DATA SCIENCE EDUCATION K-12

INDIVIDUAL PAPER: Using Entropy Analysis to Explore Student Engagement in an Online High School Data Science Course

Barnas Monteith, Center for Science Engagement ,barnas@engagescience.org
Liu Zifeng, University of Florida, liuzifeng@ufl.edu
Jie Chao, Concord Consortium, jchao@concord.org
Kenia Wiedemann, Concord Consortium, kwiedemann@concord.org
Janet Bih Fofang, University of Maryland, bihjanetshufor@gmail.com
Linlin Li, WestEd, lli@wested.org
Dexiu Ma, Texas Tech University, dema@ttu.edu
Rabab Mohamed, Texas Tech University, rabab.mohamed@ttu.edu
Anupom Mondol, Texas Tech University, a.mondol@ttu.edu
Yelee Jo, WestEd, yjo@wested.org
April Fleetwood, Florida Virtual School, afleetwood@flvs.net
Lodi Lipien, Florida Virtual School, llipien@flvs.net
Yuanlin Zhang, Texas Tech University, y.zhang@ttu.edu
Wanli Xing, University of Florida,wanli.xing@coe.ufl.edu

Abstract

Data science is revolutionizing academia and industry, creating a high demand for a workforce fluent in this field. While the availability of data science courses has increased recently, few curricula rigorously build on mathematical logic. The LogicDS Project addresses this gap by engaging high school students from rural communities in an online data science course integrating mathematics, statistics, and programming into a unified framework based on logic and reasoning. A one-week course, consisting of six lessons, was developed and 110 participants were recruited. Pre- and post-intervention data, along with students' LMS activity logs, were collected to analyze engagement. Results indicate that the Logic-Based framework effectively engages students from diverse backgrounds, with participants finding the course valuable for learning data science skills. Notably, entropy analysis of student activity logs correlated with other mixed methods analyses, providing insights into engaging K-12 students in data science education.

Objective of the Study

K-12 data science education research has increased over the past decade (Du et al., 2022; Mobasher et al., 2019). However, the interdisciplinary nature of data science presents challenges in developing high school courses. Current efforts often focus on summer camps or generalized frameworks with specific data science elements (e.g., Grover et al., 2015; Weintrop et al., 2016). Few studies integrate discrete math and programming logic into curricula, covering classical theorems and proofs (e.g., Bouhnik & Giat, 2009). LogicDS aims to engage high school students by integrating mathematics, statistics, and programming, enhancing their problem-solving skills and data science learning outcomes. Based on this approach, we aim to answer two research questions: (1) What are student engagement levels when learning an online data science curriculum? (2) Are there differences in engagement levels among students from different backgrounds?

Method

Design and Development of LogicDS

Most data science programs available to high school students are traditionally delivered in person or in informal settings. This delivery method limits systematic learning opportunities in data science, especially for rural communities and during periods when in-person meetings are challenging. We developed LogicDS, to offer a logic-based unified integration of the interdisciplinary foundations of data science under the data investigation cycles framework (Bargagliotti, 2020). The course aims to enhance students' understanding and proficiency in data science, encompassing foundational concepts in mathematics, statistics, and computing.

Participants

110 self-selected students, motivated to learn more about data science, signed up to participate in the course and research study. Table 1 shows the demographic information of the participants. This study received IRB approval (IRB #202400397) from *Anonymized*.

Procedure and Measurements

A study was conducted with 93 participating students using a custom-developed open source LMS (Learning Management System) called LARA/AP. Over one week, from April 22 to April 29, 2024, students completed six lessons, in an experimental intervention known as “Week of Data Science”. Students filled out pre-surveys and post-surveys to provide demographic information, prior experience, and motivation in learning data science (shown in Table 2), and their engagement and perceived value of the curriculum (shown in Table 3).

