Making Data Science Count In and For Education

Joshua M. Rosenberg
University of Tennessee
jmrosenberg@utk.edu

Michael Lawson
University of Southern California
lawsonm@usc.edu

Daniel J. Anderson
University of Oregon
daniela@oregon.edu

Ryan Seth Jones
Middle Tennessee State University
ryan.jones@mtsu.edu

Teomara Rutherford
University of Delaware
teomara@udel.edu

Abstract

New data sources and analytic techniques have enabled educational researchers to ask new questions and work to address enduring problems. Yet, there are challenges to those learning and applying these methods. In this chapter, we provide an overview of a nascent area of both scholarship and teaching, educational data science. We define educational data science as the combination of capabilities related to quantitative methods in educational research, computer science and programming capabilities, and teaching, learning, and educational systems. We demonstrate that there are two distinct—but complementary—perspectives on educational data science in terms of being both in education (as a research methodology) and for education (as a teaching and learning content). We describe both of these areas in light of foundational and recent research. Lastly, we highlight three future directions for educational data science, emphasizing the synergies between these two perspectives concerning designing tools that can be used by both learners and professionals, foregrounding representation, inclusivity, and access as first-order concerns for those involved in the growing community, and using data science methodologies to study teaching and learning about data science. We highlight the potential for the growth of educational data science within learning design and technology as situated with the broader data science domain and in education more broadly.

Introduction

Today, data analysts and programmers leverage data to inform and even transform many aspects of contemporary life, from policing and insurance to media and advertising (O’Neill, 2016). In education, teachers, administrators, and policymakers use data to understand educational processes and outcomes, such as student learning (Buckingham Shum et al., 2013; Datnow & Hubbard, 2015; Moore & Shaw, 2015) and the presence of inequities (Reardon et al., 2019). Parents, too, use data to guide decisions about where to live and send their children to school (Hasan & Kumar, 2019). Students use data in their classes to learn about their communities (Wilkerson & Laina, 2018), social issues (Gutstein, 2017), and scientific phenomena (Lehrer & Schauble, 2004). With all of this in mind, it is evident that the availability of data impacts both contemporary life and teaching and learning in educational systems.

The role of data is especially germane to those studying learning design and technology (LDT) and related fields, such as the learning sciences (e.g., Wilkerson & Polman, 2019) and educational technology (e.g., Hillman & Säljö, 2016). LDT scholars have been at the forefront of efforts to identify ways to address new and long-standing questions using novel data sources. For example, LDT researchers have found ways to utilize novel datasets from social media to understand the role of social media in teaching and learning (Coughlan, 2019; Greenhalgh et al., 2020; Kimmons & Smith, 2019; Romero-Hall et al., 2018), use telemetric data collected as students interact with educational technology to gain insight into students’ motivation and learning (e.g., Bernacki et al., 2015; Peddycord-Liu et al., 2018; Rodriguez et al., 2019), and use datasets from wearable devices to engage and understand the learning of K-12 students about data analysis and interpretation (Lee et al., 2015).

However, the LDT field’s creative integration of novel datasets also brings with it new challenges. For instance, issues related to student privacy and the length of time data can be stored are essential concerns to be raised, especially as new and expansive forms of data tracking are becoming commonplace (Leibowitz, 2018). Some issues are social, such as how widely-accessible data should be, and how data are used by educational stakeholders as part of improvement processes, rather than merely for evaluation (American Educational Research Association, 2015). Furthermore, when it comes to state-of-the-art uses of educational data, scholars who apply data science as a methodology often carry out their work in isolation from those who study how to help students learn to analyze new data sources. As a consequence, scholars who work on one of these different topics may talk past the other—to the detriment of both and this growing area of wor.

This chapter, then, is intended to articulate one view of the role of data in LDT and the broader field of education. We do this through the lens of *data science*, because data science, which we define in the next section, aligns with the types of research involving digital technologies that are already being carried out in LDT, and because data science can distinguish newer ways in which data is used in education from those that are commonplace. The specific aims of this chapter, then, are twofold:

- To convey a view of data science and the state of educational data science. Doing so involves being precise about what data science is—and also what it is not. We do this in the *Defining Data Science* section of this chapter.
- To articulate what we view as two common, but commonly confused, perspectives for how data science can be applied in education. The first is as a research method—what we refer to as data science in education. The second is as a context for teaching and

learning—data science for education. We do this in the *Intersections of Data Science and Education* section.

