Designing Data Science Workshops for Data-Intensive Environmental Science Research

Anonymous

Abstract

Over the last 20 years, statistics preparation has become vital for a broad range of scientific fields, and statistics coursework has been readily incorporated into undergraduate and graduate programs. However, a gap remains between the computational skills taught in statistics courses and those required for the use of statistics in scientific research. Ten years after the publication of "Computing in the Statistics Curriculum," the nature of statistics continues to change, and computing skills are more necessary than ever for modern scientific researchers. In this paper, we describe research on the design and implementation of a suite of data science workshops for environmental science graduate students, providing students with the skills necessary to retrieve, view, wrangle, visualize, and analyze their data using reproducible tools. These workshops fill a critical hole in the environmental science and statistics curricula, supporting students with opportunities to grow in their skills for computing with data. Open to faculty, staff, and the larger community, these workshops promote continued learning of the tools necessary for working with data and provide additional resources for incorporating data science into the classroom.

Keywords: data science, data visualization, data wrangling, `R`, environmental science, workshops, reproducible research

1 Introduction

Scientific fields have seen profound increases in the volume and variety of data available for analysis. Matched with the growth in computational power, today's scientific researchers are faced with computational and statistical expectations beyond those of the coursework dictated by their curriculum. In the environmental sciences, though statistics courses have been readily incorporated into undergraduate and graduate curricula, an abundance of literature suggests that these curricula fail to equip graduate students with the computing skills necessary for research in their field (Andelman et al., 2004; Green et al., 2005; Hampton et al., 2017; Hernandez et al., 2012, Mislan, Heer, & White, 2016; Teal et al., 2015; Theobold and Hancock, 2019). Only one of these studies (Theobold and Hancock, 2019), however, acknowledges the substantial role statistics courses could potentially play in students' acquisition of computational skills.

Over the last 10 years, a large number of statistics educators have echoed Nolan and Temple Lang's call to "embrace computing and integrate it fully into statistics undergraduate major and graduate programs" (Nolan and Temple Lang, 2010, p. 97; Baumer, 2015; Baumer, Horton, & Wickham, 2015; Cetinkaya-Rundel and Rundel, 2018; Cobb, 2015; Hardin et al., 2015; Horton and Hardin, 2015; Kaplan, 2018; McNamara and Horton, 2018). Indeed, the American Statistical Association Curriculum Guidelines for Undergraduate Programs in Statistical Science (American Statistical Association Undergraduate Guidelines Workgroup, 2014) reflect the increasing importance of data science skills. Despite this campaign for computing in the statistics classroom, graduate-level statistics service courses have largely been overlooked, even though their potential impact is substantial. Unlike courses designed for an undergraduate or graduate program in Statistics, these service courses often act as the sole exposure to computing with data prior to the start of a student's independent research.

The intention of this research is to (1) describe the computing skills necessary for research in the environmental sciences, (2) investigate how these skills can be infused into currently existing extracurricular workshops, and (3) understand the experiences of attendees of these workshops. We consider the following research questions:

1. What do environmental science faculty members identify as the key computing skills

graduate students require to implement statistics for research in their field?

2. How can the key computing skills identified by environmental science faculty be incorporated into currently existing workshop materials?

3. What are the experiences of individuals attending data science workshops?

To investigate these questions, we executed a three-phase design-based implementation research model (Fishman et al., 2013). In the first phase, we conducted in-depth interviews with faculty from environmental science fields regarding the computational skills they believe are necessary for graduate students to succeed in their research. Phase two then focused on adapting currently existing workshop resources to design of a series of data science workshops targeting the key computational skills distilled from these interviews. The final phase consisted of implementing the workshops and collecting survey responses from the workshop attendees regarding their experiences participating in each workshop.