Data Collection and Analysis

In addition to pre- and post- surveys from students, we collected comprehensive timestamped LMS activity log data. The log data (227,075 rows) includes 25 learning events (e.g., mouse-tracking events) for every student. Entropy analysis was utilized to analyze this data to analyze student engagement during the LogicDS lessons. This method measures complex events sequences to ascertain meaningful correlations from an otherwise high volume of chaotic data (Mai et al., 2023). In the context of analyzing student engagement in learning activities, entropy analysis can be utilized to aid in the categorization of discrete student interactions within the learning platform.

Results

Survey Results

Table 4 shows part of the post-survey results. In the pre-survey, three students reported that they knew a lot about data science, while ten students had never heard of it in the past three years. Students described data science as "the study of using mathematical and programming methods to extract and use data" or "science in technology and statistics." After completing the curriculum, 30% of students (28 out of 93) responded that they knew a lot about data science. Additionally, 75% of students agreed or strongly agreed that they enjoyed the data science lessons.

Engagement Analysis

Figure 1 shows the entropy analysis of students’ interactions across all six lessons in the LogicDS curriculum. The baseline entropy level score for the first lesson was 2.36; lessons 2 through 5 (core lessons) showed a higher entropy score than the first introductory lesson.

A one-way ANOVA was conducted to compare mean entropy scores among different groups, as shown in Table 5. The analysis revealed no significant differences based on gender, ethnicity, or locale. However, middle school students had significantly higher entropy scores than high school students ($F = 4.30$, $p = 0.04$).

Discussion and Implications

Initially, the pre-survey indicated that only three students reporting extensive knowledge of data science. Post-survey results demonstrated a significant increase in students' self-reported knowledge of data science, with 30% of students indicating a high level of understanding, potentially suggesting that these students were engaged in the course material. Furthermore, 75% of the students enjoyed the lessons, indicating that the curriculum had a positive impact on them (Zhang et al., 2024; Grover et al., 2015; Weintrop et al., 2016).

The analysis of student interactions revealed higher entropy scores across all five main lessons (lesson 2 to lesson 6) compared to the introductory lesson (lesson 1), with average entropy values exceeding 2.36. These high entropy scores indicate frequent and diverse interactions with the course material, potentially reflecting engagement in the lessons. Higher entropy may also result from discrete curriculum design issues and UI/usability factors, such as navigating to different portions of the materials to resolve confusion or confirm answers.

The one-way ANOVA results showed no significant differences in entropy based on gender, ethnicity, or location, suggesting that the students from different demographic groups exhibited similar patterns of interactions with the learning platform. However, middle school students exhibited significantly higher entropy scores compared to high school students (F=4.30, p=0.04), possibly because younger students had greater interaction with the

learning material or needed more time/help. The LogicDS curriculum was shown to actively immerse students in interactive learning of data science core concepts, with positive overall impacts.

Acknowledgments
The Learning Management System (LMS) used for this research, LARA/AP(Activity Player) is an open-source product of Concord Consortium; Concord, MA. This material is based upon work supported by the National Science Foundation under Grant No.2201393.

Figures and tables

Table 1 Student Participants Demographic

Category	**Subcategory**	**Number of Students**	**Percentage**
Total Students		93	100%
Age	11-14 years old	22	23.70%
	15-18 years old	69	74.20%
	19 years old or older	2	2.20%
Grade	Grades 6 to 8	8	8.60%
	Grades 9 to 11	77	82.80%
	Grade 12	8	8.60%
Gender	Female	41	44.10%
	Male	47	50.50%
	Not Responded	5	5.40%
Ethnicity	Hispanic or Latinx	25	27%
	White	41	44.10%
	African American or Black	19	20.40%
	Asian	21	22.60%
	Native American	1	1.10%
	Not Responded	6	6.50%
Location	Small cities, towns, or rural areas	33	35.50%

Table 2 Pre-survey Questionnaire

Construct	**Question No.**	**Question**
Demographics	Q1	How old are you?
	Q2	What grade are you in?
	Q3	What is your gender?
	Q4	Are you Hispanic or Latinx?
	Q5	Which of the following best describes you? Select one or more answer choices.
	Q6	Which of the following best describes the community you live in?
	Q7	What language do you speak at home most of the time?
	Q8	How many digital devices with screens are there in your home? (Count all the devices, including televisions, computers, tablets, e-book readers, smartphones, etc.)