Defining Data Science in Education and Data Science for Education

For our purposes, we adopt a similar definition of data science that has been used in computer science and engineering, which defines *data science* as the intersection of a) the application of mathematics and statistics, b) computer science and programming techniques, and c) knowledge about a particular discipline (Conway, 2010; Vanderplass, 2016). Here, each dimension of data science is of the same importance—the combination of any two without the third defines a related activity (such as machine learning or educational technology) that is not directly representative of data science.

Applying the above definition of data science to education can illustrate how data science is similar to and different from other types of work, as represented in Figure 7.1.

[INSERT FIGURE 7.1 ABOUT HERE]

In this description, quantitative methods are represented by the combination of disciplinary knowledge and math and statistics; machine learning is represented by the combination of computer science and math and statistics; and, educational technology is represented by the combination of disciplinary knowledge and computer science and programming. Thus, this definition lays out what data science is, as well as what it is not: From this view, machine learning is not a synonym for data science, but, instead, a part of it. In the field of educational data science, the substantive knowledge that is brought to bear upon questions and topics is specific to knowledge about teaching, learning, and educational systems. In other words, educational data science involves the application of mathematics and statistics, computer science and programming skills, and knowledge about teaching, learning, and educational systems to ask (and answer) questions and pose (and work to solve) problems.

Although we argue that these three core dimensions define educational data science, others have defined data science by the scale of the data, rather than the domains marshaled to work with data (Schutt & O'Neill, 2014). We believe this is a cause of the many data science practices arising as solutions to dealing with large and complex data sets (e.g., audio-visual data). However, the practice of data science does not necessarily involve large and complex data sets. Although many of the data science methodologies may have developed in response to *Big Data* (Gandomi & Haider, 2015), the application of data science methodologies can also extend to data of a much smaller scale—with similar benefits. Programming, for instance, is central to the practice of data science, and researchers who can use programming in the course of analyzing data are better prepared to tackle data-related challenges that emerge in their work no matter the scale. As two examples, programming and the application of computer science-related capabilities can facilitate the organization and coding of survey-based data, or make it easy to explore textual data through the use of Natural Language Processing techniques. This emphasis away from the scale of the data used (and focus toward programming) has implications for the community of data scientists. For example, there is a vast and rapidly-growing network of individuals supporting open source work in data science (Augur, 2016; Gutierrez, 2015), for which programming is helpful—or even necessary to ensure the trustworthiness and the reproducibility of analyses at any scale (Lowndes et al., 2017). Engaging in a more programmatic approach to analyzing data, regardless of its scale, can broaden participation in this community, which can contribute to a more open, transparent, and reproducible research practice (Lowndes et al., 2017). In summary, we use this definition to argue that data science can be seen as a field that extends beyond the analysis of Big Data—and includes, but is broader than, machine learning, quantitative methods in educational research, and educational technology.

The Intersections of Data Science and Education

In the previous section, we defined data science as the intersection of capabilities in three domains—mathematics and statistics, computer science and programming, and teaching, learning, and educational systems—with the disciplinary expertise being focused around knowledge about teaching, learning, and educational systems. In this section, we distinguish between two perspectives on how data science intersects with education.

The first perspective concerns the application of data science to answer educational questions or to solve educational problems, data science as a research methodology. This is what we refer to as *data science in education*. The second perspective relates to data science as a context for teaching and learning; data science as a domain, akin to science or mathematics education. We refer to this as *data science for education*. Both perspectives are described below.

Data Science in Education: Data Science as a Methodology

Data science in education, then, is specifically oriented around researching teaching, learning, and educational systems through the lens and techniques of data science. For those using data science in education as a methodology, a distinctive consideration is how the research process typically proceeds. In non-data science research, the researcher begins with a question, framed in a theoretical or conceptual framework (Booth et al., 2003). However, in the application of data science in education, scholars may begin with a theory-driven question, but allow the data to guide their inquiry through exploratory data analysis; moreover, a dataset can be new enough that its description alone can serve as a novel contribution. For example, Kimmons and Smith (2019) reported descriptive data on the accessibility of the websites of K-12 schools in the United States. There is also another way that the data can guide research activity: A compelling dataset can serve as a context for the generation of new questions. For instance, data generated as students interact with educational games highlight questions about the nature of student decision-making in a context different than in a typical face-to-face classroom (e.g., Liu et al., 2017). Thus, one consideration for those doing data science in education has to do with considering which data sources are best to answer specific questions (in a top-down manner) and what questions can be answered with preexisting data (in a more bottom-up manner). Regardless of the data science methods used, a sound educational theory must inform either the *questions* asked in top-down approaches or the *interpretation* of data-driven insights gained from bottom-up approaches.

