Using Large Data Sets for Open-Ended Inquiry in Undergraduate Science Classrooms

CATHERINE M. O'REILLY, REBEKKA D. GOUGIS, JENNIFER L. KLUG, CAYELAN C. CAREY, DAVID C. RICHARDSON, NICHOLAS E. BADER, DAX C. SOULE, DEVIN CASTENDYK, THOMAS MEIXNER, JANET STOMBERG, KATHLEEN C. WEATHERS, AND WILLIAM HUNTER

Analysis and synthesis of large and complex data sets are increasingly important components of scientific research. To expose undergraduate students to these data sets and to develop valuable data-analysis skills, a team of environmental scientists and education researchers created Project EDDIE (Environmental Data-Driven Inquiry and Exploration). Project EDDIE is a pedagogical collaborative that develops and assesses flexible modules that use publicly available, large data sets that allow students to explore a range of concepts in the biological, earth, and environmental sciences. These modules have been implemented in a range of courses, class sizes, and institutions. We assessed six modules over eight courses, which were taught to a total of 1380 students. EDDIE modules led to significant improvements in these students' competence using spreadsheet software, as well as their conceptual understanding of how to use large, complex data sets to address scientific problems. Furthermore, the students reported positive and informative experiences using large data sets to explore open-ended questions.

Keywords: Project EDDIE; quantitative literacy; inquiry-guided, active learning; environmental sensors

Our understanding of the environment is increasingly informed by the analysis and synthesis of large data sets. In many ways, the environmental sciences, including earth science and ecology, are undergoing an "informatics" revolution, with networks of sensors and people generating unprecedented amounts of data at a range of spatial and temporal scales (Benson et al. 2009, Michener and Jones 2012, Hampton et al. 2015, La Deau et al. 2016, Read et al. 2016). These large data sets may comprise long-term data collected manually or high-frequency data generated by automated sensor-based systems (Benson et al. 2009, Schimel and Keller 2015) and are often complex, containing many variables, multiple sites, missing data points, and incorrect sensor readings. Young scientists must be better prepared to manage, visualize, and analyze such large data sets; however, this training is still lacking at the undergraduate and graduate student levels (Hernandez et al. 2012, Michener and Jones 2012, Hampton et al. 2015, Read et al. 2016, Weathers et al. 2016).

Although large data sets are commonly used in research, many current undergraduate science curricula remain focused on analyzing data from small-scale studies. The use of relatively small data sets is widespread across undergraduate classrooms, in part because they are typically derived from activities designed to allow students to ask their own questions, design experiments or manipulate equipment, and generate and analyze their own data. These are important learning outcomes, but working with data sets that are limited in size or complexity does not give students the opportunity to practice data management, spreadsheet navigation skills, or hypothesis testing based on data—skills that are sorely needed (Strasser and Hampton 2012, Rubin and Abrams 2015). Students recognize that these small data sets are often not appropriate for drawing strong conclusions, and a common refrain in laboratory reports is "more research is needed."

Despite encouragement to use authentic data, and despite its public availability online (Ellwein 2014, Gould 2010), instructors face several barriers to working with large data sets in the classroom. For example, many data sets are provided in formats that students are not familiar with (e.g., csv or txt) and need to be downloaded and translated into a more user-friendly format. Inexperience with spreadsheet navigation can lead to student frustration, especially when there are hundreds to thousands of records. The real-time nature of many of these data sets means that the instructor may feel the need to identify a useful location or subset of available data that provide a clear example of the topics, and

BioScience 67: 1052–1061. © The Author(s) 2017. Published by Oxford University Press on behalf of the American Institute of Biological Sciences. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com.
doi:10.1093/biosci/bix118 Advance Access publication 8 November 2017

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

Table 1. A description of the EDDIE modules, the data sets that are included with the modules (additional data are available online in referenced data sources), and the courses in which they were used during the 2014–2015 academic year.

EDDIE module	Science concepts	Quantitative reasoning concepts	Data sets included	Courses in which the module was taught
Ice Phenology	Climate change, ice-off, phenology, physical, biological, and cultural implications of changing ice-off dates	Regression, graphing, variation, spreadsheet navigation	6 lakes, each with 100–200 records	Freshwater Ecology
Lake Mixing	Lake thermal stability, mixing regimes, climate change, seasonal variation	Graphing, variation, spreadsheet navigation	6 lakes, each with 3000–4000 records	Freshwater Ecology
Lake Metabolism	Gross primary production, respiration, eutrophication	Graphing, variation, spreadsheet navigation	5 lakes, each with 100–500 records	Freshwater Ecology
Stream Discharge	Discharge and climate, flood probability and frequency, runoff and urbanization	Probability, regression, variation, extrapolation and interpolation, spreadsheet navigation	Full data set of 28,000 records, working data set with 1000 monthly records	Hydrology, Environmental Geology
Nutrient Loading	Water quality, discharge, concentration, loading	Correlation, covariation, variation, spreadsheet navigation	23,000 records	Hydrogeology, Gen Ed Hydrology, Freshwater Ecology, Gen Ed Biology
Climate Change	Greenhouse gasses, global warming, long-term climate variation (glacial and interglacial periods)	Regression, variation, spreadsheet navigation	4 data sets with 60–3000 records	Ecology, Gen Ed Biology, Environmental Geology

Abbreviation: Gen Ed, general education courses.

there may be concerns about whether the messiness of the data may obscure the point of the lesson (Gould et al. 2014). In a practical sense, instructors face challenges that may hamper their use of large data sets, including access to computer laboratories, different student skill levels with spreadsheet software, and differences among versions of software and operating systems. Although some best practices for using large data sets with undergraduate students have been proposed, there are still multiple challenges to incorporating large data sets into curricula across the sciences and assessing their effectiveness (Langen et al. 2014).

