Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Evaluating Students' Scientific Inquiry Strategies in Large-Scale Digital Assessments

Namrata SRIVASTAVA[a*], Emma LINSENMAYER[b], Mario PIACENTINI[b] & Gautam BISWAS[a]

[a]*Vanderbilt University, USA*

[b]*OECD, France*

*namrata.srivastava@vanderbilt.edu

Abstract: As education shifts toward digital learning environments, new assessments are needed to evaluate not only what students know but also how they learn using available tools and resources. The OECD Programme for International Student Assessment (PISA) 2025 Learning in the Digital World (LDW) assessment addresses this challenge by engaging students in open-ended, interactive tasks that require them to apply learning content to solve inquiry-based problems. This paper focuses on two key scientific inquiry strategies assessed through LDW tasks: Control of Variable (CoV), which reflects how systematically students conduct inquiry experiments, and Deriving Relationship from Data (DRD), which involves interpreting relationships or patterns from the experimental results. We present a scoring rubric designed for fine-grained analysis of student performance, allowing for partial credit to be linked to degrees of applying and reasoning with the strategies to derive solutions to assigned problems. Using a learning-by-modeling task in the LDW framework, we apply this rubric to pilot data collected from 6,800 students across 63 countries. Our findings show that students' success with the CoV and DRD strategies is influenced not only by prior knowledge but also by how they engage with instructional phases and utilize digital and scaffolding tools during the task.

Keywords: computational problem-solving, scoring rules, digital learning environment

1. Introduction

As digital technologies become more prevalent in education, there is a growing recognition that assessments must evolve beyond traditional testing measures of static content knowledge (NRC, 2012; OECD, 2017; Vo & Simmie, 2025). Assessments need to evaluate how students learn, apply, and adapt their learning and understanding in interactive, real-world problem-solving contexts – key competencies emphasized globally as part of 21st century education reform (Chu et al., 2017; Griffin et al., 2012). The Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) 2025 Learning in the Digital World (LDW) assessment marks a significant advance in this area. This innovative assessment immerses students in open-ended digital tasks, requiring them to engage with instructional content and demonstrate strategic problem-solving skills using digital tools to solve problems (OECD, 2023).

Unlike traditional assessments that rely on selected responses, LDW tasks are open-ended, interactive, and simulate authentic learning. Students build and debug computational artefacts, conduct experiments, analyze data, and reflect on progress. This shift aligns with a trend towards active learning (Prince, 2004) and simulation-based assessments that better capture complex inquiry and reasoning (Baker et al., 2016; Gobert et al., 2013). It embodies the concept of "learning by doing", allowing students to build and apply knowledge through experimentation, feedback, and exploration. Consequently, LDW marks a significant shift from traditional approaches, emphasizing active processes over static recall. Students must demonstrate their development, integration of new information, testing ideas, and application of strategies while using digital tools.

This change in assessment design demands new interpretations of student responses. Traditional scoring methods that categorize answers as right or wrong in open-ended and interactive tasks, like those in LDW assessments, fail to capture partial understanding, emerging reasoning, and strategic problem-solving. Effective assessments should evaluate both students' final products and their processes. This highlights the need for scoring frameworks that provide final scores (summative assessment) and detailed feedback to foster critical thinking and problem-solving skills. Such feedback helps students and educators understand key learning processes and identify improvement areas (formative insight).

This paper introduces a scoring rubric for evaluating students' problem-solving strategies within an LDW computational problem-solving task. The rubric uses partial-credit scoring to capture the nuances of student reasoning and awards points that reflect their use of productive problem-solving methods to achieve goals. The goal is to acknowledge students for engaging in meaningful inquiry strategies and to gain insights into their problem-solving processes. We focus on two core strategies that are crucial to scientific inquiry and central to the LDW competency model: **Control of Variables (CoV)** and **Deriving Relationships from Data (DRD)**. CoV assesses how effectively students can design inquiry experiments that systematically change one independent variable while keeping other potential influencing factors constant to isolate the effect of the variable being tested. DRD evaluates how well students can interpret experimental results by analyzing generated data to identify patterns, correlations, and causal relationships between variables. We chose to focus on these strategies because both are essential for interpreting scientific processes and are consistently emphasized in science education frameworks as fundamental to inquiry and causal reasoning.

