Title: Using messy, authentic data to promote data literacy and reveal the nature of science

Citation: Schultheis, E. H. and M. K. Kjelvik (2020). "Using Messy, Authentic Data to Promote Data Literacy & Reveal the Nature of Science." The American Biology Teacher 82(7): 439–446.

Authors: Elizabeth H. Schultheis[1,2] and Melissa K. Kjelvik[1,2]

[1]BEACON Center for the Study of Evolution in Action
Michigan State University
East Lansing, MI 48824

[2]W.K. Kellogg Biological Station
Michigan State University
Hickory Corners, MI 49060

Corresponding author: Elizabeth H. Schultheis

eschultheis@gmail.com
W.K. Kellogg Biological Station
3700 E. Gull Lake Drive
Hickory Corners, MI 49060
Phone: (631) 327 - 2870
Fax: (269) 671- 2351

Keywords: data literacy, Data Nuggets, nature of science, first hand data, second hand data, scaffolding, messy data, authentic data

Abstract: Authentic, "messy data" contain variability that comes from many sources, such as natural variation in nature, chance occurrences during research, and human error. It is this messiness that both deters potential users of authentic data, and gives data the power to create unique learning opportunities that reveal the nature of science itself. While the value of bringing contemporary research and messy data into the classroom is recognized, implementation can seem overwhelming. We discuss the importance of frequent interactions with messy data throughout K-16 science education as a mechanism for students to engage in the practices of science such as visualizing, analyzing, and interpreting data. We describe strategies to help facilitate the use of messy data in the classroom, while building complexity over time. We outline one potential sequence of activities, with specific examples, to highlight how various activity types can be used to scaffold students' interactions with messy data.

Introduction

Almost two decades ago, Lynn Steen, president of the Mathematical Association of America, recognized that “the world of the twenty-first century is a world awash in numbers” (NCED, 2001). As we look to the future, data literacy is only becoming more essential as science and society increasingly rely on information found in large datasets (Steen, 1999; NRC, 2003; Manyika et al., 2011). Because science and data are tightly linked (Bowen & Roth, 2005; Speth et al.; 2010), we can weave data seamlessly through K-12 and undergraduate science education and increase student’s exposure to data.

Data literacy is defined as “the ability to understand and use data to inform decisions” (Mandinach & Gummer, 2013) and is an interdisciplinary field lying at the intersection of data science, quantitative reasoning, and authentic context (Kjelvik & Schultheis, 2019). In addition to specific abilities, data literacy is characterized by habits of mind such as curiosity, resiliency, and ethical decision-making (Box 1). Data literacy is becoming more commonplace in formal and informal K-16 education (Konold et al., 2000; Metz, 2008, Speth et al., 2010), and is addressed in K-12 and undergraduate science education reform efforts. These include the Next Generation Science Standards (NRC, 2012; NGSS Lead States, 2013), ACT College Readiness Standards (ACT, Inc., 2014), the new AP Biology Curriculum Framework (The College Board, 2013), Vision and Change (AAAS, 2015), and the American Statistical Association Curriculum Guidelines for Undergraduate Programs in Statistical Science (ASA, 2014). These initiatives highlight several scientific practices into which data can be integrated, including developing students' abilities to analyze and interpret data, use mathematical thinking, and communicate arguments based on evidence (NRC, 2012; NGSS Lead States, 2013).

Current challenges in data literacy

Despite science education reform efforts, the basic skills necessary for data literacy are not yet sufficiently taught in schools. High school graduates lack a proficiency in data use necessary to conduct contemporary research (Hernandez et al., 2012; Strasser & Hampton, 2012) and for a career that involves working with data (Finzer, 2013; Oceans of Data Institute, 2014). The result is a workforce lacking the quantitative abilities desired by employers. According to a recent report, the United States workforce faces a shortage of 1.5 million managers and analysts with the ability to interpret large datasets for the purpose of decision making (Manyika et al., 2011). As stated by Juan LaVista, Principal Data Scientist at Microsoft, "Basic skills in working with data that every person should have are not being taught in K-16 schools. Thus, they are lacking at the highest levels in the broad array of professions that are becoming increasingly data-driven" (Oceans of Data Institute, 2014). Therefore, to prepare today's students for data-intensive careers, training in data literacy needs to be incorporated throughout science education.

Outside of the workforce, students in today's classrooms are the next generation of citizens voting on pressing issues concerning science. The role of data in society is becoming more important as technological advancements continue (Schield, 2004; Wolff et al., 2017; Borges-Rey, 2017). Many global issues are informed by scientific research, and if individuals do not understand the scientific process and the role of scientific data, they will not value research funding, or information collected by the scientific community.

