Harvard Data Science Review • Issue 6.3, Summer 2024

Data Science Learning in Grades K–12: Synthesizing Research Across Divides

Joshua Rosenberg[1] **Ryan Seth Jones**[2]

[1]University of Tennessee Knoxville, Knoxville, Tennessee, United States of America,
[2]Middle Tennessee State University, Murfreesboro, Tennessee, United States of America

Published on: Jun 06, 2024

DOI: https://doi.org/10.1162/99608f92.b1233596

ABSTRACT

What do we know about data science learning at the grades K–12 (precollegiate) level? This article answers this question by using the notion of agency to provide a framework to review the diverse research agendas and learning environments relevant to data science education. Examining research on data science education published in three recent special issues, we highlight key findings from scholars working in different communities using this lens. Then, we present the results of a co-citation coupling analysis for articles published in one of three recent data science education special issues with research spanning various levels and contexts. This co-citation analysis showed that while there are some common touchpoints, research on data science learning is taking place in a siloed manner. Based on our review of the literature through the lens of agency and our analysis, we discuss how the data science education community can synthesize research across disciplinary and grade-level divides.

Keywords: data science education, grades K–12, literature review, co-citation coupling

---

What Is Known About Grades K–12 Data Science Learning? A Review Using Agency as a Lens

Vignette A: Antonio's Data Science Club

Antonio is excited because he is taking a new class at his middle school called ‘Data Science.’ This class is offered for 6 weeks, and students rotate through it, so he’s heard his friends talk about it and the ‘hackathon’ they have at the end of the class. He has also heard about people working in jobs as a ‘data scientist,’ although he doesn’t know anyone personally in these jobs. On the first day of class, the teacher told students, “In this class, we are going to solve big problems with big data.” Antonio was not very sure what the teacher meant by ‘big data,’ but it sounded exciting!

On the first day of class, the teacher asked each student to log into one of the lab computers and open a software called RStudio. The teacher told the class that RStudio is a tool for analyzing data and finding important stories in the data. The teacher also told them that this is a tool that data scientists all over the world use. Antonio was excited to explore data like a real data scientist!

Antonio and his class opened a data set about recycling in Los Angeles and followed the instructions the teacher gave to create a graph of the weight of recycled paper in 2014 and 2019. They also followed the instructions to create a map with dots for different recycling centers. The code that the teacher gave them to do these looked so cool to Antonio! He understood how to change parts of the code but didn’t understand what most of it meant. However, his friend understood most of it and helped him at times. Sometimes, the code or the data didn’t work the way it was supposed to, and Antonio wasn’t sure why. He could get frustrated sometimes, but the teacher said, “data science is messy!”

By the end of the 6 weeks, the students had used different codes in RStudio to explore the recycling data and gave presentations about what they had learned. The students in the class presented many different things from the 6 weeks, and Antonio found some of it interesting but had a hard time concentrating when other students presented. He was happy, though, because the teacher was very excited about the presentations and said he was now a data scientist!

Vignette B: Trulia’s Science Investigations

Trulia's first day in science class was different than any class she's ever experienced. Her teacher attended a professional development over the summer called "Data Science in the Science Class" and told the students on the first day of class that they were going to be "real scientists" in the class. Each Friday, the entire class conducted research on the ecosystem around the school. The students spent the first day of class talking about the different things they already knew about their schoolyard. Trulia shared that she enjoys the benches and the walking track but that sometimes students drop trash around the track, which upsets her. She was unsure how the conversation relates to science, but she enjoyed hearing what other students think about the schoolyard.

Over the course of the fall, the students spent each Friday walking around the schoolyard and making notes on the different plants and animals they noticed. They talked as a whole class about what they were noticing and thinking. In each of these conversations, the teacher ends by asking "What are you wondering about our schoolyard?" As this list grew, Trulia noticed that she and a few other students were wondering where people drop trash most often in their schoolyard and how they might reduce trash in the schoolyard ecosystem.

One Friday late in the fall, Trulia and a few friends developed a plan for collecting data to answer the question, "Does the car drop-off area or the playground area have more trash?" Trulia and her friends decided to count the trash in each location and spent an entire Friday collecting their data. They were surprised by how hard it was to count the trash, especially with the wind blowing it around to different areas of the schoolyard! They all counted in different ways and used different size areas. Because of this, it was very hard to compare the data!

In the spring semester, the teacher used a few Fridays to talk about ideas related to sampling and measurement, and Trulia's friends decided to all sample and measure in the same way in both areas. They collected data using these procedures every Friday for the next few months and regularly talked about what they thought they were learning. It was still hard to measure each time, but now they all agreed they could compare measurements across different weeks and locations, and they were surprised to find that the drop-off area had much more trash than the playground! By the end of the semester, the students had used their data and findings to suggest changes to improve their schoolyard ecosystem.

1. Introduction

Data science has rapidly grown over the past decade as a career path and program of study at the undergraduate level (National Academies of Sciences, Engineering, and Mathematics, 2018). In response, *data science education*, or teaching and learning about data science, has also grown. Data science education research and course and program development are now taking place at both the postsecondary and K–12 levels, particularly in the last 5 years (Biehler & Fleischer et al., 2022; Jiang, Lee, et al., 2022; Mike et al., 2023; Wilkerson & Polman, 2020).

As data science education is emerging at the grades K–12 and higher education levels, the research and curriculum development are highly variable and often lack a shared focus, language, or set of values. After all, are Antonio and Trulia from the above vignettes both experiencing data science learning? These vignettes highlight how data science education manifests in such diverse and dynamic ways that the experiences students have under the banner of *data science education* can often be similar in name only. What is more, methods and technologies are changing so rapidly that the educational experiences today related to data science are unlikely to resemble data science education 20 years from now. Given these conditions, what do we mean when we use the term data science education? How can data science educators discuss their work with such diverse ideas and practices? There is a need for a framework for data science learning to help to answer these questions.

Within the emerging data science education research community, many are asking, ‘What do we know about data science learning?’ To our knowledge, no such review of the literature exists. This is an especially prominent and important issue at the grades K–12 level, as educational standards are impacted by an underlying research base—one that includes ideas about how the curriculum can be horizontally (between subject areas) and vertically (across grade levels) aligned. Furthermore, K–12 educators (and researchers) can have a different emphasis than those working in higher education: K–12 educators may emphasize cultivating students’ interest and positive dispositions toward a domain (e.g., for students new to data science required to take this class), whereas in some cases these can be assumed (e.g., for students electing a data science major). Thus, we focus on the grades K–12 level, particularly in how we interpret and frame our review and implications, though we think that this work has takeaways for those interested in undergraduate data science education.

It does not take too much time in the literature, though, to recognize the emergent and diverse nature of data science as a discipline. Data science—as a newer discipline itself—is emergent and diverse, changing quickly with new digital technologies and questions (Cao, 2017; National Academies of Sciences, Engineering, and Mathematics, 2018). It is an interdisciplinary, if not a transdisciplinary, endeavor (National Academies of Sciences, Engineering, and Mathematics, 2018) that involves a range of competencies that have been traditionally taught and learned separately (Cao, 2017; Jiang, Lee, et al., 2022; H. Lee, Mojica et al., 2022; Tierney, 2012). The emergent and interdisciplinary nature of what falls under data science means that data science education faces significant challenges in synthesizing what we know and are learning across different research and development agendas (H. Lee, Mojica et al., 2022; V. R. Lee, Pimentel et al., 2022).

Because of the challenge of summarizing what we know about data science learning at the grades K–12 level, this article aims to use a framework to review the diverse research agendas and learning environments relevant to data science education. As we have suggested, the diverse communities engaged in this research have different foci, methods, and language for their data science education research, but in reviewing the data science education literature, we found that these often could be attributed to differences in who and what motivates the instructional design. In this article, we refer to these underlying motivations as *agency*, or the influences on a domain of inquiry, and argue that the variation in data science education approaches can often be described in terms of which agents are given priority (Pickering, 1995). We argue that a focus on agency helps us to illuminate underlying conceptions of data science learning.

