Student Engagement and Data Science Skills during Authentic Collaborative Inquiry: Insights from Contrasting Group Analysis

Gaoxia Zhu[1], Chin Lee Ker[1], Carol K. K. Chan[2] and Chew Lee Teo[1]
[1]Nanyang Technological University
[2]University of Hong Kong

Abstract: *We have previously developed an authentic, collaborative inquiry model to simulate students' interests using real-world context and data, engage students in data science practice, improve their data science practice through external feedback and finally, trigger them to reflect on the learning experience. This study further delves into how student groups with different data science performances demonstrate data science skills and how they engage in the data science workflow. We analyzed the video recordings of a productive and unproductive group and found that the productive group scoped their investigative questions early, followed and was mostly engaged in the data science flow. In contrast, the unproductive group focused on non-investigative questions and did not realize issues with their inquiry until receiving external feedback, which triggered their revisions and disengagement. This study provides insights into challenges K-12 students could encounter during collaborative data science inquiry and how to support their learning.*

Keywords: Data Science, engagement, collaborative inquiry, K-12, video analysis

1. Introduction

Our society is experiencing big data growth (Vesset et al., 2012), whereby there is a large amount of data around us, highlighting the importance of acquiring data science skills. Hence, developing the future generation's data science skills is critical. Data science skills include generating, collecting, cleaning, managing, analyzing, and communicating data and using the results to inform decisions or provide insights (Overton & Kleinschmit, 2022). Researchers (e.g., Dichev & Dicheva, 2017) emphasized that mastering data science skills will allow us to derive intelligent insights from the upsurge of data in the world. Governments (e.g., The Curriculum Development Council, 2017) have also stressed the importance of performing science inquiry and data science activities, such as conducting investigations, analyzing data, and interpreting data, starting from the primary school level.

Data science components have been integrated into various curriculums. However, most of the literature on data science focuses on introductory courses at the undergraduate level (Thompson & Arastoopour Irgens, 2022). In contrast, K-12 students have limited opportunities to work with data formally in school and they are often not taught how to analyze and interpret contextual, and dynamic real-world data (Finzer, 2013), which may make them see data science as less relevant to them. Therefore, developing authentic data science curricula for secondary students, teaching them data science skills, and engaging them in data science practice is a significant and pressing issue (Heinemann et al., 2018). In a previous study, we (Zhu et al., 2025) developed an authentic, collaborative inquiry model of SPIRE (**S**timulate, **P**ractice, **I**mprove and **Re**flect) and an associated 2-day out-of-school data science program for secondary school students. The program aims to stimulate students' interest by presenting real-world problems, involve them in data science practices and workflows to develop their skills, and promote continuous improvement in data science-supported investigation through feedback, such as input from experts. Additionally, the model emphasizes the importance of reflection, encouraging students to evaluate their learning experiences and address challenges to advance their investigation. Collaboration among students is central to the model and program as it mirrors the way professional data scientists work together to tackle complex problems using data.

Still, acquiring data science skills is a demanding and time-intensive process (English, 2014). In data science education, challenges associated with perceptions of evaluating students, teacher confidence, teacher professional development, the interdisciplinary and complex nature of data science have been reported (Mike, 2020). These challenges suggest the need to dive into students' data science learning process to understand how student groups with different performing levels engage or disengage in the data science learning process to provide implications for better supporting data science teaching and learning. To address the research gap, this exploratory case study examines how students engage in data science work and what skills are deployed. Specifically, we investigated differences between contrasting groups with the research question: How do productive and unproductive student groups engage in the data science workflow, what skills do they demonstrate, and what are their engagement and disengagement levels?

2. Literature Review

Education frameworks and standards increasingly emphasize data science in K–12, yet key concepts are often introduced late and with limited opportunities for meaningful engagement (Drozda et al., 2022; Finzer, 2013). For instance, the Common Core State Standards for High School: Statistics and Probability (2022) highlight interpreting data, drawing justified inferences, and making data-driven decisions. In practice, however, relevant ideas tend to appear only briefly in later high school (Drozda et al., 2022), and data science is usually embedded superficially within mathematics, science, or computer science rather than as a distinct curriculum strand. Most research also focuses on introductory undergraduate courses (Thompson & Arastoopour Irgens, 2022), leaving limited understanding of how secondary students engage in data science processes and what challenges they face. This gap constrains efforts to design and support deeper learning at this level.

