Learning to Reason with Data: How Did We Get Here and What Do We Know?

Andee Rubin
TERC, Cambridge, MA

Author Note
Correspondence concerning this article should be addressed to Andee Rubin, TERC, Cambridge, MA, 02140. Contact: andee_rubin@terc.edu

To appear in the Journal of the Learning Sciences, January, 2020 as part of a special issue on situated data.

Abstract

In "50 Years of Data Science," Donoho claims that most of the core activities of data science are what statisticians have been doing routinely for their entire careers. Traditional statistics education, however, has primarily focused on decontextualized problems and mathematical proofs, rather than actually reasoning with data the way statisticians do. In spite of the recent uproar about how the current crop of students is totally unprepared to deal with the data deluge of the present and future, teaching students to reason with data is not a totally new enterprise, and a small but insistent statistics education community has been studying the process for decades. In this commentary, I emphasize five critical aspects of working with data that have emerged from this body of work: context, variability, aggregate, visualization, and inference. I believe these will remain relevant in spite of the addition of new techniques to our arsenal of methods for making meaning using data and can form the basis for ongoing collaborations among learning scientists, statistics educators and data science educators.

> We introduce students to good literature in their early years. We do not reserve great literature until they are older – on the contrary, we encourage them to read it or we read it to them. Similarly, we can give young students experience with real mathematical processes rather than save the good mathematics for later. Through collecting and analyzing real data, students encounter the uncertainty and intrigue of real mathematics. (Russell & Corwin, 1989, p. 1)

This quote is from the authors' preface to a set of an elementary school math modules called *Used Numbers*, which provided K-5 students with a rich set of data activities. I begin with it for several reasons. First of all, I love the notion that data collection and analysis are the analogue of good literature, that reasoning with data is more authentic than the ubiquitous "word problems" in the same way that literature is more authentic than Dick and Jane stories. Second, I appreciate the importance of "uncertainty" and "intrigue" in mathematics, which contrast starkly with the "certainty" and bloodlessness with which mathematics is generally introduced to students. Finally, I note the date, 30 years ago. In spite of the recent uproar about how the current crop of students is totally unprepared to deal with the data deluge of the present and future, teaching students to reason with data is not a totally new enterprise, and a small but insistent statistics education community has been studying the process for decades. This commentary provides an introduction to some major themes of that research, in order to provide common ground for conversations between learning sciences researchers and those who have been studying how students learn to reason with data.

While the *Used Numbers* modules were only in print for a few years, many of the activities were subsequently incorporated into the *Investigations in Number, Data, and Space* (2017) K-5 curriculum, which is still in wide use. Why, then, is "data science" considered a new field and "data science education" a topic about which little is known? A recent paper by Donoho (2017) helps explain this potential contradiction. In "50 Years of Data Science," Donoho claims that most of the core activities of data science ("the collection, management, processing, analysis, visualization and interpretation of vast amounts of heterogeneous data", p. 745) are what statisticians have been doing routinely for their entire careers. Traditional statistics education, however, has primarily focused on decontextualized problems and mathematical proofs, rather than actually reasoning with data the way statisticians do. Most datasets used in such statistics classes are small and "clean," often constructed to illustrate a particular mathematical point. The recent emergence and visibility of data science as a field has brought attention to aspects of working with data that have mostly been ignored in statistics classes. Large amounts of data are now available to the general public, so there is increased opportunity—and need—for non-statisticians to engage in data-based reasoning.

There have been exceptions to the decontextualized approach to statistics education: *Used Numbers* (1989) was a prominent example, as was the Exploring Data module of the 1995 *Quantitative Literacy* Series (Landwehr & Watkins, 1995), high school math modules focusing on data and co-written by members of the American Statistical Association and the National Council of Teachers of Mathematics. These curricula, which more accurately reflected statistical practice, have fallen out of favor in the US in the past twenty years. Yet, internationally, there has continued to be a robust statistics education movement and a growing statistics education research community. This community has studied how students from primary school through university, as well as adult workers, reason with data. Their findings are documented in the *Statistics Education Research Journal* and in the *International Handbook of Research in*

Statistics Education (Ben-Zvi, Makar & Garfield, 2018). Of particular relevance are two chapters in that book: Zieffler, Garfield and Fry (2018) provide an insightful description of the history of statistics education, while Petocz, Reid, and Gal (2018) trace how research on the teaching, learning, understanding and using of statistics in both formal and informal contexts has developed over the past 40 years.