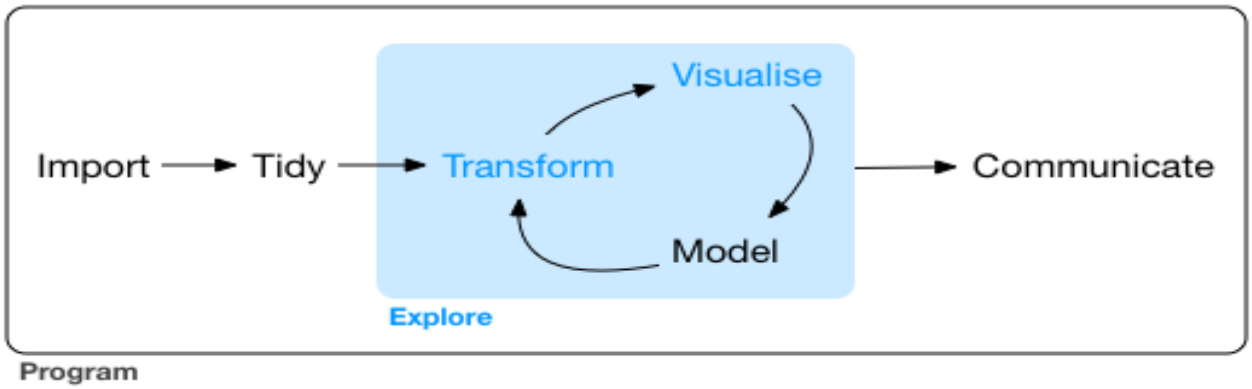

Figure 1: Data Analysis Cycle, Wickham, H. & Grolemund, G. (2017) *R for Data Science*. Sebastopol, California: O'Reilly.

For this research, the collection of disciplines who perform research across a variety of environmental science fields are captured under the term "environmental science." At our institution, these are the fields of Ecology, Land Resources Environmental Sciences, Plant Sciences, and Animal and Range Sciences, whose students are required or highly recommended to complete graduate-level statistics coursework for a masters or doctoral degree. In this paper, the "data analysis cycle" consists of all stages in the data analysis

process, from data importation to data exploration to the communication of results, where data modeling is but one component (Figure 1). The "data science skills" necessary to engage in this cycle may include general programming concepts such as loops, user-defined functions, or conditional statements. However, the cornerstone of data science skills differs fundamentally from general programming skills, with a focus on data rather than computer architecture, design, and applications.

We begin by outlining areas of research that address the computational and statistical training of graduate students in the environmental sciences and the potential for extracurricular workshops to fill in the gaps. Next, we outline the design-based implementation research methodology used to design and implement a suite of data science workshops tailored to environmental science graduate students. Section 4 summarizes the first phase of research, which outlined the computational skills faculty members identified as necessary for graduate students to succeed in their independent research. Next, Section 5 discusses how these identified skills were interwoven into existing data science workshop materials for researchers in the environmental sciences. Section 6 summarizes the backgrounds and experiences of the workshop attendees during the 2018-2019 academic year, and describes the research conducted on the implementation of the workshops. We then outline future research plans for a second iteration of this design work. To close, we revisit the current climate of computing in the statistics curriculum for service courses and describe how these types of extracurricular workshops can assist in further integration of computing into these classrooms.

2 The Current Climate of Statistics and Computing in the Environmental Sciences

Due to the substantial growth in the volume and variety of available data over the last two decades, the practice of environmental science has changed dramatically. Advances in technology have made computationally heavy applications of data science techniques—such as management and coalition of large data sets, high frequency spatial and temporal data visualization, and hierarchical Bayesian modeling—essential understandings for environmen-

tal science research. This flood of data has "challenged the research community's capacity to readily learn and implement the concepts, techniques, and tools" (Hampton et al., 2017, p. 546) necessary for data-intensive environmental science research, creating a crucial need to re-evaluate how our educational system can better prepare current and future generations of researchers (Green et al., 2005; Hampton et al., 2017).

2.1 Computing in the Environmental Sciences Curriculum

Arising from a decade of mumblings (Andelman et al., 2004; Dodds et al., 2007, 2008; Eglen, 2009; Green et al., 2005; Hastings et al., 2005; Kelling et al., 2009; Wilson, 2006; Wilson et al., 2008; Wing, 2006), 2010 brought two studies on the computational ill-preparation of environmental students by their curriculum. First, Strasser and Hampton found that undergraduate students were not being prepared with the data management tools necessary to engage in environmental science research, as fewer than 20% of instructors were including key data management topics in their courses, such as workflows, databases, and reproducibility. The importance of these skills, however, was affirmed by the majority (77%) of instructors. Yet, instructors largely stated that "data management should be taught in a different course" (p. 10). The results of this study suggested that—across institutions—"data management education is not currently a priority for ecology instructors" (p. 10). That same year, an environmental science graduate student led a large scale study of the computational experiences of future environmental scientists (Hernandez et al., 2012, p. 1068). In a survey of environmental science graduate students across the United States, the authors found that over 74% of the students surveyed reported they had no skills in any programming language—including `R`—and only 17% reported basic skill levels in any programming language. Hence, a large number of students may be leaving their graduate programs without the data science skills necessary for data-intensive research in their field. Hernandez et al. suggested that student-focused workshops could bridge this gap, by "providing intensive environments" where students could learn "particular methods or technologies" (p. 1075). The authors also noted that developing and offering these workshops would be simpler than developing new courses to organize and implement.