Manipulating, analyzing, and interpreting large data sets in the context of open-ended exploration can have substantial benefits for students (Ellwein et al. 2014). Working with large, messy, heterogeneous data sets may motivate students to develop and rely on conceptual frameworks or mental models, and interpretation of such data prompts students to focus on discerning pattern and process rather than on "correct" answers (Gould et al. 2014). Large data sets also help students explore the stochastic nature of environmental and earth systems, potentially improving their understanding of uncertainty, randomness, and variation (Brewer and Gross 2003, Gougis et al. 2016). Moreover, students build critical computer-, spreadsheet-, and data-management skills when they work with large data sets (Strasser and Hampton 2012, Carey and Gougis 2017, Klug et al. 2017). Using authentic data sets from online repositories along with open-ended questions reinforces the need and rationale for this skill development and can help students develop an appreciation for large complex data sets associated with basic environmental monitoring (Ellwein et al. 2014). Finally, data sets with high spatial resolution allow students to find place-based data that are meaningful to them; furthermore, real-time data allow students to see what is currently happening in the world around them. In addition to helping students understand that data exploration is part of the scientific method, visualizing and interpreting large data sets could make students better able to produce and evaluate data presented in public formats.

The Project EDDIE (Environmental Data-Driven Inquiry and Exploration) team developed modules that use large, authentic data sets to explore a range of concepts in the biological, earth, and environmental sciences (table 1). Our modules primarily link to existing online public data sets and allow flexibility for the instructor to choose data to focus on specific locations, time periods, or content. Modules are designed to be adaptable and scalable across different skill levels, both within and across different types of institutions and courses, and focus broadly on undergraduate courses, although they can also be used with graduate courses. We explored the gains and challenges associated with incorporating these large data set activities and assessed the effectiveness of these modules specifically in terms of students' (a) spreadsheet skill development and (b) their conceptual understanding of how large data sets can be used. Furthermore, by working closely with instructors at seven institutions and across a range of course levels, we qualitatively and quantitatively assessed both student and instructor experiences using these large data-set modules.

EDDIE modules description

To develop curricular materials that use large data sets, a team of environmental scientists from a range of disciplines

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

Table 2. The spreadsheet functions involved in EDDIE modules and their corresponding skills in Excel. These typically involve using menu commands and/or keyboard shortcuts.

Function	Skills
Importing and formatting data	· Opening non-Excel files and selecting through formatting options · Copying Web text and pasting (or using paste special) into Excel · Using *Text to Columns* · Removing header rows · Removing columns · Formatting date columns
Graphing	· Creating plots (X–Y scatter time-series, bar plots) · Formatting plots · Creating plots with subsets of the data
Data selection	· Selecting columns of data using shortcuts · Identifying coordinates on a graph by hovering over a data point
Data manipulation	· Using formulas for calculations linking to other cells · Pasting formulas down-column using shortcuts (filling down) · Sorting data
Analyses	· Adding trend lines · Selecting option to view equation and R2 · Using Excel formulas to calculate average and standard deviation · Using Excel formulas to find maximum or minimum values

(hydrology, freshwater ecology, biology, and geosciences) and education researchers created Project EDDIE to collaboratively write modules and then use and assess these modules in their classrooms. Functionally, we define a *large data set* as a data set that cannot be viewed in a single screen of a spreadsheet program without scrolling, which makes computational tasks such as summarizing and plotting the data necessary. We used data sets ranging from approximately one hundred to tens of thousands of records (table 1), requiring a wide range of spreadsheet competence skills (table 2). Notably, many of these skills are beyond those that a student might use when collecting their own data. The EDDIE modules used in this study are listed in table 1 and available online at *projecteddie.org*. Each module consists of an instructor's manual, introductory lecture slides, student handouts, preclass readings, data sets, homework questions, and answer keys.

The EDDIE module format was designed so that instructors could adapt the modules as desired for their own classroom. Each module has a flexible A–B–C structure that follows the 5E learning cycle (Bybee et al. 2006, Carey et al. 2015, Carey and Gougis 2017). Part A *engages* students in initial data exploration and skill development using simple analysis that bypasses some of the technical challenges associated with the manipulation of data. Part B asks students to *explore* and *explain* through a more detailed analysis that requires them to independently discuss and decide which analyses are appropriate for the data or explain the implications of data variability. In Part C, students *expand* on the developed ideas by exploring data from sites of their choosing to address questions that they have developed. For the final part of the 5E learning cycle, students *evaluate* their learning by participating in class discussions and completing homework assignments. Introductory classes may only teach parts A and B in a class, whereas more advanced classes may teach part B in class and then assign part C for homework. The instruction time for a complete module varies but is targeted for a 3- to 4-hour laboratory period.

Open-ended questions are incorporated into each module, requiring the students to choose their own data by selecting a subset of data from different locations or time periods. The goal of these sections is for students to grapple with the inherent issues of spatial and temporal variability within the system and to think carefully about what the data represent. For example, students make their own decisions about how to split up a temporal data set before and after human activity to compare climate trends in one module or to look for the impact of urbanization on river flooding in another. In a different module, students determine how their results might change if they had a shorter time series. These questions prompt students to confront how their interpretations are influenced by data availability and variation.