Related to the usefulness of new datasets, a common way that researchers have carried out data science in education is by *combining disparate sources of data* to explore novel research questions. As examples, Kelchen, Rosinger, and Ortagus (2019) demonstrated how data on state-level educational policies in the United States could be joined to data on student outcomes to compare the effects of different policies between states, and Rosenberg et al. (2016) combined data on how many public school teachers were employed in each state in the United States with social media data to understand the activity of participants in one of 47 state-based educational Twitter hashtags. These new combined datasets can give rise to exploratory bottom-up analyses as well; differences between states in participants' activity sparked a study of the activity of regularly-occurring Twitter chats (Greenhalgh et al., 2020). Such dataset combinations may require new types of skills from the dimensions of computer science and statistics to integrate and analyze the data correctly, and will likely also require the addition of researchers with more diverse knowledge bases relating to teaching, learning, and educational systems than are needed by single-data-type studies.

Using new data sources presents opportunities but also challenges, and these challenges may necessitate the development of new methods to suit the data at hand. For example, Anderson et al. (2020) used Natural Language Processing techniques on the text of state-wide science standards to provide content-related validity evidence for science education assessment items. The development of these types of new methods is especially relevant for data sources that have traditionally been analyzed using qualitative methods: Analyzing audio and visual data, for instance, requires deciding not only what data to model and how to model it, but also to decide what the unit of analysis in audio-visual data is and how to create variables (Bosch et al., 2018; D'Angelo et al., 2019). Further, such methods present challenges regarding the nature of how algorithms to process language data, especially of those from marginalized groups--care must be taken in specifying training data and creating algorithms that do not themselves reproduce existing inequities (Mayfield et al., 2019; Zou & Schiebinger, 2018).

In addition to the aforementioned methodological challenges, there are more foundational challenges presented by ready access to data. Social media provides an example of this tension. Although social media can meet the professional learning-related needs of educators (Greenhalgh & Koehler, 2017; Trust et al., 2016), teach students to write (Galvin & Greenhow, 2020), and facilitate communication between those enrolled in graduate programs (Romero-Hall, 2017; Rosenberg et al., 2016), it also can represent the exploitation of users' data. These issues are not distinct to social media platforms; Morris and Stommel (2017) describe how the terms of service for the popular plagiarism-detection service TurnItIn allows the company to own the license for all of the student papers submitted to it, and Rubel and Jones (2016) raise key questions for researchers and analysts using administrative (e.g., student grades, test scores) and learning management system data in light of increasingly ubiquitous applications of learning analytics in post-secondary educational institutions that may bely students' reasonable expectations of privacy. These questions are pressing, and scholars are working to address them through, for instance, developing values-driven learning analytics approaches (Chen & Zhu, 2019), ethical uses of artificial intelligence and machine learning that recognize the potential for ingrained biases (Greene et al., 2019), and examining how teachers can prepare students to protect themselves online (Krutka et al., 2019); those using data science methodologies in education should consider these issues and nascent solutions to them in the course of carrying out their work.

In summary, data science in education is the perspective of educational data science that concerns applying the dimensions of data science to educational research: studies about teaching, learning, and educational systems. This area of work is relatively new, but can be characterized by describing compelling datasets, combining different data sources to create new (and useful) sources of data, and developing new research methods that are suited to the kinds of data—such as text and audiovisual data—increasingly brought to bear upon educational questions and problems. Although we are optimistic about this growing application of data science, we also describe how the use of large, often unobtrusively collected data sources highlights the importance of considering privacy of those from whom the data is collected as well as broader ethical and equity questions about how new methods are developed and applied.

Data Science for Education: Data Science as a Teaching and Learning Context

Data science for education pertains to the teaching and learning of data science and the concepts, people, and resources that support it. The use of the word ‘education’ here might instill visions of K-12 classrooms, but we take the view that data science education is not restricted to any particular educational context, but, instead, is defined by the development of an individual’s work with data (Wise, 2019). From this perspective, data science education is an expansive domain that includes teaching and learning data science in different contexts (e.g., K-12, post-secondary, industry, online, and informal settings), and examples of data science education include graduate seminars on data science methods (Schneider et al., 2020), workshops and training (Anderson & Rosenberg, 2019), and K-12 courses that engage students in working with data. These K-12 courses are often situated in mathematics and science content or classes (e.g., Hancock et al.,1992; Lehrer & Schauble, 2004; Lee et al., 2015), but occasionally in other content areas (e.g., social studies; Drier & Lee, 2008; Lehrer & Romberg, 1996).