2. Agency in Data Science Education Research

We draw on Pickering's (1995) scholarship to describe three types of agency that provide a framework for discussing the variation in data science education learning environments and research: material, personal, and disciplinary. These ideas about agency come from Pickering's studies on the performance of science. Within the philosophy of science, Pickering focused on the material aspects of how science works. From this perspective, scientific work is a performance that emerges in real time without the knowledge of the final form that will result. This means practices, methods, machines, explanations, and knowledge are built without knowing the final results or the implications of their development.

We believe that the idea of agency helps us think about data science education by foregrounding the often-invisible factors that motivate teaching and learning in this domain. *Material agency* in data science education illuminates the ways the material world and the machines we build to control the world influence the types of learning that we value in data science education. The world is not passive as we study it, but is active in constraining and sometimes even resisting the ways we want to observe it. This means we cannot see global migration patterns (Kahn, 2019) or movie ratings and revenue (Fergusson & Pfannkuch, 2022), for example, with our natural senses alone. We need to build machines to "capture, seduce, download, recruit, enroll, or materialize" (Pickering, 1995, p. 7) the material world. These machines, then, create another form of material agency as they generate new forms of data and questions we did not anticipate while building them. This is one reason why the field of data science is motivated by new problems created by innovative digital technologies and the worlds and digital trace data they create (Fischer et al., 2020; Welser et al., 2008).

The material agency of the natural world and the machines we build to capture it exert agency on humans, but humans also exert agency on the world and these machines. Humans exert *personal agency* in an attempt to capture and control material agency. These forms of agency interact with one another, as material agency resists being captured, and personal agency accommodates the plans for capture and thus modifies these plans to where the world's agency is sufficiently pinned down. However, humans rarely act alone when exerting their agency. Pickering also attends to the social aspect of scientific work by describing a third form of agency when communities shape practices, ideas, and machines. These kinds of *disciplinary agency* support humans to capture material agency but also constrain personal agency as practitioners are held to disciplinary norms and conventions in order to participate in particular communities. The agency of these community norms then interacts with material and personal agency. These three forms of agency are summarized in Table 1.

Table 1. Material, personal, and disciplinary agency.

	Definition	**Examples**

Material Agency	The material world and the machines we build to control the world influence the types of learning that we value in data science education	• Organismal features, such as growth or movement patterns, influence sampling and measurement procedures (material world) • Algorithms and software design influence how questions can, or cannot, be asked of data (machines we build)
Personal Agency	Personal processes by which humans attempt to capture and control material agency	• A scientist changes sampling and measurement procedures to make them more manageable • Students discuss how to ask a question to guide an investigation about their community
Disciplinary Agency	Social processes by which communities shape practices, ideas, and machines to develop norms and expectations about how to capture material agency	• A widely used piece of software in a discipline influences how data are analyzed • Specific statistical methods become expected within a discipline creating conventions disciplinary experts assume

We note that the purpose of using agency as a frame is not to wedge research into a particular group; instead, it is to reveal the relative areas of emphasis and to highlight opportunities for future design, research, and development. Instead, our approach was to understand how each of the articles we reviewed *prioritized* the forms of agency differently; all three forms are always at play, but how authors prioritized them makes a difference and led us to review them within a particular section. Further, different from Mike et al.’s (2023) review published in *Harvard Data Science Review*, our review is focused more narrowly on the topic of data science learning (as viewed through the idea of agency), rather than data science education, more broadly. We also note that learning was not an explicit overarching topic (supercluster) or more narrow topic (cluster) in Mike et al.’s (2023) systematic review of more than 1,000 articles related to data science education, justifying our delimited focus on data science learning herein.

Before proceeding to the review, we acknowledge that we come to this review with our own positionality. This article is coauthored by two educational researchers who were, prior to becoming faculty members, teachers at public high schools in the United States. Both coauthors work in a College of Education, but the second teaches courses in a doctoral program that includes students with strong subject areas (e.g., the physical and life sciences) backgrounds. The first coauthor taught science and the second co-taught mathematics. While the first coauthor has developed more of a body of scholarship that is explicitly about using and teaching about data science methods, the second has a deep foundation in the statistics education discipline. These backgrounds bring the author team a broad vantage to try to summarize what is known about data science learning. These backgrounds can also help the reader to begin to understand the influences that shape the research we selected to review and how we reviewed this research.

3. Research Emphasizing Material Agency

Material agency refers to the ways the natural world and human-engineered worlds are active in scientific work by creating phenomena that humans seek to understand, and by resisting human efforts to examine phenomena as we—as researchers—would prefer. Both the natural world and the worlds that emerge from human-engineered tools and systems are constantly variable, which creates challenges for those seeking to understand it. After all, how can we answer a question about the world when the world is constantly changing? The statistics field grew in response to a variable world (Cobb & Moore, 1997). As we suggested in the previous section, data science, too, has grown out of a need to manage new forms of variable data created by rapid growth in computing power (Cao, 2017; H. Lee, Mojica, et al., 2022; V. R. Lee & Wilkerson, 2018). Data science, then, is fundamentally an effort to parse out and explain variation in ways that are meaningful for the questions at hand. Statistics education and science education communities have made productive student engagement with the variation and uncertainty generated from material agency a focus of both research and design.

The American Statistical Association’s Guidelines for Assessment and Instruction in Statistics Education (GAISE) reports have been especially prominent. The *Pre-K-12 GAISE II* report grounds all statistical inquiry as a practice of making sense of variable data (Bargagliotti et al., 2020; Cobb & Moore, 1997). While the K–12 report (Bargagliotti et al., 2020) structures its guidelines in a developmental levels manner suitable for the gradual learning of younger students, the college report (GAISE College Report ASA Revision Committee, 2016) sets broader recommendations that reflect the expectations and requirements of the higher education context. Still, the focus on variation and variable data is present in both reports, and this emphasis is sometimes referred to as engaging students with the *context* of a data set (e.g., Rubin, 2020). This means that students should have opportunities to develop data creation practices to understand how choices about sample and measurement influence the variability in data (Hardy et al., 2020). When carefully designed, this type of engagement with variability can support students to better understand both practices related to data science and the phenomenon under investigation (Ford & Forman, 2006; Manz, 2015). This is especially true when students are positioned to parse variation that informs their question from variation due to random noise in the data (Hardy et al., 2020).

In addition to the material agency coming from the natural world, new digital technologies are creating types of material agency that influence data science work in new ways and that create new forms of variability to describe, explain, or predict. One example is the use of digital measurement tools, such as the probeware commonly used in middle and high school science classes to measure temperature, pH, and gas pressure (Lee & Wilkerson, 2018). Students can be supported to engage with these tools in ways that help them reflect on the agency of the tool and how to use it to accomplish their own goals (Hardy et al., 2020). When given opportunities to grapple with challenges associated with data creation, students are better positioned to manage material agency when cleaning and managing the 'messy data' that results when they carry out investigation procedures (Hammett & Dorsey, 2020; Kjelvik & Schultheis, 2019; Rosenberg et al., 2020; Schanzer et al., 2022).

Research has shown, however, that it is important to carefully consider the forms of variability that students encounter. Variation from measurement error alone has proven powerful and accessible for early experiences with data generation (Konold & Lehrer, 2008; Konold & Pollatsek, 2002). In contrast, variability caused by natural differences, production errors, and random fluctuations are more challenging to reason through. It is clear that the source of variability is highly consequential, but little is known about how students reason about measurement error when using opaque measurement tools such as digital probes. This is likely a consequential difference for students since visible measurement methods, such as using a ruler to measure the circumference of a tree trunk, provide accessible ways for students to reason about error due to measuring trees at different heights, reading the rulers with degrees of precision, or creating gaps or overlaps when iterating the ruler around the trunk. But it is less clear what resources students can use to reason about sources of error when using digital probes, such as during investigations that involve recording the temperature or pH of a solution. The probes could be calibrated imprecisely, students might record data with a lag or at periodic intervals, or there might be slight differences between manufacturers of these instruments. These considerations may be less visible to students, and more work is needed to understand how to support students well with these tools.