Furthermore, developing data science skills and statistical literacy is challenging due to several reasons (English, 2014). The interdisciplinary nature of data science makes it difficult to design a comprehensive course within limited time, while addressing both domain-specific and data ethics issues and accommodating students' diverse backgrounds (Mike, 2020). Teachers often lack sufficient training in incorporating data into their teaching (Merk et al., 2020). Students typically have limited prior knowledge (Heinemann et al., 2018) and can experience a high cognitive load during the implementation of interdisciplinary data science programs.

Specifically, students may experience cognitive, emotional, behavioral, and social engagement and disengagement during data science investigations (Inel-Ekici & Ekici, 2022). Engagement is widely viewed as a multifaceted construct involving passion and dedication, involvement in learning activities, and effort, resilience, and persistence when facing challenges (Fredricks et al., 2004). Synthesizing prior work, Wong et al. (2024) identified three core dimensions of engagement: cognitive, affective (emotional), and behavioral and found that each is positively associated with academic achievement and subjective well-being. In collaborative settings, a fourth dimension, social engagement, is also emphasized, focusing on interpersonal interaction, caring, and building on others' ideas (Fredricks et al., 2016).

Cognitive engagement occurs when learners are committed, thoughtful, and willing to invest effort in understanding complex concepts and developing challenging skills (Fredricks et al., 2004). Emotional engagement refers to learners' positive or negative reactions during learning, which influence their academic achievement and attendance (Wara et al., 2018). Behavioral engagement involves active participation in learning activities, such as sustained effort and attention (Fredricks et al., 2004). Social engagement, meanwhile, encompasses coordinating efforts and building mutual understanding among group members to solve problems, all of which contribute to the co-construction of knowledge and enhanced learning (Fredricks et al., 2016).

In contrast, cognitive disengagement may arise when students struggle to select effective strategies or develop shared understandings of tasks and content (Barron, 2003). Emotional challenges emerge through disruptive affective responses; such as hopelessness or boredom (Ahmed et al., 2013). Behavioral challenges include loss of focus, reduced contribution, off-task conversations, or shifts to unrelated activities, often occurring when group members lose sight of shared goals (Järvelä & Järvenoja, 2011). Social challenges arise when communication breaks down, when attempts to engage peers are not reciprocated, or when disagreements emerge regarding group organization, workflow, or norms (Hussein, 2021). Research examining students' engagement and disengagement throughout the data science learning process remains scarce. This highlights the importance of understanding how student groups with different levels of data science skills engage or disengage in data science practices to better support their learning and guide curriculum development.

3. Methods

3.1. Participants and Context

Participants included 43 secondary students (M_{age}=13.91, 23 female students) from Singapore and Hong Kong joining a two-day data science program in June 2023. Among them, 24 were in secondary 1 (Grade 7), 4 in secondary 2 (Grade 8), 12 in secondary 3 (Grade 9), and 3 in secondary 4/5 (Grade 10). Participants had different data science backgrounds, some having prior data science experience because of attending Science Club and co-curricular activities in school, whereas some having little to no experience interacting with data.

The participants formed 7 groups based on their shared inquiry interests around the "Data Science and Sustainability" theme, with around five to ten members in each group. The program consisted of four key components designed based on the SPIRE model: 1. **Stimulate** students' interest by connecting data science to real-world challenges and demonstrating its relevance through expert talks and videos. 2. Engage students in **practice** data science workflows through hands-on activities, from questioning to data analysis and interpretation. 3. Help students **improve** and refine questions, analyses, and approaches through expert feedback. 4. Support students to **reflect** metacognitively, deepening understanding and recognizing data science's value.