We tested our scoring rubric using data collected during a PISA LDW pilot study in 63 countries with 6800 students. Our findings show that students' success with CoV and DRD strategies is shaped not only by prior knowledge but also by how they engage with instructional phases and use available tools during the task.

2. Background

Inquiry-based learning (IBL) facilitates active exploration and investigation of topics that interest students (Pedaste et al., 2015). It encourages students to ask questions, gather evidence, draw conclusions, and construct knowledge, fostering deeper understanding and critical thinking skills similar to the practices followed by professional scientists (Keselman, 2003). This method is often regarded as supporting the development of problem-solving skills (Pedaste & Sarapuu, 2006). In a comprehensive review of the literature, Pedaste et al. (2015) proposed a synthesized framework for IBL consisting of five distinct inquiry phases: *Orientation, Conceptualization, Investigation, Conclusion*, and Discussion. These phases form a cyclical and iterative cycle of learning and problem solving and can be further divided into sub-phases. For instance, the Investigation phase is divided into three sub-phases: *Exploration, Experimentation*, and *Data Interpretation*, involving tasks such as generating hypotheses, planning and designing experiments, and analyzing data to draw conclusions.

A central aspect of the Experimentation and Data Interpretation sub-phases is the ability to design and conduct valid experiments. In science education, this centers on two crucial inquiry skills: **control of variables (CoV)** and **deriving relationships from data (DRD)** strategies, which are the focus of the current study. The CoV strategy refers to a student's ability to manipulate one independent variable at a time while keeping others constant to establish relationships between the independent and outcome variables. This skill is essential for ensuring that observed effects can be attributed to the variables being tested, rather than being confounded by uncontrolled factors (Chen & Klahr, 1999). A substantial body of research has explored how this strategy develops (Kuhn, 2010; Zimmerman, 2007) and how educators can best support its development (see Schwichow et al. (2016) for meta-analysis). However, findings remain mixed, and only a few empirical studies have examined how students enact this strategy using process-level data. For instance, Schwichow et al. (2016) conducted a comprehensive meta-analysis of 72 intervention studies focused on CoV instruction, reporting an overall positive effect of teaching CoV (mean effect size = 0.61). However, they also found substantial variability across studies, influenced by factors such as the type of instructional support, the assessment method, and whether feedback or demonstrations were included in the intervention. These contrasting outcomes suggest that students' performance on CoV tasks may depend on additional contextual or behavioral factors, highlighting the need to go beyond static outcome measures and explore how students engage with tasks at the process level.

Complementing CoV is the DRD strategy, emphasizing students' ability to interpret data patterns and infer relationships between variables. Although DRD is a critical aspect of the inquiry process, it has attracted comparatively less attention in the literature. For instance, prior studies have shown that students struggle to apply DRD strategies such as recognizing relevant trends, distinguishing between causal and correlational patterns, and expressing findings through models or graphs (Donnelly-Hermosillo et al., 2020; Masnick & Klahr, 2003). These challenges highlight the need for targeted assessments that can capture students' developing inquiry skills.

Digital assessments like the PISA LDW offer significant potential for exploring students' understanding of variable reasoning and data interpretation skills. The interaction data, along with process data captured through digital logs, can provide deeper insights not only into whether students solve tasks correctly or incorrectly but also into how they approach, engage with, and navigate the investigative phases of inquiry. This study aims to provide a process-oriented perspective on how students develop inquiry skills in digital environments by analyzing these strategies through log-based analysis and rubric-aligned scoring.

