -1.24$, $SE = 0.30$, $p < .001$) after controlling for pre-test performance. Mathematical pre-test score was also a positive predictor ($B = 0.37$, $SE = 0.12$, $p < .01$). This pattern likely reflects the *Learning Phase* post-test's stronger emphasis on single-variable effects, which were explicitly addressed in CVS++ instruction but not foregrounded in GPS.

In the *Near Transfer Phase*, group performance converged (CVS++: $M = 33\%$, $SD = 20\%$; GPS: $M = 32\%$, $SD = 24\%$). The overall regression model was not significant ($F(2, 94) = 2.46$, $p = .09$, $R^2 = .05$). Instructional group did not predict post-test performance ($B = -0.01$, $SE = 0.36$, $p = .98$), while mathematical pre-test score remained a significant positive predictor ($B = 0.31$, $SE = 0.14$, $p = .03$).

A similar pattern was observed in the *Far Transfer Phase*. Post-test performance was comparable across groups (CVS++: $M = 37\%$, $SD = 19\%$; GPS: $M = 36\%$, $SD = 21\%$), and the overall model was not significant ($F(2, 79) = 0.81$, $p = .45$, $R^2 = .02$). Neither instructional group ($B = 0.03$, $SE = 0.40$, $p = .94$) nor mathematical pre-test score ($B = 0.21$, $SE = 0.16$, $p = .21$) significantly predicted post-test performance.

3.2 Differences in strategies application between groups (RQ2)

To examine behavioral differences between instructional groups, we quantified students' use of both multivariate reasoning strategies based on logged action sequences. Figure 4b (left) shows GPS strategy use, measured as the percentage of *GPS_explore* actions, across phases. In the *Learning Phase*, students in the GPS group used the GPS strategy significantly more than students in the CVS++ group ($U = 208.5$, $p < .001$). A small proportion of GPS-labeled actions also appeared in the CVS++ group, which is expected because students could unintentionally record at least three states with identical outcome values. A similar pattern was observed in the *Near Transfer Phase*, where the GPS group continued to use the GPS strategy significantly more often than those in the CVS++ group ($U = 752.5$, $p < .01$). The slight increase in GPS-labeled actions in the CVS++ group may reflect simulation characteristics, as repeated absorbance values were easier to obtain when exploring the newly introduced wavelength variable. In the *Far Transfer Phase*, no significant group differences were observed, and overall GPS strategy use was very low in both groups ($M = 3\%$, $SD = 8\%$), indicating difficulties in transferring GPS to a new context.

Figure 4b (right) presents CVS++ strategy use across phases. In the *Learning Phase*, students in the CVS++ group used the strategy significantly more than students in the GPS group ($U = 2657.0$, $p < .001$). Some CVS++ behavior was also observed in the GPS group, primarily reflecting proportional exploration of individual variables. This pattern persisted in the *Near Transfer Phase*, with the CVS++ group again showing significantly greater strategy use ($U = 1710.0$, $p < .001$). However, in the *Far Transfer Phase* group differences disappeared, and overall CVS++ use was low in both groups ($M = 6\%$, $SD = 7\%$). Across the *Near* and *Far Transfer Phases*, most students primarily engaged in proportional exploration of single variables, while explicitly labeled *CVS++* actions constituted only a small fraction of overall behavior. This pattern likely reflects the increased difficulty of applying CVS++ to novel and less familiar variables in the transfer simulations.

3.3 Performance-related inquiry patterns (RQ3)

We applied sequence mining to identify strategic action patterns distinguishing high performers (HP) from low performers (LP). Results were robust to variations in the median split procedure, including alternative HP and LP definitions (HP $\geq$ median vs. HP $>$ median) and exclusion of students with median scores.

Learning Phase. This phase included 38 LP and 62 HP students, with a significantly higher proportion of CVS++–instructed students among HP students ($\chi^2 = 7.73$, $p < .01$). Sequence mining yielded 60 candidate sequences of length $2 - 10$. We compared overall sequence usage distributions between HP and LP and identified the five sequences with the largest Jensen-Shannon divergence (Table 1). These sequences predominantly contained *GPS_ explore* actions and were more characteristic of LP students, consistent with **RQ1** findings linking GPS instruction during the *Learning Phase* to lower post-test performance. All five subsequences showed significant group differences in binary usage. In each case, LL students were more likely than HL students to use the subsequence: `record-GPS_explore` ($\chi^2 = 10.41$, $p < .01$); `record-analysis` ($\chi^2 = 9.62$, $p < .01$); `GPS_explore-GPS_explore-record-GPS_explore` ($\chi^2 = 10.55$, $p < .01$); `GPS_explore-GPS_explore` ($\chi^2 = 9.50$, $p < .01$); and `GPS_explore-record -GPS_explore` ($\chi^2 = 8.32$, $p < .01$). In contrast, no significant group differences were observed in usage frequency conditional on subsequence occurrence (all $U = 740.0 - 841.5$, all $p \geq .99$).