In this commentary, I will emphasize five critical aspects of working with data that have emerged from this body of work. I believe these will remain relevant in spite of the addition of new techniques to our arsenal of methods for making meaning using data and can form the basis for ongoing collaborations among learning scientists, statistics educators and data science educators. For shorthand, I designate these topics: context, variability, aggregate, visualization, and inference.

Context

Data are contextualized. George Cobb and David Moore, venerable statistics educators, famously wrote, "There is more than just content that distinguishes statistical thinking from mathematics. Statistics requires a different *kind* of thinking, because *data are not just numbers, they are numbers with a context* (Cobb & Moore, 1977, p. 801, emphasis in original). While I won't argue with this characterization, the papers in this issue push me to ask: "contextualized for whom?" For the people who designed the data collection methods and decided on the purpose of collecting the data, the context is often obvious. But these days, people are more likely to be presented with data someone else has collected, often for a different purpose. Making meaning from someone else's data requires being aware of the measurement process that generated the data – asking, "Who collected these data? When? Where? How? Why?" When confronted with data we didn't collect ourselves, we need to be data journalists, making sure we dig into the circumstances surrounding data collection.

In the papers in this volume, the issue of context plays a significant role in several papers. In Lee and Dubovi (2020/this issue), for example, families brought their intimate knowledge of the context of blood sugar and insulin level measurements to the categorization of test results. A blood sugar level that might be considered "high" in one circumstance might be considered "nothing to worry about" in another, depending on its timing with respect to a meal and insulin dosing. In such a case, the meaning of the number "200" relied on knowing situational details about the measurement, including, but not limited to, the measurement technique (finger prick or constantly-monitoring pump?). The context of data collection was paramount in Kahn (2020/this issue) as well, as students working with a variety of open data sets noted that income data collected by different agencies was clumped differently, making comparisons across datasets difficult, if not impossible. This same paper also highlighted how working with data values requires constant reference back to the context. Was $25,000 the value of a home or a household income? It made a big difference to the story the students were constructing.

Variability

Data always exhibit variability. Going back to Cobb and Moore, we find this succinct statement: "The need for [the discipline of statistics] rises from the *omnipresence of variability*." (Cobb & Moore, 1977, p. 801, emphasis in original). Cobb and Moore go on to note that not only do individuals vary with respect to any measurement, but even repeated measurements on the same individual vary. Data are best considered as distributions, collections of values that are distributed across a set of possible values. Some values are more common, others less so and still others may not occur at all. Statistical methods help us ferret out relationships in these variable data. The process of data analysis can generally be regarded as an attempt to parcel out the variability in a set of measurements, attributing some of the differences in values to discernable causes, while other differences are unexplained. I appreciated the perspective Hardy, Dixon and Hsi (2020/this issue) took on this idea; they pointed out that some portions of data reflect a phenomenon ("fact"), while others are an outgrowth of particularities of data collection ("artifact"). While the idea of considering a single value to have multiple parts with different causes may seem baffling to a pure mathematician, it is fundamental to working with data. The ubiquity of variability also means that conclusions in most data-based reasoning are expressed in probabilistic terms; the words "never" and "always" rarely occur.

Probabilistic thinking, however, is difficult for many people. Kahneman and Tversky's work on people's struggles with probabilistic reasoning (Kahneman, Slovic, & Tversky, 1982) started a line of inquiry that continues today (Kahneman, 2011) and has been an important component of research on statistical reasoning. Kahneman and Tversky (1992) originally identified two heuristics that affect how people approach probabilistic situations: the *representativeness* heuristic (which can lead people to judge the likelihood of an event based on its similarity to the population from which it was drawn) and the *availability* heuristic (which can lead people to overestimate the chance of an event occurring if they recently observed a particularly salient example). Even in an era of "big data," knowing how to handle uncertainty in data will continue to be important. In a review of research on uncertainty in statistics education, Pratt and Kazak (2018) suggest that a modelling perspective on probability, using appropriate technology, holds promise as a way to counteract some of the struggles research has identified. This could be an important area of conversation between the learning sciences, where work on modeling is common, and statistics education research.

Several papers in this volume use data that vary over space and/or time, noting that technology has made these data both more common and more accessible. Certainly, data that students collect by studying their own lives or their own communities are likely to vary over both space and time. Such data can be difficult to analyze, as both temporal and spatial data are auto-correlated, that is, data that are close to one another in time (or space) are more likely to have similar values than those that are far apart in time (or space). Data scientists are still developing methods for dealing with the complex time- and space-dependent information (e.g., climate data) that gives us insight into the future. Data science education needs to keep track of these developments and figure out how to help students approach analysis of temporal and spatial data in statistically valid ways.