3. Methodology

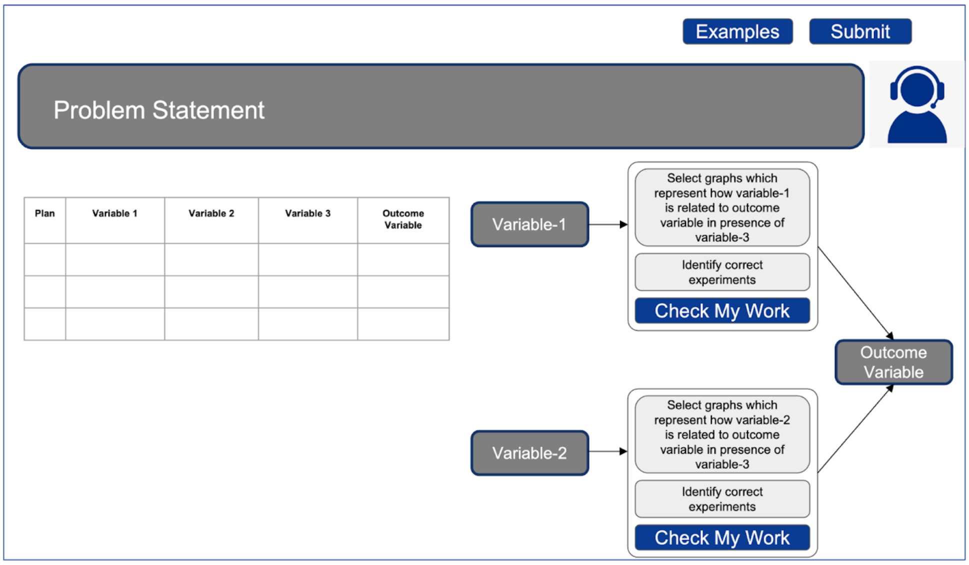
3.1 LDW Unit Description

To assess students' application of CoV and DRD strategies in a digital inquiry setting, we analyzed student interaction data for the "Increasing Tomato Yield" Unit within the PISA LDW framework. The Example Unit is a 30-minute interactive assessment task that follows the four-phase structure of LDW: *Show, Learn, Apply*, and *Reflect*.

1. **Intro and Show phase (Pre-test):** The unit begins with a static introductory page that outlines the unit's overall goals and illustrates a real-world scenario related to variable relationships. This is followed by the Show phase, which includes four pre-test items designed to assess students' prior understanding of core concepts. These items measure students' ability to design controlled experiments, interpret variable relationships in graphical form, and draw inferences from visual data. They establish a baseline before students engage in scaffolded learning activities.
2. **Learn phase:** During this phase, students receive guided instruction from a virtual tutor as they complete a series of scaffolded tasks. These tasks help familiarize students with the experimental interface and provide hands-on experience using the CoV strategy to design controlled experiments. Additionally, students learn to apply DRD strategies, such as generating graphs to interpret their experimental results. Scaffolding is provided through example solutions and correct answers after each task. This approach allows students to practice selecting variables, conducting controlled experiments, and matching outcome patterns to graphs. The goal of this phase is to ensure that all students have the foundational knowledge needed to engage productively with the remaining tasks in this unit.
3. **Apply phase (Comprehensive Challenge task):** In this phase, students encounter an open-ended task in a new but related context, where they must independently apply the strategies they learned earlier. This phase focuses on three key objectives:
   i) Designing valid experiments using the CoV strategy.

ii) Applying the DRD strategy to choose the graph that accurately represents the relationship between variables.

iii) Identifying the subset of experiments that support the relationships between independent and outcome variables.



Plan	Variable 1	Variable 2	Variable 3	Outcome Variable

Select graphs which represent how variable-1 is related to outcome variable in presence of variable-3

Identify correct experiments

Select graphs which represent how variable-2 is related to outcome variable in presence of variable-3

Identify correct experiments

Figure 1. A prototype of the Challenge task.

In contrast to the Learn phase, students do not receive real-time guidance or access to final solutions during the Challenge task. However, they can refer back to the annotated solutions from the Learn phase for help. Additionally, a "Check My Work" feature is available that provides feedback on whether the graphs and experiments they have chosen accurately represent the underlying relationships (see Figure 1). The overall assessment adopts “a low threshold, high ceiling” approach that enables all students to make some progress while allowing more advanced learners to demonstrate deeper reasoning and mastery of problem-solving using the CoV and DRD strategies.