Data science for education spans not only a diverse set of contexts but also a different set of ideas and practices. Supporting students in growing their understanding and competency in what professional data scientists know and do is a complicated endeavor; this endeavor is further complicated by the need to consider the developmental trajectories for learners along and across each of the dimensions that make up data science. There is past research that is focused on one or more of the components of data science education—especially quantitative methods (the intersection of math and statistics education)—but there is less research that recognizes and encompasses the intersection of all three components.

As an example of the complexity of data science for education, consider two extreme ends of a hypothetical learning progression (Alonzo & Gotwals, 2012) for creating models of data: learning about algorithms as step-by-step instructions to carry out a classroom task in the elementary grades, and building a machine learning-based classification model to predict water quality in an undergraduate-level class. To progress along this path, the student's knowledge of algorithms in early grades needs to grow into ideas about statistics, probability, and modeling (Lehrer & English, 2018). These ideas will need to be coordinated with ideas about computer hardware and software, and proficiency within particular technologies and even programming languages to leverage the power of computing for analyzing large data sets and with domain knowledge specific to the study of water quality.

What is more, students' engagement with these topics needs to productively resemble the ways professionals engage with them (e.g., Jones et al., 2017). Other fields, such as mathematics, have applied ideas regarding professional socialization and mathematics practices to the development of standards for learners (Cuoco et al., 1996; National Council for Teachers of Mathematics, 2000). Disciplinary identity is a necessary prerequisite to such a practice. This is why data science education must be conceived as a meta-discipline with a disciplinary identity beyond its component parts. This will involve coordinating research at the K-12 and undergraduate levels and between mathematics education, statistics education, science education, and computer science education. This will also include creating (and researching) opportunities for data science learners to use statistical and data science-related tools that are designed not only for learning but also for professional data science practice (McNamara, 2019; Rosenberg et al., 2020), even if, at first, learners must use tools designed for professionals in a more constrained way. Learners must also be socialized into the conventions of data scientists that go beyond tools: conventions such as dispositions toward open science and privacy and ethical issues.

This socialization can start early—even young children can learn to work with data in ways that engage the synergies between the three dimensions of data science and that reflect the professional dispositions of data scientists. Lehrer and Schauble (2004) describe an instance in which late elementary students investigate plant growth through *data modeling* as a way to understand the statistical concepts of variation and distribution. In such situations, data modeling can serve as an organizing set of practices for engaging in inquiry in science and mathematics learning (Lehrer & Schauble, 2015). As learners encounter and generate data, they can be supported to see and use data visualization, statistics, and models as tools to create new knowledge about the natural world (e.g., Arnold et al., 2018; Konold & Pollatsek, 2002; Lehrer & Romberg, 1996). In these examples, students' work with data is used to support the learning of domain-specific content; however, meaningful engagements with data can themselves be their own end, operating as part of a data science education that can be applied beyond specific disciplinary content. This goal aligns with and complements the increasingly relevant constructs undergirding computational thinking, a set of competencies and dispositions that leverage the affordances of computational processes to solve problems and express ideas (Papert, 1996; Wing, 2006). Becoming proficient in working with data can provide learners with an increasingly in-demand capability, as the number of occupations, from education to entrepreneurship, that demand or involve taking action based on data skyrocket (Wilkerson & Fenwick, 2017). Additionally, becoming data fluent can be personally empowering, because of the parts of our lives—from paying energy bills to interpreting news articles—that use data. Although these examples suggest fruitful points of coordination for integrating developmental trajectories related to data science education, much more work is needed to envision what a productive data science education might look like.

In summary, data science for education is a perspective focused on teaching (and learning) how to analyze data in ways akin to how data scientists make sense of data. Teaching and learning data science is challenging, in part because there are three distinct sets of capabilities comprising data science: math and statistics, computer science and programming, and knowledge of a specific domain. Moreover, although there are examples of data science-related research at the K-12 and postsecondary levels, much of the existing research is grounded in other disciplines (e.g., statistics or science education). Establishing data science education as a scholarly discipline in and unto itself will be necessary for the practices and dispositions of data scientists to proliferate. This need presents both challenges and opportunities for those researching this area. Given their disciplinary knowledge of teaching, learning, and educational systems, those actively engaged in *data science in education* may be uniquely positioned to communicate how their research practices can be applied to *data science education*.

Discussion: Three Synergies and Future Directions for Educational Data Science

In this chapter, we have sought to address two aims, defining educational data science through its focus on capabilities related to computer science and programming, math and statistics, and teaching, learning, and educational systems, and articulating two perspectives on data science within the field of education, data science in education and data science for education. As we conclude this chapter, we would like to consider how these two perspectives on data science within the field of education may work better together than in isolation. In particular, we believe three synergies that propel educational data science forward.