Interestingly, this volume's papers don't actually confront these complexities of time and space, in spite of examining data that could be impacted by these quantities. In Van Wart, Lanouette, and Parikh's work (2020/this issue), students collected data at different locations in order to compare air quality at each. Their analysis, however, treated the two transit stations just as "different," without considering their actual locations, how far they were from one another, or whether pollution at one might migrate to the other.

Aggregate

There is a constant and difficult-to-resolve tension between a case-based view of data and an aggregate view of data. This has been a recurring theme in statistics education research and – not surprisingly – it surfaces in several papers in this volume. Many researchers have pointed out that students find it relatively easy to figure out how to interpret a graph if the points represent people and, in particular, if they can find themselves in the graph, as one of the cases. This is a great entry point, but students can also get "stuck" focusing on individual cases and find it difficult to move beyond this case-based view.

Several of the authors referred to a paper that I consider central to the field of statistics education: Konold, Higgins, Russell and Kahlil (2015). This paper delineates four lenses people take on data: 1) a pointer to the data collection activity; 2) individual case values; 3) classifier, grouping values together but without attention to their relationship to the whole; and 4) aggregate, considering values and their frequency. Konold et al. argue that sophisticated thinking about data requires the fourth lens: understanding how a distribution of values can be considered as an aggregate, summarized as a single or small set of values. While finding an aggregate value (how to calculate a mean or median) is often the first statistical calculation students learn, conceptual understandings of *why* one might want to compute an aggregate value, *what* it does – and doesn't – reflect about the distribution it summarizes, and *how* to use an aggregate value to generalize beyond the data at hand are often not addressed.

Roberts and Lyons (2020/this issue) referred to Konold et al.'s case/aggregate continuum in documenting CoCensus visitors' use of personal pronouns in describing the data displays they created. Particularly interesting was their contrast of visitors focusing on individual data points, as in "Alright, so I'm in Manhattan" with clusters of data points, as in "There's a lot of me right there." This characterization of the data seems to me more of a "classifier" usage, rather than a true aggregate that reflects variability. The visitors were noticing the high frequency of a particular data value, but weren't making an explicit connection to other data values or other locations on the map. Konold et al.'s continuum could be helpful going forward in making subtle distinctions in how people regard data in which they locate themselves.

Lee and Dubovi's (2020/this issue) families found themselves in a setting in which aggregate values would seem to be particularly useful, as they were faced with data that changed constantly, from minute to minute and day to day. In spite of this, most of the families reasoned using pre-determined ranges and cut-offs, informed by in-depth contextual knowledge of their child's circumstances, rather than, say, a moving average. Calculating such aggregates is certainly a deterrent, but even when aggregate displays were readily available, few of the families used them.

The Type 1 Diabetes work reminds me of a related body of research about measurement and aggregate thinking in a health context, led by Hoyles and Noss (2002), which examined how pediatric nurses used quantitative reasoning in their jobs. In one study, they observed how these nurses decided when a patient's blood pressure was "outside of the normal range," a judgment that required considering the ideas of average, normal variability and abnormal variability. Hoyles and Noss reported that the nurses had developed a notion of an individual child's average blood pressure, including an awareness of non-worrisome factors (like instrument error) that could lead to unusual readings. This wasn't a mathematical average per se, but rather some idea of a baseline tendency. Yet, it was just what was needed for these nurses to be successful at their job of determining when a child needed medical attention. Interestingly, these nurses felt that a *population* average wasn't a useful idea because they were always judging readings against an individual child's baseline. Further investigation of how aggregate thinking intersects with the need to make individual diagnostic decisions could be an intersection of interest for statistics research and the learning sciences.

Visualization

Easy access to large quantities of data begs for visualization tools, and most of the papers in this issue included at least one example of a data display facilitated by a computer tool. As a contributor to the design of several data visualization tools, though, I wanted to see more examples of ways participants created and interacted with visual displays of data. We know that the specific affordances of individual tools can have a profound effect on how we interpret data. Hardy et al. (2020/this issue) noted that something as seemingly trivial as the way axes scale can radically affect meaning-making. In that paper, when Hope looked at the CO_2 levels over time in her two conditions using a tool that auto-scales axes, she saw the graphs as close to identical and, therefore, decided her experiment had failed to give her insight into plant death. When she visualized them using a different tool, however, she realized that, while the shapes were similar, the values were quite different – and she reversed her conclusion. What other "interface" decisions in data visualization tools might have this kind of unintended consequence?

