International Society of the Learning Sciences

Navigating the Conceptual and Critical Demands of Data Literacy in High School Spaces using Firsthand Data

Noora F. Noushad, Jooeun Shim, Susan A. Yoon
noora@gse.upenn.edu, jsim@upenn.edu, yoonsa@upenn.edu
University of Pennsylvania

Abstract: Curriculum efforts to promote data literacy continue to evolve in K-12 spaces to meet the larger need of developing data-informed citizens. The learning objectives of these efforts tend to shift from conceptual data literacy objectives such as demonstration of adequate statistical reasoning to critical data literacy objectives such as empowering students to design, collect and analyze data to serve their needs. However, most studies do not delineate between these objectives, often giving an incomplete picture of student understanding of data, and the effectiveness of the curriculum in meeting these two data literacy objectives. In this paper, we study the outcomes that emerged from a curriculum unit designed to empower students to collect, analyze and present data from their neighborhood. We specifically evaluate the conceptual and critical data objectives of 14 high school students. Our findings indicate that students struggled to demonstrate equal competence in conceptual and critical data literacy goals.

Introduction

Among the many needs to make educational content relevant for application in the knowledge economy, there is a growing effort to develop teaching and learning practices that promote data literacy in K-12 classrooms (Frank et al., 2016; Gebre, 2018; Wilkerson & Polman, 2020). In the current knowledge economy, big data is ubiquitous, with corporations and governments using data from individual's everyday behaviors (e.g., shopping habits, social media interactions, and voting preferences) —a development that is only going to increase in coming years (Borges-Rey, 2016; Deahl, 2014; Pangrazio & Selwyn, 2019). However, individuals continue to be limited in their ability to understand the nature of data and often perceive it as an objective statistical measure, rather than understanding how data is collected, manipulated, and used, often without their permission (Borges-Rey, 2016; Pangrazio & Selwyn, 2019). This effort to develop a data-informed citizenry has ignited heightened interest in educational research to curate curricula that enable students to actively engage, reason, and critique activities and products involving data (Wilkerson & Polman, 2020; Wolff et al., 2016; Lee et al., 2021). This is evident in emerging standards and coordinated curricular efforts to promote "data science for everyone" (Lee et al., 2021; Ridgway, 2016; Wilkerson & Polman, 2020).

However, current curricular efforts often fail to capture the conceptual data objectives (e.g., ability to draw accurate inferences, interpret central tendencies, read graphs) and critical data perspectives (e.g., ability to engage in data production process, critique inherent bias in data) needed to develop data informed citizenry - often simplifying the objectives or measuring one over the other, as opposed to both (Irgens et al., 2020; Philip et al., 2018; Lee et al., 2021). Consider for example, the use of mobile sensors and firsthand data. Most data literacy curricula that use sensors and firsthand data value activities that involve learners posing questions for data analysis, designing experiments and visiting sites for data collection (Hardy et al., 2020; Manz, 2012). These studies measure students' ability to engage meaningfully in data literacy practices such as sampling and measurement, developing a sense of agency and ownership over data, and drawing inferences based on their knowledge of the context, which data literacy objectives that are largely critical in nature (Hardy et al., 2020; Stornaiuolo et al., 2020). Conversely, less emphasis is provided in these studies to evaluate students' ability to statistically or quantitatively reason with data which includes traditional statistical skills such as reading a graph, identifying variability, boundaries of drawing inference (Lee & Wilkerson, 2018; Rubin, 2020; Irgens et al., 2020). Importantly, there exists conflicting evidence on the effectiveness and the value addition of using firsthand data. Studies have shown that student's proximity with data sets interferes with their ability to draw accurate inferences, often discounting the variability they observe during the data collection process (Hug & McNiell, 2008; Konold, 2015; Lee & Wilkerson, 2018). There are very limited studies that have evaluated both conceptual objectives and critical data literacy perspectives in K-12 interventions (Irgens et al., 2020; Philip et al., 2018; Pfannkuch et al, 2018).