Our first synergy comes from considering together the software tools for conducting data science research and those for teaching and learning data science. Historically, scholars have found these to be separate (Gould et al., 2018). For example, tools for professionals, such as R, have emphasized their performance (R Core Team, 2020). In contrast, those for learners, such as the Common Online Data Analysis Platform (CODAP), have emphasized their ease-of-use

(Common Online Data Analysis Platform, 2014). This delineation contributes to an issue: Learners eventually require functionality that the tool they have used does not provide, whereas the tools used by professionals remain challenging to begin to use. McNamara (2019) recommends that developers of statistical tools recognize that individuals analyzing data are likely to use different tools over time. So it is necessary to "build (either technically or pedagogically) an onramp toward the next tool" (p. 382). In this way, those designing (and studying the impacts of) tools for learners can be informed by the high-performing software used by professional statisticians and data scientists. Also, those developing (or improving) tools such as R can expand their user base by considering how tools for learners make use of their lower barriers to entry. The tidyverse set of R packages is an example of a statistical software tool that is both accessible and performant (Wickham et al., 2019). In the realm of programming and computer science, Scratch (Resnick et al., 2009) is another example of a low-barrier, but high-ceiling, tool. Although tidyverse and Scratch are promising examples, more work is needed to develop a smooth pathway of learning from entry-level to sophisticated, professional tools; such a path is necessary to support the consistent identity of data scientists across the learning trajectory.

The second synergy concerns a focus on representation, inclusivity, and access. Issues of equity are deeply entwined with issues of education. Similarly, the use of data and the practice of data science is inherently a political one (Green, 2018). As such, the community of educational data science must be representative of the students it serves. As an emerging field, data science has the opportunity to build a culture that emphasizes representation from the start; indeed, there have been calls to prioritize diversity within data science more broadly (e.g., Berman & Bourne, 2015). However, as data science draws heavily on its component dimensions, starting "from scratch" is mostly an illusion--the fields of math and statistics and computers and programming are already overwhelmingly male and white (Fisher et al., 1997; Lewis et al., 2019), and this likely spills over into data science. As a response, members of marginalized groups have organized to improve diversity in specific data science platforms (e.g., R-Ladies Global[1] and pyladies[2]) and the use of data for racial justice (e.g., Data for Black Lives[3]). Data science cannot rest on its status as a new field to absolve itself of marginalizing individuals from non-dominant groups. In essence, steps are needed to increase representation for a robust data science community.

To build an inclusive and representative data science, there must be broad access to developing expertise in data science and its component domains. Within educational data science, the idea of access includes enabling those with a deep grounding in educational disciplinary knowledge to develop expertise within the other data science components. Some educational graduate students have strong statistics-related capabilities. Still, there is substantial variability across sub-fields: Students in curriculum and instruction and teacher education, for example, may have fewer requirements and expectations related to statistics and quantitative methods than those in education policy or educational psychology. Educational graduate students may also have had limited experience with (and formal educational experiences in) programming. For educational researchers and others using data science methods in education—including data analysts, administrators, and educators—access to data science requires opportunities to learn how to program and apply programming skills in the context of using quantitative methods to ask education-related questions and solve education-related problems.

[1] https://rladies.org/

[2] https://www.pyladies.com/

[3] http://d4bl.org/

Learning to program may be most fruitful if learning opportunities are created either by those with experience and expertise in education (e.g., Anderson's [2020] courses; Bovee et al.'s [2020] book, *Data Science in Education Using R*), or through collaboration and joint training opportunities (e.g., university Data Science centers, such as that at the University of Delaware,[4] Nosek et al.'s [2019] proposed *STEM Education Research Hub* focusing on building the capacity of educational researchers to use new research practices, many of which involve programming). In sum, for educational data science to successfully expand as a discipline, those already involved in it must think carefully about who is welcomed into it, and how to recognize and invite the expertise of all of those who wish to be involved.

A final synergy concerns the application of data science to itself as a discipline to understand how data science is taught and learned. For example, the tidycode R package (McGowan, 2019) is a data science tool that can be used to analyze the R code of those learning about data science: It could be used to understand, for example, how the breadth of the code someone writes (e.g., code not only for creating visualizations but also to prepare data and to use statistical models) expands over the semester for a data science class. As another example, much of the research on how data science is taught and learned uses qualitative research methods (Lehrer & Schauble, 2015); audio and visual data from data science education classes or workshops could also be analyzed using Natural Language Processing techniques to better understand the experiences of teachers and learners and to improve how data science is taught and learned. As with all data science research, and as we have argued above, such bottom-up and novel methods should be interpreted in light of theory and insights gleaned from prior and concurrent research using more traditional methods.