Antonio experienced material agency by studying the real-world context of recycling and by experiencing the opportunities and constraints of data science tools, like RStudio. However, the experience of the context and tools minimized the material agency in an investigation like this. Antonio did not take part in creating the data, and so is unlikely to experience much of the challenges associated with measurement and data collection. In addition, Antonio's teacher gave him the code to analyze the data, which means the students were not afforded opportunities to grapple with the tools to experience how their initial ideas get shaped by the capabilities of the digital tools. Trulia, on the other hand, had a much closer experience with the context of their investigation. She knew her schoolyard well, and generated research questions based on what she observed there. As she and her friends tried to count trash in the yard, they experienced first-hand the challenges associated with even the most straightforward questions and how things like wind in the material world created problems for their data collection. It took multiple months, but Trulia and her friends used their experiences of challenges created by the material world to make sense of their data and to inform their answers to the research questions.

Although it is important to engage students with material agency, students are often not given these opportunities (Hardy et al., 2020; Miller et al., 2018). This happens when disciplinary agency is prioritized because students’ responses to material agency rarely replicate disciplinary norms. If a teacher has a disciplinary norm as the primary goal of the investigation, they often restrict engagement with the challenges of material agency because they are likely to complicate students’ conceptual interpretations of the investigation (Hardy et al., 2020). In addition, reasoning about variability is challenging for most people (Torok & Watson, 2000). Careful design, then, is needed to support teachers to engage students productively with material agency.

4. Research Emphasizing Personal Agency

In this context, personal agency refers to the efforts humans make to capture material agency in ways that help us control, observe, and explain it. As the natural and human-engineered worlds create variable phenomena, humans have developed strategies, representations, statistics, models, and computational systems to structure the variability in ways that allow us to make observations and inferences that cannot be made by direct observation alone. However, these methods are always human-generated artifacts, which means that data science methods and claims are always driven by our ideas and theories (Hardy et al., 2020). Methods are never neutral and always represent the perspectives, goals, and values of the people and communities that develop them. Learning sciences communities and statistics education communities often focus on understanding the personal agency of learners in research on data science learning (e.g., Lee & Wilkerson, 2018).

When students are given personal agency to make decisions about their data science investigations, they can develop an understanding of and competency in using their agency to make claims about the world around them. One result is that students expand their notions of what counts as data to include information and artifacts from their own personal lives (Stornaiuolo, 2020). For learners, then, data is not information generated by someone else, but potentially information about themselves and their lives, as data is increasingly collected from or about students. Students can also develop competency in crafting *data stories* that are both empirically grounded and personally meaningful (Kahn, 2019; Lee & Dubovi, 2019; Roberts & Lyons, 2020; Stornaiuolo, 2020; Wilkerson & Laina, 2018; Wilkerson et al., 2021). These data stories position the students against the data and the wider society the data represent, and students often use their agency and resources around them to better understand their own history and community (Kahn, 2019; Roberts & Lyons, 2020). This can also help students understand that data and claims made with data are not objective but are always stories told by people from a particular perspective (Rubin, 2020). This can support them to critically interrogate the personal agency behind data science systems and claims that others have developed, sometimes referred to as critical data literacies (Sander, 2020; Stornaiuolo, 2020). This critical stance on data science positions students to be not only objects of investigation but also active agents in generating, interpreting, and critiquing data and decisions made with data.

Research has also shown that students’ personal agency can have strong epistemic congruence with disciplinary norms and values. By epistemic congruence, we mean that students' data science work rarely replicates disciplinary work, but, rather, the motivation behind the students' approaches and the ways they use the representations, statistics, models, and computational systems they develop resemble the ways disciplinary tools are used. Student-generated data visualizations might not replicate the rules of a histogram or dot plot, but they can often use principles like scale, order, and frequency as tools to communicate something important about their data (Konold et al., 2015; Lehrer & Schauble, 2007; Petrosino et al., 2003). This work can then support students in viewing data in terms of the aggregate shape created by representations, which is important for thinking about descriptive statistics to index center and variability (Watson & Mortiz, 2003). Research has shown that students can invent innovative statistics that attend to distributional characteristics in meaningful ways and that with thoughtful design and support, they use and reason about these statistical inventions in ways that are similar to professional statistical work (Jones et al., 2017; Lehrer & Kim, 2009). Students can then use these ideas and practices as epistemic tools to make inferences with data in innovative ways (Konold, 2002). Finally, research has shown that students can exert personal agency to create, compare, revise, and use probability models to make inferences and develop machine learning algorithms (Lehrer & English, 2018; Zimmerman-Neifeld et al., 2019).

Antonio's data science class was designed to help him develop a personal connection to data science and to identify as someone who can participate in data science. By engaging him and his classmates in the development of a claim about recycling in Los Angeles, and by having them communicate their findings to the class, the teacher helped Antonio experience the personal agency involved in making claims with data. However, Antonio was not given opportunities to use his agency to shape the question, data creation, or analysis strategies. But Trulia's class heavily emphasized the personal agency involved in data science by supporting students to investigate their local community and to develop questions that were meaningful to them. The focus on trash in their schoolyard came from their personal interest in improving their schoolyard, and created an investigation in which they used their personal agency to develop measurement protocols, discuss variation in measurement and strategies for shared protocols, and generate claims about their schoolyard for others to critique.

5. Research Emphasizing Disciplinary Agency

Disciplinary agency concerns the norms and practices that individuals in a discipline adopt—and explicitly and implicitly coerce others to adopt. Of course, such forms of agency are malleable and can change over time, as is the case with the emerging domain that is data science. Statisticians and computer scientists have been prominent in writing about data science learning from a perspective that highlights the norms and practices of their disciplines. This work is taking place in communities that are likely familiar to statisticians and computer scientists, especially those who carry out research in postsecondary settings.

Indeed, one of the most important papers that highlights disciplinary agency is Nolan and Temple Lang's (2010) article calling for a greater role for computation in the statistics curriculum, as this work was not primarily motivated by issues related to personal or material, but rather by changes in the statistics (and burgeoning data science) discipline. Though at the undergraduate level, its recommendations are pertinent to K–12 data science education. This article called for greater breadth and depth regarding students' use of computational methods when they are learning statistics—as well as the importance of using computational methods in the context of working with data. Disciplinary norms are privileged by considering the six topics Nolan and Temple Lang (2010) recommended for integration into the undergraduate curriculum, including scientific computing with data (i.e., programming the steps to be taken in an analysis), computational statistics, and the use of integrated development environments (applications for editing, writing, and debugging code). Several of these are contiguous with traditional statistics topics; others—such as advanced computing, including the use of distributed/high-performance computing systems—are less often covered in traditional statistics education.

Many articles highlight the importance of programming to the statistics and data science disciplines (Çetinkaya-Rundel et al., 2022; Çetinkaya-Rundel & Ellison, 2020; Dogucu & Çetinkaya-Rundel, 2020; Fergusson & Pfannkuch, 2022 Hardin et al., 2015; Heinzman, 2022; Kim & Hardin, 2021; Kim & Henke, 2021). Also, an emphasis on programming has been an emphasis for researchers who approach data science through the lens of computing education and the role of data in the work of computer scientists (Dryer et al., 2018; Schanzer et al., 2022).

This emphasis on programming is not limited to these articles; Nolan and Temple Lang (2010) pulled no punches when writing about their importance, claiming that "computational literacy and programming are as fundamental to statistical practice and research as mathematics" (p. 96). Providing further evidence for the centrality of programming, Schwab-McCoy et al. (2021) conducted a survey of introductory data science course instructors and found that RStudio (a commonly used integrated development environment for R) and Jupyter Notebooks (a type of document for data scientists to use Python) were the most commonly used software for introductory, college-level data science courses. Mirroring this finding, the most commonly used programming languages were R and then Python, respectively. Further, there is a clear synergy between this work and that of computer scientists, especially those taking a data-centric approach (e.g., Schanzer et al., 2022; Krishnamurti & Fisler, 2020). Notably, two prominent K–12 data science curricula, the high school-focused *Introduction to Data Science* (Gould et al., 2018) and the middle and high school–focused *Bootstrap: Data Science* (Schanzer et al., 2022), both emphasize programming.