Additionally, the ability to use data for personal decision making is an important skill. Data inform all aspects of everyday life (Mayes et al., 2014), including decisions regarding courses of medical treatment, financial investments or savings strategies, voting and political actions, and food and material consumption. Further, the ability to interpret data, and use them to construct arguments based on evidence, gives individuals the option to advocate for themselves and their communities. Learning how data can be used to persuade, or to identify cases of misuse or misrepresentation, empowers an individual to think freely, question the arguments of others, and make decisions for themselves (Lutskey, 2008; Mayes et al., 2014). Therefore, making these abilities ubiquitous in the general public may help fight inequality in society.

These deficiencies in data and scientific literacy ultimately result in a workforce without the necessary quantitative skills necessary for modern jobs and a public unable to use data in their everyday lives (Steen, 1999; NCES, 2005). Here we discuss why the use of authentic data throughout science education may be a remedy to these challenges. We hypothesize that the strongest learning experiences surrounding data and science literacy arise when students have frequent opportunities to work with authentic, messy data (Schultheis & Kjelvik, 2015). This is due to the inherent qualities of messy data, and their ability to engender unique learning opportunities not found in other resources. However, messy datasets can be quite complex, creating a potential barrier for classroom use (Kjelvik & Schultheis, 2019). To break down this barrier, we highlight techniques to scaffold messy data usage and propose an activity sequence that provides students with repeated practice working with various types of messy data, with increasing complexity over time.

Learning opportunities from the use of messy, authentic data in the classroom

Authentic data result from scientific observations and investigations. These datasets are collected in a variety of ways including by scientists, citizen scientists, sensors and other automated processes, or generated through modeling and simulations. Authentic data are always attached to a context, and the connection a student feels to data may differ based on their ability to find and understand its relevance (Langen et al., 2014). Working with authentic data is engaging for students (Langen et al., 2014) because it allows students to take on the role of a scientist, which may lead to the same sense of awe felt through exploration of unanswered questions and learning something new about the way the world works (Gould et al., 2014). Alternatively, if context is removed by having students explore patterns or trends without meaning, data lose their power to capture the interest and engagement of students (Konold & Higgins, 2003) and students are deprived of the journey of exploring the unknown (Gould et al., 2014). This often occurs when students work with heavily curated examples of data with messiness removed, or fake data generated to illustrate a specific scientific or mathematical concept.

Throughout this paper we use the term “messy data” to represent a particular type of authentic data (Kjelvik & Schultheis, 2019). A key element in messy datasets is variability. The source of this variability comes from both natural variation and systematic or precision error (Gould et al., 2014). These datasets may have missing values due to events that took place during a study, and could contain outliers, unexpected trends, or lack significant results. The interpretation of messy data may or may not support original hypotheses and predictions, but has the potential to inspire additional scientific questions beyond those initially conceived when the study began.

Nature of science

Science is a way of understanding the natural world, and is both an accumulation of knowledge and a way of knowing (NGSS Lead States, 2013 - Appendix H). The overarching goal of science is to investigate the unknown, and the interpretation of authentic, messy data plays an important role in this process. To those unfamiliar with the nature of science, messiness in scientific data, or unexpected results, may lead to distrust in scientific findings; however, it is in fact these aspects that gives science its power. For example, messy data provide unique opportunities to engender connections between a student and the data; a missing data point or outlier in a table can come to life when used for a discussion surrounding failed experimental trials and the personal story that the researcher went through when collecting data. Similarly, results that run contrary to predictions deepen our curiosity about how the world works and motivates scientists to pursue unanticipated research paths and ask new questions. Therefore, an important outcome of science education should be for students to come away with an understanding of the nature of scientific knowledge as not a fixed truth, but something constantly being updated to include recent discoveries (Duschl, 1990; Dasgupta et al., 2014; Strode, 2015).

Research has shown that students benefit from explicit instruction concerning the nature of science (Moss, 2001; Khishfe & Abd-El-Khalick, 2002; Schwartz et al., 2004) and that promoting a student's curiosity from an early age can lead to increased achievement in math and reading (Shah et al., 2018). Educators can use authentic, messy data to introduce the nature of science and promote associated habits of mind (Box 1). For example, highlighting the non-linear, cyclical process of science can help students understand that scientists must often reexamine and revise their thinking about a system before fully understanding it. Additionally, by exploring when to remove outliers from a dataset based on statistical parameters or their biological relevance, instructors can bring up issues of data ethics. Finally, instructors can emphasize the value of focusing on what data as evidence tells us, over trying to confirm previously held beliefs (Hogan & Maglienti, 2001). These types of discussions may lead students to think scientifically and can help normalize the messy aspects of research, resulting in a classroom culture that values uncertainty.