Module implementation and assessment

A set of EDDIE modules was implemented during the 2014–2015 academic year across eight courses teaching ecology, biology, or hydrology at seven institutions of higher education in the United States (tables 1 and 3). These courses were spread across a range of institution types (4-year liberal-arts college to R1 universities), course levels (nonscience freshmen to upper level), and class sizes (10 to 1200 students), and two courses had coenrolled graduate students (table 3). These courses incorporated at least one of six EDDIE modules, with six courses using only one module and two courses using two modules. One course was a large general-education biology course with four lecture sections, and these students completed an EDDIE module in their laboratory sections, each of which had 25 students. Because of the time demands associated with coding the open-ended responses in our assessment tool, we only used responses from one lecture section of this general-education biology course; the students in this lecture section were distributed

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

*Table 3. Institutional Carnegie designations (*http://carnegieclassifications.iu.edu*) and course descriptions in which EDDIE module implementation and assessment occurred.*

Institution code	Carnegie classification description	Course description(s)	Class size	*N**	Unique code
	Public research university (highest research activity)	Hydrology; seniors and graduate students	45	22	R1-B
R 1	Public research university (highest research activity)	a. Freshwater ecology; junior and senior science majors	30	a. 8	R1-A1
		b. Freshwater ecology; science graduate students	10	b. 4	R1-A2
R 2	Public research university (higher research activity)	General education biology	1200**	70	R2
Master's 1	Public master's colleges and universities (larger programs)	Ecology; sophomore and junior science majors	17	16	M1-A
Master's 1	Private master's colleges and universities (larger programs)	Introductory biology; sophomore–senior nonscience majors	30	6	M1-B
Master's 3	Public master's colleges and universities (smaller programs)	Hydrogeology; sophomore and junior science majors	23	17	M3
Bac/A&S	Private baccalaureate colleges—arts and sciences	Hydrology; freshmen and sophomore science and nonscience majors	25	10	Bac

Note: The course size and the sample size of students are included. We created unique codes for each course on the basis of institution codes and whether there were multiple courses from that institution.
*The number of students who completed enough sections of both the pre- and postquestionnaires for the data to be used in this study.
**The overall course enrollment; the module was taught in each of 54 laboratory sections that had 25 students.

across the 54 laboratory sections associated with the course. In total, across all courses in this study, 1380 students completed at least one EDDIE module. Although a few instructors employed computer laboratories, most of the instructors had students use their personal laptops, sometimes working in pairs.

We collected both quantitative data and qualitative information about the modules. The students were quantitatively assessed using a questionnaire that was administered both prior to and after using any modules. To recruit students for pre- and postmodule questionnaires, the instructors showed a 4-minute video in class and provided a link to complete the optional assessment online. Approximately 1 week after the module, the instructors provided a link to complete the optional postmodule assessment. The questionnaires took about 20 minutes, and in some cases, the students were offered minimal course credit (i.e., less than 2% of their final course grade) for completing both the pre- and postmodule assessments. Qualitative data about using the modules were collected from the students using end-of-semester course evaluations and from the instructors through a phone interview immediately after using the module and subsequent discussions.

Spreadsheet competence. To assess whether working with large data sets improved the students' facility with spreadsheets (we used Microsoft Excel, hereafter *Excel*), the students ranked their comfort level using Excel on a scale from 0 to 4, choosing one of the following options: 0, *I don't know how to do anything in Excel*; 1, *I only know how to do a few things in Excel*; 2, *I know how to do several things in Excel well*; 3, *I feel very competent in Excel but would not feel comfortable teaching others how to use Excel*; or 4, *I feel very competent in Excel and would feel comfortable teaching others how to use Excel*. The students then used a similar scale (0–4) to rank their ability in performing several functions in Excel (calculate an average, calculate a median, calculate a standard deviation, calculate variation, perform a correlation, find a maximum value in a data array, draw a trendline, analyze an equation for a trendline, create a bar graph, and create a line graph). The scores were aggregated to yield a spreadsheet competence score that had a possible range from 0 to 60.

Conceptualizing how large data sets are used. To examine students' conceptual understanding of how to use large data sets and their value when solving problems, we created questions following the Experimental Design Ability Test (Sirum and Humburg 2011), in which an environmental problem was presented, an online large data set was suggested for solving the problem, and the students wrote a narrative response. The environmental problems that were presented were related to the topic of the module that the students had completed, as were the types of online data sets mentioned in the questions. The open-ended narrative responses were grouped into three different types of categories: 0, the student would seek out other information either on the Internet or by interviewing experts; 1, the student would go collect his or her own data; or 2, the student would use the data set in their solution. Details of the coding procedures and exemplars of each category of responses are provided in the supplemental material.

Because the sample sizes per class and per institution were small, we pooled data across institutions for analyses of pre- and postmodule assessment scores. We used paired *t*-tests