Several people in the learning sciences community, including myself, have recently been using CODAP (Common Online Data Analysis Platform, 2019), a web-based tool for students in grades 6 and up developed by Concord Consortium (Wilkerson et al., 2018). CODAP has several design characteristics I have come to regard as necessary for tools that support productive thinking with data. CODAP can display data in tables, graphs and maps; a single CODAP document often has a single table, but several graphs and/or maps. The critical feature is that these displays are linked, so that if a user highlights a case or set of cases is in one display, it is highlighted in all the others. Not only does this feature reinforce the idea of a case, but it provides a kind of unity to the different manifestations of a case, a way of remembering that these displays are just lenses on the same set of measurements. Linked representations also provide users with ways to select meaningful subsets of data in one representation (e.g., all the points in a particular region of a map), then examine that group in a linked graph. The study of the learning affordances of CODAP and other data visualization tools could be an arena where learning scientists and statistics education researchers have complementary perspectives that could enrich one another's work.

Computer visualization tools are central to reasoning with data, partly because they automate a lot of the work of designing and producing data representations. On the flip side, though, they hide the mapping process of data values to representational forms that forms the logical basis of data visualization. They also limit users' choices of representational elements to those already included in the software. In order to focus more on the representational process and to give students more opportunities for self-expression, non-computer approaches that position students as creators of data displays can be useful. I consider myself an early adopter and enthusiast of *Dear Data* (Lupi & Posavec, 2016), so I was struck by Stornaiuolo's (2020/this issue) use of the book as an inspiration and a model for her students. *Dear Data* is the story (in the form of an illustrated book) of two graphic artists who crafted a friendship from weekly exchanges of representations of data they collected about their lives during the week, with personal topics ranging from "times I touched my phone" to "clothes in my closet." The book positions the authors – and, by extension, its readers – as producers of both data and data representations. This positioning also runs through several of the papers in this issue. Seeing data as being produced (either by oneself or by other humans with methodologies, purposes, and biases) rather than being collected like so many coins or stamps is an important paradigm shift that will need to be part of effective data education.

The *Dear Data* book also highlights the importance of a legend or key that specifies how data values are mapped to elements of a visualization. Figuring out what a case is, what the associated attributes are and creating a key that communicates clearly to an audience are significant intellectual tasks. We get a glimpse of how students approached these issues in Stornaiuolo's paper, but there is more work to be done to understand how these skills develop.

A related skill that we should be studying is the ability to discern which choices of data representation mapping best reveal patterns in the data.

Inference

In traditional statistics education, inference is a dreaded topic, the bugaboo of many students' experience. The probabilistic, counter-factual nature of statistical inference is notoriously difficult to teach and learn. There are new approaches that have the potential to make statistical inference more accessible, but there is a related basic habit of mind, both more straight-forward and more varied, that we need to include in discussions of data. If we learn anything from examining some data, we must ask: does what we have learned apply beyond this particular dataset? How generalizable is it? What cautions must we take when we apply our conclusions to new data or a new context?

Even before we consider generalizing beyond the data we have, it is worthwhile noting that drawing conclusions that do *not* go beyond the data is complicated as well. For example, students in Kahn's paper (2020/this issue) appeared to approach the census data with a purpose in mind – to create a story that justified their family's move from Origin Place A to Destination Place B. Kahn reported, "when youth encountered similarities, or undesirable differences (i.e., Destination Place B was worse off or no better than Origin Place A), they were typically hesitant to include such comparisons." So, two students looking at the very same data could draw different conclusions, based on the personal perspective and purpose they brought to the task. Not only are data themselves not context-free, but the observations we make of the data and the conclusions we draw are not made without the intrusion of purpose, whether or not it is explicit or conscious.

Some of the biggest and most visible users of data these days are commercial giants such as Amazon or Netflix who are constantly involved in inferential statistics, predicting what each of us wants to buy or watch based on our past choices. Understanding how this process works and how it can go wrong is a critical part of data science education and figuring out how to introduce young students to notions of inference could be a place for statistics educators and learning scientists to collaborate. As a starting point, I offer the framework of "informal statistical inference" that has developed in the statistics education community as a way of describing and studying statistical inference that does not require the complex mathematical machinery of p-values. Makar and Rubin (2018) describe a three-part framework for "informal statistical inference," which has proven useful in working with students as young as elementary school:

- *Inference involves a claim that goes beyond the data at hand.* Inferences are predictive; they make a claim about what we'll find when we look at additional data. While the classic situation involved making a prediction about an entire population given information about a sample, many inferences are actually predictions about the future.
- *Inferences are expressed with a degree of uncertainty.* Because inferences are made on the basis of incomplete information, we cannot make them with certainty. In informal statistical inference, these expressions of uncertainty are not necessarily formal probabilistic statements, but may be less precise or even qualitative.
- *Inferences use data as evidence.* Especially for young students, the relationship between data and conclusion can be elusive. Explicitly tying evidence to conclusions is a skill that needs practice and nurturing.