3.2. Data Collection and Analysis

The data sources for this study mainly include (1) the artifacts/ final products (i.e., slides) produced by the seven student groups (2) the video recordings of two groups (productive and unproductive groups) selected based on the data science skills demonstrated in their final products to illuminate different kinds of student engagement and data-science. The video recordings captured the practice and improvement components when students were working in their small groups. The activities were designed around Formulation of Investigative Questions, Data Search, Data Analysis, Consultation, and Refinement.

We applied the coding scheme we previously designed (Zhu et al., 2025) to analyze the data science skills demonstrated in the final products of all student groups. As shown in Table 1, the coding scheme includes ten skills such as scoping questions and explaining rationales. Each skill includes three levels: 0 suggests that a group is not competent in that skill, 1 indicates that the group is moderately competent, whereas 2 indicates that the group is highly competent. Two raters first independently coded the final products using the coding scheme and then compared and discussed their allocated scores for each product until an agreement was reached.

Table 1

Coding scheme to analyze students' data science skills in the final products

Data science skill	Level 0	Level 1	Level 2
Scope question	Scope was wide, objective was absent	Scope was wide, objective was unclear	Scope was narrow, objective was clear
Explain rationale	Explanation was absent	Explanation was partial	Explanation was complete
Identify variables	Identification was absent	Identification was vague	Identification was clear
Data search	Unable to find necessary data	Found data online	Collected own data
Align data with questions	Data is unrelated to any variable	Data is partially related to some variables	Data is completely related to all variables
Appropriate analysis	Analysis was absent	Analysis was not appropriate	Analysis was appropriate

Comprehensive analysis	Analysis was absent	Visualizations were created	Visualizations were created and statistical analyses were performed
Interpret results	No interpretation	Interpretation is partially accurate and relevant	Interpretation is accurate and relevant
Answer the question	Answer was absent	Answer did not relate to analyses conducted	Answer was related to analyses conducted
Consider other factors	Unaware of other factors	Aware that there is one relevant factor	Aware that there are more than one relevant factors

Among the seven groups, Groups 1 and 6 achieved the highest data science score (16 points), followed by Group 5 (13 points), Group 7 (8 points), Group 2 (7 points), and Groups 3 and 4 (6 points). We selected Group 6 to present the productive group and Group 2 to represent the unproductive group. Group 2 was selected over groups 3 and 4 as their data science processes were fully video recorded in the two-day data science program.

We adapted the group collaborative learning analysis method outlined by Järvenoja et al. (2019) to examine the video data of the productive and unproductive groups. First, a researcher divided the recordings into 30-second segments to organize the analysis and preserve a clear chronological sequence of group interactions (Sinha et al., 2015). A brief summary of events was written for each segment. Next, each segment was coded in terms of group-level engagement, mixed engagement, or disengagement. Each interval was also coded in terms of what data science skills were demonstrated, or "Unsure" if it was difficult to tell the specific data science skill being demonstrated (e.g., typing out a note on Knowledge Forum), or "Unrelated" if off-task activities (e.g., turning on the laptop, discussing events not related to the investigation) were observed. If an utterance did not occur, the researcher used other cues, such as behavior data to code whether students were engaged or not.

4. Results

4.1. Data Science Skills Demonstrated in the Final Products

We first describe the final products of the contrasting groups to illustrate different kinds of data-science skills and engagement. As shown in Table 2, the coding results show that the productive group showed high-level data science skills in all the other categories except for explaining the rationales of the investigation and considering other factors beyond those included in their investigative questions. In contrast, the unproductive group showed low to medium levels of data science skills in all the ten categories. The results suggested the need to investigate their data science learning process to understand what might have contributed to the difference.

