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

Barnas Monteith[1], Zifeng Liu[1], Jie Chao[1], Kenia Wiedemann[1], Janet Bih Fofang[1], Linlin Li[1], Dexiu Ma[1], Rabab Mohamed[1], Anupom Mondol [1], Yelee Jo[1], April Fleetwood[1], Lodi Lipien[1], Yuanlin Zhang[1], and Wanli Xing[1]

[1]Affiliation not available

August 30, 2024

Abstract

Data science is revolutionizing academia and industry, and there is a strong demand for a workforce fluent in data science. The availability of such courses has increased substantially in recent years. However, there are few rigorous curricula built on mathematical logic as the foundation of data science. The LogicDS Project aims to engage high school students from rural communities in an online data science course that integrates mathematics, statistics, and programming concepts into a single unified framework based on logic and reasoning. We developed a one-week data science course consisting of six lessons (a sampling of the full course offering) and recruited 110 participants. We collected pre- and post-intervention data and students' LMS activity log data to analyze their engagement. Results indicate that our Logic-Based framework for data science education effectively engages students from a variety of backgrounds; it was shown that they perceived the course as valuable for learning data science skills/concepts. Notably, our focus on the entropy analysis of student activity logs correlated to our other mixed methods analyses. This study provided insights for engaging K-12 students learning data science.

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

[Barnas] [Monteith]*, Center for Science Engagement ,barnas@engagescience.org
[Zifeng] [Liu]*, 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, and there is a strong demand for a workforce fluent in data science. The availability of such courses has increased substantially in recent years. However, there are few rigorous curricula built on mathematical logic as the foundation of data science. The LogicDS Project aims to engage high school students from rural communities in an online data science course that integrates mathematics, statistics, and programming concepts into a single unified framework based on logic and reasoning. We developed a one-week data science course consisting of six lessons (a sampling of the full course offering) and recruited 110 participants. We collected pre- and post-intervention data and students' LMS activity log data to analyze their engagement. Results indicate that our Logic-Based framework for data science education effectively engages students from a variety of backgrounds; it was shown that they perceived the course as valuable for learning data science skills/concepts. Notably, our focus on the entropy analysis of student activity logs correlated to our other mixed methods analyses. This study provided insights for engaging K-12 students learning data science.

Objective of the Study
K−12 data science education research has gained increased attention in the past decade (Du et al., 2022; Mobasher et al., 2019). However, the interdisciplinary nature of data science has presented challenges in developing high school courses. Most current efforts focus on summer camps, after-school programs, or generalized frameworks that include specific data science elements (e.g., Grover et al., 2015; Mobasher et al., 2019; Weintrop et al., 2016). A few studies have introduced discrete math and programming logic into high school curricula, covering topics like classical theorems, properties, and proofs (e.g., Bouhnik & Giat, 2009). However, these opportunities are still inaccessible to most high school students, particularly those from rural communities, due to the number of disciplines involved and limited work in data science curriculum. To address these challenges, we developed an six-lesson foundational data science course named LogicDS. At the core of this course is a novel, natural yet deep integration of the interdisciplinary foundations of data science in math, computing, and statistics using mathematical logic which is a well-studied area in mathematics and computer science. The goal of LogicDS is to engage high school students in learning data science by integrating mathematics, statistics, and programming—the core components of data science—to help students learn and develop problem-solving skills and benefit their data science learning outcomes.

This study details the development and preliminary results of a high school data science foundation course based on this approach, addressing the following research questions: (1) What are student engagement levels when learning an online data science curriculum? And (2) Are there differences in engagement levels among students from different backgrounds (e.g., gender, grade level, location)? We focus on engagement because this new experimental course approach is relatively untested in the field, and as we did not require/enforce any prerequisite courses or subject knowledge, we do not yet have evidence as to whether or not the material may be too challenging for students. Inversely, entropy analysis may allow us to identify learning difficulties/challenges, in addition to engagement. The authors intend to further explore this in subsequent analyses.

* These authors contributed equally to this work.

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). Specifically, we propose an innovative approach to integrating the foundations of data science into LogicDS using mathematical logic. Mathematical logic provides a language for us to develop precise definitions of statistical concepts to explicate the logic and reasoning underlying these concepts, and thus promote deeper sense making and reasoning by using real-life datasets. UF and Texas Tech University will collaborate on the study. 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. Overall, 77 students were from high school, and 33 students (30%) come from underrepresented regions (i.e., small cities, and rural areas). This study received IRB approval (UF IRB #**202400397**) to conduct mixed methods research with all students, and feedback interviews with teachers.

Procedure and Measurements

We conducted a study with 110 students using a custom-developed open source LMS (Learning Management System) called LARA (Lightweight Activities Runtime & Authoring; it is also presently known as AP, or Activity Player, GitHub link: https://github.com/concord-consortium/lara). 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 (pre-survey, shown in Table 2), and their engagement and perceived value of the curriculum (post-survey, shown in Table 3).