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

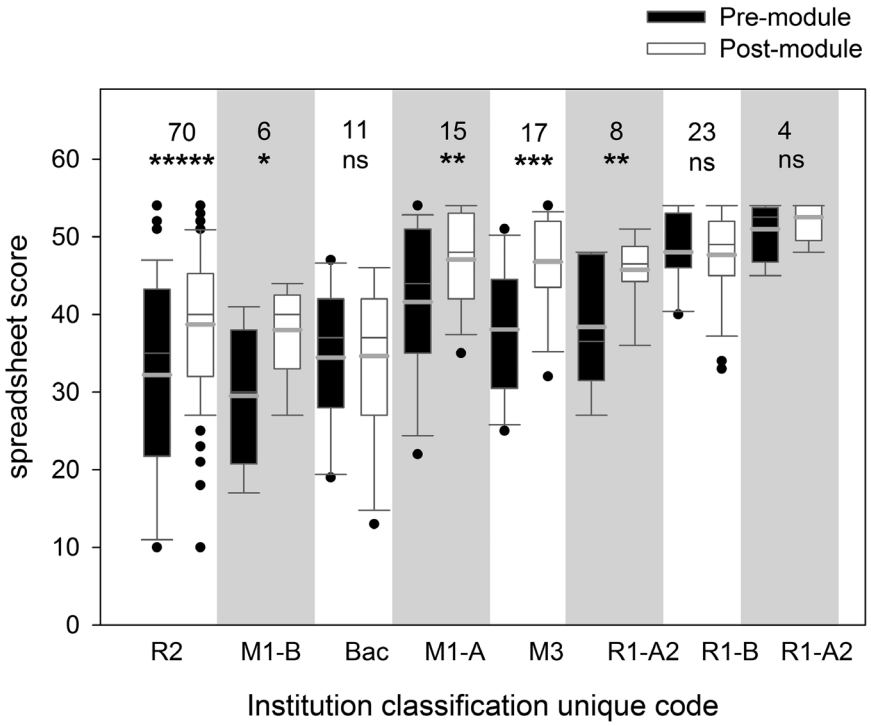

Figure 1. Box and whiskers plots showing pre- and postmodule spreadsheet competence assessment scores for each of the different courses, organized from lowest to highest level (left to right). The thick, light gray line represents the mean. Above the boxes, the numbers represent the sample size, and the asterisks indicate a significant difference between the pre- and postmodule scores, where $p < .05$ is *, $p < .01$ is **, $p < .001$ is ***, and $p < 1 \times 10^{-16}$ is **. The institution-classification unique codes are described in table 2.***

to determine differences in spreadsheet competence and Wilcoxon signed-rank tests to determine the differences in student response categories for the question related to using large data sets.

Developing spreadsheet competence

Using EDDIE modules across this range of disciplines and levels led to statistically significant improvements on the students' self-reported Excel competence across courses (paired *t*-test, $t(153) = -8.21$, $p < .00001$, premodule = 37 ± 1, postmodule = 42.1 ± 0.7). Despite the small sample sizes, this difference was detectable in many of the individual courses as well (figure 1). In general, lower-level courses (those on the left in figure 1) tended to demonstrate low initial spreadsheet competence scores and exhibited the greatest increases in self-reported Excel competence, a pattern that was also found among three institutions using the EDDIE lake modules suite (Klug et al. 2017). Two courses in which students already had high initial scores on the premodule assessment (and therefore had little room to score higher) did not experience a significant change; these R1-B and R1-A2 courses were upper-level courses that included graduate students who were presumably already comfortable and proficient in Excel, making it less likely that working with the module would yield substantial gains in Excel ability. The other course with no significant change in spreadsheet competence scores was predominantly freshmen and sophomores at a private liberal-arts college (Bac). In this case, it was possible that these nonscience majors were not motivated to improve their spreadsheet abilities, found the spreadsheet activities unrewarding, retained a phobia about mathematics and data, or had overestimated their initial competence in Excel.

The clear gains in ability to use spreadsheets for analysis of data were accompanied by particular challenges that were common across the courses. Both the students and instructors commonly became frustrated with the *Excel barrier*, with the students feeling unfamiliar with the program and the instructors annoyed that more time than they had expected was devoted to procedural details that were disconnected from overarching learning goals. Some of the students became visibly bothered by their struggle to get past basic spreadsheet tasks before they could get to the actual data manipulation or analysis. EDDIE modules require a set of skills that are substantially different from those a student would use when working with data they had collected and inputted themselves (table 2). It is possible that the different skills needed for these modules provided a greater set of challenges for the students than the typical use of Excel. These additional challenges are probably related to (a) the fact that large data sets can initially appear overwhelming to students because they cannot easily see patterns just by looking at the numbers themselves, (b) frustration associated with trying to use existing skills on such large volumes of data (e.g., scrolling down to select a column of data), and (c) the acquisition of new skills needed to properly work with the data (e.g., using keyboard shortcuts to select a column of data).

The instructors in this study addressed the Excel barrier in various ways, which are incorporated into a compilation of best practices (box 1) and teaching tips (box 2). Some of the instructors modeled an example analysis using a sample data set; their students then repeated the analysis using a new data set. Some of the instructors provided a reference handout with Excel basics that detailed how to use equations and how to make a graph. Others wrote common navigational Excel keyboard shortcuts on the board. Some of the instructors dedicated time to work through separate Excel tutorials (e.g., several are available from the Science Education Resource Center at *http://serc.carleton.edu/index.html*), and others had more experienced students help their peers after they had completed the activities. All the

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

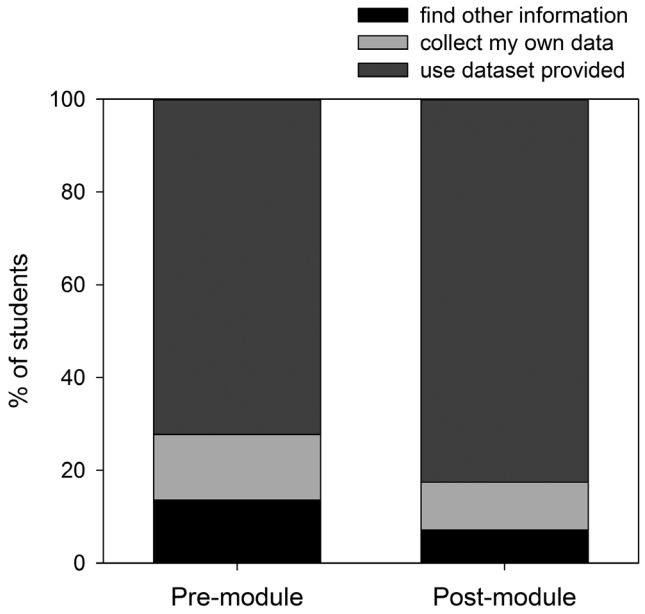

Figure 2. The percentage of students providing responses in each category when asked to solve a hypothetical environmental problem and told about a relevant large data set.