Inquiry investigations are a mechanism for students to put scientific habits of mind into action with first-hand experience collecting data. However, without previous experience with messy data, students may not be familiar with many of the skills and concepts necessary for working with complex data, and therefore may become frustrated if they face them all at once during their first inquiry experience (Kanari & Millar, 2004; Langen et al., 2014). Without prior exposure to messy data and the process of science, students may be led to the misconception that they have "messed up" when they see variation around their sample means or collect data that go against their predictions and does not support their hypothesis (Séré et al., 2001). This leads students to not trust the data they have collected, and leaves them unable to challenge what is accepted in the field or critique the findings of others (Holmes et al., 2015). Students often believe that the data they have collected are of lower quality than those collected by experts in the field (Allie et al., 1998), when in fact data collected by scientists are often as messy as student-collected data (Gould et al., 2014). However, when given opportunities to practice working with messy data before conducting inquiry investigations, students have greater confidence in data they collect themselves and are more likely to challenge an accepted model based on their findings (Holmes et al., 2015).

First- and second-hand data

Scientists use a variety of data types, including data from their own research, collected by their collaborators, and archived in online repositories. Similarly for students, authentic data will ideally come from many sources, including data they collect themselves during inquiry projects;

guided use of online data repositories; reading peer-reviewed journals; or classroom activities designed to scaffold students as they work with data.

These data sources fall into two general categories: first-hand data collected by students directly, and second-hand data obtained by students or teachers from outside sources (NRC, 1996; Palincsar & Magnusson, 2001; Magnusson et al., 2004). Using a variety of data sources during instruction can deepen student understanding of science content (Duschl, 1990). Therefore, when selecting data-centric activities for the classroom, it is important to consider that first- and second-hand data may lead to two different learning experiences for students, and the use of both in the classroom may be complementary (Hug & McNeill, 2008).

When collecting first-hand data, students are better able to question the strengths and weaknesses of the dataset, having directly experienced where uncertainty and variability entered during data collection (Kastens et al., 2015). When working with data they collected themselves students are more likely to see how the source and quality of data are important for what claims can be made, discuss limitations such as measurement error, and cite the sources from which the data came (Hug & McNeill, 2008). In addition, students may feel a personal connection to first-hand data, better understand the real-world significance behind the values, and be able to more easily visualize what the variables represent in the natural world (Hug & McNeill, 2008). First-hand data may, therefore, be particularly helpful when students are learning to be critical users of data. However, first-hand data also come with limitations, such as the types of phenomena that can be studied in a classroom setting and the amount of time required to conduct in-depth investigations (Hug & McNeill, 2008).

When working with second-hand data, students have the opportunity to extend beyond what is possible when working with their own data (Palincsar & Magnusson, 2001). For example, they can explore long-term environmental patterns like climatic variations, or a diverse set of genomes from DNA sequences. These second-hand datasets can supplement first-hand investigations by serving as models of data organization and methods used for data collection (Palincsar & Magnusson, 2001). However, the use of this broader pool of data has some potential drawbacks. For example, when students work with large datasets from online repositories they may lack a full understanding of the variables without proper metadata. Or, students may distrust second-hand data without proper identification of the interest groups and methods behind its collection (Langen et al., 2014; Kastens et al., 2015). Therefore, both first- and second-hand data provide rich opportunities for students, but it is important to explicitly guide students' interactions with various forms of data to draw out the most productive experiences.

Using authentic data in the classroom

The importance of practice and scaffolding

The use of messy data can be a challenge for students of all ages, especially those that have few inquiry or research experiences of their own. To build student comfort and confidence, educators can provide opportunities for repeated exposure to messy data and the research process in multiple settings (Germann & Aram, 1996). A study by Holmes and colleagues (2015) emphasized the importance of repetition - students who were repeatedly asked to make decisions using data showed increased sophistication in their reasoning, were better prepared to identify limitations in data or study designs, and were more likely to propose changes to improve their own investigations. With numerous experiences working through diverse datasets, students will be able to develop the tools and habits of mind to independently use and interpret data (Konold & Higgins, 2003).