[4] https://dsi.udel.edu/

Conclusion

In this chapter, we sought to elucidate the importance of educational data science (how it counts) by defining it in terms of the intersection of math and statistics, programming and computer science, and teaching, learning, and educational systems, and articulating two (related) perspectives, data science in education (data science as a distinctive research methodology characterized by considering top-down and bottom-up research approaches and new sources of data and methods), and data science for education (data science as complex teaching and learning context characterized by a diverse set of ideas and practices and the need to establish a new field of study).

As being able to understand and work with data continues to grow as a source of power (and empowerment) in our society, researchers in LDT and the broader field of education have a responsibility and great potential to advance the field of data science. Accordingly, we described synergies concerning the creation of tools that can be used by both learners and professionals, representation, inclusivity, and access as first-order concerns for those involved in educational data science, and turning data science as a methodology upon itself to study teaching and learning about data science.

Inquiring about and using data is not only something done by researchers or data analysts, but also comprises a set of practices being taken up more broadly by citizens to inform decision making (O’Neill, 2016). Increasingly, those who hold the data hold power; data scientists are key players in social and educational change. Researchers in LDT and the broader field of education, we believe, have a unique role in applying data to compelling social issues and in understanding and molding the knowledge of future data scientists.

References

Alonzo, A. C., & Gotwals, A. W. (Eds.) (2012). *Learning progressions in science: Current challenges and future directions.* Rotterdam: Sense Publishing.

American Educational Research Association (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448-452.

Anderson, D. J., & Rosenberg, J. M. (2019). *Transparent and reproducible research with R.* Workshop carried out at the Annual Meeting of the American Educational Research Association. Toronto, Canada.

Anderson, D. (2020). *Data science specialization for UO COE.* Retrieved from: https://github.com/uo-datasci-specialization.

Anderson, D. J., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). Evaluating content-related validity evidence using a text-based, machine learning procedure. *Educational Measurement: Issues and Practice.*

Arnold, P., Confrey, J., Jones, R. S., Lee, H. S., & Pfannkuch, M. (2018). Statistics learning trajectories. In *International handbook of research in statistics education* (pp. 295-326). Springer, Cham.

Augur, H. (2016). *The future of big data is open source.* Retrieved from: https://dataconomy.com/2016/06/the-future-of-big-data-is-open-source/.

Berman, F. D., & Bourne, P. E. (2015). Let's make gender diversity in data science a priority right from the start. *PLoS Biology, 13*(7).

Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. (2015). Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning, 10*(1), 99-117.

Booth, W., Colomb, G., & Williams, J. (2003). *The craft of research*. Chicago: University of Chicago Press.

Bosch, N., Mills, C., Wammes, J. D., & Smilek, D. (2018). Quantifying classroom instructor dynamics with computer vision. In Carolyn Penstein Rosé et al (Eds.) *International Conference on Artificial Intelligence in Education* (pp. 30-42). Springer, Cham.

Bovee, E. A., Estrellado, R. A., Motsipak, J., Rosenberg, J. M., & Velásquez, I. C. (2020). *Data science in education using R*. London, England: Routledge.

Buckingham Shum, S., Hawksey, M., Baker, R. S., Jeffery, N., Behrens, J. T., & Pea, R. (2013). Educational data scientists: a scarce breed. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 278–281). Leuven, Belgium.

Chen, B., & Zhu, H. (2019). Towards value-sensitive learning analytics design. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 343-352). Tempe, AZ.

Common Online Data Analysis Platform [CODAP Computer software]. (2014). The Concord Consortium. Concord, MA. Retrieved from: https://codap.concord.org/

Conway, D. (2010). *The data science venn diagram*. Retrieved from: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Coughlan, T. (2019). The use of open data as a material for learning. *Educational Technology Research and Development*. DOI: 10.1007/s11423-019-09706-y.

Cuoco, A., Goldenberg, E. P., & Mark, J. (1996). Habits of mind: An organizing principle for mathematics curricula. *The Journal of Mathematical Behavior, 15*(4), 375-402.

D’Angelo, C. M., Smith, J., Alozie, N., Tsiartas, A., Richey, C., and Bratt, H. (2019). Mapping individual to group level collaboration indicators using speech data. In *Proceedings of the Computer-Support Collaborative Learning Conference*. Lyon, France.