To meet the larger goal of developing data-informed citizens, it is imperative that studies on data literacy evaluate both conceptual and critical data objectives (Irgens et al., 2020). If both objectives are not attended to with equal emphasis, there is a risk of developing student knowledge and skills of statistics alone without the ability to critically assess the purposes of data use that can preserve individual agency and well-being. Likewise, without a core understanding of how to access or read data products, being able to make accurate critiques of data applications will be sacrificed (Irgens et al., 2020; Lee et al., 20201; Philip et al., 2018). This study aims to address the above concerns. We first present the design of a curriculum and instruction intervention on the topic of bioinformatics that has one of its central aims as promoting students' conceptual

and critical competencies in data literacy using firsthand data. We then evaluate the conceptual and critical data objectives of 14 high school students who engaged with the curriculum. We analyze students' conceptual and critical understanding of data literacy using Rubin's (2020) data literacy framework and Hardy et al.'s (2020) data production framework asking the following research questions 1) In what ways did students demonstrate an accurate understanding of conceptual objective and critical data perspectives? 2) What constructs of conceptual objectives and critical perspectives did they struggle to demonstrate accurately?

Theoretical framework

Learning scientists have defined data literacy in multiple ways, often with the primary objectives fluctuating between the conceptual and critical focus (Hardy et al., 2020; Kjelvik & Schultheis, 2020; Rubin, 2020). For example, Rubin (2020) conceptualizes data literacy as the ability to work with statistical elements of data that include computing aggregates, accounting for variability, and drawing accurate inferences among others. Notice that the primary objective of data literacy with the conceptual focus here is on quantitative reasoning which is the ability to engage with numerical information found in graphs, equations, and descriptive statistics to critically engage with data collected from real-world phenomena. On the other hand, Hardy et al. (2020) define data literacy as the ability for students to understand data as produced, which is through the interaction between human intent, disciplinary tools such as sensors, and the limitations of the context within which data is collected. Here, a critical focus of data literacy is on student's ability to engage with data production practices, i.e., to be able to make active decisions in the design, collection, analysis and communication of data with the broader intent of turning data into knowledge that benefits the student or their larger community. While there is overlap across both definitions on the ability to read data critically and use authentic data sets, they differ in the granular focus of their primary objectives. These objectives, if taken separately, would reflect different learning goals when translated into classroom settings. (Irgens et al., 2020; Lee et al., 2021; Wolff et al., 2016).

Defining conceptual and critical data objectives

While multiple frameworks that speak to conceptual needs of data literacy exist (Kjelvik & Schultheis, 2020; Rubin, 2020), in this paper, we use Rubin (2020)'s five parameters to study the conceptual understanding of students' data literacy because of the framework's focus on the contextual factors that shape and influence the nature of data. The framework includes five critical objectives on working with data, which can be briefly explained as follows: 1) *context,* the "who, where, what, why and how" of data collection that shows its purpose and surrounding circumstances; 2) *variability,* the nature of data that varies in value over time and space; 3) *aggregate*, summarizing data as a set of values and understanding what it does and doesn't reflect about its distribution; 4) *visualization,* visual displays of data, deciding how the data is displayed and understanding what choices are made in the process; and 5) *inference*, drawing conclusions based only on the data without using outside assumptions. These are the constructs we measure as a parameter of conceptual understanding of data literacy in this study.

There exist three interconnected views of critical data literacy: 1) data literacy as comprehension; 2) data literacy as critique; and 3) data literacy as participation (Irgens et al., 2020). The first two views conceptualize critical understanding of data as the ability to not only comprehend the variability and human bias within the data production process but also critique how data is inherently politically, socially, and racially constructed. In this paper, we focus on the third perspective of critical data literacy which is data literacy as participation. The primary focus here is to actively engage with data production practices, which include making active decisions about the design, collection, analysis, and communication of data with the broader intent of turning data into knowledge that benefits the student or their larger community (Hardy et al., 2020; Irgens et al., 2020). Hardy et al. 's (2020) data production framework provides a lens to study student's engagement as data agents by analyzing three core data perspectives; a) *material:* the ability to understand that data is produced as a result of negotiations between instruments such as sensors and the material world, b) *disciplinary:* an understanding of the disciplinary nature of tools and inquiry practices, and c) *human*: an understanding that data products cannot be delineated from human intentions that drive the production and design of data. In this paper, critical data literacy is measured by analyzing the degree to which students developed these perspectives. Combining these two lenses, which speak to both conceptual and critical data objectives, helps us measure student outcomes that speak to both their statistical reasoning abilities and their ability to critically engage with data production practices.