4. **Reflect phase:** This phase consists of three brief self-assessment tasks. First, students evaluate whether they could not accomplish, partially accomplished, or fully accomplished each of the three sub-goals in the Challenge task. Second, they reflect on their emotional state during the problem-solving (for example, feelings of confusion or boredom). Third, they indicate how difficult they found the task to be, and the amount of effort required to solve it. These reflections provide valuable insights into students' self-regulation and perceived performance.

While each phase of the task contributes to the overall learning experience, this paper specifically evaluates student performance during the Challenge task. In this task, students independently demonstrate their problem-solving strategies. Additionally, students’ responses to the Show phase questions were scored to assess their prior knowledge, which is included in the analysis to provide context for their performance in the Challenge task.

3.2 Rubric Development

The framework for the challenge task was developed using a partial-credit system that evaluated students' use of the CoV and DRD strategies. Instead of assigning binary scores, we awarded points for both complete and partially correct responses.

The rubric is structured around three sub-goals of the Challenge task: (1) selecting correct relationships and graphs (DRD), (2) identifying relevant experiments (CoV), and (3) assessing the quality and coverage of the designed experiments (CoV). For each sub-goal, the rubric criteria were defined to reflect levels of correctness, completeness, and strategic execution. A summary of these components, illustrated with an example task – “Increasing

Tomato Yield," which includes three variables of interest (sunlight intensity, soil type, and amount of water used) is provided in Table 1.

To assess DRD strategies, the rubric evaluated whether students correctly identified if each independent variable (sunlight intensity, soil type, and amount of water used) influenced the outcome (tomato yield) and selected the appropriate graph to represent that relationship. Partial credit was awarded to students who recognized the general direction of the relationship but overlooked more nuanced patterns. Additionally, students received points for selecting a subset of experiments that clearly demonstrated the effect of a variable, reflecting their skill in interpreting and choosing informative data.

For CoV, the rubric assessed both the quality and coverage of students' experiments. Quality was evaluated based on whether students systematically held one variable constant while varying another, which indicates strategic planning. Coverage was determined by the number of unique experiments conducted by the students. Full credit was awarded for complete, well-controlled experiments across both soil conditions, while partial credit was given for incomplete or partially controlled experimental designs.

Table 1. *Rubric for scoring students' inquiry tasks*

Goal	Sub-goals	Example ***Design experiments to investigate how light and water affect the number of tomatoes produced, and check whether the relationships depend on soil type.***	Points Awarded	Max points
DRD	Choosing the correct relationship and graph (variable-1)	Tomato yield is not impacted by soil type (partial credit for choosing "no")	1	8
		Tomato yield increases from low to normal sunlight intensity conditions, but drops for high light intensity conditions (e.g., non-linear pattern)	2	
	Choosing the correct relationship and graph (variable-2)	Tomato yield is influenced by the amount of water used (partial credit for choosing "yes")	1	
		Tomato yield increases linearly with the amount of water used for a soil type	2	
	Choosing the correct experiments (variable-1)	Selecting experiments that show the correct water use-tomato yield relationship	1	
	Choosing the correct experiments (variable-2)	Selecting experiments that show the correct sunlight intensity-tomato yield relationship	1	
CoV	Coverage of experiments	3–5 unique experiments conducted; may cover only part of a variable set	1	3
		6–11 experiments conducted; possibly covers 2 conditions for 1 variable	2	
		12 experiments conducted; covers all combinations of 2 variables while holding one constant	3	
	Quality of experiments	**Unpaired Incomplete:** Held water use or sunlight intensity constant for only one soil condition and only one other variable	1	3
		Unpaired Complete: Held water or light constant for only one soil condition, but did this for both light and water	2	
		CoV Complete: Exhaustively held soil type constant while varying both water and light	3	

3.3 Data Collection

This data was collected as part of an OECD PISA 2025 LDW pilot study. The pilot was administered in 63 countries and involved 6,800 students. Students completed a series of LDW prototype units under standardized assessment conditions using a digital platform that automatically recorded detailed interaction log data. Ethics approval and data handling protocols adhered to OECD standards and local requirements in the participating countries.