Out of these calls for high-quality resources for scientific computing, emerged the Carpentries project (The Carpentries, 2019). Housing Data, Software, and Library Carpentry,

the Carpentries comprises "communities of instructors, trainers, maintainers, helpers, and supporters" all sharing the mission to "teach foundational computational and data science skills to researchers." These workshops address the need for "good training resources for researchers looking to develop skills that will enable them to be more effective and productive" (The Carpentries, 2019, p. 135). Furthermore, these workshops are necessary, because "training in data and computing skills is still largely absent from undergraduate and graduate programs," so "most or all of what [researchers] know about data management, analysis, and sharing has been learned piecemeal, or not learned at all" (p. 136). Data Carpentry provides domain-specific training for researchers in the "core data skills for efficient, shareable, and reproducible research practices" (Data Carpentry, 2020). As part of this mission, the Carpentries collaboratively develops publicly available lessons for specific populations of researchers, which do not assume that attendees have any prior knowledge before attending the workshops. Teal and colleagues acknowledge that, while the Data Carpentry workshops "will not be able to teach researchers all of the skills they need in two days," the workshops "are a way to get started," lowering the activation energy required and empowering researchers "to be able to conduct the analyses necessary for their work in an effective and reproducible way" (p. 143). The success of these workshops can be viewed as a "symptom of the current curriculum's shortcomings" (Hampton et al., 2017, p. 547), as there continues to exist a "paucity of systematic training within university programs to equip students with the computational skills they need to conduct data-intensive research" (p. 547).

Disappointingly, none of these conversations have acknowledged the substantial role students' statistics education potentially plays in their attainment of the data science skills necessary for research. Hampton and fellow environmental science researchers claim that "three decades ago, environmental scientists were ill-prepared to use statistics in their research, and now statistics preparation is considered vital." In fact, in ecology today, it is "extremely difficult to publish a manuscript without any statistical testing" (Hampton et al., 2017, p. 547). Furthermore, with the invention of the RStudio integrated development environment (IDE) (RStudio Team, 2015b), the software used throughout the data analysis cycle by environmental science researchers has changed. In 2017, the number of studies in environmental science journals reporting the use of `R` as the "primary tool reported in data analysis" was 58%, compared to 11.4% in 2008 (Lai et al., 2019, p. 1). Moreover, the pre-

ponderance of environmental science graduate students are now required to produce code as part of their research (Mislan et al., 2016). The clear need for data science proficiency in environmental science research requires a transformation of the graduate curriculum similar to that which infused statistics preparation into the required graduate coursework.

2.2 Computing in the Statistics Curriculum

Changes in the digital age have also had "a profound impact on statistics and the nature of data analysis" (Nolan and Temple Lang, 2010, p. 97), with today's skills differing substantially from what was needed but five to ten years ago. In the year following the publication of "Computing in the Statistics Curriculum" (Nolan and Temple Lang, 2010), the Mckinsey Report (Manyika et al., 2011) was published. The McKinsey report stated that, by 2018, "the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions" (p. 3). With calls to transform the undergraduate statistics curriculum resounding nationally, the 2014 American Statistical Association (ASA) President, Nathaniel Schenker, convened a workgroup to update the association's guidelines for undergraduate programs. These new guidelines included an increased emphasis on data science skills and real applications, specifically students' ability to "access and manipulate data in various ways, use a variety of computational approaches to extract meaning from data, program in higher-level languages" (American Statistical Association Undergraduate Guidelines Workgroup, 2014, p. 7).