Methodology

Context

In this section, we provide a brief overview of the NSF-funded project within which this smaller study was housed. The larger project focused on developing a curriculum and instruction for teaching bioinformatics in high school biology classrooms. The curriculum spanned approximately 16 lessons, or about 20 hours of instruction, and guided students in

problem-based learning (PBL) investigation on the topic of air quality and asthma in urban environments. For this smaller study, we focused on the data literacy components of the broader bioinformatics unit. We mapped the data literacy unit activities onto the instructional recommendations of Rubin (2020) and Hardy et al. (2020) to ensure the conceptual and critical data objectives were being met (Table 1). These activities were introduced across three phases intended to build student familiarity with 1) data tools, 2) data collection and analysis, and 3) real-world data production. Students collected air quality measures (carbon monoxide (CO) and particulate matter (PM)) with portable sensors and phones in different neighborhood locations. Students then analyzed the class dataset to compute averages and compared their findings to Environmental Protection Agency (EPA) data that was stored on a project-constructed website. This website allowed students to make visualizations of their own data and view visualizations of air quality data around the U.S. for comparison. Based on their data analyses students presented solutions to address the disparities that they saw in their urban areas as compared to other urban, suburban, or rural areas.

Table 1

Curricular activities that align with recommendations from Rubin (2020) and Hardy (2020)

Curricular Recommendations	Curricular Activities
Variability: Use data sources that are likely to vary both in space and time (R).	- Students collect air quality data from four different locations or from some location at four different times of the day and discuss how the measured data changes over time/location.
Aggregate: Engage in discussions about the intent of computing aggregates in statistics and reflect on what aggregate values communicate and what it doesn't (R).	- Reflect on the rationale of computing mean and median. - Debate if mean or median is a better representative of a sample and why? - Conduct a t-test to compare averages between two locations.
Visualization: Engage in self-expression and revealing patterns (R).	- Visualize data using bar graphs. - Discuss alternate visualizations such as infographics and pie charts.
Inference: Discuss what we have learned applies beyond this particular dataset? (R).	- Teacher engages in-class discussion on how generalizable the data collected from their specific data sample is. - Students report findings, their analysis, and possible recommendations.
Material: Provide students with the opportunity to manipulate data collection and analysis tools (H).	- Follow procedures to operate the mobile sensor for data collection and accumulate data on Google Sheets. - Practice collecting data using mobile sensors in and outside school. - Analyze the data accumulated from the sensors and compute the mean of all readings collected using data analysis tools.
Humanistic: Provide students choices in setting purposes for data production (H).	- Select the problem they want to explore using data on the issue of asthma. - Select a variety of neighborhood areas to investigate the issue of asthma.
Disciplinary: Provide students with the opportunity to design data collection and analysis in a goal-directed manner reflective of most science-inquiry processes (H).	- Set research questions for the data investigation process and present rationale for expert (teacher) approval. - Conduct rehearsals of data collection planning for contingencies. - Revise the data collection plan based on the trials with the sensors. - Students report findings, their analysis, and possible recommendations.

Participants

For this study, we report the findings from one biology teacher, Sam, who had 17 years of teaching experience and taught at a public high school in the northeastern U.S. He was nominated as a desirable teacher by the director of science in the school district to select an ideal population (IES & NSF, 2013) and designated a model teacher. This study reports on data collected from 14 students from his class. There were 11 11th graders and 3 12th graders of which, 10 were female and 4 were male students, who self-reported to be 42.86% Asian or Pacific Islander, 35.71% White, 7.14% for African American, Hispanic, and others respectively. Most of the students did not have prior experience working with data.