In short, programming—especially in R at the undergraduate level (Schwab-McCoy et al., 2021)—is a disciplinary tool that has a strong influence on data science. This research teaches us that novices can learn to code in the context of a one-semester course. We recognize that this research is primarily at the undergraduate level, but the students in such courses often bring similar degrees of preparation as students at the high school level—especially older high school students. Further, the research around the two K–12 data science curricula also emphasizes programming—and there is some evidence for how even young students can learn to program and find programming to be valuable to them (Heinzman, 2022; Schanzer et al., 2022). We do note that far from all research on data science learning (even at the undergraduate level) emphasizes programming. Many scholars have called for an approach that emphasizes data literacy at least at the very outset of students learning data science (Burckhardt et al., 2021; Gould, 2017; Kjelvik & Schultheis, 2019; Mike & Hazzan, 2022; Wise, 2020). These studies either provide web-based tools for students to gradually become familiar with code (Burckhardt et al., 2021) or do not use programming at all (Mike & Hazzan, 2022).

Programming is not the sole essential skill. Another element of research emphasizing disciplinary agency in undergraduate course contexts and particular designs emphasizes what Horton and Hardin (2021) refer to as *creative structures*. Again, though these are at the undergraduate level, the course designs and technologies could be widely relevant to grades K–12 data science education, and so we review this research here with an eye toward whether and how it may be so. These software tools have the potential to engage students with the goals and values of the disciplines that created them, exerting disciplinary agency, as students use them to make sense of data. The creative structures that have been introduced at the undergraduate level include deliberate choices about aspects of the course beyond the programming language that is (often) used: the course website, nature of assignments and projects, and means for teachers and students of giving and receiving feedback, among others. As others have pointed out, this work often takes the form of case studies (for examples of case studies, see Schwab-McCoy et al., 2021). Çetinkaya-Rundel and Ellison's (2020) paper is emblematic of this approach. Other papers extend this sophisticated infrastructure to emphasize reproducibility through the use of the server tool Docker (Çetinkaya-Rundel & Rundel, 2018) and other technically sophisticated technologies and programming tools (Burckhardt et al., 2021; Dogucu & Çetinkaya-Rundel, 2020; Kim & Henke, 2021). Still other work bridges between tools used primarily by K–12 learners (e.g., the Common Online Data Analysis Platform, CODAP) and programming tools (Biehler & Fleischer, 2021). Thus, a second focus is on creative, thoughtfully designed course structures, especially at the undergraduate level, with potential for this work to be drawn upon by scholars, curriculum developers, and teachers concerning grades K–12 data science education. This work suggests that in order for students to be successful, instructors need to consider the instructional design of their courses: how the course goals, instruction, assessments, opportunities for practice and help, and even the technologies used to manage the submission of assignments align and work together to support learners to know about and do data science.

A final finding highlighting disciplinary agency concerns scaffolds for learning new and highly valued analytic and modeling techniques in data science disciplines (Horton & Hardin, 2021), such as machine learning. For example, Fergusson and Pfannkuch (2022) show how the core tenets of machine learning can be taught to K–12 students when designed and taught in a particular manner; namely, using a particular ('informal') approach that emphasizes visualizations, a potentially relevant data set (movie ratings), and a browser-based environment for students to run R code. Other papers emphasize machine learning (e.g., Jiang, Nocera, et al., 2022; Zimmermann-Niefield et al., 2019) and even artificial intelligence (Druga & Ko, 2021), as well as Bayesian approaches (Erickson, 2017; Kazak, 2015; Rosenberg, Kubsch, et al., 2022; Warren, 2020), developing statistical software (Reinhart & Genovese, 2021), web scraping using social media data (Boehm &

Hanlon, 2021; Dogucu & Çetinkaya-Rundel, 2020), and using git and GitHub (Adams et al., 2021; Beckman et al., 2021; Kim & Henke, 2021). This work shows that learners can develop the capacity to use new analytic and programming tools with deliberately designed courses.

Antonio's data science experience was strongly influenced by disciplinary agency. His teacher structured the exploration around a data analysis software valued among professionals, RStudio. As Antonio explored the data on recycling, his investigation and thinking were both supported and constrained by the features of this tool, and since the teacher gave the students code for their analysis, the students had limited agency in deciding how or why to use the analytic techniques. However, the techniques were motivated by their widespread use in data science disciplines. Trulia's investigation, on the other hand, had much less influence from disciplinary agency. The teacher was motivated by the discipline of ecology to engage students in the schoolyard investigations, but the students had the majority of the agency to determine what questions to ask, methods for sampling and measurement, and analytic strategies. This meant their measurement and analysis deviated in many ways from disciplinary norms.

6. How Divided Are Data Science Education Communities? A Co-Citation Coupling Analysis

A premise of our above review using the idea of agency was that data science learning takes place across many different communities. For example, the special issue of the *Journal of Statistics and Data Science Education* (Horton & Hardin, 2021) was focused on reflecting on a key article on data science teaching and learning at the undergraduate level, Nolan and Temple Lang's (2010) call for statistics educators to introduce computing concepts and computational skills to students learning statistics. While this paper was clearly influential on the authors writing in this special issue, three other special issues on data science education only cited Nolan and Temple Lang's (2010) work in one article. These special issues were in the *Journal of the Learning Sciences* (Wilkerson & Polman, 2020), the British *Journal of Educational Technology* (Jiang, Lee, et al., 2022), and the *Statistics Education Research Journal* (Biehler et al., 2022). This disconnect highlights the distinctiveness across the various disciplines studying data science education, as the authors of these articles motivated their research based on their discipline-specific research—not the need to integrate computing into statistics that Nolan and Temple Lang (2010) highlighted.

Because of this, we became interested in what connections between scholars privileging different forms of agency existed—or did not exist—across research communities. While our primary focus was on the different forms of agency emphasized in research taking place in distinct communities, an ancillary focus was on how research focusing on different levels (grades K–12 and higher education, as well as learning in informal settings) was also emerging across divides.

6.1. Methods for Exploring Connections Across Research Communities

A way we surmised that these siloed bodies of research could be interrogated is by examining scholarship published in recent, prominent special issues about data science education. A recent set of special issues on data science education make this task more tractable. Namely, there have been special issues of the *Journal of the Learning Sciences*, *British Journal of Educational Technology*, and *Statistics Education Research Journal* on data science education, with different emphases regarding the three forms of agency we have discussed—as we unpack in the following paragraphs.[1] Furthermore, while the first two special issues had a grades K–12 focus, the latter featured research predominantly at the higher education level, making these special issues ripe for interrogation. We note that like in our above review of prior research, the purpose of using agency as a frame is not to wedge research into a particular group; instead, it is to reveal the relative areas of emphasis and to highlight opportunities for future design, research, and development.

The eight articles in the *Journal of the Learning Sciences* special issue emphasized personal agency; this is in line with the title of the special issue, *Situating Data Science: Exploring How Relationships to Data Shape Learning*. Articles used complex data sets to explore their families' migration stories (Kahn, 2019) or their personal diabetes data (V. R. Lee & Dubovi, 2019), for instance.