Without proper guidance, students often feel overwhelmed when left to independently perform data-centric activities, but too much structure can cause students to lose motivation and the curiosity that originally inspired them (Konold & Higgins, 2003). Scaffolding strategies can be used to support students as they develop their understanding of data-centric practices. Scaffolding is defined as instructional techniques that guide students to greater independence and understanding of concepts and processes. The "fading", or gradual removal, of these scaffolds can build students' abilities to perform tasks on their own (McNeill et al., 2006). Faded scaffolding can help students perform tasks independently and make connections across contexts, and has been shown to be more effective than providing a scaffold and removing it all at once (McNeill et al., 2006). Examples of faded scaffolding strategies for authentic data include (1) providing decision making tools to help students identify appropriate statistics for analyzing data or the selection of the appropriate graph type for data representation (Angra & Gardner, 2016), (2) initially providing, and then slowly removing, graph features when helping students construct graphs (Schultheis & Kjelvik, 2015), or (3) providing a structure for student explanations, ensuring they include all necessary evidence and elements (McNeill et al., 2006).

Features of data-centric activities and example lesson sequence

To help educators categorize and compare qualities of data-centric activities we previously identified a list of features that can be varied to increase complexity in classroom activities using data - selection, curation, scope, size, and messiness (Kjelvik & Schultheis, 2019). In this paper, we focus on the feature of "messiness" and describe a potential sequence of classroom activities to demonstrate one way in which various data-centric activities can be used to scaffold students' interactions with messy data (Table 1). Although explicit instruction is needed to move students from simple to complex interactions with data, there are many diverse paths educators can take. In this section we describe one potential sequence of classroom activities and associated focal topics for each (Activity Type A-E, Table 1).

To begin, students can be introduced to the interpretation of simple data tables or visualizations by examining datasets that have already been curated or graphed for classroom use. These are commonly found in textbooks, lectures, or other educational activities (Activity Type A). These tasks can be woven in to supplement other course activities by using a dataset to make a clear point connecting data to scientific content. Although the dataset may not contain messiness, the use of these simplified datasets can increase awareness of how data are used to disseminate research results and support scientific principles. An example of this type of activity is having students work with data already summarized in a simple table, or visualized for them in a graph. These materials can be used for a lesson designed to hone in on data interpretation. Teachers looking for this type of resource can use Data Nuggets, resources designed to scaffold student abilities when graphing. Each Data Nugget comes in three graphing levels, where the simplest provides the graph to students as a way to practice data interpretation (Box 2).

Following these curated examples, teachers can introduce lessons designed to involve students in some aspects of dataset curation and summarization. These datasets can leave some data curation steps to the students, such as summarizing data by calculating averages across groups (Activity Type B). Typically, data summarization simplifies messiness within a dataset and eases interpretation. For example, simplified datasets used in textbooks, scientific journals, and news and media sources have likely gone through some level of data summarization. However, data summarization can also be used to hide messiness in a dataset to mislead or misrepresent results. By practicing components of data summarization, students can learn how changes can be made to datasets to illustrate particular concepts. In general, Data Nuggets (Box 2) are designed to highlight messiness and reveal some of the iterative research components of the scientific process. As a part of this process, some Data Nuggets provide a data table, but leave some form of mathematical calculation (e.g. means, total counts, converting to ratios or percentages) to the students to perform data summarization. These resources, which pair interpretation of messy data and the stories about unexpected or unclear results, have proven useful for teachers (Schultheis & Kjelvik, 2015). Students can use the same dataset to compare several different ways the data can be represented and how that might affect interpretation. Giving students opportunities to practice data summarization in a variety of contexts can give them insight into how data are presented and to think through how the data may have been modified to produce the variables displayed.

Next, students can move from examining well-structured problems to more complex inquiry investigations (Activity Type C). From our own experience, we've found that scaffolding inquiry experiences by sharing the true stories and messy datasets behind scientific research has given students a better understanding of how unanticipated results are a common occurrence in the process of science. Additionally, to help students move into inquiry, Data Nuggets can be used to introduce students to a scientific topic and study system. After students examine the dataset provided by the scientist, they are asked to generate their own questions that resulted from analyzing the highlighted data. This can be a way to provide a base for students to launch their own inquiry questions.

As students begin to ask their own questions and consider different ways to collect data, inquiry will require more creativity on their part (Konold et al., 2000; Kastens et al., 2015). By transitioning to inquiry, students can begin to step out into the unknown by collecting their own data and engaging in the practices of science (Activity Type C). Importantly, first-hand data collection and inquiry projects are often how students are first introduced to the various ways messiness can enter a dataset. Whether through natural variability or experimental error, students often must grapple with unexpected results. The sources of variability and the nature of scientific investigations is an important discussion topic at this step. Students must be guided to think through variation and that it represents more than “human error” during data collection. Having prior experience with messy datasets that resulted from scientists’ research can help students realize that messiness is a key part of how researchers learn about the world.