Data Sources and analysis

We collected and analyzed data from three sources: 1) students' final project reports, 2) student focus group interviews, and 3) classroom observation notes. *Students' final project reports* included detailed write-ups on the data inquiry process. Students were provided with a project template and a rubric to include their group's research questions, their description of the data collection sites, the inferences they made about the data, and the rationale behind their reasoning. They worked in groups to complete the project. There were five project reports in total. *Student's focus group interviews* included student reflections captured from two student groups. The interviews probed insights into the data collection and design process where students discussed the problems they encountered while using the sensors and Google Sheets. Finally, the *classroom*

observation notes captured detailed descriptions of the classroom interaction and teachers' instructional practices, activities taking place including materials distributed to students, and conversations that occurred during the data collection process.

Students' final projects and interview transcripts were used as the primary data source for our analysis, while observation notes provided additional evidence to triangulate our claims. To address the research questions, we reviewed students' final projects and interview transcripts to identify parts that illustrated an understanding of data. These segments of data were then extracted to a spreadsheet and analyzed for instances that demonstrated critical and conceptual data objectives using an internally developed coding manual. The coding manual included eight categories, five reflective of conceptual and three reflective of critical data literacy. The definitions which were derived from our two theoretical frameworks. After reviewing the segments of data, we adapted the coding manual to better capture students' understanding of data as demonstrated in the curriculum. There were 182 segments of data that emerged from this process. The interrater reliability test from the larger study was conducted by two raters with 38 of the 182 segments (21% of the data) yielding an alpha score of .85 which is an acceptable level of agreement.

From these segments, for this paper, we extracted 41 segments from Sam's students and categorized the segments of data across the eight categories of data literacy using the coding manual. In addition, the segments of data were further examined by first and second authors to determine if they were reflective of an accurate understanding of the category it was coded for. For example, the following segment of data "This place is very important to this project because every day this chemical plant would let out pollution, 386,000 tons of hazardous air pollutants into the atmosphere each year." was coded as evidence of an accurate understanding of 'context' because students' choice of data sites reflected intentionality and alignment with their larger research question. We coded segments of data as inaccurate understanding of the 'context' if it did not include detailed information of data collection site, description of the sites is random information that does not align with the research question. Table 2 shows accurate and inaccurate categorization manuals for critical data literacy.

Table 2
Categorization manual for conceptual and critical data literacy (Category and Definition)

Accurate	Inaccurate
Context: When student descriptions of data collection sites reflect intentionality and alignment with the research question for their projects.	**Context:** Student description of data collection sites are random information that does not align with the research question.
Variability: Student reflections refer to interferences caused by context, sensor etc. to the data collection process.	**Variability:** Student does not report any fluctuations experienced during the data collection process
Aggregate: When making references to central tendencies students compare results to a universal standard in order to draw conclusions. They refer to and explain outliers.	**Aggregate:** Reports only indicate the mean and median values pooled from different data collection sites.
Inferences: When drawing conclusions, students refrain from generalizing their findings beyond the scope of data and refrain from using "always" or "never." The reports mention the variability they observed during the data collection.	**Inferences:** The student reports draw conclusions without accounting for the variability observed during the data collection process.
Material: Students discuss the troubleshooting strategies employed during the data collection process.	**Material:** Students only refer to issues faced during the process of data collection or analysis.
Humanistic: Students refer to the factors that pertain to community or personally relevant factors that motivated them to pursue certain data questions and sites over others.	**Humanistic:** Students make no reference to community or personally relevant factors that motivated the design of the data production process. The description appears generic.
Disciplinary: By referring to the research questions, parameters of data collection site etc. These are often instances laid out during the planning stage.	**Disciplinary:** There is no reference to the steps that went into planning for the data collection or analysis.

Findings

Across the five final projects, the segments of data that were coded for accurate critical data perspective were slightly more than those coded for accurate conceptual data objectives. Out of 41 segments of data coded as accurate, 40.74% were assigned to conceptual objectives while 59.26% reflected critical data objectives (Table 3). The findings are organized to report accurate and inaccurate demonstrations of conceptual and critical data.