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record, 117*(4), 1-26.

Drier, H. S., & Lee, J. K. (1999). Learning about climate: An exploration in geography and mathematics. *Social Studies and the Young Learner, 12*(1), 6-10.

Fisher, A., Margolis, J., & Miller, F. (1997). Undergraduate women in computer science: experience, motivation and culture. In *Proceedings of the 28th SIG-CSE Technical Symposium on Computer Science Education* (pp. 106-110). ACM Press: San Jose, CA.

Galvin, S., & Greenhow, C. (2020). Writing on social media: A review of research in the high school classroom. *TechTrends, 64,* 57-69.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Gould, R., Wild, C. J., Baglin, J., McNamara, A., Ridgway, J., & McConway, K. (2018). Revolutions in teaching and learning statistics: A collection of reflections. In *International Handbook of Research in Statistics Education* (pp. 457-472). Springer, Cham.

Green, B. (2018). *Data science as political action: grounding data science in a politics of justice*. arXiv preprint arXiv:1811.03435.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Greenhalgh, S. P., & Koehler, M. J. (2017). 28 days later: Twitter hashtags as “just in time” teacher professional development. *TechTrends, 61*(3), 273-281.

Greenhalgh, S. P., Koehler, M. J., Rosenberg, J. M., & Staudt Willet, B. (in press). Considerations for using social media data in Learning Design and Technology research. In E. Romero-Hall (Ed.), *Research Methods in Learning Design & Technology*. Routledge.

Greenhalgh, S. P., Rosenberg, J. M., Staudt Willet, K. B., Koehler, M. J., & Akcaoglu, M. (2020). Identifying multiple learning spaces within a single teacher-focused Twitter hashtag. *Computers and Education.* https://doi.org/10.1016/j.compedu.2020.103809

Gutierrez, D. (2015). *Open source software fuels a revolution in data science*. Retrieved from: https://insidebigdata.com/2015/03/16/open-source-software-fuels-a-revolution-in-data-science/.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337-364.

Hasan, S., & Kumar, A. (2019). Digitization and divergence: Online school ratings and segregation in america. SSRN. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265316.

Hillman, T., & Säljö, R. (2016). Learning, knowing and opportunities for participation: Technologies and communicative practices. *Learning, Media and Technology, 41*(2), 306-309.

Jones, R. S., Lehrer, R., & Kim, M. J. (2017). Critiquing statistics in student and professional worlds. *Cognition and Instruction, 35*(4), 317-336.

Kelchen, R., Rosinger, K. O., & Ortagus, J. C. (2019). How to create and use state-level policy data sets in education research. *AERA Open*. https://doi.org/10.1177/2332858419873619.

Kimmons, R., & Smith, J. (2019). Accessibility in mind? A nationwide study of K-12 web sites in the United States. *First Monday, 24*(2). DOI: 10.5210/fm.v24i2.9183.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289.

Krutka, D., Manca, S., Galvin, S., Greenhow, C., Koehler, M., & Askari, E. (2019). Teaching "against" social media: Confronting problems of profit in the curriculum. *Teachers College Record, 121*(14).

Lee, V. R., Drake, J., & Williamson, K. (2015). Let's get physical: K-12 students using wearable devices to obtain and learn about data from physical activities. *TechTrends, 59*(4), 46-53.

Leibowitz, A. (2018, September 6). Could monitoring students on social media stop the next school shooting? *New York Times*. Retrieved from: https://www.nytimes.com/2018/09/06/us/social-media-monitoring-school-shootings.html

Lehrer, R., & English, L. (2018). Introducing children to modeling variability. In *International Handbook of Research in Statistics Education* (pp. 229-260). Springer, Cham.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*(1), 69-108.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Education Research Journal, 41*(3), 635-679.

Lehrer, R. & Schauble, L. (2015). Developing scientific thinking. In L. S. Liben & U. Müller (Eds.), *Handbook of Child Psychology and Developmental Science: Cognitive Processes* (Vol. 2, 7th ed., pp. 671-174). Hoboken, NJ: Wiley.

Lewis, C., Shah, N., & Falkner, K. (2019). Equity and diversity. In S. Fincher & A. Robins (Eds.), *The Cambridge Handbook of Computing Education Research* (pp. 481-510). Cambridge, UK: Cambridge University Press.

Liu, Z., Cody, C., Barnes, T., Lynch, C., & Rutherford, T. (2017). The antecedents of and associations with elective replay in an educational game: Is replay worth it? In *Proceedings of the 10th International Conference on Educational Data Mining*. Wuhan, China.