With this curricular momentum, in 2015, *The American Statistician* produced a special issue on "Statistics and the Undergraduate Curriculum," to encourage submissions of broader topics in the statistics curriculum. Articles in the special issue ranged from detailing how computing should be included throughout the Statistics curriculum (Green and Blankenship, 2015; Tintle et al., 2015; Hesterberg, 2015), to presenting thoughts on how data science topics should be integrated into undergraduate statistics courses, (Nolan and Temple Lang, 2015; Grimshaw, 2015; Baumer, 2015; Hardin et al., 2015). In the same issue, George Cobb provocatively stated that the statistics curriculum needed to be rebuilt "from the ground up" (2015), as "what we teach lags decades behind what we practice" and "the gap between our half-century-old curriculum and our contemporary statistical practice continues to widen"

(p. 268). Moreover, despite the issue's focus on the broader statistics curriculum, authors continued to lament that the current Introductory Statistics curriculum teaches a snapshot of the entire data analysis cycle, "wherein challenges with data computational methods, and visualization and presentation are typically elided" (Baumer, 2015, p. 336).

The following year, however, brought the revised GAISE college report (Revision Committee, 2014), creating a push for reform in the Introductory Statistics curriculum. The six recommendations originally outlined by the committee in 2005 continued, but the authors suggested two new emphases for the first recommendation (teach statistical thinking), which reflect the modern practice of statistics. First, statistics educators should "teach statistics as an investigative process of problem-solving and decision making," and we should "give students experience with multivariable thinking" (2014, p. 3). These recommendations reiterate the sentiments heard throughout the statistics community, that students should emerge from our courses with the understanding that data analysis "isn't just inference and modeling, it's also data importing, cleaning, preparation, exploration, and visualization" (Cetinkaya-Rundel, 2018). Yet, the inclusion of these topics in the Introductory Statistics curriculum is still a heated discussion. Many educators continue to believe (1) that it is not possible to teach statistical concepts and programming in just one course, (2) that teaching programming takes up valuable time which could be used towards teaching important statistical concepts, or (3) students are not interested in learning to program (Cetinkaya-Rundel, 2018). Thus, many students leave their Introductory Statistics course without "a set of practices and attitudes about data that are immediately applicable to their lives" (Gould, 2010, p. 309).

Amidst these conversations, `R` packages were being created, which would fundamentally changing how users interact with `R`. These `R` packages, universally known as the "`tidyverse`," have created user friendly `R` tools which "share an underlying design philosophy, grammar, and data structures" (Wickham, 2017). Statistics educators have begun to leverage these tools in the Introductory Statistics classroom to teach reproducibility (Baumer et al., 2014), data management (Baumer et al., 2015), dynamic data (Hardin, 2018), and big data (Wang et al., 2017). While there exists a growing momentum to incorporate these new `R` tools into the Introductory Statistics classroom, attention has yet to be paid to alternative statistics service courses, such as those taken by environmental science graduate students. These courses, like Introductory Statistics, serve graduate students from a variety of scientific

backgrounds. However, unlike an undergraduate Introductory Statistics course, students are expected to emerge from their statistics coursework with the ability to complete the analyses required for their research.

The frustrations echoed by environmental science educators (Hampton et al., 2017; Teal et al., 2015) suggest that, despite the inclusion of statistics coursework into these graduate programs, students continue to leave the statistics classroom without the data science skills necessary to participate in the data analysis cycle in their own research. The fundamental question raised ten years ago by Nolan and Temple Lang still applies today: do our students leave the statistics classroom able to "compute confidently, reliably, and efficiently?" (2010, p. 100). An in-depth study of environmental science graduate students' experiences acquiring the computing knowledge necessary for their research answered this question with a resounding 'no' (Theobold and Hancock, 2019). Like the hypothesis of Teal and colleagues (2015), these students did not attribute their acquisition of the data science skills necessary for their research to the statistics courses they took for their degree. Rather, students gained the data science skills necessary to engage in the entire data analysis cycle through independent research experiences, an "all-knowing" past or current graduate students, and peer networks. Ten years after the publication of "Computing in the Statistics Curriculum," we continue to assume that "students will 'pick up' the skills they need" to participate in the data analysis cycle outside of their statistics coursework (Gould, 2010, p. 309).

2.3 Extracurricular Workshops to Bridge the Gap

Reiterated by both statistics education and environmental science researchers alike (Nolan and Temple Lang, 2010; Teal et al., 2015), this lack of training in computational skills impedes the progress of scientific research, sends the signal to students that computing is not of intellectual importance, and is laden with hidden costs. Students may pick up bad habits, misunderstandings, or the wrong concepts, learn just enough to get what they need done, spend weeks or months on tasks that could be done in hours or days, and they may be unaware of the reliability and reproducibility—or lack there of—of their results (Nolan and Temple Lang, 2010, p. 100; Teal et al., 2015, p. 136). But why are these skills still so rarely included in these service courses when the need for them is widely recognized?