Finally, educators can transition students from datasets that they can interact with and summarize by hand, to ones where digital tools are necessary, such as online visualization platforms or statistical programs (Activity types D, E). Working with smaller datasets at first can help students buy-in to the activity, providing them with the motivation to use a larger second-hand dataset to answer additional follow-up questions (Schultheis & Kjelvik, 2015). To facilitate this process we’ve created Digital Data Nuggets where students can start by working with the pencil and paper activity, and then move onto a digital platform to explore both larger versions of the same data, and bring in new variables (Box 2). This process accurately represents the nature of science and how scientists often begin their own investigations by looking at summarized data published in studies, and working with data hosted in online repositories. Another way to scaffold this transition for students, a potential strategy is to “nest” student-collected datasets within larger online datasets. This is a common strategy when engaging students in citizen science projects where they collect small amounts of data themselves, but then contribute these data into a larger pool that they can then analyze and interpret. This scaffolding step can be used to help students see their first-hand data as part of a bigger picture, which could help support a strengthened connection to what might otherwise be an overwhelming dataset (Lehrer & Schauble, 2004).

Within the context of online learning platforms, there are opportunities for students to be more involved in the scientific question of interest, selection of the variables to be explored, and curation of the dataset and resulting visual representation. Students can start by working in platforms designed specifically for educational settings, such as Digital Data Nuggets, that provide guidance and direction (Activity Type D). As students gain familiarity with digital tools, they can progress towards processing and pulling larger datasets out of online repositories or even building their own dataset that brings together several sources of data (Activity type E). By moving students to a digital data environment, students will be given opportunities to explore and discuss messiness at a different scale. The ability to examine large datasets will provide students with opportunities to apply what they have learned about messiness from the previous activity types that had much more limited datasets.

Conclusion

Experiences working with messy data provide opportunities to increase students' content knowledge while simultaneously increasing their understanding of the nature of science and the scientific enterprise (Mourad et al., 2012; Langen et al., 2014). While student data literacy is currently low (Steen, 1999; NCED, 2001; Wilkins, 2010; Manyika et al., 2011; Oceans of Data Institute, 2014), it can improve when given opportunities to interact with authentic data (Duschl, 1990; Gould et al., 2014; Kastens et al., 2015). This value has been recognized by educators and curriculum reform efforts, which comes at a perfect time to tap into the resources made available through freely available datasets and educational resources (Picone et al., 2007; Metz, 2008;

Gould et al., 2014; Schultheis & Kjelvik, 2015; Harsh & Schmitt-Harsh, 2016; Angra & Gardner, 2017).

Just as the messiness and complexity of authentic datasets makes their use intimidating for students and teachers, it also has the potential to bring about learning opportunities not possible when messiness is hidden. Because messy data are a product of true scientific endeavors, they have the potential to immerse students in the practice of science and the habits of mind of a scientist. Science is about exploring the unknown, and this often results in surprising results. Datasets from scientific research contain artifacts from study methodology and true variability that hints to the complexity of our world. Students working with these data will be given a window to see how science works and may hopefully feel inspired and confident in their own ability to ask questions and tap into their desire to understand "why?".

Acknowledgements

Many ideas expressed in this article are related to our experiences developing Data Nuggets at Michigan State University (MSU), currently funded by NSF DRK-12 1503211. Support for Data Nuggets has been provided by the MSU Kellogg Biological Station (KBS) NSF GK-12 Project (DGE 0947896), the NSF Long-term Ecological Research Program (LTER) at KBS (DEB 1637653), MSU AgBioResearch, the LTER at KBS K-12 Partnership (NSF DEB 1027253), and BEACON Center for the Study of Evolution in Action (NSF DBI 0939454). Special thanks to Julie Morris, Ariel Cintron-Arias, Laurel Hartley, Kristin Jenkins, Robert Mayes, Louise Mead, Paul Strode, Molly Stuhlsatz, Gordon Uno, and Jeremy Wojdak for thought-provoking conversations at the Data Nugget National Institute for Math and Biology Synthesis (NIMBioS) working group (NSF DBI 1300426). Thank you to Cheryl Hach, May Lee, Paul Strode, and the anonymous reviewers who provided useful feedback on earlier versions of this manuscript. This is KBS Publication #2137.

References

American Association for the Advancement of Science (AAAS) (2015). Vision and change in undergraduate biology education: Chronicling change, inspiring the future. AAAS Vision and Change in Undergraduate Biology Education Initiative. Ed. D. Smith. Washington, DC. http://www.visionandchange.org/

ACT, Inc. (2014). ACT College and Career Readiness Standards: Science. www.act.org/standard/planact/science

Allie, S., Buffler, A., Campbell, B., Lubben, F. (1998). First-year physics students' perceptions of the quality of experimental measurements. *International Journal of Science Education*, *20*(4), 447–459.