Table 3
Summary of data coded as accurate or inaccurate

	Con* # (%)	Var* # (%)	Agg* # (%)	Vis* # (%)	Inf* # (%)	Total # (%)	Mat** # (%)	Dis** # (%)	Hum** # (%)	Total # (%)

Accurate	7 (63.64)	0 (0)	3 (27.27)	1 (9.09)	0 (0)	11 (100)	1 (5.88)	10 (58.82)	6 (35.29)	17 (100)
Inaccurate	0 (0)	3 (23.08)	2 (15.38)	4 (30.77)	4 (30.77)	13 (100)	0 (0)	0 (0)	0 (0)	0 (0)

*Conceptual perspectives: Con = Context, Var = Variability, Agg = Aggregate, Vis = Visualization;
**Critical objectives: Inf = Inference, Mat = Material, Dis = Disciplinary, Hum = Human nature of data

Students' accurate and inaccurate demonstration of conceptual data objectives

Out of the total 11 accurate references made to conceptual objectives, 68.29% references were made to context, and 31.71% were made to aggregate. This is indicative of the fact that students understood the context-dependent nature of data - choosing to be intentional in aligning the selection of the data sites to the research questions they wanted to answer and exhibited confidence in being able to curate, analyze and present data that serve their personal interests and community needs.

Among the five conceptual data literacy concepts, the most frequently and accurately referenced construct was the concept of 'context'. These included rich descriptions of data collection sites communicating the intentional selection of these sites. These descriptions indicated that students understood the context-dependent nature of data, often communicating in the final report important details that shaped their data collection and analysis process. For example, in the excerpt below, students describe the three data sites from where data was collected, providing details of the site that could influence the research question they were investigating which was whether the air quality is better in areas that were less crowded and had more greenery such as the example below,

> Using the sensor and the app on our phone we collected data at five locations around [School]: 1) 17th and [Blue] St.: There were no cars passing by [...], 2) Community College Underpass: There were a few smokers nearby [...] There was also a bus docking and lots of cars passing by, 3) Outside Dunkin Donuts: [...] but there were a lot of cars passing by [...]

In the above excerpt, students signal to context-related information such as the number of people, the number of cars, and the quality of greenery at the sites which were present during the process of data collection. This reveals that students understood the need to capture and provide as much information present across data sites that may be relevant to the nature of inquiry the data collection process was being conducted for.

The second category of conceptual data literacy that appeared more frequently in the student's final data reports was the 'aggregate' code. An 'aggregate' view of data signals a conceptual understanding where students are able to view aggregate values as numbers that summarize trends as opposed to individual data values that are an exact representation of the population (Rubin, 2020). 'aggregate' code was assigned to segments of data where students made references to outliers in their data while reporting the mean scores of the data collected from various sites or when students compared their mean score of Air Quality Index (AQI) to the standards mentioned in the National AQI table while interpreting the data. For example, in the excerpt below, students reference how they interpreted the average calculated across three data sites while referencing the standards of the National AQI table and accounting for an outlier.

> Our PM 2.5 data, in comparison to the AQI, is good, with average levels of 11[...] Our CO data, in comparison to the Air Quality Index, is very poor, at an average of about 18. However, if you leave out one of the data points that is very high in comparison to the others, the CO level average is just 12.84, which falls into the poor section.

In the above excerpt, students not only point to an outlier in the data set but also detail how the omission of the datapoint could change the inference drawn from the data set of the AQI being average to poor. This indicates that students have a nuanced understanding of central tendencies that go beyond just the ability to calculate the mean or median.

The conceptual constructs of 'variability,' 'visualization,' and 'inference' were not referenced accurately in the student's final data reports. Out of 13 inaccurate references, 23.08% was referenced to 'variability,' while 30.77% was referenced to 'visualization' and 'inference' respectively. Errors included the absence of visual representation of data, overgeneralization of inferences beyond the data collected by students, and the omission of variance in the reports. Despite the use of mobile sensors in the curriculum, the segments of data that referred to the 'variability' and 'inference' construct accurately were 0. The low scores indicate that students struggled to critically account for the fluctuations that arose during the data collection process caused by sensors, and the data collection sites in their findings or analysis. For example, when reflecting on the data collection process, a group in their focus group interview referred to how the malfunctioning mobile air sensors posed a problem,

> We were nervous at times that we didn't actually get some data points with the sensors. The Bluetooth connection sometimes would go on and off. So, when we weren't sure, we ended up leaving some sites and collected 7 data points from some sites to make up for the sites we couldn't collect data from.