Near Transfer Phase. This phase included 25 LP and 69 HP students, with no significant differences in instructional group distribution ($\chi^2 = 0.66$, $p = .42$). Sequence mining identified 38 subsequences used by at least 35% of either group. The five sequences with the highest Jensen–Shannon divergence are shown in Table 1. Most contained GPS-related actions but were embedded within broader inquiry cycles involving exploration, recording, and analysis. The most divergent sequence, `record-explore-record-analysis`, reflects a systematic inquiry loop independent of the type of variable manipulation. All top five sequences were more characteristic of HP students. However, none showed significant group differences in binary usage after correction for multiple comparisons (all $\chi^2 = 2.55 - 4.37$, all $p = .11$). In contrast, frequency usage analyses revealed clearer

Table 1: Top five sequences by Jensen-Shannon divergence (JSD), with the percentage (%) of HP and LP using the sequence at least once and the average frequency (#) of usage among HP and LP along with statistical test results. Markers on % and # columns indicate significant differences between HP and LP for that metric (†p $<$ .10, *p $<$ .05, **p $<$ .01).

Phase	**Sequence**	**JSD**	**%HP**	**%LP**	**#HP**	**#LP**
Learning	*'record', 'GPS_explore'*	0.099	50	84^{**}	2.1	3.5
	'record', 'analysis'	0.097	58	89^{**}	2.0	3.3
	'GPS_explore', 'GPS_explore', 'record', 'GPS_explore'	0.094	35	71^{**}	1.0	1.7
	'GPS_explore', 'GPS_explore'	0.091	52	84^{**}	2.2	4.2
	'GPS_explore', 'record', 'GPS_explore'	0.076	42	74^{**}	1.4	2.8
Near Transfer	*'record', 'explore', 'record', 'analysis'*	0.061	42	16	0.5	0.2^{*}
	'GPS_explore', 'record', 'explore'	0.050	49	24	0.8	$0.4^{\dagger}$
	'record', 'explore', 'record'	0.041	64	40	1.8	$1.4^{\dagger}$
	'explore', 'explore', 'explore'	0.035	45	24	0.9	$0.7^{\dagger}$
	'analysis', 'GPS_explore'	0.035	45	24	5.6	$1.1^{\dagger}$
Far Transfer	*'explore_proportional', 'record', 'explore'*	0.092	37	8^{*}	0.4	0.1^{*}
	'record', 'explore_proportional'	0.078	40	12^{*}	0.6	0.2^{*}
	'explore_proportional', 'record'	0.077	46	16^{*}	0.8	0.2^{*}
	'explore', 'explore', 'explore'	0.055	89	100	17.4	10.6
	'record', 'explore', 'record'	0.054	40	$16^{\dagger}$	0.8	0.3^{*}

distinctions: HP students used `record-explore-record-analysis` significantly more often than LP students (U = 1093.0, p = .045). The remaining four sequences were also used more frequently by HP students, with differences reaching trend-level significance after correction: `GPS_explore-record-explore` (U = 1072.5, p = .06), `record-explore-record` (U = 1026.5, p = .08), `analysis-GPS_explore` (U = 1009.5, p = .08), and `explore-explore-explore` (U = 1013.5, p = .08).

Far Transfer Phase. This phase included 25 LP and 57 HP students, with no significant differences in instructional group distribution (χ^2 = 0.04, p = .85). Sequence mining identified 59 subsequences used by at least 35% of either group. The five most divergent sequences (Table 1) largely reflected inquiry cycles involving exploration and recording. Unlike earlier phases, analysis actions no longer appeared among the most divergent patterns, likely indicating reduced perceived need for explicit analysis. GPS-related actions disappeared from the top sequences, reflecting their minimal overall use. Instead, *explore_proportional* actions emerged, which align more closely with CVS++–type reasoning though can be enacted by students from both instructional groups. Four of the five sequences were predominantly associated with HP students. The remaining sequence (`explore-explore-explore`) was used frequently by both groups, with slightly higher usage among LP students. Chi-square and Mann-Whitney tests revealed significant group differences for several highly divergent sequences. `explore_proportional-record-explore` was used more often by HP students in terms of binary usage (χ^2 = 5.81, p = .02) and frequency (U = 919.0, p = .01). Similarly, `record-explore_proportional` showed higher usage among HP students both in binary terms (χ^2 = 5.21, p = .02) and frequency (U = 916.5,

$p = .01$). The sequence `explore_proportional-record` also differed between groups in both binary ($\chi^2 = 5.35$, $p = .02$) and frequency usage ($U = 934.5$, $p = .01$). `record-explore-record` showed a trend-level difference in binary usage ($\chi^2 = 3.63$, $p = .06$) and was significantly more frequent among HP students ($U = 878.5$, $p = .03$). In contrast, `explore-explore-explore` did not differ between groups for either binary usage ($\chi^2 = 1.50$, $p = .22$) or frequency ($U = 764.0$, $p = .30$).