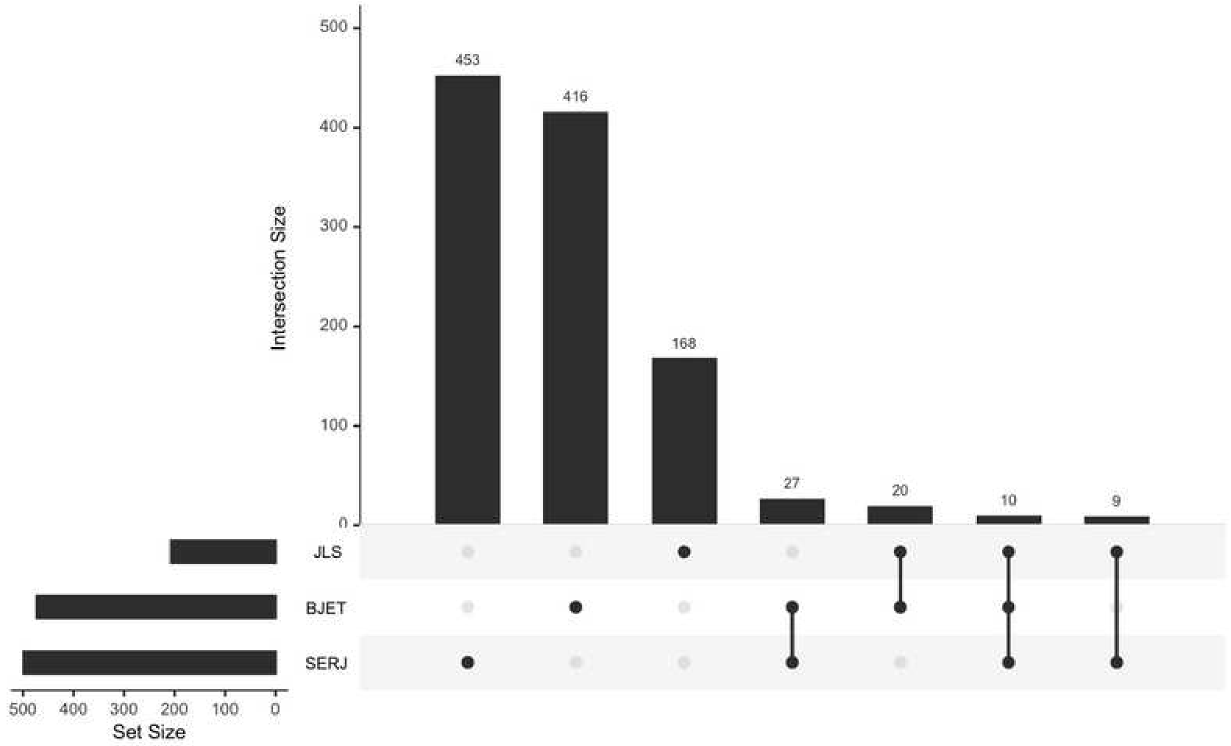
The nine articles in the *British Journal of Educational Technology* special issue focused on disciplinary agency as it plays out in different disciplines, like social studies (Shreiner & Guzdial, 2022) and art education (Matuk et al., 2022). The title of this special issue is also in accordance with this focus: *Data Science Education Across the Disciplines*.

The 11 articles in the *Statistics Education Research Journal* special issue focused on material agency, particularly in the context of undergraduate statistics and data science courses, but also as it pertains to grades K–12 learners. For instance, Mike and Hazzan (2022) described how to teach machine learning in an accessible manner to undergraduates and H. Lee, Mojica, et al. (2022) used interviews with practicing data scientists to characterize the nature of their work to be able to inform teacher education and curriculum development.

To examine the divides across research communities, we conducted a co-citation coupling analysis. This involves examining co-citations by articles. In our case, we examined co-citations of articles for the articles in the three special issues on data science education—28 in total. For this analysis, we created a spreadsheet with columns for the *citing article* and the *cited article*. For instance, Kahn (2020) cited Lee and Wilkerson (2018), so Kahn was recorded as the citing article, and Lee and Wilkerson as the cited article. This process repeats for each article Kahn (2020) cited. H. Lee, Mojica, et al. (2022) also cited Lee and Wilkerson (2018), meaning that Kahn and H. Lee, Mojica, et al. (2022) would be 'coupled' through (at least) one co-citation. We carried out this process of identifying all the cited articles for each of the 28 special issue articles and then counting the number of co-citations among all 28 articles. From this data, we can examine the nature of the co-citations among and between the special issues to understand whether there is segregation between communities emphasizing different forms of agency.

6.2. Results of Co-Citation Coupling Analysis

To present the results of the co-citation coupling analysis, we examined how many articles were cited by articles in the *Journal of the Learning Sciences* (*JLS*), *British Journal of Educational Technology* (*BJET*), and *Statistics Education Research Journal* (*SERJ*) special issues, the number cited by articles in pairs of special issue articles, and the number cited by articles an all three. The results of this analysis are depicted in Figure 1—what is called an *upset* plot created using the UpSetR R package (Conway et al., 2023). The vertical bars indicate the number of articles falling into one of seven categories. The first three bars represent the number of articles uniquely cited by one or more articles from a single special issue. For example, the first bar—with a value of 453—indicates that 453 unique articles were cited across the 11 articles in the *Statistics Education Research Journal* special issue. These 453 articles were not cited by articles in either of the other two special issues. Of these, 25 articles were cited by two special issue articles, and three or more special issue articles cited a total of six articles.



Note. The vertical bars are arranged by the number of articles cited by articles in one, pairs of two, or all three special issues.

Figure 1. The results of the co-citation coupling analysis, indicating the number of articles cited by articles in the *Journal of the Learning Sciences* (*JLS*), *British Journal of Educational Technology* (*BJET*), and *Statistics Education Research Journal* (*SERJ*) special issues (black dots) on data science education, the number cited by articles in two special issues (lines connecting two black dots), and in all three (line connecting three black dots).

Following this interpretation, it can be seen that 416 unique articles were cited by one or more of the nine *British Journal of Educational Technology* special issue articles. We think it is noteworthy that none of these 416 cited articles were cited by more than one article, as it suggests the diversity of ideas that are drawn upon for the articles in a single issue.

Far fewer articles—168—were cited exclusively in one or more of the eight *Journal of the Learning Sciences* articles. Of these, four articles were cited by two special issue articles, and one was cited in four of these articles. Note that we have interpreted *Journal of the Learning Sciences* to emphasize personal agency, *British Journal of Educational Technology* to emphasize disciplinary agency, and *Statistics Education Research Journal* material agency. The findings we have just described suggest greater cohesion among the articles (in terms of co-citation coupling) among the articles in the *Statistics Education Research Journal* than those in the other special issues.

We can also examine which specific articles were cited the most exclusively within the special issues. The top-three most-cited articles that were cited only within each of the special issues follows.

Top-three most-cited articles only cited in *Statistics Education Research Journal* special issue articles:

- De Veaux et al. (2017), cited by five articles
- Biehler and Schulte (2017), cited by four articles
- Wild and Pfannkuch (1999), cited by four articles

Top-three most-cited articles only cited in *Journal of the Learning Sciences* special issue articles:

- Konold et al. (2015), cited by four articles
- Gutiérrez and Jurow (2016), cited by two articles
- Harré (2003), cited by two articles

Note that we do not name any articles in the *British Journal of Educational Technology* because none that were exclusively cited by articles in this special issue were cited more than once.

These suggest the nature of the work which is uniquely cited by scholars publishing in communities emphasizing particular forms of agency.

The next two bars in Figure 1 indicate the number of articles cited by one or more articles in two of the special issues. The first of these three has a value of 27 and it indicates that 27 unique articles were cited by *British Journal of Educational Technology* and *Statistics Education Research Journal* articles. The second of these three bars, with a value of 20, indicates that 20 unique articles were cited by *British Journal of Educational Technology* and *Journal of the Learning Sciences* articles. The next, with a value of 10, indicates that this many unique articles were cited by articles in all three special issues. Finally, nine unique articles were cited by articles in the *Journal of the Learning Sciences* and *Statistics Education Research Journal* articles. These findings suggest that there is notable co-citation coupling between *British Journal of Educational Technology* articles and those in the other two special issues, and less between those in the *Journal of the Learning Sciences* and the *Statistics Education Research Journal*.

Like earlier, we can examine which specific articles were cited the most among the special issues, as follows.