However, in the same group's final report the students did not mention how the interference caused by the erroneous functioning of sensors or the omission of data points from some sites and the addition of data sources from other sites, influenced the conclusion they drew about the air quality near food trucks. Their final report stated the following finding conclusively, "Our data shows that all of the food distributors have good PM 2.5 levels, but the CO levels are not so good [...] Using the averages of CO and PM 2.5 levels, it is shown that halal trucks have the best air quality." This excerpt indicates that while students experienced material resistance during the process of data collection, they did not account for how these interferences shaped the conclusions they drew about the air quality near food trucks. In their final report, the group did not reference the omission of data sites. This is indicative of the fact that students found it challenging to account for the interferences that may have arisen from the variance in context or sensors. In the final inferences, they made about the data students chose to draw general inferences about the data instead of using a language of uncertainty to communicate the findings or state the limitations of their data collection process. It may also be the case that students may not have had the conceptual understanding of data to understand that the omission of specific data sites cannot be compensated with additional data collected from other sites.

Students accurate and inaccurate demonstration of critical data objectives

Out of the 17 total references made to critical data perspectives, 35.29% was coded as 'human perspective,' 58.82% was given to 'disciplinary,' while only 5.88% was given to 'material perspective.' This is indicative of the fact that students understood the inherently personal nature of data production process and also understood the inquiry driven nature of data production. However, they struggled with troubleshooting the issues that came up during the process of data collection.

Out of the total 17 references made to 'critical perspective, 35.29% was given the 'human perspective' code. This is indicative of students' confidence in being able to curate, analyze and present data that serve their needs. The presence of 'human perspective' indicated that students understood that one can collect, use and analyze data for purposes that may vary from being personally relevant to self or to others. The segments of data coded for 'human perspective' included students' reflections on the personal relevance of the data production process. These reflections varied from broader impacts of community such as clean air, more green space to more personally relevant issues such as quality of air in places that students frequented after school or investigating the correlation between unconventional smoking mediums such as hookah bars and vaping to educate their peers on harmful effects of smoking. For example, a group from Sam's class mentioned their rationale to design their data production process to study the quality of air in a hookah bar as follows.

> But people (peers) think vaping is not as harmful as smoking, and hookah too. They just think smoking is more harmful, but in reality, they are all kind of related to each other. They have the same effect [...] So, we were just thinking of something that would be kind of unique but also have the same idea. So, we came up with hookah […] Going to a hookah lounge is important for our experiment as this place is relevant to our community because it has become a popular hangout spot for our friends.

In this excerpt, students chose to investigate research questions that helped answer a question that was relevant to people in their immediate community. The group chose to investigate a question that would help improve a problem (i.e., increase awareness about the harmful effects of Hookah). As opposed to the narrow belief that data is produced primarily as evidence in evaluating claims, the human perspective promotes the idea that goals and purposes of data production can vary from being conceptual, playful, personally relevant to self or others (Hardy et al., 2020). These segments of data, therefore, demonstrate students' understanding of a human perspective or the variant nature of goals of data produced for humans. We know the more students personalize the goals for producing data, the higher the likelihood that students adopt the mindsets of being able to produce and critically engage with data (Stornaiuolo et al., 2020; Wilkerson & Polman, 2020).

Another code that appeared more frequently in the student reports and interviews were 'disciplinary perspectives' (58.82%). These codes included references that indicated a systematically designed goal-directed inquiry. For example, references made to research questions, the number of trials planned for data collection and use of mobile sensors. The disciplinary codes largely captured the degree to which students systematically planned their data production process prior to stepping out in the field. For instance, in the excerpt below students' detail methods and procedures they engaged in to plan their data collection process.