For this paper, we specifically focus on data from one prototype unit – the Increasing Tomato Yield Unit – designed to assess students' ability to conduct controlled experiments and derive relationships from data. The dataset contains a varying number of log events across countries, with an average of 34,275 events (Minimum = 16,536 events, Maximum = 91,462 events, and standard deviation = 12,465 events). Regarding student participation, the number of students ranged from 52 to 239, with a mean of 108 students per country.

We extracted students' interaction logs from the Show phase and the Apply phase to answer questions. We preprocessed them to capture prior knowledge (via pre-test questions) and to analyze the development and application of CoV and DRD strategies in the Challenge task (Apply phase).

3.4 Data Preprocessing

We first transformed the raw interaction log data from the Challenge task into a structured format to facilitate rubric-based scoring. This process involved extracting key events from each student's log file, including (1) experimental trials (i.e., selected values for sunlight intensity, soil type, and the amount of water used); (2) corresponding outcome values (e.g., number of tomatoes); (3) chosen graphs; (4) subsets of experiments selected to justify variable relationships; and (5) the use of the "Check My Work" feature. We then parsed and aggregated these events to identify unique experiments and evaluate whether students systematically varied one variable while keeping others constant, which is a key criterion for scoring the CoV strategy. For scoring DRD, we analyzed students' graph selections and assessed whether the chosen subsets of experiments provided sufficient evidence for determining the variable relationships, and if the chosen relation was correct. This processing was conducted using custom scripts developed in Python. The final dataset included individual student scores for each rubric item, along with metadata such as country and pre-test scores.

3.5 Data Analysis

Following preprocessing and rubric-based scoring, we conducted descriptive and inferential analyses to examine student performance patterns during the Challenge task and identify factors that may influence performance, such as prior knowledge and task engagement.

First, we calculated summary statistics, including means, standard deviations, and score distributions for total performance scores in the Challenge task, individual DRD and CoV components of the scores, and pre-test scores across all students to identify overall trends. Next, to analyze learning transitions, we divided the students into low and high-performance bands using a median split for both pre-test scores and Challenge task scores. This categorization resulted in four transition groups:

- Group-1: LowPre-test → LowTotal-score
- Group-2 HighPre-test → LowTotal-score
- Group-3: LowPre-test → HighTotal-score
- Group-4: HighPre-test → HighTotal-score.

These groups allowed us to explore how students' performance changed from the pre-test to the final task. For instance, we identified how many low pre-test scorers' performances improved in the final task (Group 3) and how many high pre-test scorers' performances declined (Group 2) in the final task. Finally, we investigated the role of potential mediating factors such as students' engagement during the Learn phase and Challenge task. Specifically, we compared metrics such as the amount of time spent in each phase and the use of digital tools (e.g., graph or feedback buttons) across the four groups (Groups 1–4). To ensure robust analysis of timing-related metrics, we applied a 98% winsorization to mitigate the influence of spurious outliers and removed any missing values to maintain data quality.

4. Results and Findings

4.1 Descriptive statistics

Challenge task: Figure 2 (left panel) shows the distributions of students' total scores and their individual DRD and CoV scores (bottom panel) during the Challenge task. In terms of students' overall performance in the task, we found that most students scored between 0 and 5 (Mean = 3.62, SD = 3.5, out of a maximum of 14). Additionally, the right-skewed distribution of the total score suggests that only a small number of students achieved high scores. At the same time, a large proportion demonstrated either partial understanding or incomplete execution of strategies.