Educational Background and Expectations	Q9	What math classes have you taken so far (including those you are taking now)? Select one or more answer choices.
	Q10	What computing classes have you taken so far (including those you are taking now)? Select one or more answer choices.
	Q11	Which of the following do you expect to complete? (Please select all that apply)
	Q12	What job(s) do you want to have in the future?
Experience	Q13	Are you familiar with Data Science?
	Q14	If you have heard of Data Science in the past 3 years, from where? Select all that apply.
	Q15	In your own words, describe Data Science. If you have never heard of it, just guess what it means.
Motivation	Q16	What motivated you to sign up for these Data Science lessons?

Table 3 Post-survey Questionnaire

Construct	**Question No.**	**Question**
Engagement	Q1	The lessons grabbed my attention.
	Q2	I couldn't focus on the lessons.
	Q3	I enjoyed the lessons.
	Q4	I didn't like the lessons.
	Q5	I chose to spend extra time on the lessons.
	Q6	I did only what I was told to do and nothing extra.
	Q7	Because of the lessons, I started to think a lot about data science.
	Q8	I wasn't thinking about the content very much during these lessons.
	Q9	I really enjoyed the learning I did in these lessons.
	Q10	What I learned in these lessons is fascinating.
	Q11	What I learned in these lessons is boring.
	Q12	I looked forward to taking the lessons each day.
Perceived Value	Q13	I am sure I will use this knowledge again.
	Q14	There is no point in learning all of this.
	Q15	I could relate what I learned from the lessons to real life.
	Q16	The things I studied in these lessons are important to me.

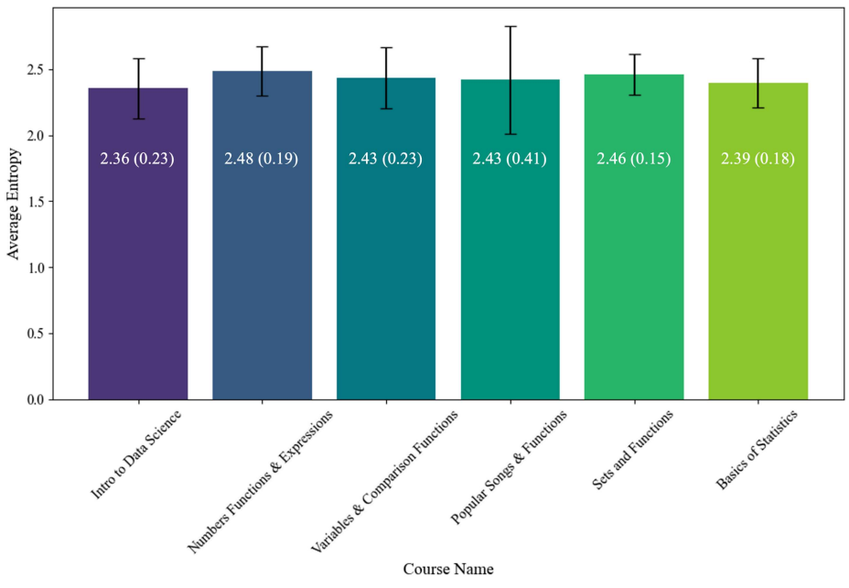
Note: All questions are on a five-point Likert scale.