Environmental science educators have reiterated the challenges in integrating comput-

ing into the curriculum outlined by Nolan and Temple Lang. These barriers can be boiled down to "attempting to fit more material into already-full courses and curriculum, which are taught by people who do not feel prepared to address topics relevant to big data and data-intensive research" (Hampton et al., 2017, p. 547). These hurdles are potentially even greater for graduate-level statistics service courses. Instructors of these courses are explicitly told the statistical content students are expected to learn, and are implicitly assumed to be teaching students the data science skills necessary for them to participate in the entire data analysis cycle. Claiming these graduate students ought to take additional, data science specific courses to obtain these skills is infeasible for many, as graduate programs leave little room for additional coursework.

Until computing has been meaningfully integrated into these service courses, extracurricular workshops hold the potential to address the gap between the computational preparation of students by their coursework and the computational requirements of their research. Extracurricular workshops are not a direct substitute for the prolonged instruction of these skills that occurs in a course. But, short, intensive workshops, such as those provided by the Carpentries, are able to teach immediately useful skills that can be taught and learned quickly, keep learners active by using live coding and formative assessment, work with a learners from a variety of backgrounds, and build learners' self efficacy (Word et al., 2017). Additionally, because workshops are able to thrive outside of university curricula, they hold the ability to "adapt materials rapidly and remain on the leading edge of technological development" (Hampton et al., 2017, p. 547). Furthermore, workshops offer the opportunity for a wide variety of researchers, not just students, to acquire the data science skills necessary for data-intensive research, supporting the broader community of researchers.

3 Methodology

Improving environmental science graduate students' access to "powerful, effective learning opportunities" (Fishman et al., 2013, p. 137) necessitates understanding the skills required for these students to be successful in their research. Design-based implementation research (DBIR) (Cobb et al., 2003; Fishman et al., 2013; O'Neill, 2012) "offers a model for the design and testing of innovations within the crucible of classrooms and other contexts for learning"

(Fishman et al., 2013, p. 140). Rather than creating workshops covering content outside parties believe are important, DBIR uses collaboration with members of the community to develop "evidence-based improvements" (p. 143) to teaching innovations—situating community members as "co-designers of solutions to problems" (Fishman et al., 2013, p. 140) rather than bystanders. This collaboration is critical when developing resources for researchers in the broader scientific community, as the discipline of Statistics was developed to support research in other scientific disciplines to evaluate evidence obtained from data.

This paper describes the results of the first iteration of this DBIR, consisting of three phases. Section 4 summarizes the first phase of this research, investigating the computing skills necessary for environmental science research by graduate students. As the direct supervisors of graduate students, environmental science faculty members are potentially aware of the computing skills that are vital to researchers in their respective fields. Thus, interviews with faculty members from these fields allow for us to gain an understanding of the essential skills required of environmental science graduate students. Phase two of this research, described in Section 5, details how the skills identified during phase one were used to tailor currently existing Data Carpentry (Data Carpentry, 2020) and Software Carpentry (Software Carpentry, 2020) lessons to meet the needs of graduate students in the environmental sciences. Finally, Section 6 chronicles the final phase of this research, implementing and evaluating these workshops. This final evaluation phase focuses on the backgrounds and experiences of workshop attendees, rather than the workshop content or learning outcomes of attendees, which are described as directions for future research.

4 Outlining the Computing Skills Necessary for Environmental Science Research

In the spring of 2017 and fall of 2018, faculty members from diverse fields within the environmental sciences were invited to participate in a one-hour interview. All faculty members currently overseeing a graduate student from the Ecology, Land Resources Environmental Science, Animal & Range Sciences, and Plant Sciences & Plant Pathology departments were emailed requesting their participation in this research. While some faculty enthusiastically

agreed to participate, others declined for three main reasons—they hadn't directly overseen a graduate student recently, they deemed themselves to be weak in statistics, or they were unavailable to meet. Table 1 outlines the number of faculty requested for participation and the number of faculty interviewed, by department affiliation.