When examining DRD and CoV scores separately, we found that most students struggled with the DRD strategy. The average DRD score was relatively low (Mean = 1.57, SD = 2.23, out of a maximum of 8), with many students (50%) scoring 0 points. These findings suggest that students had difficulty recognizing patterns in the experimental table, selecting correct graphs, or determining whether a variable had an effect. In contrast, students performed relatively better on the CoV strategy (Mean = 2.04, SD = 1.59, out of a maximum of 6). While a few students achieved full points for implementing the CoV strategy, a larger proportion scored between 1 and 3 points, demonstrating partial understanding of designing controlled experiments.

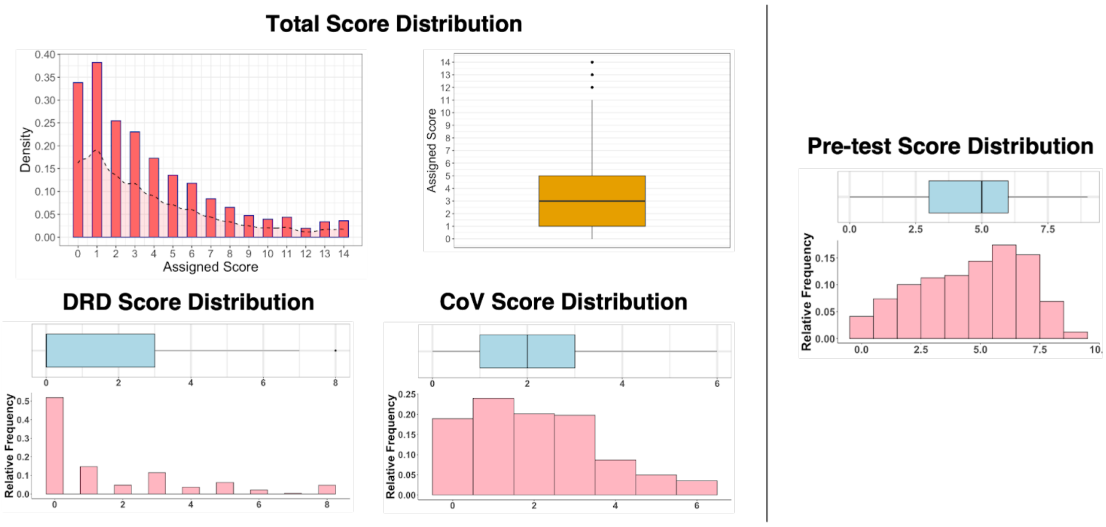


Figure 2. Distribution of total scores for the Challenge task (top-left panel), individual DRD score and CoV score distribution (bottom-left panel), and pre-test scores (right panel)

Pre-test scores: Figure 2 (right panel) displays the distribution of students' pre-test scores. On average, students scored 4.59 points (SD = 2.26) out of a maximum of 9 points, indicating that most approached the Challenge task with some preliminary understanding of CoV and DRD concepts.

4.2 Relationship between prior knowledge and Challenge task performance

Figure 3 presents two complementary views of the relationships among pre-test scores, total score in the Challenge task, and individual CoV and DRD scores. Treating the score as a ranked variable, we conducted a Spearman rank-order correlation to determine the relationship between students' prior knowledge and their total score in the Challenge task.

Correlation Analysis: The correlation matrix (left panel) reveals a strong positive correlation between CoV and DRD scores (ρ = 0.62), indicating that students who effectively designed control experiments were also more likely to accurately identify and represent relationships from the data. We also found moderate correlations between students' pre-test scores and their performance on the Challenge task: CoV (ρ = 0.51), DRD (ρ = 0.47), and total scores (ρ = 0.54). These results suggest that students with greater prior knowledge were generally more successful in applying inquiry strategies during the task. However, the moderate strength of these correlations implies that prior knowledge alone does not fully explain students' success in applying CoV and DRD strategies, reinforcing the importance of process-based assessment.