Table 4 Post-survey Results

Construct	**Question No.**	**Response**	**Frequency**	**Percentage (%)**
Engagement	Q1	Quite a bit like me	16	17.78
		Very much like me	5	5.56
	Q2	A little bit like me	17	18.89
		Not at all like me	12	13.33
	Q3	Quite a bit like me	51	56.67
		Very much like me	9	10
	Q4	Not at all like me	52	57.78
		A little bit like me	18	20

	Q5	Quite a bit like me	14	15.56
		Very much like me	7	7.78
	Q6	A little bit like me	50	55.56
		Not at all like me	4	4.44
	Q7	Quite a bit like me	50	55.56
		Very much like me	13	14.44
	Q8	Not at all like me	58	64.44
		A little bit like me	20	22.22
	Q9	Agree	57	63.33
		Strongly agree	18	20
	Q10	Agree	62	68.89
		Strongly agree	13	14.44
	Q11	Disagree	54	60
	Q12	Strongly agree	54	60
		Agree	22	24.44
Perceived Value	Q13	Strongly agree	49	54.44
		Agree	28	31.11
	Q14	Strongly disagree	38	42.22
		Disagree	29	32.22
	Q15	Agree	60	66.67
		Strongly agree	12	13.33
	Q16	Agree	60	66.67
		Strongly agree	18	20

Note: Only part of the optional results is presented here.



Note: The numbers on the bar chart (e.g., 2.36 (0.23)) represent the average entropy score and its corresponding standard deviation.

Figure 1 Lesson Average Entropy

Table 5 Average Entropy Score of Different Groups

Background Attribute	**Groups (n)**	**Average Entropy Score (SD)**	**One-way ANOVA**
Grade level	Middle school level (n=7)	2.65 (0.23)	$F = 4.30$, $p = 0.04$**
	High school level (n=82)	2.54 (0.13)	
Gender	Female (n=38)	2.54 (0.16)	$F = 0.69$, $p = 0.50$
	Male (n=46)	2.57 (0.12)	
Hispanic	Hispanic (n=21)	2.60 (0.18)	$F = 2.07$, $p = 0.13$
	Non-Hispanic (n=60)	2.53 (0.11)	
Ethnicity	White & Asian (n=48)	2.56 (0.12)	$F = 0.33$, $p = 0.57$
	Others (n=41)	2.54 (0.16)	
Location	Rural (n=17)	2.55 (0.11)	$F = 0.10$, $p = 0.91$
	Suburban (n=64)	2.55 (0.14)	
	Urban (n=8)	2.57 (0.15)	

Note: Only those students who reported their background were included.

References

Bouhnik, D., & Giat, Y. (2009). Teaching high school students applied logical reasoning. *Journal of Information Technology Education*. Innovations in Practice, 8, 1.

Du, H., Xing, W., Pei, B., Zeng, Y., Lu, J., & Zhang, Y. (2022, April). A Descriptive and Historical Review of STEM+ C Research: A Bibliometric Study. In *International Conference on Computer Supported Education* (pp. 1-25). Cham: Springer Nature Switzerland.

Grover, S., Pea, R., & Cooper, S. (2015). Designing for deeper learning in a blended computer science course for middle school students. *Computer science education, 25*(2), 199-237.

Mai, T. T., Crane, M., & Bezbradica, M. (2023). Students' Learning Behaviour in Programming Education Analysis: Insights from Entropy and Community Detection. *Entropy (Basel, Switzerland)*, *25*(8), 1225. https://doi.org/10.3390/e25081225

Mobasher, B., Dettori, L., Raicu, D., Settimi, R., Sonboli, N., & Stettler, M. (2019). Data science summer academy for chicago public school students. *ACM SIGKDD Explorations Newsletter, 21*(1), 49-52.

Starnes, D. S., & Tabor, J. (2018). The practice of statistics. New York: WH Freeman.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of science education and technology, 25*, 127-147.

Zhang, Y., Du, H., & Xing, W. (2024). A new approach to high school data science: Set theory and logic. In *Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024* (pp. 2101-2102). International Society of the Learning Sciences.