4 Discussion and Implications

In this study, we taught students two multivariate reasoning strategies, CVS++ and GPS, in inquiry-based learning through a rule-based AI chatbot. Across *Learning*, *Near Transfer*, and *Far Transfer Phases*, we tracked inquiry behaviors in interactive simulations and related them to conceptual understanding of combined variable effects. Sequence mining further allowed us to identify behavioral patterns distinguishing high and low performers.

Conceptual post-test results showed a significant difference between instructional groups only in the *Learning Phase*, with CVS++ group outperforming GPS, however no differences emerged in transfer phases. This pattern is best explained by the composition of the *Learning Phase* post-test, which included a high number of questions on single-variable effects due to the specific inquiry task. Because the GPS strategy does not emphasize isolating single-variable effects, GPS students were less prepared for these items immediately after instruction. Importantly, this disadvantage disappeared in subsequent phases, suggesting that students were able to expand their strategic approaches over time by integrating elements from multiple strategies. This aligns with prior findings indicating that flexible strategy use, rather than strict adherence to a single strategy, supports learning across contexts [27, 29].

Behavioral analyses showed that students largely followed the strategy they were taught in the *Learning* and *Near Transfer Phases*, but strategy use declined sharply in the *Far Transfer Phase*, particularly for GPS. This likely reflects difficulties in transferring procedural strategies to a novel context [9], as well as the potentially higher demands of GPS. Interestingly, we observed instances of unintended strategy use in the non-instructed group, but they were most probably influenced by simulation affordances. For example, GPS-like behavior emerged more frequently in the *Near Transfer Phase* due to properties of the wavelength variable, whereas it nearly disappeared in the *Far Transfer* simulation where identical outcome values were harder to obtain. These findings highlight that students' inquiry behaviors and strategy usage can be shaped not only by instruction, but also by task context, simulation characteristics, and other external factors.

Finally, we examined whether and how strategy-related behaviors related to conceptual learning by comparing action sequences of high and low performers. In the *Learning Phase*, sequence mining confirmed earlier results while adding important nuance: GPS students, who scored lower on the post-test despite high strategy adherence, showed exploration sequences characteristic of low performers, whereas CVS++ exploration aligned with high performers. Crucially, exploration actions rarely appeared in isolation, typically embedded within sequences included recording and, less frequently, analysis actions. This suggests that strategic exploration alone is insufficient, and learning depends on coordination across phases of inquiry. In the *Near Transfer Phase*, this pattern became even clearer: high performers' sequences reflected full inquiry cycles iterating between exploration, recording, and analysis [7], and GPS-related actions became more common among high performers, even if initially trained on CVS++. In the *Far Transfer Phase*, where GPS usage largely vanished, proportional exploration emerged as the dominant differentiating behavior. Even when full multivariate strategies were no longer applied, students who continued to explore variables proportionally and record their observations performed better on the conceptual post-test. This finding reinforces the idea that when strategy transfer breaks down, more basic but structured forms of quantitative inquiry can still support learning [19].

Overall, these findings suggest that multivariate quantitative reasoning is supported less by persistent specific strategy use and more by engagement in systematic inquiry. Rule-based AI guidance can effectively scaffold strategy use and ensure consistent instructional exposure, but sustained learning and transfer appear to depend on students' capacity to adapt and recombine strategic elements across contexts. For future research, this points to the importance of designing AI-supported inquiry environments that not only teach specific strategies, but also explicitly support abstraction, reflection, and strategic flexibility. From an instructional perspective, scaffolds that emphasize inquiry cycles and gradual transition from proportional to fully multivariate reasoning may be particularly effective in supporting robust transfer.

Acknowledgments. This project was substantially financed by the Swiss State Secretariat for Education, Research and Innovation SERI. Thanks to Peter Bühlmann and Tanya Nazaretsky for their valuable input.