Table 2

Data science skills demonstrated by the two small groups in the final products

	Scope questions	Explain rationales	Identify variables	Data search	Align data with questions	Appropriate analysis	Comprehensive analysis	Interpret results	Answer the question	Consider other factors
UG	1	1	1	1	1	0	0	1	1	0
PG	2	0	2	2	2	2	2	2	2	0

Note. UG: Unproductive group; PG: Productive group

4.2. Data Science Skills Deployed and Engagement or Disengagement in Video Analysis

Figures 1 and 2 show the engagement, disengagement, and mixed engagement frequencies (x-axis, represented by colors) as well as data science skills (y-axis) across five major activities designed based on the SPIRE model. The productive group was engaged in the data science workflow most of the time except that they had some disengagement in scoping investigative questions and data search in the beginning. In contrast, it was observed that the unproductive group had higher frequencies of disengagement than the productive group, and all the disengagement was observed in the Consultation, and Refinement phases after external feedback was provided by the assigned experts. Furthermore, the unproductive group did not follow the data science workflow as compared to the productive group, and they went back and forth among various data science skills to rationalize and support their investigation in the Consultation and Refinement phases.

As shown in Figure 1, the productive group started the data science process by focusing on Scoping Investigative Questions, Explaining the Rationale of the Investigation, and performing Data Search. They discussed the possibility of studying how AI could be used to reduce pollution and clean up oceans, considering that trash is released into oceans every day. They briefly searched for information online regarding statistics on ocean pollution. Towards the end of the first phase, the group found that their question was not very investigative and too broad, leading to a few instances of disengagement and mixed engagement. They also realized that the data they needed regarding the use of AI in oceans could not be found online. Upon realizing that it was also impossible to invent AI robots to collect ocean trash within the program, they reframed and narrowed the scope of their questions in the second phase. A group member suggested that since there were different AI characters online, they could chat with each character and find out how long it takes for each AI character to forget what they had originally said to them. They revised their investigative questions as “How many messages does it take for an AI to forget what you said?” and “Does the popularity of the AI affect its memory?”. They briefly mentioned how their research could relate to sustainability. The group identified their variables as “number of messages”, “type of AI character” and “popularity of AI character”. Being aware there was no existing data to answer their questions, the group drafted a data collection plan and began collecting data in experimental trials, where each member talked to different AI characters. In phases Data Analysis to Refinement, the group spoke to different AI characters and tested their memory using two different methods—asking the AI a question or changing the plot suddenly. The group was heavily engaged in Data Search and Appropriate Data Analysis but also had a few instances of mixed engagement, for example, when a group member got too excited updating his fellow group mate about his interaction with the AI character, leading to brief distractions in both members; another group member expressed defeat, showed doubts and uncertainty in plotting a graph.

Figure 2 shows the analysis results of the unproductive group. The group started the data science process by scoping their research question concerning the mechanism of a solar-powered electric car, emphasizing the rationale of their investigation, and searching for information online about solar panels. They were largely engaged in the first three phases but did not engage in appropriate data science skills in the Data Analysis stage, when they were expected to do so. Then, during the Consultation phase, the group realized that their question was not investigative after an expert provided feedback. Even though there were attempts to narrow the scope of the investigative questions, the group did not succeed in doing so nor identified existing data consisting of the relevant variables. Still, their finalized question, “How to make electric cars use sustainable energy?” was not solvable using data. During the Refinement phase, they spent a large portion of time creating and filming a skit to emphasize the rationale of their investigation and researched the mechanism of solar-paneled cars, but they had challenges engaging with Appropriate Data Analysis, resulting in a large frequency of disengagement.

Figure 1

Data science skills and engagement or disengagement across the workflow by the productive group