Data Collection and Analysis

In addition to pre- and post- surveys from 93 students, we collected comprehensive timestamped activity log data throughout. The log data (227,075 rows of data) includes 25 learning events (e.g., mouse-tracking events) for every student, during their course interactions. Entropy analysis was utilized to analyze this data to analyze students' 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. The use and interpretation of entropy are relatively novel in the edtech field. It is not fully known how to accurately normalize the entropy and whether this data can infer a higher or lower level of engagement, or course difficulty level. Additional data normalization processes and alternate interpretations must be considered; the authors will explore this further in future works.

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, 30.11% 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. Each lesson has an average entropy value above 2.36, which is the entropy score of the introduction lesson. This indicates that the main content of the curriculum lesson 2 to lesson 5 have higher entropy scores. This means students frequently interacted with various parts of the lessons, reflecting active engagement and curiosity during the learning.

We examined the engagement of students from diverse backgrounds in learning LogicDS using one-way ANOVA to compare the mean entropy scores among different groups. Table 5 presents these results. The analysis shows that students from different demographic groups exhibited similar patterns of interactions with the learning platform. with no significant differences based on gender, ethnicity, or locale However, the entropy scores for middle school students were significantly higher than those for high school students ($F = 4.30$, $p = 0.04$), suggesting greater engagement among younger students.

Discussion and Implications

The survey results reveal significant insights into students' knowledge and perceptions of data science before and after participating in the LogicDS curriculum. Initially, the pre-intervention survey indicated a small number of students possessed significant understanding of data science, with three students reporting extensive knowledge and ten students having no prior exposure. Post-survey results demonstrated a significant increase in students' self-reported knowledge of data science, with 43% of students indicating a high level of understanding, potentially suggesting that the notions that these students were engaged in the course material. Furthermore, 75% of the students enjoyed the lessons, suggesting that the curriculum had a positive impact on participants (Zhang et al., 2024; Grover et al., 2015; Weintrop et al., 2016). The entropy analysis of student interactions revealed high engagement across all six lessons, with an average entropy value exceeding 2.30. This high entropy score indicates frequent and diverse interactions with the course material, possibly reflecting active engagement and curiosity. Higher entropy may also be caused by discrete curriculum design issues and UI/usability factors (for instance, students must navigate different portions of the curricular materials to resolve confusion, check answers, etc.). The one-way ANOVA results showed no significant differences in entropy based on gender, ethnicity, or location, suggesting that the curriculum was equally effective across diverse student groups. However, middle school students exhibited significantly higher entropy scores compared to high school students ($F=4.30$, $p=0.04$), indicating that younger students have greater interaction with the content. Overall, the LogicDS curriculum was shown to be effective in enhancing students' knowledge, engagement, and interest in data science.

In conclusion, our study demonstrates the potential of a well-structured data science curriculum to engage and educate high school students. By addressing the interdisciplinary nature of data science and incorporating practical, real-world applications, the LogicDS curriculum offers a promising model for future educational programs aimed at preparing students for careers in data science and related fields.

Acknowledgments
The Learning Management System (LMS) used for this research, LARA/AP(Activity Player) is an open-source product of Concord Consortium; Concord, MA.

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.

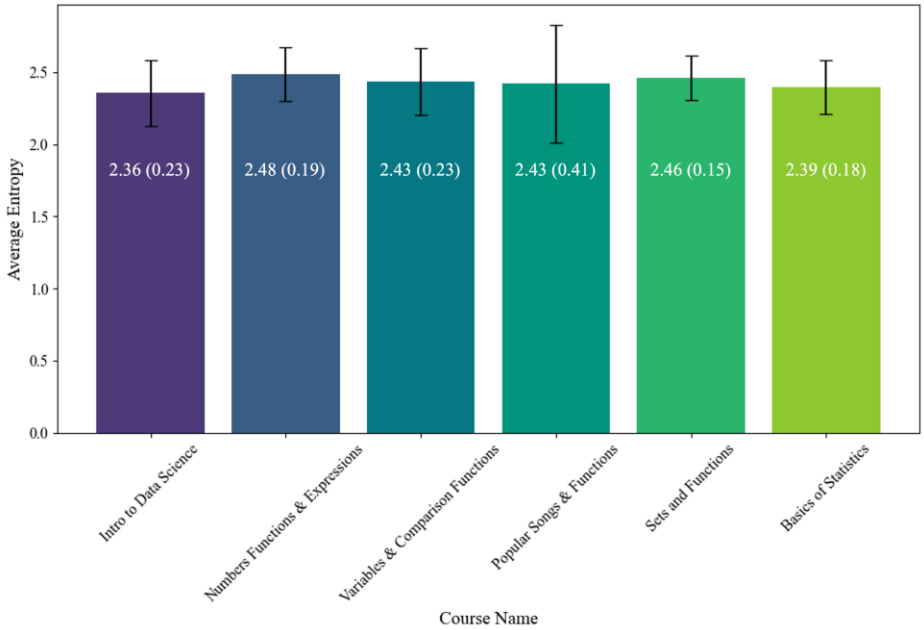
Note: All questions are on a five-point Likert scale.

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.