American Statistical Association (ASA). (2014). "Curriculum Guidelines for Undergraduate Programs in Statistical Science," available at http://www.amstat.org/education/curriculumguidelines.cfm.

Angra, A. & Gardner, S. M. (2016). Development of a framework for graph choice and construction. *Advances in Physiology Education, 40,* 123-128.

Angra, A. & Gardner, S. M. (2017). Reflecting on Graphs: Attributes of Graph Choice and Construction Practices in Biology. *CBE Life Sciences Education, 16*(3).

Borges-Rey, E. L. (2017). Data Literacy and Citizenship: Understanding 'Big Data' to Boost Teaching and Learning in Science and Mathematics. In Handbook of Research on Driving STEM Learning With Educational Technologies (pp. 65-79). IGI Global.

Bowen, G. M. & Roth, W. M. (2005). Data and Graph Interpretation Practices among Preservice Science Teachers. *Journal of Research in Science Teaching 42*(10), 1063-1088.

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sciences Education*,*13*(2), 265-284.

Duschl, R. A. (1990). *Restructuring science education: The importance of theories and their development*. Teachers College Press.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, *7*(2).

Germann, P. J., & Aram, R. J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of research in science teaching*,*33*(7), 773-798.

Gould, R., Sunbury, S., & Dussault, M. (2014). In praise of messy data. *The Science Teacher*, 81(8), 31.

Harsh, J. A. & Schmitt-Harsh, M. (2016). Instructional Strategies to Develop Graphing Skills in the College Science Classroom. *The American Biology Teacher, 78*(1), 49-56.

Hernandez, R. R., Mayernik, M. S., Murphy-Mariscal, M. L., & Allen, M. F. (2012). Advanced technologies and data management practices in environmental science: lessons from academia. *BioScience*,*62*(12), 1067-1076.

Hogan, K. & Maglienti, M. (2001). Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. *Journal of Research in Science Teaching*, *38*(6), 663-687.

Holmes, N. G., Wieman, C. E., & Bonn, D. A. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences*, *112*(36), 11199-11204.

Hug, B. & McNeill, K. L. (2008). Use of First-hand and Second-hand Data in Science: Does data type influence classroom conversations? *International Journal of Science Education 30*, 1725-1751.

Kanari, Z. & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748-769.

Kastens, K. A., Krumhansel, R., & Baker, I. (2015). Thinking Big - Transitioning your students from working with small, student-collected data sets towards "big data". *The Science Teacher, 82*(5), 25-31.

Khishfe, R. & Abd-El-Khalick, F. (2002). Influence of explicit and reflective versus implicit inquiry-oriented instruction on sixth graders' views of nature of science. *Journal of Research in Science Teaching*, 39(7), 551-578.

Kjelvik, M. K. & E. H. Schultheis (2019). Getting Messy with Authentic Data: Exploring the Potential of Using Data from Scientific Research to Support Student Data Literacy. *CBE Life Sciences Education,* 18(2), es2.

Konold, C. E., Coulter, R., & Feldman, A. (2000). Engaging students with data. *Learning & Leading with Technology*, 28(3), 50-55.

Konold, C. E. & Higgins, T. L. (2003). "Reasoning about data", In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), A research companion to Principles and Standards for School Mathematics, Reston, VA: National Council of Teachers of Mathematics, pp.193-215.

Langen, T. A., Mourad, T., Grant, B. W., Gram, W. K., Abraham, B. J., Fenrnandez, D. S., Carroll, M., Nuding, A., Balch, J. K., Rodriguez, J., & Hampton, S. E. (2014). Using large public datasets in the undergraduate ecology classroom. *Frontiers in Ecology and the Environment*, *12,* 362-363.

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

Lutskey, N. (2008). Arguing with Numbers: Teaching Quantitative Reasoning through Argument and Writing. Calculation vs. Context: Quantitative Literacy and Its Implications for Teacher Education. Eds. Madison, B. L. & Steen, L. A. Mathematical Association of America, 59-74.

Magnusson, S. J., Palincsar, A. S., Hapgood, S., & Lomangino, A. (2004). How should learning be structured in inquiry-based science instruction?: investigating the interplay of 1st-and 2nd-hand investigations. In *Proceedings of the 6th international conference on Learning sciences* (pp. 318-325). International Society of the Learning Sciences.