> How does a chemical plant affect air pollution in the environment surrounding it? Chemical plant Honeywell in Philadelphia, Procedure: First we will test the air in three different locations (School A, B and Chemical Plant Honeywell in [City]). Next, we will record the data (CO, temperature, PM 2.5, and the Humidity) for all locations. The data will be recorded three times for each location for about 30 seconds. Later, we will take all the data and compare the results for all three locations. This experiment will be performed on October 15, 2019.

Here, the group intentionally selected three data collection sites; one site that was closest to a chemical plant and two school sites that were located at a safe distance from chemical plants, to specifically investigate how the chemical plants shape the quality of air in their neighborhood. This is indicative of the fact that students understood the goal-directed nature of the data production process, choosing to be intentional in aligning the selection of the data sites to the question they wanted to answer. They also refer to the number of times and the duration of data collection at each site.

The critical perspective code that appeared the least in the student reports were 'material perspectives.' These included instances from student interviews and reports where students discussed ways in which they troubleshooted the issues they faced during the data collection process. For instance, in the below excerpt,

> Student A: I stayed in the same place and when we were back there, the door was open, but I asked her if she could close it for the experiment, and I did it every five minutes to give it some time. [...] There was no one that was coming in at the time, because I didn't want the air to get out, to affect it. So, I did it when everyone was seated, and people were smoking hookah.

In the above excerpt students refer to the issue of changing conditions at their data collection site and how they troubleshooted this variability by trying to time the collection of air quality data only when the doors were closed. These instances, however, were low. While there were references made to issues of malfunctioning sensors or contextual variance in the student interviews, there were less references made to how these issues were troubleshooted. Also, these variances reported in the student interviews were not referenced during the reporting of the data which explained the low score on accurate 'inferences' discussed earlier in the conceptual data section. The low scores on material perspective indicates that while students were able to note aspects of material setting (e.g., sensors and details of data collecting sites) that would influence the data production process they weren't able to troubleshoot or account for the fluctuations they noted during the data collection process when making inferences about the data.

Discussion

In this study, we examined how conceptual and critical data literacy objectives manifested in a curriculum that used firsthand data to promote both. Our findings indicate that if we want to broaden data literacy from its traditional focus on quantitative reasoning and statistical literacy to include critical reasoning with data production practices, we need to ensure that the rigor of conceptual understanding is also measured. Despite the use of firsthand data students struggled to demonstrate equal competence in conceptual and critical data literacy goals, with most students demonstrating an understanding of critical perspectives of data literacy over conceptual goals.

Our findings on students' conceptual outcomes replicate observations from earlier studies that have shown that student's proximity with data sets interferes with their ability to draw accurate inferences, often discounting the variability they observe during the data collection process (Hug & McNeill, 2008; Knold, 2015; Lee & Wilkerson, 2018). Our findings on critical outcomes also align with previous studies have shown that giving students the agency to decide the nature of data production and investigation is more likely to promote an understanding of data as produced (Hardy, 2020; Stornaiuolo et al., 2020). This variance in student understanding of conceptual and critical elements of data literacy is problematic as the primary goal of critical data literacy is to actively engage with data production practices, i.e., to empower students to make active decisions in the design, collection, analysis, and communication of data with the broader intent of turning data into knowledge that benefits the student or their larger community (Hardy et al., 2020; Irgens et al., 2020). However, in order for students to be empowered as critical agents of data, they need to demonstrate an accurate understanding of the conceptual aspects of data. The lack of competence in conceptual data literacy concepts is more likely to lead to a critical understanding that is not conceptually sound (Jiang & Khan, 2019). The inaccurate understanding of conceptual concepts such as 'variability' and 'inference' is more likely to lead to overgeneralization of data and the drawing of inaccurate inferences (Rubin, 2020), a trend we observed in our findings. Despite the inaccurate conceptualization of 'variability,' in our study, the student groups scored high on critical perspectives aspects of data. This means that students demonstrated the intention to produce data for themselves or for others while being unaware of their misconception of core conceptual data literacy principles. As demonstrated in the result section, while students accurately made note of the variables in the data collection sites during the data collection process, most groups did not make the connection between the variance they noticed during the data collection process and the inferences they drew about the data instead choosing to draw conclusive inferences. Therefore, the efficiency of enactments of critical data literacy is called into question due to the incomplete or inaccurate conceptual understanding of the student groups. Our study adds value to the current debates about teaching data literacy from either a conceptual or critical perspective. We suggest that in order for students to become completely data literate, curriculum and instruction should focus on knowledge and skills that combine both perspectives.