Top-three most-cited articles only cited by articles in *British Journal of Educational Technology* and *Statistics Education Research Journal* special issues[2]:

- Bargagliotti et al. (2020), cited by five articles
- D’Ignazio and Klein (2020), cited by five articles
- Wilkerson and Polman (2020), cited by five articles

Top-three most-cited articles only cited by articles in *British Journal of Educational Technology* and *Journal of the Learning Sciences* special issues:

- Bhargava et al. (2015), cited by four articles
- Enyedy and Mukhopadhyay (2007), cited by four articles
- Taylor and Hall (2013), cited by four articles

Top-three most-cited articles only cited in all three special issues:

- Lee and Wilkerson (2018), cited by nine articles
- Finzer (2013), cited by six articles
- Philip et al. (2016)[3], cited by four articles

Top-three most-cited articles only cited in *Statistics Education Research Journal* and *Journal of the Learning Sciences* special issues[4]:

- Hardin et al. (2015), cited by four articles
- Segel and Heer (2010), cited by three articles

These suggest the nature of the work uniquely cited by articles in pairs of special issue articles and the nature of the work cited by articles across all three special issues.

7. Discussion

On the surface, the differences in Antonio and Trulia’s data experiences in the vignettes we led off with suggest that the nature of data science learning may be difficult to reconcile. However, from the lens of agency, we can see that personal, material, and disciplinary agencies all motivated the investigations, interacting with each other and being prioritized in different ways. For Antonio, disciplinary agency was more highly prioritized, reducing the influence of material and personal agency. For Trulia’s teacher, personal and material agency were a high priority, which reduced the influence of disciplinary agency. Teachers at the grades K–12 (and those supporting these teachers as curriculum designers, professional development providers, or researchers) can consider these emphases when teaching. There will not be one correct form of agency for teachers at this level to emphasize; instead, teachers can consider what their goals are when teaching particular students specific data science capacities or ideas. For learners new to data science, it may be critical to emphasize personal agency to encourage students to develop an initial interest in it.

We finish by suggesting a few takeaways in relation to the questions we posed at the beginning of this article. We offer these in the spirit of takeaways to work and build on as an emerging discipline.

What is it that we mean when we use the term data science education? Data science education researchers need to be more explicit about what we prioritize and what motivates our work. Although many are doing work under the label data science, our work highlights the importance of explicit attention to how we conceptualize both data science and learning, and how material, personal, and disciplinary agency influence our work with students. At a minimum, this is a question of definitions. For example, while computing is generally considered to be a core part of what sets data science apart from statistics (Breiman, 2001; Donoho, 2017; National Academies of Sciences, Engineering, and Medicine, 2018), its role in data science learning is not entirely settled. Namely, some data science curricula—even at the undergraduate level—deemphasize programming (Burckhardt et al., 2021; Mike & Hazzan, 2022). This is the case at the K–12 level, too; few papers in either the aforementioned *Journal of the Learning Sciences* or the *British Journal of Educational Technology* special issues involved programming (though many used computers to access, process, visualize, and model data), although high school (*Introduction to Data Science*) and middle and high school (*Bootstrap: Data Science*) curricula both emphasize programming. So, how do these forms of material and disciplinary agency influence our work?

Also, given that many definitions of data science include a discipline-specific component (cf. V. R. Lee et al., 2022), a related question concerns what about learning data science applies across disciplines and what is specific to the topics, data, methods, and phenomena that characterize specific disciplines, a question asked by data science education scholars (Finzer, 2013; Jiang, Lee, et al., 2022). As teachers in different subject areas may have different degrees of experience teaching with data (Rosenberg, Schultheis, et al., 2022), efforts to support data science learning across disciplines may need to carefully consider what discipline-specific supports, tools, and resources are needed for teachers and learners alike. But the issue is more than just a question of definition, as educational work is always motivated by issues related to how we conceptualize and what we prioritize in learning. For example, is this an effort primarily in workforce development or one of cultivating an educated citizenry? How do we design our data science education projects so that can exert personal agency in their learning?

How can data science educators discuss their work across such diverse ideas and practices? A focus on agency in a paper about work done across diverse communities places the question of who has a voice—and whose voices are valued—at the forefront of data science education. Data, models, and claims are never neutral and always a product of human activity. Asking questions about personal agency forces us to ask how we are supporting students' voices in data science education. Are they getting opportunities to choose investigations and make claims that impact their local communities? In an emerging, interdisciplinary data science world, questions about disciplinary agency require us to reflect on which disciplines are being privileged and whose voices have had a place in shaping those disciplines?

While there is doubtless value in coming to similar findings despite using different research approaches, there is also value in different approaches to researching data science learning. We have found it beneficial to consider essential research on data science learning at the grades K–12 levels by scholars not only with backgrounds in computer science and statistics education, for example, but also in the arts (Matuk et al., 2022), humanities (Vance et al., 2022), social studies (Shreiner & Guzdial, 2022), and everyday life and experiences (Gebre, 2022; Radinsky & Tabak, 2022; Vacca et al., 2022). Other scholars have argued for the importance of a broadly humanistic approach to data science education, one that recognizes that questions about who uses data overlap with questions about who has value and power (V. R. Lee et al., 2021; Philip et al., 2016). As the field develops, it is important to continuously reflect on who shapes what we know about data science learning.

Acknowledgments

The authors wish to thank Conrad Borchers, Amanda Garner, Spencer Greenhalgh, Matthew Koehler, Jeanne McClure, and Sarah Narvaiz for feedback on an earlier version of this article.

This material is based upon work supported by the National Science Foundation under Grant No. 1942770 and No. 2239152.

Disclosure Statement

An earlier version of this work was shared as a commissioned paper for the National Academy of Sciences, Engineering, and Medicine *Foundations of Data Science for Students in Grades K-12: A Workshop*.

References

Adams, B., Baller, D., Jonas, B., Joseph, A.-C., & Cummiskey, K. (2021). Computational skills for multivariable thinking in introductory statistics. *Journal of Statistics and Data Science Education, 29*(Suppl. 1), S123–S131. https://doi.org/10.1080/10691898.2020.1852139

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education. *American Statistical Association*.

Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). Implementing version control with git and GitHub as a learning objective in statistics and data science courses. *Journal of Statistics and Data Science Education, 29*(Suppl. 1), S132–S144. https://doi.org/10.1080/10691898.2020.1848485

Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., & Shoup, N. (2015). *Beyond data literacy: Reinventing community engagement and empowerment in the age of data* [White paper]. Data-Pop Alliance. https://datapopalliance.org/wp-content/uploads/2015/11/Beyond-Data-Literacy-2015.pdf

Biehler, R., & Fleischer, Y. (2021). Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. *Teaching Statistics, 43*(S1), S133–S142. https://doi.org/10.1111/test.12279

Biehler, R., De Veaux, R., Engel, J., Kazak, S., & Frischemeier, D.(2022). Editorial: Research on data science education. *Statistics Education Research Journal, 21*(2), Article 1. https://doi.org/10.52041/serj.v21i2.606

Biehler, R., & Schulte, C. (2017). Perspectives for an interdisciplinary data science curriculum at German secondary schools. In R. Biehlr, L. Budde, D. Frischemeier, B. Heinemann, S. Podworny, C. Schulte, & T. Wassong (Eds.), *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts* (pp. 2–14). https://www.telekom-stiftung.de/sites/default/files/files/PaderbornSymposiumDataScience2017_0.pdf#page=7

Boehm, F. J., & Hanlon, B. M. (2021). What is happening on Twitter? A framework for student research projects with tweets. *Journal of Statistics and Data Science Education*, *29*(Suppl. 1), S95–S102. https://doi.org/10.1080/10691898.2020.1848486

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Burckhardt, P., Nugent, R., & Genovese, C. R. (2021). Teaching statistical concepts and modern data analysis with a computing-integrated learning environment. *Journal of Statistics and Data Science Education*, *29*(Suppl. 1), S61–S73. https://doi.org/10.1080/10691898.2020.1854637

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, *50*(3), Article 43. https://doi.org/10.1145/3076253

Çetinkaya-Rundel, M., Dogucu, M., & Rummerfield, W. (2022). The 5Ws and 1H of term projects in the introductory data science classroom. *Statistics Education Research Journal*, *21*(2), Article 4. https://doi.org/10.52041/serj.v21i2.37

Çetinkaya-Rundel, M., & Ellison, V. (2020). A fresh look at introductory data science. *Journal of Statistics Education*, *29*(Suppl. 1), S16-S26. https://doi.org/10.1080/10691898.2020.1804497

Çetinkaya-Rundel, M., & Rundel, C. (2018). Infrastructure and tools for teaching computing throughout the statistical curriculum. *The American Statistician*, *72*(1), 58–65. https://doi.org/10.1080/00031305.2017.1397549

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823. https://doi.org/10.2307/2975286

Conway, J. R., Lex, A., Gehlenborg, N. (2023). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., ... Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

D'ignazio, C., & Klein, L. F. (2020). *Data feminism.* MIT press.