Mandinach, E. B. & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. Educational Researcher, 42(1), 30-37.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. www.mckinsey.com/insights/ business_technology/big_data_the_next_frontier_ for_innovation. Viewed 8 March 2017.

Mayes, R. L., Forrester, J. H., Christus, J. S., Peterson, F. I., Bonilla, R., & Yestness, N. (2014). Quantitative reasoning in environmental science: A learning progression. *International Journal of Science Education*, *36*(4), 635-658.

McNeill, K. L., Lizotte, D. J., Krajcik, J. S., & Marx R.W. (2006). Supporting Students' Construction of Scientific Explanations by Fading Scaffolds in Instructional Materials. *The Journal of the Learning Sciences,*15(2):153-191.

Metz, A. M. (2008). Teaching statistics in biology: using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. *CBE Life Sciences Education, 7*(3), 317-326.

Moss, D. M. (2001). Examining student conceptions of the nature of science. *International Journal of Science Education, 23*(8), 771-790.

Mourad, T., Grant, B. W., & Gram, W. K. (2012). Engaging undergraduate students in ecological investigations using large, public datasets. *Teaching Issues and Experiments in Ecology*. Vol 8.

National Center for Education Statistics (NCES) (2005). Science: The nation's report card. https://nces.ed.gov/nationsreportcard/pdf/main2005/2006451.pdf

National Council on Education and the Disciplines (NCED) (2001). Mathematics and Democracy: The Case for Quantitative Literacy.

National Research Council (NRC) (1996). National science education standards. Washington, DC: National Academies Press.

National Research Council (NRC) (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists. Washington, DC: National Academies Press.

National Research Council (NRC) (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

NGSS Lead States. (2013). Next Generation Science Standards: For states, by states. Washington, DC: National Academies Press.

Oceans of Data Institute. (2014). Profile of the big-data-enabled specialist - executive summary.

Palincsar, A. S. & Magnusson, S. J. (2001). The interplay of first-hand and second-hand investigations to model and support the development of scientific knowledge and reasoning.In S. Carver & D. Klahr (Eds.), Cognition and instruction: Twenty-five years of progress (pp. 151–193). Mahwah, NJ: Lawrence Erlbaum Associates.

Picone, C., Rhodes, J., Hyatt, L., & Parshall, T. (2007). Assessing gains in undergraduate students' abilities to analyze graphical data.*Teaching Issues and Experiments in Ecology, 5*, 1-54.

Schield, M. (2004). Information literacy, statistical literacy and data literacy. In IASSIST QUARTERLY (IQ).

Schultheis, E. H., and Kjelvik, M.K. (2015). Data Nuggets: Bringing Real Data into the Classroom to Unearth Students' Quantitative and Inquiry Skills.*The American Biology Teacher 77*(1), 19-29.

Schwartz, R. S., Lederman, N. G., and Crawford, B. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education, 88*(4), 610-645.

Séré, M. G., Fernandez-Gonzalez, M., Gallegos, J. A., Gonzalez-Garcia, F., De Manuel, E., Perales, F. J., & Leach, J. (2001). Images of science linked to labwork: A survey of secondary school and university students. *Research in Science Education*, *31*(4), 499-523.

Shah, P. E., Weeks, H. M., Richards B., & Kaciroti, N. (2018). Early childhood curiosity and kindergarten reading and math academic achievement. *Pediatric Research*

Sorgo, A. (2010). Connecting Biology and Mathematics: First Prepare the Teachers. *Life Science Education, 9*, 196-200.

Speth, E. B., Momsen, J. L., Moyerbrailean, G. A., Ebert-May, D., Long, T. M., Wyse, S., & Linton, D. (2010). 1, 2, 3, 4: infusing quantitative literacy into introductory biology. *CBE Life Sciences Education, 9*(3), 323-332.

Steen, L. A. (1999). Numeracy: The New Literacy for a Data-Drenched Society. *Educational Leadership 57*(2), 8-13.

Strasser, C. A. & Hampton, S. E. (2012). The fractured lab notebook: undergraduates and ecological data management training in the United States. *Ecosphere*, *3*(12), 1-18.

Strode, P. K. (2015). Hypothesis Generation in Biology: A Science Teaching Challenge & Potential Solution. *The American Biology Teacher*, *77*(7), 500-506.

The College Board (2013). AP Biology: Course and Exam Description: Effective Fall 2012: Revised Edition.www.collegeboard.org.

Wilkins, J. L. M. (2010). Modeling quantitative literacy. *Educational and Psychological Measurement 70*(2): 267-290.

Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2017). Creating an understanding of data literacy for a data-driven society. *Journal of Community Informatics, 12*(3), 9-26.