References

1. Abdi, A.: The effect of inquiry-based learning method on students' academic achievement in science course. Univers. J. Educ. Res. (2014)
2. Adair, A., Pedro, M.S., Gobert, J., Segan, E.: Real-time ai-driven assessment and scaffolding that improves students' mathematical modeling during science investigations. In: Proc. AIED (2023)
3. de Araujo, A., Papadopoulos, P.M., McKenney, S., de Jong, T.: A learning analytics-based collaborative conversational agent to foster productive dialogue in inquiry learning. J. Comput. Assist. Learn. (2024)
4. Caspari, I., Graulich, N.: Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. Int. J. Phys. Chem. Educ. (2019)
5. Chang, J., Park, J., Park, J.: Using an artificial intelligence chatbot in scientific inquiry: Focusing on a guided-inquiry activity using inquirybot. Asia-Pac. Sci. Educ. (2023)
6. Chi, M., Dohmen, I., Shemwell, J.T., Chin, D.B., Chase, C.C., Schwartz, D.L.: Seeing the forest from the trees: A comparison of two instructional models using contrasting cases. In: Proc. AERA (2012)

7. Cock, J.M., Roll, I., Käser, T.: Consistency of inquiry strategies across subsequent activities in different domains. In: Proc. AIED (2023)
8. Conlin, L.D., Hallinen, N.R., Schwartz, D.L.: Supporting middle schoolers' use of inquiry strategies for discovering multivariate relations in interactive physics simulations. In: Proc. ICLS (2014)
9. Dixon, R.A., Brown, R.A.: Transfer of learning: Connecting concepts during problem solving. J. Technol. Educ. (2012)
10. Hou, X., Forsyth, C., Andrews-Todd, J., Rice, J., Cai, Z., Jiang, Y., Zapata-Rivera, D., Graesser, A.: An llm-enhanced multi-agent architecture for conversation-based assessment. In: Proc. AIED (2025)
11. Jones, S.R.: Multivariation and students' multivariational reasoning. J. Math. Behav. (2022)
12. Keselman, A.: Supporting inquiry learning by promoting normative understanding of multivariable causality. J. Res. Sci. Teach. (2003)
13. Kirschner, P., Sweller, J., Clark, R.E.: Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. Educ. Psychol. (2006)
14. Kuhn, D.: Reasoning about multiple variables: Control of variables is not the only challenge. Sci. Educ. (2007)
15. Kuhn, D., Pease, M., Wirkala, C.: Coordinating the effects of multiple variables: A skill fundamental to scientific thinking. J. Exp. Child Psychol. (2009)
16. Lee, J., An, T., Chu, H.E., Hong, H.G., Martin, S.N.: Improving science conceptual understanding and attitudes in elementary science classes through the development and application of a rule-based ai chatbot. Asia-Pac. Sci. Educ. (2023)
17. Lesperance, R., Kuhn, D.: Breadth and relevance of multivariable inquiry supports deep understanding of science practice. Sci. Educ. (2023)
18. Lolinco, A.T., Holme, T.A.: Developing a curated chatbot as an exploratory communication tool for chemistry learning. J. Chem. Educ. (2023)
19. Mayes, R.L., Forrester, J.H., Schuttlefield Christus, J.D., Peterson, F., Walker, R.: Quantitative reasoning learning progression: The matrix. Numeracy (2014)
20. Minner, D.D., Levy, A.J., Century, J.: Inquiry-based science instruction—what is it and does it matter? results from a research synthesis years 1984 to 2002. J. Res. Sci. Teach. (2010)
21. Panorkou, N., Germia, E.F.: Integrating math and science content through covariational reasoning: The case of gravity. Math. Think. Learn. (2021)
22. Pedaste, M., Mäeots, M., Siiman, L.A., De Jong, T., Van Riesen, S.A., Kamp, E.T., Manoli, C.C., Zacharia, Z.C., Tsourlidaki, E.: Phases of inquiry-based learning: Definitions and the inquiry cycle. Educ. Research. Rev. (2015)
23. Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D., Yee, N., Roll, I.: Identifying productive inquiry in virtual labs using sequence mining. In: Proc. AIED (2017)
24. Ramful, A., Narod, F.B.: Proportional reasoning in the learning of chemistry: Levels of complexity. Math. Educ. Res. J. (2014)
25. Rodriguez, J.M.G., Bain, K., Towns, M.H., Elmgren, M., Ho, F.M.: Covariational reasoning and mathematical narratives: Investigating students' understanding of graphs in chemical kinetics. Chem. Educ. Res. Pract. (2019)
26. Saba, J., Kapur, M., Roll, I.: The development of multivariable causality strategy: Instruction or simulation first? In: Proc. AIED (2023)
27. Star, J.R., Rittle-Johnson, B.: Flexibility in problem solving: The case of equation solving. Learn. Instr. (2008)

28. Teplá, M., Distler, P.: The impact of long-term inquiry-based science education on students' motivation and knowledge acquisition: the role of gender, subject, and level of inquiry. Humanit. Soc. Sci. Commun. (2025)
29. Xu, L., Liu, R.D., Star, J.R., Wang, J., Liu, Y., Zhen, R.: Measures of potential flexibility and practical flexibility in equation solving. Front. Psychol. (2017)