Dogucu, M., & Çetinkaya-Rundel, M. (2020). Web scraping in the statistics and data science curriculum: Challenges and opportunities. *Journal of Statistics and Data Science Education, 29*(Suppl. 1), S112–S122. https://doi.org/10.1080/10691898.2020.1787116

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Druga, S., & Ko, A. J. (2021). How do children's perceptions of machine intelligence change when training & coding smart programs? In M. Roussou & S. Shahid (Eds.), *IDC '21: Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (pp. 49–61). ACM. https://doi.org/10.1145/3459990.3460712

Dryer, A., Walia, N., & Chattopadhyay, A. (2018). A middle-school module for introducing data-mining, big-data, ethics and privacy using RapidMiner and a Hollywood theme. In T. Barnes & D. Garcia (Eds.), *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (pp. 753–758). ACM. https://doi.org/10.1145/3159450.3159553

Enyedy, N., & Mukhopadhyay, S. (2007). They don't show nothing I didn't know: Emergent tensions between culturally relevant pedagogy and mathematics pedagogy. *The Journal of the Learning Sciences, 16*(2), 139–174. https://doi.org/10.1080/10508400701193671

Erickson, T. (2017). Beginning Bayes. *Teaching Statistics, 39*(1), 30–35. https://doi.org/10.1111/test.12121

Fergusson, A., & Pfannkuch, M. (2022). Introducing high school statistics teachers to predictive modelling and APIs using code-driven tools. *Statistics Education Research Journal, 21*(2), Article 8. https://doi.org/10.52041/serj.v21i2.49

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2). https://doi.org/10.5070/T572013891

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education, 44*(1), 130–160. https://doi.org/10.3102/0091732X20903304

Ford, M. J., & Forman, E. A. (2006). Chapter 1: Redefining disciplinary learning in classroom contexts. *Review of Research in Education, 30*(1), 1–32. http://dx.doi.org/10.3102/0091732X030001001

GAISE College Report ASA Revision Committee. (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

Gebre, E. (2022). Conceptions and perspectives of data literacy in secondary education. *British Journal of Educational Technology, 53*(5), 1080–1095. https://doi.org/10.1111/bjet.13246

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Gould, R., Machado, S., Johnson, T. A., & Molynoux, J. (2018). *Introduction to Data Science v 5.0*. UCLA Center X.

Gutiérrez, K. D., & Jurow, A. S. (2018). Social design experiments: Toward equity by design. In M. Cole, W. Penuel, K. O'Neill (Eds.), *Cultural-historical activity theory approaches to design-based research* (pp. 79–112). Routledge.

Hammett, A., & Dorsey, C. (2020). Messy data, real science. *The Science Teacher, 87*(8), 40–49. https://www.nsta.org/science-teacher/science-teacher-aprilmay-2020/messy-data-real-science

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. T., & Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "Think with data." *The American Statistician, 69*(4), 343–353. https://doi.org/10.1080/00031305.2015.1077729

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. https://doi.org/10.1080/10508406.2019.1678164

Harré, R. (2003). The materiality of instruments in a metaphysics for experiments. In H. Radder (Ed.), *The Philosophy of scientific experimentation* (pp. 19–38). University of Pittsburgh Press.

Heinzman, E. (2022). "I love math only if it's coding": A case study of student experiences in an introduction to data science course. *Statistics Education Research Journal, 21*(2), Article 5. https://doi.org/10.52041/serj.v21i2.43

Horton, N. J., & Hardin, J. S. (2021). Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking. *Journal of Statistics and Data Science Education*, *29*(Suppl. 1), S1–S3. https://doi.org/10.1080/10691898.2020.1870416

Jiang, S., Lee, V. R., & Rosenberg, J. M. (2022). Data science education across the disciplines: Underexamined opportunities for K-12 innovation. *British Journal of Educational Technology, 53*(5), 1073–1079. https://doi.org/10.1111/bjet.13258

Jiang, S., Nocera, A., Tatar, C., Yoder, M. M., Chao, J., Wiedemann, K., Finzer, W., & Rosé, C. P. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. *British Journal of Educational Technology, 53*(5), 1114–1133. https://doi.org/10.1111/bjet.13253

Jones, R. S., Lehrer, R., & Kim, M.-J. (2017). Critiquing statistics in student and professional worlds. *Cognition and Instruction*, *35*(4), 317–336. https://doi.org/10.1080/07370008.2017.1358720

Kahn, J. (2019). Learning at the intersection of self and society: The family geobiography as a context for data science education. *Journal of the Learning Sciences*, *29*(1), 57–80. https://doi.org/10.1080/10508406.2019.1693377

Kazak, S. (2015). A Bayesian inspired approach to reasoning about uncertainty: 'How confident are you?' In K. Kraimer & N. Vondrova (Eds.), *CERME9-Ninth Congress of the European Society for Research in Mathematics Education* (pp. 700–706). HAL Archives.

Kim, A. Y., & Hardin, J. (2021). “Playing the whole game”: A data collection and analysis exercise with Google Calendar. *Journal of Statistics and Data Science Education*, *29*(Suppl. 1), S51–S60. https://doi.org/10.1080/10691898.2020.1799728

Kim, B., & Henke, G. (2021). Easy-to-use cloud computing for teaching data science. *Journal of Statistics and Data Science Education*, *29*(Suppl. 1), S103–S111. https://doi.org/10.1080/10691898.2020.1860726

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), Article es2. https://doi.org/10.1187%2Fcbe.18-02-0023

Konold, C. (2002). Teaching concepts rather than conventions. *New England Journal of Mathematics*, *34*(2), pp. 69–81.

Konold, C., & Lehrer, R. (2008). Technology and mathematics education. In L. D. English (Ed.), *Handbook of international research in mathematics education* (pp. 49–69). Routledge. https://www.srri.umass.edu/sites/srri/files/Konold%20%26%20Lehrer%202008/index.pdf

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal of Research in Mathematics Education*, *33*(4), 259–289. https://doi.org/10.2307/749741

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics*, *88*(3), 305–325. https://doi.org/10.1007/s10649-013-9529-8

Krishnamurthi, S., & Fisler, K. (2020). Data-centricity: A challenge and opportunity for computing education. *Communications of the ACM*, *63*(8), 24–26. https://doi.org/10.1145/3408056

Lee, H., Mojica, G., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*, *21*(2), Article 3. https://doi.org/10.52041/serj.v21i2.41

Lee, V. R., & Dubovi, I. (2019). At home with data: Family engagements with data involved in type 1 diabetes management. *Journal of the Learning Sciences*, *29*(1), 11–31. https://doi.org/10.1080/10508406.2019.1666011

Lee, V. R., Pimentel, D. R., Bhargava, R., & D'Ignazio, C. (2022). Taking data feminism to school: A synthesis and review of pre-collegiate data science education projects. *British Journal of Educational Technology*, *53*(5), 1096–1113. https://doi.org/10.1111/bjet.13251

Lee, V. R., & Wilkerson, M. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects* [Commissioned paper]. National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6–12.

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. https://doi.org/10.3102/0013189X211048810

Lehrer, R., & English, L. (2018). Introducing children to modeling variability. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 229–260). Springer. https://doi.org/10.1007/978-3-319-66195-7_7

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, *21*(2), 116–133. https://doi.org/10.1007/BF03217548

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In *Thinking with data* (pp. 163–190). Psychology Press.