Department	Faculty Invited	Faculty Interviewed
Animal & Range Sciences	7	2
Ecology	15	8
Land Resources Environmental Sciences	24	8
Plant Sciences & Plant Pathology	15	5

Table 1: Number of faculty members requested for participation and interviewed, by department.

4.1 Data Collection

Faculty agreeing to participate were interviewed regarding (1) the computational skills they believe are necessary for masters and doctoral students to implement statistics for research in their field, and (2) how they believe graduate students acquire these necessary skills. The full interview protocol is included in the GitHub repository associated with this manuscript [1].

Based on faculty's responses, the interviewer asked follow-up questions to further explore why the faculty believe the computational skill(s) in question are necessary. For instance, if a faculty voiced the need for students to be able to build a data workflow, further information was sought regarding what specific computing skills this would require. Alternatively, when the response from faculty consisted of the statistical understandings necessary for graduate student researchers, follow-up questions were asked to delve further into what computing skills a student may require to successfully implement this type of statistical analysis with their data. Not only did these interviews provide valuable feedback on *what* content the workshops should include, they also added insight into *why* workshops form an ideal mode of delivery for this needed training.

[1] Materials associated with this manuscript are available at https://github.com/atheobold/data-science-workshops-jse

4.2 Data Analysis

The primary author lead a three-stage data analysis process (Miles, Huberman, Saladaña, 2014). During the first stage, the interviews for every faculty member were transcribed verbatim. Following this process, the primary author read the transcripts independently, highlighting excerpts where computing skills were discussed. The author then created descriptive codes for the skills faculty identified as necessary in each of these excerpts. At the close of this stage, the author examined these codes for specific references to computing skills currently addressed in Data Carpentry's *Data Analysis and Visualization in `R` for Ecologists* lesson (Michonneau et al., 2019).

Following this process, the primary author began the second stage of analytical coding. This stage acts as a method of synthesizing descriptive summaries, tying together "different pieces of data into a recognizable cluster," demonstrating how the data are instances of a general concept (Miles et al., 2014, p. 95). During this stage, skills were linked thematically, and categories that held across multiple interviews were retained. For example, every faculty voiced students' need to work with data in `R`. These themes were initially categorized as "working with data," with additional categories of data wrangling, and data visualization created. Next, the author searched through these themes to uncover how each theme related to the others. Through this process it was determined that certain categories captured similar constructs, and were merged into a single category, whereas other constructs were voiced independently, and separate categories were formed. For example, while every faculty voiced students' need to work with data in `R`, these sentiments were voiced alongside students' need to preform other data wrangling operations, such as reorganize data, filtering out rows of data, selecting columns, creating new variables, or modifying existing variables. Hence, the themes of "working with data" and "data wrangling" were merged into the single theme of "working with data." Alternatively, while reproducibility is a key aspect to working with data in `R`, the skills identified by faculty which this theme captures were not voiced alongside a specific software. However, when these faculty commented on the need for students' work to be reproducible, `R` was continually mentioned as the vehicle to support this need.

In the final stage of the analysis, the primary author searched the faculty transcripts for evidence supporting the emerging themes, scrutinizing whether each identified skill fit

into the existing themes. Following this validation process, the first and second authors met to discuss the rationale for each code and inspect the skills identified by faculty in the context of the emergent themes. These final themes ground the theory for creating an effective intervention promoting the acquisition of computing skills necessary for graduate-level environmental science research. reproducible research

4.3 Skills Identified by Environmental Science Faculty

While some faculty had difficulties disentangling the statistical methods students use in research from the computing required to implement those methods, many were able to express the expectations they held for graduate students in their field throughout the entire data analysis cycle. A substantial overlap was seen between faculty expectations and the components of "data acumen" outlined by NAS (National Academies of Sciences, Engineering, and Medicine, 2018), with faculty expectations falling into three categories: (1) working with and wrangling data, (2) data visualization, and (3) reproducibilty.

4.3.1 Working with Data

All of the faculty interviewed believed that students' experiences in the statistics classroom do not adequately prepare students to work with and organize large, messy datasets. The need to manipulate large datasets is not unique to the environmental sciences. In fact, a faculty member stated that it is "not uncommon to be analyzing half a million records, but I think it's uncommon to be doing it effectively or efficiently." As graduate students perform their research, they are required to assemble datasets for analyses. This requires students to think about "storing data, managing data, matching data, and collating data," potentially merging a variety of data types into one meaningful data set. Every faculty member emphasized that students need to know how to "organize their data and get it in a way that can be used by `R`."