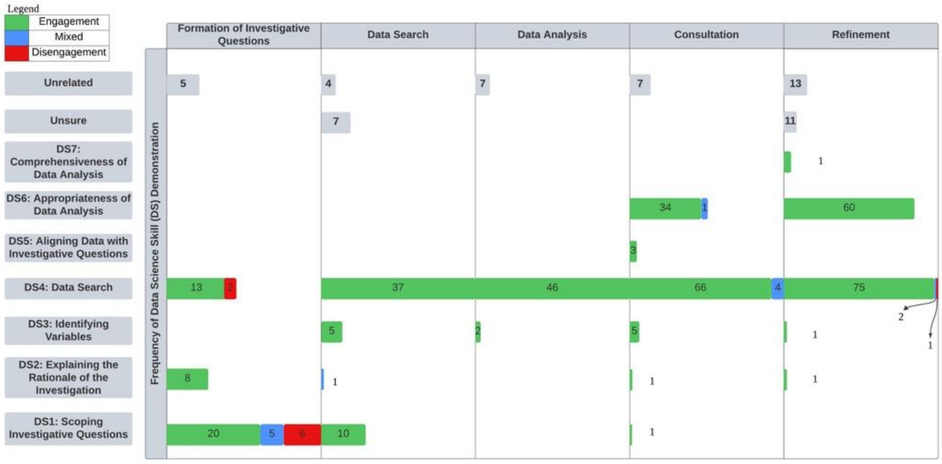

Figure 2

Data science skills and engagement or disengagement across the workflow by the unproductive group

5. Discussion and Conclusion

Considering the importance of cultivating students' data science skills in the current society and the challenges associated with data science education, this study used contrasting group analysis to examine how high- and unproductive groups engaged in data-science workflow and deployed data science skills. The in-depth qualitative analysis and mapping of data science skills and engagement (or disengagement) over the data science workflow uncover what challenges secondary school students can encounter and provide implications for support that should be provided during data science learning.

We found that both the unproductive and productive groups had challenges in narrowing the scope of their investigative questions. Having non-investigative research questions (e.g., too broad, does not direct data analysis) is a typical problem encountered in data science learning (Ratan et el., 2019). A good investigative question should not be too broad nor too narrow and must be researchable using data science (Ratan et el., 2019). With similar challenges in the beginning, however, the two groups eventually performed differently. The following reasons could have explained the difference in performance. Firstly, the productive group had the patience to navigate through the challenge, brainstormed and discussed their other ideas, and changed their robot-design question to investigative questions that could be answered using data and were more feasible to work on within two days. In contrast, the unproductive group was very excited and kickstarted the information search without pausing to reflect on whether their question was investigative, even though reflection was emphasized in the SPIRE mode (Zhu et al., 2025) and in our program design. The group searched for information related to the mechanisms of an electric car, as well as the current statistics of electric car users. Secondly, the productive group's practice is more aligned with the data science workflow, which provides their work structure and sequences and ensures a smoother and more comprehensive investigation (Gupta, 2022). The unproductive group, on the other hand, spent too much time trying to explain the rationale of using electric cars but did not engage in appropriate data analysis until late phases, and they had a high frequency of disengagement.

The analysis of the unproductive group also shows the importance of providing external feedback to help students identify issues and limitations with their data science investigation, which may not be explicitly observable to students themselves. Similarly, it has been found that it is typically challenging for secondary school students to source and search for information online due to their limited skills in adopting appropriate keywords and identifying relevant information (Wu et al., 2014). Monitoring and feedback from instructors or capable others should be provided earlier to keep students on track of the desired data science workflow before it is too late.

This study used two contrasting groups and conducted learning products and video analyses to provide in-depth analysis and insights. However, there are some limitations. Firstly, our sample size is small, making it difficult to generalize the findings. In the future, we plan to extend the analysis to other groups to examine whether similar findings can be found and if new patterns will emerge. Secondly, this study was conducted in a two-day out-of-school setting. The extent to which the findings can be generalized to school contexts needs further investigation.

References

Ahmed, W., van der Werf, G., Kuyper, H., & Minnaert, A. (2013). Emotions, self-regulated learning, and achievement in mathematics: A growth curve analysis. *Journal of Educational Psychology, 105*(1), 150–161.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–359.

Common Core Standards. (2022). "Statistics & probability." Statistics & Probability | Common Core State Standards Initiative. http://www.corestandards.org/Math/Content/SP/

The Curriculum Development Council (2017). *Science education key learning area curriculum guide (Primary 1 - secondary 6).* https://www.edb.gov.hk/attachment/en/curriculum-development/kla/science-edu/SEKLACG_ENG_2017.pdf

Dichev, C., & Dicheva, D. (2017). Towards data science literacy. *Procedia Computer Science, 108*, 2151-2160.