instructors provided one-on-one support. To some extent, helping these students learn Excel was made more challenging by the different operating systems and versions the students had on their laptops. This can be alleviated by completing the modules in a computer laboratory; however, the trade-off of a more uniform experience may be that students are not able to develop the skill sets on their own computers. Importantly, for the instructors who taught modules again the year after this study (data not reported here), Excel provided less frustration during those classes. This indicates that instructors can rapidly adapt the module or how they teach to better resolve the Excel barrier once they have a better understanding of the issue.

The students' comments reflected positively on using Excel, and the students seemed to appreciate the opportunity to work with large data sets. One student said, "I liked how we took raw data and manipulated it in such a way that was easy to understand/interpret by making graphs," and another commented, "The EDDIE modules made me think about how to use my computer and its data analysis capabilities in a totally different way." With respect to manipulating and graphing the data, a student said, "These modules made me think more about all the work that goes into manipulating data. I have developed a greater appreciation for all the work that goes into this." In response to a general prompt about what they liked about using the modules, the students spoke specifically about their gains from using Excel, saying, "[I liked] learning how to use certain Excel functions more effectively and sifting through data" and "Excel was something I've always wanted to learn and I'm so glad I finally know how to use it." Across all modules, students often commented that their favorite part was making the graphs to visualize the data and interpreting these graphs themselves; one student noted, "My favorite part was when the students had to make the graphs so we could see for ourselves the change in climate."

Conceptualizing how large data sets are used in science

Working with large data sets in an EDDIE module appeared to influence how willing the students were to work with a large data set to solve the hypothetical environmental problem posed in the assessment instrument. Even before engaging in an EDDIE module, 69% of the students discussed how they would use the large data set in their response (figure 2). Of the remaining students who did not choose to use the large data set in the premodule assessment, 13% chose to collect data themselves, and 18% chose to find other information (figure 2). When they stated that they would collect data themselves, the students' explanations sometimes indicated that they would collect the data for the same variables that were contained in—or could be calculated using—the provided data set. This suggests that students may not be able to conceptualize how these large data sets were structured or how the data could be used, nor recognize that they themselves might be capable of using the data set.

After working with large data sets in an EDDIE module, significantly more of the students (83%) chose to use the large data set to address the hypothetical environmental problem posed in the assessment instrument. Across institutions, the number of students who chose to use the large data sets increased by 10% on the postmodule assessment. The distribution of postmodule responses differed significantly compared with that of the premodule responses, with significant shifts toward choosing to use the large data set rather than collect data themselves or find a solution elsewhere (Wilcoxon paired signed rank, $p = .03$; figures 2 and 3). Over half of the students (65%) who initially did not choose to use the data set to solve the problem stated they would use it on the postmodule assessment (figure 3). This increase suggests increased student awareness of the value of large data sets and their role in science.

Interestingly, even after completing the module, 17% of the students still did not choose to use the large data set to address the hypothetical scientific problem posed in the assessment instrument, instead stating that they would collect data themselves (10%) or seek information from the Internet or an expert (7%; figure 2). Of the students who initially chose to seek information elsewhere, 24% of these students later chose to collect data, which we interpreted as a sign of the students' increased confidence in data itself (figure 3). It was surprising that even the students who initially chose to use the data set in the premodule survey subsequently chose to seek information or collect their own data (figure 3). Exploring large data sets may have deterred some of the students, perhaps because the data remained or became confusing or intimidating or because they believed

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

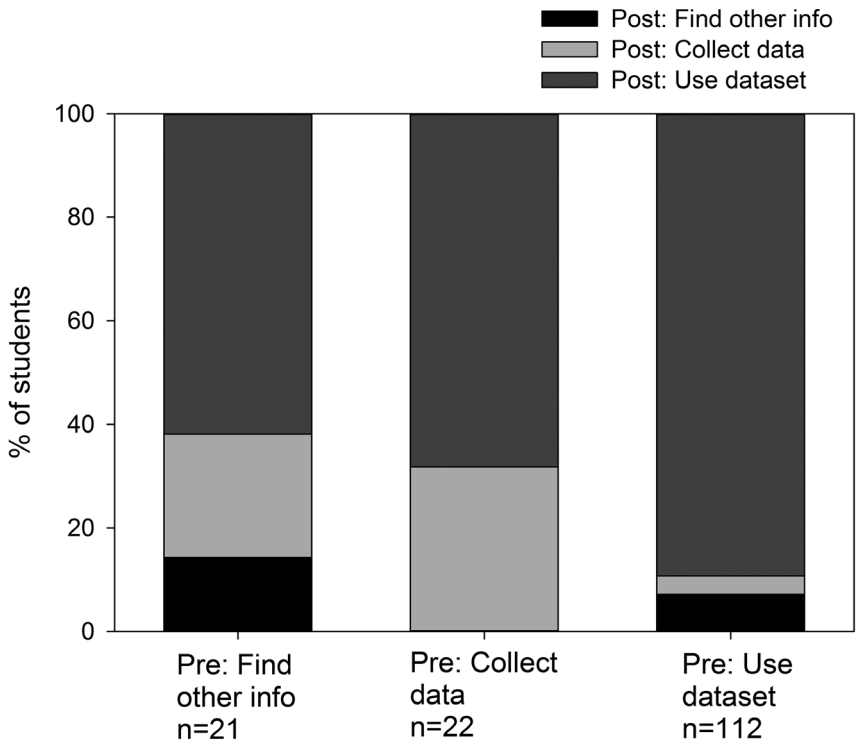

Figure 3. The percentage of students' postmodule responses in each category relative to their premodule responses across those categories, with respect to coded responses to a hypothetical environmental problem.