References

Borges-Rey, E. (2016). Unravelling data journalism: A study of data journalism practice in British newsrooms. *Journalism Practice*, *10*(7), 833-843.
Deahl, E. (2014). Better the data you know: Developing youth data literacy in schools and informal learning environments. *MIT media press,* 9–115.
Frank, M., Walker, J., Attard, J., & Tygel, A. (2016). Data Literacy - What is it and how can we make it happen? *Journal of Community Informatics*, *12*(3), 4–8.
Gebre, E. H. (2018). Young adults' understanding and use of data: Insights for fostering secondary school students' data literacy. *Canadian Journal of Science, Mathematics and Technology Education*, *18*(4), 330–341.
Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L., & Zanontian, L. (2016). Teaching data science to students. *Proceedings of the roundtable conference of the International Association of Statistics Education (IASE), Berlin, Germany, 6-11 May 2016.*
Hardy, L., Dixon, C., & Hsi, S. (2020). From data collectors to data producers: Shifting students' relationship to data. *Journal of the Learning Sciences*, *29*(1), 104-126.
Hug, B., & McNeill, K. L. (2008). Use of first-hand and second-hand data in science: Does data type influence classroom conversations?. *International Journal of Science Education*, *30*(13), 1725–1751.
IES & NSF. (2013). *Common guidelines for education research and development*. Washington, DC: Authors.
Irgens, G. A., Simon, K., Wise, A., Philip, T., Olivares, M. C., Van Wart, S., Vakil, S., Marshall, J., Parikh, T. S., Lopez, M. L., & Wilkerson, M. (2020). Data literacies and social Justice: Exploring critical data literacies through sociocultural perspectives. *Proceedings of the International Society of Learning,* 406–411.
Jiang, S., & Kahn, J. B. (2019). *Data Wrangling Practices and Process in Modeling Family Migration Narratives with Big Data Visualization Technologies.* Proceedings of Computer Supported Collaborative Learning Conference.
Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy. *CBE—Life Sciences Education*, *18*(2), es2.
Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics, 88*(3), 305-325.
Lee, V. R., & Wilkerson, M. H. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects.* Commissioned Paper for the National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12. Washington, D.C.
Lee, V. R., Wilkerson, M. H., & Lanouette, K. (2021). A call for a humanistic stance toward K–12 data science education. *Educational Researcher*, *50*(9), 664-672.
Manz, E. (2015). Representing student argumentation as functionally emergent from scientific activity. *Review of Educational Research*, *85*(4), 553-590.
Pangrazio, L., & Selwyn, N. (2019). Personal data literacies: A critical literacies approach to enhancing understanding of personal digital data. *New Media & Society*, *21*(2), 419–437.
Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM*, *50*(7), 1113-1123.
Philip, T. M., Gupta, A., Elby, A., & Turpen, C. (2018). Why ideology matters for learning: A case of ideological convergence in an engineering ethics classroom discussion on drone warfare. *Journal of the Learning Sciences*, 27(2), 183–223.
Ridgway, J., Arnold, P., Moy, W., & Ridgway, R. (2016, July). Deriving heuristics from political speeches for understanding statistics about society. In *Promoting understanding of statistics about society. Proceedings of the IASE Roundtable Conference*.
Rubin, A. (2020). Learning to reason with data: How did we get here and what do we know? *Journal of the Learning Sciences*, *29*(1), 154–164.
Stornaiuolo, A. (2020). Authoring data stories in a media makerspace: Adolescents developing critical data literacies. *Journal of the Learning Sciences*, *29*(1), 81–103.
Wilkerson, M. H., & Polman, J. L. (2020). Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences*, *29*(1), 1–10.
Wolff, A., Gooch, D., Montaner, J. J. C., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, *12*(3).

Acknowledgement

This work was funded by the U.S NSF DRK-12 (DRL#1812738)