Drozda, Z., Johnstone, D., & Van Horne, B. (2022). *Previewing the national landscape of K-12 data science implementation*. National Academies of Science, Engineering, and Medicine.

English, L. D. (2014). Promoting statistical literacy through data modelling in the early school years. In Chernoff, E.J. & Sriraman, B. (Eds.), *Probabilistic thinking: Presenting plural perspectives*. Advances in Mathematics Education (pp. 441–457). Springer.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2), 1-9.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research.*, *74*(1), 59–109.

Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction, 43*, 1-4.

Gupta, S. (2022, May 16). Data Science Process: A Beginner's Guide in Plain English. Springboard. https://www.springboard.com/blog/data-science/data-science-process/

Heinemann, B., Opel, S., Budde, L., Schulte, C., Frischemeier, D., Biehler, R., … & Wassong, T. (2018, November). Drafting a data science curriculum for secondary schools. In *Proceedings of the 18th Koli calling international conference on computing education research* (pp. 1-5).

Hussein, B. (2021). Addressing collaboration challenges in project-based learning: The student's perspective. *Education Sciences*, *11*(8), 434-.

Inel-Ekici, D., & Ekici, M. (2022). Mobile inquiry and inquiry-based science learning in higher education: advantages, challenges, and attitudes. *Asia Pacific Education Review*, *23*(3), 427–444.

Järvelä, S., & Järvenoja, H. (2011). Socially constructed self-regulated learning and motivation regulation in collaborative learning groups. *Teachers College Record*, *113* (2), 350– 374.

Järvenoja, H., Näykki, P., & Törmänen, T. (2019). Emotional regulation in collaborative learning: When do higher education students activate group level regulation in the face of challenges? *Studies in Higher Education*, *44*(10), 1747-1757.

Mike, K. (2020, August). Data science education: Curriculum and pedagogy. In *Proceedings of the 2020 ACM conference on international computing education research* (pp. 324-325). ACM.

Merk, S., Poindl, S., Wurster, S., & Bohl, T. (2020). Fostering aspects of pre-service teachers' data literacy: Results of a randomized controlled trial. *Teaching and Teacher Education*, *91*, 103043-.

Overton, M., & Kleinschmit, S. (2022). Data science literacy: Toward a philosophy of accessible and adaptable data science skill development in public administration programs. *Teaching Public Administration, 40*(3), 354-365.

Ratan, S., Anand, T., & Ratan, J. (2019). Formulation of research question - Stepwise approach. *Journal of Indian Association of Pediatric Surgeons, 24*(1), 15–20.

Sinha, S., Rogat, T. K., Adams-Wiggins, K. R., & Hmelo-Silver, C. E. (2015). Collaborative group engagement in a computer-supported inquiry learning environment. *International Journal of Computer-Supported Collaborative Learning, 10*, 273-307.

Wara, E., Aloka, P. J., & Odongo, B. C. (2018). Relationship between emotional engagement and academic achievement among Kenyan secondary school students. *Academic Journal of Interdisciplinary Studies, 7*(1), 107-118.

Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of Educational Psychology, 116*(1), 48–75.

Thompson, J., & Arastoopour Irgens, G. (2022). Data detectives: A data science program for middle grade learners. *Journal of Statistics and Data Science Education, 30*(1), 29-38.

Vesset, D., Woo, B., Morris, H. D., Villars, R. L., Little, G., Bozman, J. S., ... & Yezhkova, N. (2012). Worldwide big data technology and services 2012–2015 forecast. *IDC report*, 233485.

Wu, C. H., Hwang, G. J., & Kuo, F. R. (2014). Collab-Analyzer: An environment for conducting web-based collaborative learning activities and analyzing students' information-searching behaviors. *Australasian Journal of Educational Technology, 30*(3). 356–374.

Zhu, G., Teo, C. L., Ong, A. K. K., Yuan, K. G., Ker, C. L., & Yang, Y. (2025). Towards an authentic collaborative inquiry model for cultivating data science skills and attitudes: effects of SPIRE on secondary school students. *Education and Information Technologies*, *30*(14), 19933-19959.