Box 1: Example data literacy learning objectives and habits of mind

Below we have detailed just some of the opportunities for rich learning experiences, conversations that can be had with students, and some of the dispositions that may develop when using authentic data and research in the classroom.

Student outcomes and quantitative reasoning learning objectives for using messy data in the classroom:

- Discuss sources of variation found in data (natural, experimental)
- Differentiate instances when data does or does not provide support for a hypothesis
- Analyze and interpret results beyond what may have been expected from predictions
- Explain that science is an iterative process and does not follow a linear methodology
- Apply mathematical thinking to answer scientific questions
- Understand that there are limitations to scientific studies and data collection, often impacting the design of research studies
- Critique whether a dataset is appropriate evidence to answer a scientific question
- Construct a claim that is supported by data as evidence

Habits of mind that characterize data literacy:

- Belief and capability: Confidence in one's ability to perform data skills such as analyzing data, interpreting trends and patterns, and critically reviewing claims supported by evidence.
- Resiliency: Understand that the process of science is not complete with a single study and have the associated persistence to continue to seek out answers. Acknowledge and accept that often a study yields more questions than answers.
- Humility: Awareness of the limits of scientific knowledge; what we know today can always be overturned by new data.
- Ethical: Removal of personal bias. Self-awareness regarding potential assumptions.
- Flexibility: Comfort with messiness, uncertainty, and the ability to accept failure. Open to the challenging of beliefs and able to place trust in the scientific process.
- Inventiveness: Develop testable questions and creative ways to find solutions.
- Curiosity: Drive for knowledge and understanding that leads to an inquiry mindset. Seeing creative possibilities and new ways to represent data.
- Critical Thinking: Ability to connect scientific principles to the numbers and patterns found in datasets. Actively question data and the evidence used to support claims.

Box 2: Data Nuggets provide opportunities for repeated practice with authentic data

Data Nuggets are K-16 classroom activities, co-designed by scientists and teachers, designed to bring contemporary research and authentic data into the classroom (Schultheis & Kjelvik, 2015; http://datanuggets.org). Within each activity, students engage in the practices of science as they read scientific text, visualize and interpret data, construct explanations based on evidence, and ask questions. Each activity is written by the scientists themselves and provides the story of the people behind the research and what first inspired them to ask questions and pursue their passion. Because the authenticity of the research process is maintained, students often face unexpected results, including messy data that do not support original hypotheses.

Data Nuggets can be used to scaffold students as they build confidence in their quantitative abilities. Each activity is assigned a Content Level (1-4) according to the difficulty of the reading, vocabulary, and scientific concepts. Additionally, each Data Nugget activity is available in three Graph Types (A-C), according to the graphing skills required. Type A activities provide the graph for the students, allowing a focus on interpretation and using data to support scientific explanations. Type B activities provide scale and axis labels, but requires students to graph the data. Type C provides an unlabeled grid on which students create their own visualization of the data, allowing more flexibility and opportunities to determine appropriate representations.

To further the quantitative skills and abilities represented in Data Nuggets, we created Digital Data Nuggets to scaffold student data literacy abilities. Students can explore smaller datasets by hand using Data Nuggets, and then move onto Digital Data Nuggets where datasets are larger and need tools to help with visualization and curation. These activities are built in collaboration with existing online data visualization platforms that are designed to allow students to easily explore large datasets, construct graphs, do statistics, and more. Using these data visualization platforms allow students to visualize and explore big data, while not requiring them to develop data science skills simultaneously.

Table 1

Table 1. Potential sequence of classroom activities that advance in the complexity and sophistication of students' interactions with authentic messy data.

Data Characteristics	Potential Focal Topics
Activity Type A. Simplified second-hand data, summarized, curated to display a clear trend.	Easily illustrate a specific scientific concept (e.g. NGSS Disciplinary Core Idea) and how scientific results are disseminated.
Activity Type B. Second-hand data that includes some level of messiness (e.g. variation, outliers, does not follow predictions) or curation by student.	Introduce students to statistical concepts, curation, how to interpret data with variation and unexpected results, how data can be modified and displayed in different ways.
Activity Type C. First-hand data collected from classroom labs or inquiry projects.	Asking scientific questions, how to quantify variables, importance of experimental design (e.g. replicates, controls), give students ownership and a personal connection to data.
Activity Type D. Large, second-hand online datasets with guided instruction.	Introduction to computational tasks and data visualization techniques, examines variability at a larger scale.
Activity Type E. Large, second-hand online datasets open to student inquiry investigation.	Organizing data, finding and selecting appropriate variables, building knowledge from multiple sources.