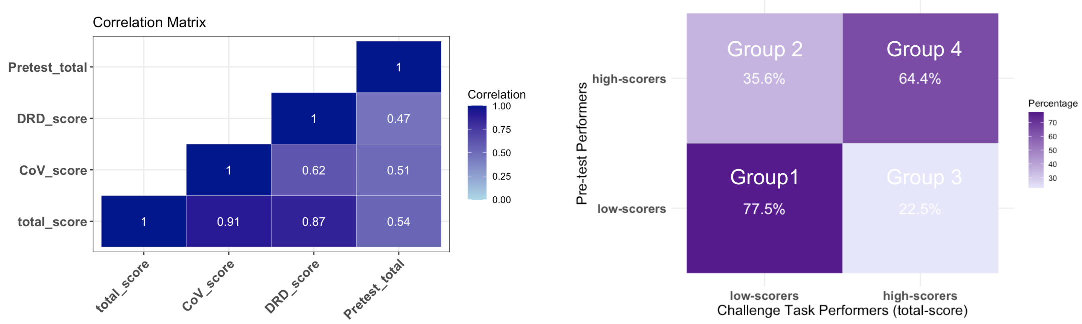


Figure 3. Correlation Matrix (left panel) and Score Distribution Plot across pre-test score and total score (right panel)

Transition groups: The heatmap in Figure 3 shows the distribution of students in four performance transition groups (Groups 1-4 defined in Section 3.5). Results indicate that students with higher prior knowledge were more likely to achieve high scores in the Challenge task, with 64.4% of high pre-test scorers remaining in the high-performance band (Group 4). In contrast, 77.5% of students with low prior knowledge stayed in the low-performance band (Group 1), suggesting limited improvement. Additionally, 22.5% of students with low prior knowledge performed better in the Challenge task (Group 3), while 35.6% of high prior knowledge students showed declining performance (Group 2). These transitions highlight variability in students' performance and raise questions about factors affecting their success or struggle.

4.3 Exploratory Analysis

Table 2 summarizes the CoV and DRD scores, time spent in task phases, and digital tool usage across four transition groups. Kruskal-Wallis tests showed significant differences among groups in score and time metrics (p < .001), confirmed by post-hoc analyses. Chi-squared tests also indicated significant differences in tool usage (Graph-button use: $\chi^2(3)$ = 2507.60, p < .001; Check-My-Work button: $\chi^2(3)$ = 688.09, p < .001). Students who improved (Group-3: Low→High) and those who remained high performers (Group-4: High→High) had higher CoV and DRD scores, spent more time in the Learn and Challenge phases, and utilized digital tools such as the graph and check-my-work buttons more frequently. Meanwhile, students who remained low performers (Group 1) showed the least engagement across all dimensions.

Table 2. *Summary of score-metrics, time-duration metrics, and tool-use metrics for the four transition groups*

	Score-related metrics		Time-related metrics (in seconds)			Tool-related Metrics	
Transition Group	Average CoV	Average DRD Score (SD)	Average Time spent in Show	Average Time spent in	Average Time spent in	% who used	% who used Check-

	Score (SD)		Phase (SD)	Learn Phase (SD)	Challenge task (SD)	Graph Button	My-Work Button
Group-1: Low→Low	0.99 (0.89)	0.18 (0.45)	219 (177.8)	194 (193.0)	67.5 (75.4)	30.1%	28.6%
Group-2: High→Low	1.50 (1.05)	0.29 (0.55)	298 (175.8)	266 (194.8)	91.8 (75.0)	52.3%	37.3%
Group-3: Low→High	2.91 (1.15)	2.79 (1.74)	505 (96.6)	570 (201.0)	223 (109.3)	91.3%	62.3%
Group-4: High→High	3.74 (1.29)	4.10 (2.34)	539 (73.1)	655 (176.3)	271 (109.6)	96.2%	62.5%

5. Discussion, Limitations and Future Work

In this study, we examined how students applied two key scientific inquiry strategies – CoV and DRD – during a digitally administered assessment. Combining rubric-based scoring with log data provided insights not only into student outcomes but also into how engagement and tool use shaped performance.

Students generally performed better on CoV than DRD, aligning with prior evidence that CoV is more teachable and frequently emphasized (Schwichow et al., 2016). DRD remained challenging, with nearly half of students scoring zero, underscoring persistent difficulties in recognizing patterns and casual relationships from data (Donnelly-Hermosillo et al., 2020; Masnick & Klahr, 2003). Importantly, our transition analysis showed that performance was not determined by prior knowledge alone: while many students remained within their initial performance bands, a subset improved or declined significantly, highlighting diverse learning trajectories.