that solutions are more readily available online; about half of these students did have negative gains in their self-reported spreadsheet scores. Students can be challenged by the conceptual and process skills necessary to locate and navigate data sets, generate and test hypotheses, perform statistical analyses, and generate data visualizations (Langen et al. 2014). If they do not enjoy these tasks, students may seek quick solutions from the Internet if they believe the answers are available, especially given that the data themselves were obtained from the Internet. It is also possible that students do not trust the reliability of data contained in open data sets and thus would rather collect data themselves (Langen et al. 2014).

Instructors' and students' perspectives

Many of the students appeared to be more comfortable with large data sets and better at working with them than their instructors anticipated. The students were also more competent than expected at tasks such as downloading data files in unfamiliar formats (e.g., txt or csv) and importing them into Excel. The instructors found that the students focused on the common aspects of the different data sets and became adept at using the header row's column names to find and organize the portions of the data set they needed to address their question. Similarly, the students were resourceful in accessing the data they needed if the instructions did not work as anticipated. For example, to find US Geological Survey (USGS) streamflow data from various gauging stations in the EDDIE Stream Discharge module, many of the students recognized that a browser search with the station number was a more direct path to the data portal than navigating to the data through the USGS Web interface.

At the same time, some aspects of large data sets were more complex than the students anticipated. The students were expecting data sets to be clean, with no missing data points and no outliers or questionable data. The instructors found that data variability, especially in sensor data sets, provided an excellent opportunity for discussion about data collection and quality assurance or quality control. In one of the modules, the instructor moderated a discussion on why data points were missing from a lake sensor data set. The students were initially frustrated by the gaps in the data set; however, after learning that the missing data points were due to the sensors being struck by lightning, the students were more understanding of gaps in data and the exclusion of extreme outliers from the analysis. It is possible that students' preconceived ideas about randomness, variation, sampling, and the scientific method are challenged in new ways when working with existing large data sets. In this case, the students are asked to identify and exclude outliers using a consistent defensible approach based on some quantitative aspect of the data; this may be more difficult than excluding data that they had collected themselves, which may be easier for them to personally justify discarding for a variety of their own specific reasons.

To engage the students with the data, we found that the instructors took advantage of the histories associated with these large and long-term data sets, putting a story behind the data collection. In the EDDIE Ice Phenology module (table 1), the students were introduced to the concept of lake ice-off through a story about a church recording the ice-off date for Lake Constance because parishioners annually walked across a lake to exchange a statue of the Madonna with a parish on the other side (Magnuson et al. 2000). The ice-off dates for Lake Sunapee, New Hampshire, came from generations of a single family who recorded when they could get their boat from one end of the lake to the other. The students were introduced to the EDDIE Climate Change module through a discussion on Svante Arrhenius, who, depressed after the end of his marriage in 1894, spent a winter obsessively working through the calculations that came to define the greenhouse effect (Pearce 2003), and of Charles David Keeling's persistence in collecting atmospheric carbon-dioxide data in the 1950s. The Vostok ice core data in the EDDIE Climate Change module were presented as

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

Box 1. Instructor best practices for teaching EDDIE modules.

1. Set realistic goals about what you can and cannot cover within a semester and course period. As an instructor, accept the fact that you will not be able to cover as much material when you replace lecture with active learning. This transition is not a bad thing if students are better able to retain information gained with guided-inquiry active learning (e.g., Vanags et al. 2013).
2. Use class time to situate the content of the module in the rest of your course by providing the big-picture context. This can include an introduction, using the human stories behind the data before the module, and a debriefing following the module.
3. Manage your expectations and those of your students. Remember that you are asking students to master new software *and* new scientific concepts and that struggling is an important part of learning and the scientific process. To keep students from feeling overwhelmed, allow students to master some relevant computer skills before challenging them with broad, open-ended scientific questions.
4. Remember that your students may begin the module with vastly different computer skill levels. Assess student skill level beforehand, ideally in an explicit rather than a general way (e.g., "Can you make a scatter plot in the spreadsheet program such as Excel?" versus "have you used Excel before?"). Provide mechanisms that allow students to stay on track during the exercise.
5. Before teaching a module in your classroom, go through it yourself. Make sure you can access the data you need. Some online data may not always be available, and you may want to download data sets in advance as a backup.
6. Make time for discussion during the module, and identify ways to prompt discussion. Depending on class size, discussion can occur with the entire class, as a think–pair–share activity, in small groups, or with clickers.
7. If students are bringing their own laptops into the classroom, remind them in advance to make sure they are charged or to bring in power strips.

the results from the students' imaginary research trip to Antarctica, one in which they also get to attend the Metallica concert that occurred there in 2014 (shown in online photos and video; Coleman 2013). Stories about data can also be linked to current or local events, such as introducing the EDDIE Water Quality module by discussing the 2014 drinking water quality crisis in Toledo, Ohio, or discussing a local flooding event to highlight the relevance of the EDDIE Stream Discharge module.