Lowndes, J., Best, B., Scarborough, C., Afflerbach, J., Frazier, M., O'Hara, C., Jiang, N., & Halpern, B. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution, 1*(6), 160-167.

Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., & Black, A. W. (2019). Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 444-460).

McGowan, L. D. (2019). *tidycode: Analyze lines of R code the tidy way* (R package version 0.1.0). Retrieved from: https://CRAN.R-project.org/package=tidycode.

McNamara, A. (2019). Key attributes of a modern statistical computing tool. *The American Statistician, 73*(4), 375-384.

Moore, R., & Shaw, T. (2017). *Teachers' use of data: An executive summary*. ACT. Retrieved from: http://www.act.org/content/dam/act/unsecured/documents/R1661-teachers-use-of-data-2017-12.pdf

Morris, S. M., & Stommel, J. (2017). *A guide for resisting edtech: The case against TurnItIn.* Hybrid Pedagogy. Retrieved from: http://hybridpedagogy.org/resisting-edtech/.

National Council for Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Nosek, B. A., Ofiesh, L., Grasty, F. L., Pfeiffer, N., Mellor, D. T., Brooks, R. E., III, … Baraniuk, R. (2019). Proposal to NSF 19-565 to create a STEM education research hub. https://doi.org/10.31222/osf.io/4mpuc

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown: New York, NY.

Papert, S. (1996). An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning, 1*(1), 95-123.

Peddycord-Liu, Z., Harred, R., Karamarkovich, S. M., Barnes, T., Lynch, C., & Rutherford, T. (2018). Learning curve analysis in a large-scale, drill-and-practice serious math game: Where is learning supported? In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, UK.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. Retrieved from: http://www.R-project.org/.

Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2019). Gender achievement gaps in US school districts. *American Educational Research Journal*, *56*(6), 2474-2508.

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., & Kafai, Y. B. (2009). Scratch: Programming for all. *Communications of the ACM, 52*(11), 60-67.

Rodriguez, F., Yu, R., Park, J., Rivas, M. J., Warschauer, M., & Sato, B. K. (2019). Utilizing learning analytics to map students' self-reported study strategies to click behaviors in STEM courses. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 456-460).

Romero-Hall, E. (2017). Posting, sharing, networking, and connecting: Use of social media content by graduate students. *TechTrends, 61*(6), 580-588.

Romero-Hall, E., Kimmons, R., & Veletsianos, G. (2018). Social media use by instructional design departments. *Australiasian Journal of Educational Technology, 34*(5), 86-98.

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.

Rosenberg, J. M., Terry, C. A., Bell, J., Hiltz, V., & Russo, T. E. (2016). Design guidelines for graduate program social media use. *TechTrends, 60*(2), 167-175.

Rubel, A., & Jones, K. M. (2016). Student privacy in learning analytics: An information ethics perspective. *The Information Society, 32*(2), 143-159.

Schneider, B., Reilly, J, & Radu, I. (2020). Lowering barriers for accessing sensor data in education: Lessons learned from teaching multimodal learning analytics to educators. *Journal for STEM Education Research*. Retrieved from: https://link.springer.com/article/10.1007/s41979-020-00027-x

Schutt, R., & O'Neil, C., (2014). *Doing data science: Straight talk from the frontlines*. O'Reilly Media: Sebastopol, CA.

Shaffer, D. (2017). *Quantitative Ethnography*. Cathcart Press: Madison, WI.

Trust, T., Krutka, D. G., & Carpenter, J. P. (2016). "Together we are better": Professional learning networks for teachers. *Computers & Education, 102*, 15-34.

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media: Sebastopol, CA.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, D., Francois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Muller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to tidyverse. *The Journal of Open Source Software, 4*(43), 1686.

Wilkerson, M., & Fenwick, M. (2017). Using mathematics and computational thinking. In C. Schwarz, C. Passmore, & B. Reiser (Eds.) *Helping students make sense of the world using next generation science and engineering practices* (pp. 181-204). NSTA Press: Arlington, VA.

Wilkerson, M.H., and V. Laina. 2018. Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM, 50*(7): 1223–1235.

Wilkerson, M., & Polman, J. (2019). Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences*. DOI: 10.1080/10508406.2019.1705664.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33-35.

Wise, A. F. (2019). Educating Data Scientists and Data Literate Citizens for a New Generation of Data. *Journal of the Learning Sciences*. DOI: 10.1080/10508406.2019.1705678.

Zou, J. & Schiebinger, L. (2018). AI can be sexist and racist - it's time to make it fair. *Nature, 559*, 324–326.