Manz, E. (2015). Resistance and the development of scientific practice: Designing the mangle into science instruction. *Cognition and Instruction*, *33*(2), 89–124. https://doi.org/10.1080/07370008.2014.1000490

Matuk, C., DesPortes, K., Amato, A., Vacca, R., Silander, M., Woods, P. J., & Tes, M. (2022). Tensions and synergies in arts-integrated data literacy instruction: Reflections on four classroom implementations. *British Journal of Educational Technology*, *53*(5), 1159–1178. https://doi.org/10.1111/bjet.13257

Mike, K., & Hazzan, O. (2022). Machine learning for non-majors: A white box approach. *Statistics Education Research Journal*, *21*(2), Article 10. https://doi.org/10.52041/serj.v21i2.45

Mike, K., Kimelfeld, B., & Hazzan, O. (2023). The birth of a new discipline: Data science education. *Harvard Data Science Review*, *5*(4). https://doi.org/10.1162/99608f92.280afe66

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options.* National Academies Press.

Nolan, D., & Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, *64*(2), 97–107. https://doi.org/10.1198/tast.2010.09132

Noll, J., Kazak, S., Zapata-Cardona, L., & Makar, K. (2023). Introduction to rethinking learners' reasoning with nontraditional data. *Teaching Statistics*, *45*(S1), S1–S4. https://doi.org/10.1111/test.12350

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning*, *5*(2–3), 131–156. https://psycnet.apa.org/doi/10.1207/S15327833MTL0502&3_02

Philip, T. M., Schuler-Brown, S., & Way, W. (2013). A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning*, *18*(3), 103–120. https://doi.org/10.1007/s10758-013-9202-4

Philip, T. M., Olivares-Pasillas, M. C., & Rocha, J. (2016). Becoming racially literate about data and data-literate about race: Data visualizations in the classroom as a site of racial-ideological micro-contestations. *Cognition and Instruction*, *34*(4), 361–388. https://doi.org/10.1080/07370008.2016.1210418

Pickering, A. (1995). *The mangle of practice: Time, agency, and science*. University of Chicago Press.

Radinsky, J., & Tabak, I. (2022). Data practices during COVID: Everyday sensemaking in a high-stakes information ecology. *British Journal of Educational Technology*, *53*(5), 1221–1242. https://doi.org/10.1111/bjet.13252

Reinhart, A., & Genovese, C. R. (2021). Expanding the scope of statistical computing: Training statisticians to be software engineers. *Journal of Statistics and Data Science Education*, *29*(Suppl. 1), S7–S15. https://doi.org/10.1080/10691898.2020.1845109

Roberts, J., & Lyons, L. (2020). Examining spontaneous perspective taking and fluid self-to-data relationships in informal open-ended data exploration. *Journal of the Learning Sciences*, *29*(1), 32–56. https://doi.org/10.1080/10508406.2019.165131

Rosenberg, J. M., Edwards, A., & Chen, B. (2020). Getting messy with data: Tools and strategies to help students analyze and interpret complex data sources. *The Science Teacher, 87*(5), 30–34. https://doi.org/10.2505/4/tst20_087_05_30

Rosenberg, J. M., Kubsch, M., Wagenmakers, E.-J., & Dogucu, M. (2022). Making sense of uncertainty in the science classroom: A Bayesian approach. *Science & Education*, *31*, 1239–1262. https://doi.org/10.1007/s11191-022-00341-3

Rosenberg, J. M., Schultheis, E., Kjelvik, M., Reedy, A., & Sultana, O. (2022). Big data, big changes? The technologies and sources of data used in science classrooms. *British Journal of Educational Technology, 53*(5), 1179–1201. https://doi.org/10.1111/bjet.13245

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. https://doi.org/10.1080/10508406.2019.1705665

Sander, I. (2020). What is critical big data literacy and how can it be implemented? *Internet Policy Review, 9*(2). https://doi.org/10.14763/2020.2.1479

Schanzer, E., Pfenning, N., Denny, F., Dooman, S., Politz, J. G., Lerner, B. S., Fisler, K. & Krishnamurthi, S. (2022). Integrated data science for secondary schools: Design and assessment of a curriculum. In L. Merkle & M. Doyle (Eds.), *SIGCSE 2022: Proceedings of the 53rd ACM Technical Symposium on Computer Science Education* (Vol. 1, pp. 22–28). ACM. https://doi.org/10.1145/3478431.3499311

Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2021). Data science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics and Data Science Education, 29*(Suppl. 1), S40–S50. https://doi.org/10.1080/10691898.2020.1851159

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

Shreiner, T. L., & Guzdial, M. (2022). The information won't just sink in: Helping teachers provide technology-assisted data literacy instruction in social studies. *British Journal of Educational Technology, 53*(5), 1134–1158. https://doi.org/10.1111/bjet.13255

Stornaiuolo, A. (2020). Authoring data stories in a media makerspace: Adolescents developing critical data literacies. *Journal of the Learning Sciences*, *29*(1), 81–103. https://doi.org/10.1080/10508406.2019.1689365

Taylor, K. H., & Hall, R. (2013). Counter-mapping the neighborhood on bicycles: Mobilizing youth to reimagine the city. *Technology, Knowledge and Learning*, *18*, 65–93. https://doi.org/10.1007/s10758-013-9201-5

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, *12*(2), 147–169. https://doi.org/10.1007/BF03217081

Vacca, R., DesPortes, K., Tes, M., Silander, M., Matuk, C., Amato, A., & Woods, P. J. (2022). "I happen to be one of 47.8%": Social-emotional and data reasoning in middle school students' comics about friendship. In S. Barbosa & C. Lampe (Eds.), *CHI 2022: Proceedings of the CHI Conference on Human Factors in Computing Systems*. Article 316. ACM. https://doi.org/10.1145/3491102.3502086

Vance, E. A., Glimp, D. R., Pieplow, N. D., Garrity, J. M., & Melbourne, B. A. (2022). Data science in 2020: Computing, curricula, and challenges for the next 10 years. Integrating the humanities into data science education. *Statistics Education Research Journal, 21*(2), Article 9. https://doi.org/10.52041/serj.v21i2.42

Warren, A. R. (2020). Impact of Bayesian updating activities on student epistemologies. *Physical Review Physics Education Research, 16*(1), Article 010101. https://doi.org/10.1103/PhysRevPhysEducRes.16.010101

Watson, J. M., & Mortiz, J. B. (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgments. *Journal for Research in Mathematics Education, 34*(4), 270–304. https://doi.org/10.2307/30034785

Welser, H. T., Smith, M., Fisher, D., & Gleave, E. (2008). Distilling digital traces: Computational social science approaches to studying the internet. In *The SAGE handbook of online research methods* (pp. 116–141). SAGE. https://doi.org/10.4135/9780857020055

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Wilkerson, M., Finzer, W., Erickson, T., & Hernandez, D. (2021). Reflective data storytelling for youth: The CODAP story builder. In M. Roussou & S. Shahid (Eds.), *IDC '21: Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (pp. 503–507). ACM. https://doi.org/10.1145/3459990.3465177

Wilkerson, M. H., & Laina, V. (2018). Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM – Mathematical Educaiton, 50*(7), 1223–1235. https://doi.org/10.1007/s11858-018-0974-9

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. https://doi.org/10.1080/10508406.2019.1705664

Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*(1), 165–181. https://doi.org/10.1080/10508406.2019.1705678

Zimmermann-Niefield, A., Turner, M., Murphy, B., Kane, S. K., & Shapiro, R. B. (2019). Youth learning machine learning through building models of athletic moves. In J. A. Fails (Ed.), *IDC '19: Proceedings of the 18th ACM International Conference on Interaction Design and Children* (pp. 121–132). ACM. https://doi.org/10.1145/3311927.3323139

Footnotes

1. We note there have been other special issues—recently Noll et al.’s (2023) special issue on *Rethinking Learners' Reasoning With Nontraditional Data* that was published after we began this manuscript and analysis. ↩

2. We note that *Journal of the Learning Sciences* special issue articles were published in the first issue of 2020, which may explain why these were not cited by any of the articles in it. ↩

3. Another paper by Philip and colleagues—Philip et al. (2013)—was also cited by four articles. ↩

4. Because there were seven articles tied for the third most-cited, we only included the two most-cited articles in this figure. ↩