The students' comments suggested that they enjoyed working with authentic data and broadened their appreciation and understanding of the power of large data sets. One student said, "The EDDIE modules made me think about how important it is to get large data because patterns can really only be seen when you take into account days, nights, different months, seasons, etc." Another said that after working with these data, they wanted to know more about "how the data were collected, who collected it, and why it was collected," highlighting the importance of sharing stories about the data. The students' comments showed their appreciation for "working with real data" and for ownership over how the data set was used with comments such as being thankful for "actually getting to arrive at the conclusions ourselves." The students clearly moved beyond just manipulating the data to using the data to support a conclusion: "My favorite part about this activity was actually seeing the data itself. I was able to physically see the numbers rise and fall and it just made climate change that much more evident to me." These comments suggest that working with these modules gives students an appreciation for large data sets, as well as new insights into the nature of science (Miller et al. 2010), and it could encourage the open science mindset of collaborative data sharing (Hampton et al. 2015).

Conclusions

Working with large data sets using the EDDIE modules leads to learning outcomes that range from an appreciation of large data sets to the specific skills that students must attain to be able to excel as well-informed citizens or as scientists. Even after only one module, the students showed substantial self-reported gains in spreadsheet skills and were more likely to use large data to solve a science-related problem. These significant outcomes occurred despite the diversity of courses, skill levels, instructors, and instructor adaptations of the modules, all of which can influence expected impact (Chase et al. 2013). To help engage the students with the modules, the instructors used a variety of approaches. Our compilation of instructor best practices contains practical suggestions that will improve the overall experience of working with the modules by reducing technical frustrations and enhancing the conceptual scientific context (box 1). We also found that these instructors developed a range of creative approaches to help students develop technical competency and, in particular, to encourage their students to conduct open-ended exploration (box 2); these "pro tips" further facilitated classroom discussion and student engagement with the concepts and problems.

It was clear from the assessment items on the course tests and from the student comments that working with the modules helped cement scientific concepts. This aspect of the modules can be further enhanced by allowing exploration of

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

Box 2. "Pro" tips for teaching EDDIE modules.

1. Don't be restricted to the particular study sites referenced in the modules; add in data sets from other sites, and let students find data from other sites to explore places of interest to them.
2. Pair field or lab activities with the module to help students understand what the data represent. For example, pair the lake mixing module with a field trip to collect temperature profile data from a nearby lake, or pair the stream discharge module with a field trip to a local stream to measure discharge.
3. Ask students to predict possible outcomes and draw their own plots and graphs, which you can then use to identify weaknesses in their conceptual understanding at the outset, and adjust your pace accordingly.
4. When multiple students are teamed up to work on the same computer, have them switch regularly or check in at certain break points in the modules to ensure that everyone gets to use the computer.
5. Pair students into teams on the basis of their computer operating system (e.g., PCs with PCs, Macs with Macs).
6. Pair students with unequal spreadsheet skill levels, and then ensure that both take turns operating the computer.
7. Write Excel shortcuts on the board.
8. Allow students who finish early to roam around the room as peer teaching assistants.
9. Prompt class discussion by highlighting students' activity through projecting individual students' computer screens.
10. Be an undercover expert—you can be the research assistant, and the class can give you instructions on how to address the questions. This eliminates the Excel barrier because you are making the graphs and manipulating the data, but it still allows students to engage with the key questions and concepts (i.e., make decisions about how to use data).

the data to be motivated by unanswered questions that are either posed by the instructor or generated by the students themselves after their initial data exploration. For example, in the EDDIE Climate Change module, the students chose which periods to analyze in the records of temperature and carbon-dioxide concentration, and comparisons between historical and recent rates of change were shocking to students. In the EDDIE Ice Phenology module, the students were also surprised by the rate at which ice-off dates for lakes were changing. Despite variation around the trend line, observing significant changes in ice-off dates within the students' lifetimes was eye opening to the students and a powerful example that climate change is observable. Scaling across levels and class sizes, the EDDIE modules effectively incorporated many aspects that contributed to learning and engagement—authentic data, open-ended and guided-inquiry learning, skill development—and exposed these students to new approaches in science.

Acknowledgments

We thank the students who participated, EDDIE collaborators Randall Fuller, Lucas Nave, Catherine Gibson, and NEON for hosting EDDIE workshops. Our project was funded by the National Science Foundation's Transforming Undergraduate Education in Science, Technology, Engineering, and Mathematics (TUES) program no. 1245707 and sponsored by the National Association of Geoscience Teachers (NAGT), with administrative and Web support from the Center for Mathematics, Science, and Technology (CeMaST) at Illinois State University and the Science Education Resource Center (SERC) at Carleton College.

Supplemental material

Supplementary data are available at *BIOSCI* online.

References cited

Benson BJ, Bond BJ, Hamilton MP, Monson RK, Han R. 2009. Perspectives on next-generation technology for environmental sensor networks. Frontiers in Ecology and the Environment 8: 193–200.

Brewer CA, Gross LJ. 2003. Training ecologists to think with uncertainty in mind. Ecology 84: 1412–1414.

Carey CC, Gougis RD. 2017. Simulation modeling of lakes in undergraduate and graduate classrooms increases comprehension of climate change concepts and experience with computational tools. Journal of Science Education and Technology 26: 1–11. doi:10.1007/s10956-016-9644-2

Carey CC, Gougis RD, Klug JL, O'Reilly CM, Richardson DC. 2015. A model for using environmental data-driven inquiry and exploration to teach limnology to undergraduates. Limnology and Oceanography Bulletin 24: 32–35.

Chase A, Pakhira D, Stains M. 2013. Implementing process-oriented, guided-inquiry learning for the first time: Adaptations and short-term impacts on students' attitude and performance. Journal of Chemical Education 90: 409–416.