Engagement emerged as a key mediating factor. Students who invested more time in the Learn and Challenge phases generally outperformed their peers, particularly those who remained in the low-performance group. This suggests that scaffolded practice, where students first observe worked examples and then apply strategies independently, supports deeper understanding and transfer of CoV and DRD skills. The use of digital tools also differentiated higher-performing groups: frequent interaction with the graphing feature promoted data interpretation, while the "Check-My-Work" button provided timely formative feedback, enabling students to refine experimental designs. These results align with prior research indicating that structured opportunities for practice, combined with feedback, enhance inquiry processes (Hattie & Timperley, 2007; Schwichow et al., 2016).

Despite these insights, several limitations should be acknowledged. First, our analyses were exploratory and limited to two strategies within a single LDW unit; task-specific effects, therefore, constrain generalizability. Future work should expand to additional inquiry components (e.g., hypothesis generation, error analysis), apply sequential or temporal methods (e.g., Markov models) to capture strategy development, and explore cross-national differences in performance. Second, engagement was proxied primarily by time, which may also include off-task behaviors. Richer interaction measures are needed to distinguish productive form unproductive engagement. Third, while our rubric was grounded in theory, formal validation of its psychometric reliability and transferability across contexts remains an important next step. Lastly, this study did not explore motivational or self-regulated learning (SRL) factors. Future research could investigate how students' motivation, goals, and self-regulated learning (SRL) behaviors contribute to performance differences.

Overall, this study demonstrates how open-ended digital assessments, combined with process-level analysis, can provide rich insights into not only what students learn but also how they engage with and learn core scientific inquiry strategies. These findings offer valuable implications for designing learning environments that more effectively scaffold students' use of CoV and DRD strategies in authentic, inquiry-driven contexts.

Acknowledgements

We thank the OECD team, participating researchers, students, and teachers for their contributions and support in this research. This work was partially supported by funding from the OECD.

References

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

Chu, S. K. W., Lee, C. W. Y., Notari, M., Reynolds, R. B., & Tavares, N. J. (2017). *21st Century Skills Development Through Inquiry-Based Learning: From Theory to Practice* (1st ed. 2017). Springer Singapore : Imprint: Springer. https://doi.org/10.1007/978-981-10-2481-8

Donnelly-Hermosillo, D. F., Gerard, L. F., & Linn, M. C. (2020). Impact of graph technologies in K-12 science and mathematics education. *Computers & Education*, *146*, 103748. https://doi.org/10.1016/j.compedu.2019.103748

Griffin, P. E., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, *40*(9), 898–921. https://doi.org/10.1002/tea.10115

Kuhn, D. (2010). *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (1st ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444325485

Masnick, A. M., & Klahr, D. (2003). Error Matters: An Initial Exploration of Elementary School Children's Understanding of Experimental Error. *Journal of Cognition and Development*, *4*(1), 67–98. https://doi.org/10.1080/15248372.2003.9669683

NRC (Ed.). (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

OECD. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD. https://doi.org/10.1787/9789264281820-en

OECD. (2023). *Innovating Assessments to Measure and Support Complex Skills* (N. Foster & M. Piacentini, Eds.). OECD. https://doi.org/10.1787/e5f3e341-en

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Pedaste, M., & Sarapuu, T. (2006). Developing an effective support system for inquiry learning in a Web-based environment. *Journal of Computer Assisted Learning*, *22*(1), 47–62. https://doi.org/10.1111/j.1365-2729.2006.00159.x

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

Vo, D. V., & Simmie, G. M. (2025). Assessing Scientific Inquiry: A Systematic Literature Review of Tasks, Tools and Techniques. *International Journal of Science and Mathematics Education*, *23*(4), 871–906. https://doi.org/10.1007/s10763-024-10498-8

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001