Food for Thought

The papers in this issue make me confident that the learning sciences research community can make significant contributions to a field I have considered critical for decades. I am also convinced that the statistics education research community, while it has been somewhat under the radar in the learning sciences community, has developed frameworks and insights that learning scientists could find useful in their studies of people producing, representing and analyzing data for personal and community purposes. Some additional questions arose for me as I considered this collection of research, and I include them here as questions we might discuss as a field. Given that over half of the papers reported on out-of-school experiences and two focused on third spaces, what are the implications for the inclusion of data science education in schools? What is the relationship of data science education to mathematics education? What can we learn from these studies about the data-related concepts people find most difficult to comprehend? What can we learn about promising approaches to engaging people productively with these same concepts? What are the most important next research and development steps towards nurturing the growth of a data-empowered citizenry?

References

Ben-Zvi, D., Makar, K. & Garfield, J. (Eds.) (2018). *International Handbook of Research in Statistics Education*. Springer International Handbooks of Education, Springer, Cham.

Cobb, G. & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*:9, 801-823, DOI: 10.1080/00029890.1997.11990723

CODAP (Common Online Data Analysis Platform) [Computer software]. (2019), Retrieved from https://concord.codap.org

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

Hardy, L. Dixon, C. & Hsi, S. (2020/this issue). From data collectors to data producers: Shifting students' relationship to data. *Journal of the Learning Sciences, 29*.

Hoyles, C & Noss, R. (2002). Problematising statistical meanings: A sociocultural perspective. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics (ICOTS 6)*. Cape Town: International Association for Statistics Education. Accessed from https://iase-web.org/documents/papers/icots6/2e3_hoyl.pdf

*Investigations in Number, Data, and Space® (*2017), 3rd ed. Glenview, IL: Pearson.

Kahn, J. (2020/this issue). Learning at the intersection of self and society: The family geobiography as a context for data science education. *Journal of the Learning Sciences, 29.*

Kahneman, D. (2011). *Thinking Fast and Slow*. Farrar, NY: Straus and Giroux.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge, England: Cambridge University Press.

Konold, C., Higgins, T., Russell, S.J., & Kahlil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics*, 8, 305-325.

Landwehr, J. & Watkins, A. (1995). *Exploring Data* (Revised Edition). Palo Alto, CA: Dale Seymour Publications.

Lee V. R., & Dubovi, I. (2020/this issue). At home with data: Family engagements with data involved in type 1 diabetes management. *Journal of the Learning Sciences*, *29* doi: 10.1080/10508406.2019.1666011

Lupi, G. & Posavec, S. (2016). *Dear Data.* Hudson, NY: Princeton Architectural Press.

Petocz, P., Reid, A. & Gal, I. (2018). Statistics education research. In Ben-Zvi, D., Makar, K. & Garfield, J. (Eds.). *International Handbook of Research in Statistics Education*. Springer International Handbooks of Education, Springer, Cham.,71-99.

Russell, S.J. & Corwin, R. B. (1989) *Statistics: The Shape of the Data*. Palo Alto, CA: Dale Seymour Publications.

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In Ben-Zvi, D., Makar, K. & Garfield, J. (Eds.). *International Handbook of Research in Statistics Education*. Springer International Handbooks of Education, Springer, Cham., 261-294.

Pratt, D. & Kazak, S. (2018). Research on uncertainty. In Ben-Zvi, D., Makar, K. & Garfield, J. (Eds.). *International Handbook of Research in Statistics Education*. Springer International Handbooks of Education, Springer, Cham., 193-227.

Roberts, J. & Lyons, L. (2020/this issue). Examining spontaneous perspective taking and fluid self-to-data relationships in informal open-ended data exploration, *Journal of the Learning Sciences*, 29. doi: 10.1080/10508406.2019.1651317

Stornaiuolo, A. (2020/this issue). Authoring data stories: Adolescents developing critical data literacies. *Journal of the Learning Sciences, 29*.

Used Numbers (1989). Palo Alto, CA: Dale Seymour Publications.

Van Wart, S., Lanouette, K., & Parikh, T. (2020/this issue). Third spaces for data science education using participatory digital mapping. *Journal of the Learning Sciences, 29*.

Zieffler, A., Garfield, J., & Fry, E. (2018). What is statistics education? In Ben-Zvi, D., Makar, K. & Garfield, J. (Eds.). *International Handbook of Research in Statistics Education*. Springer International Handbooks of Education, Springer, Cham., 37 – 70.