Coleman M. 2013. Metallica play a dome in Antarctica. Rolling Stone. (8 September 2017; *www.rollingstone.com/music/news/metallica-play-a-dome-in-antarctica-20131209*)

Ellwein AL, Hartley LM, Donovan S, Billick I. 2014. Using rich context and data exploration to improve engagement with climate data and data literacy: Bringing a field station into the college classroom. Journal of Geoscience Education 62: 578–586.

Gougis RD, Stomberg JF, O'Hare AT, O'Reilly CM, Bader NE, Meixner T, Carey CC. 2016. Post-secondary science students' explanations of randomness and variation and implications for science learning. International Journal of Science and Mathematics Education 15: 1039–1056. doi:10.1007/s10763-016-9737-7

Gould R. 2010. Statistics and the modern student. International Statistical Review 78: 297–315.

Gould R, Sunbury S, Dussault M. 2014. In praise of messy data. Science Teacher 81: 31–36.

Hampton SE, et al. 2015. The Tao of open science for ecology. Ecosphere 6: 1–13.

Hernandez RR, Mayernik MS, Murphy-Mariscal M. 2012. Advanced technologies and data management practices in environmental science: Lessons from academia. BioScience 62: 1067–1076.

LaDeau S, Han B, Rosi-Marshall E, Weathers KC. 2016. The next decade of big data in ecosystem science. Ecosystems 20: 274–283. doi:10.1007/s10021-016-0075-y

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018

Klug JL, Carey CC, Richardson DC, Gougis RD. 2017. Integrating high-frequency and long-term data analyses into undergraduate ecology classes improves quantitative literacy. Ecosphere 8 (art. e01733).

Langen T, et al. 2014. Using large public datasets in the undergraduate ecology classroom. Frontiers in Ecology and the Environment 12: 362–363.

Michener WK, Jones MB. 2012. Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology and Evolution 27: 85–93.

Miller MCD, Montplaisir LM, Offerdahl EG, Cheng F, Ketterling GL. 2010. Comparison of views of the nature of science between natural science and nonscience majors. CBE–Life Sciences Education 9: 45–54.

Pearce F. 2003. Land of the midnight sums. New Scientist. (7 September 2017; *www.newscientist.com/article/mg17723795-300-land-of-the-midnight-sums*)

Read, EK, O'Rourke, M, Hong, GS, Hanson, PC, Winslow, L, Crowley, S, Brewer, CA, Weathers, KC. 2016. Building the team for team science. Ecosphere 7 (art. e01291). doi:10.1002/ecs2.1291

Rubin SJ, Abrams B. 2015. Teaching fundamental skills in Microsoft Excel to first-year students in quantitative analysis. Journal of Chemical Education 92: 1840–1845.

Schimel D, Keller M. 2015. Big questions, big science: Meeting the challenges of global ecology. Oecologia 177: 925–934.

Strasser, CA, Hampton SE. 2012. The fractured lab notebook: Undergraduates and ecological data management training in the United States. Ecosphere 3: 1–18.

Vanags T, Pammer K, Brinker J. 2013. Process-oriented guided-inquiry learning improves long-term retention of information. Advances in Physiology Education 37: 233–241.

Watson JM, Kelly BA, Callingham RA, Shaughnessy JM. 2003. The measurement of school students' understanding of statistical variation. International Journal of Mathematical Education in Science and Technology 3: 1–29.

Weathers KC, et al. 2016. Frontiers in ecosystem ecology from a community perspective: The future is boundless and bright. Ecosystems 19: 753–770.

Catherine M. O'Reilly (oreilly@ilstu.edu) is an associate professor in the Department of Geography, Geology, and the Environment and Rebekka D. Gougis is with the School of Biological Sciences and the Center for Mathematics, Science, and Technology at Illinois State University, in Normal, Illinois. Jennifer L. Klug is a professor in the Biology Department at Fairfield University, in Connecticut. Cayelan C. Carey is an assistant professor in the Department of Biological Sciences at Virginia Polytechnic and State University, in Blacksburg. David C. Richardson is an associate professor in the Biology Department at the State University of New York at New Paltz. Nicholas E. Bader is an assistant professor in the Department of Geology at Whitman College, in Walla Walla, Washington. Dax C. Soule is a lecturer affiliated with the School of Earth and Environmental Sciences at the City University of New York at Queens College. Devin Castendyk is a senior geochemist with Hatch, at Fort Collins, Colorado. Thomas Meixner is a professor in the Department of Hydrology and Atmospheric Sciences at the University of Arizona, in Tucson. Janet Stomberg is a master's of science student in the School of Biological Sciences and the Center for Mathematics, Science, and Technology at Illinois State University, in Normal, and is part of the instructional faculty in the Department of Biology at Lincoln College, in Illinois. Kathleen C. Weathers is a senior scientist at the Cary Institute of Ecosystems Studies, in Millbrook, New York. William Hunter is the director of the Center for Mathematics, Science, and Technology at Illinois State University, in Normal.

CMO, RDG, JLK, CCC, DCR, NEB, DC, and TM developed modules and taught them in their courses for this study. RDG and JS designed the assessment and collected the data. CMO and RDG analyzed the data. CMO, RDG, and DCR developed the figures. CMO, RDG, JLK, CCC, DCR, NEB, DCS, DC, TM, and JS developed the tables and boxes. All the authors contributed to manuscript editing and writing.

Downloaded from https://academic.oup.com/bioscience/article-abstract/67/12/1052/4582209
by Virginia Tech user
on 31 January 2018