Often included in these skills for working with data are tasks that require reorganizing data formats from wide to long or vice versa—a skill which every faculty member griped is not acquired through the standard curriculum. "Most of them, when you're like 'long form, wide form, samples as rows, variables as columns,' they kind of look at you like 'what?'."

Standard examples in statistics courses provide students with data which are the product of cross-tabulation, so students are never forced "to figure out how to get the cross-tabulation you need, so that you can bring it into R and do your regression." These concerns reiterate the importance of "data management and curation" detailed by NAS, who stated that "at the heart of data science is the storage, preparation, and accessing of data" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 26).

4.3.2 Data Visualization

The importance data visualization plays in every stage of students' research was emphasized by every faculty member interviewed. Faculty affirmed that students should possess the ability to create visualizations of their data early, both for checks of data quality and to explore relationships. These expectations align with the the facility outlined by NAS (2018), who stated that students need to have the ability to "present data in a clear and compelling fashion" (p. 26). One faculty member declared that students' ability to look at their data in different ways dramatically shapes their research potential, and the tools available today allow for researchers to create visualizations precisely tailored for each investigation. Many faculty voiced the usefulness of the `ggplot2` (Wickham, 2016) package for students' knowledge of producing data visualizations, lowering the barrier for students to learn "how to visualize [their] data to explore and understand it."

4.3.3 Reproducibility

Every faculty member emphasized the usefulness of "manipulating data in ways that are repeatable," through using such programs as R. Across environmental science disciplines, faculty concurred that many students do not use R for data wrangling. Instead, students rely on Excel because "they are not comfortable enough with the code or [R] is kind of a black box" or that when they "don't have that instant connection with [their] data, I think it fundamentally boils down to fear." Concern was raised for the students using Excel to wrangle their data, as "they would never find [their] way back to what the original data set would have been" and that their advisers would have no way to understand why data are missing. These advisers encourage students to avoid these brute force Excel manipulations, yet students may not have the computing skills necessary to perform the same data wrangling

task in a scripted and reproducible manner, as their courses provide little to no exposure in carrying out the full data analysis cycle. These faculty concerns parallel the "workflow and reproducibility" acumen outlined by NAS, who stated that students need to "be exposed to the concept of workflows" and gain experience with the "software software systems that enable building workflows (e.g., `R` and Python) and how to document what they do (e.g., R Markdown and Jupyter Notebook)" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 28).

4.3.4 How Students Gain Computational Skills

When asked why students are not acquiring computing skills in the courses required for their degree, a faculty member stated, "We don't really have anyone to teach that. It's not that it isn't valuable, but there is no one to teach it." When pressed as to why other faculty feel uncomfortable teaching computing, this same faculty member stated, "[they believe] most graduate students come in knowing more about the tools one might use to manipulate data than their advisers do." Other faculty bemoaned the gaps between the computational skills of their graduate students and their own training, "I think that more and more in our field, my generation is sort of just catching up the next generation." These gaps impact the assistance faculty can provide to their students, as "increasingly faculty feel that they're not at the forefront of their programming abilities, so their students are being self-taught and are often computationally ahead of them." Many faculty lamented their own deficient computational abilities, with some stating that they "feel personally out of touch, because [students] work in `R` and I haven't taken the time to learn `R`, because of my training and my age." These faculty understand that often students are required to learn these skills on their own because "there is definitely a gap between the code I can help them with."

Although faculty feel that there may exist gaps between their own knowledge of working in `R` and their students', every faculty member affirmed the importance of students acquiring the computational abilities needed to perform data-intensive research. Indeed, it is not necessary for every student to be an expert, but faculty underscored the value of resources for students to acquire these computational skills necessary for their research. The majority of faculty voiced that it is often assumed that graduate students should be able to analyze their data because they've taken a statistics course. Yet, faculty members acknowledge the

poor computational preparation of their students—even after taking a statistics course—and thus "encourage [students] to use anything they can find to get more tools in [their] tool box."

5 Designing Data Science Workshops for Graduate Students in the Environmental Sciences

The second phase of this research attended to the development of a suite of data science workshops targeted to graduate students in the environmental sciences. Skills identified through faculty interviews were incorporated into a set of four 3-hour workshops covering (1) the basics of programming in `R`, (2) intermediate programming tasks in `R`, (3) creating appropriate and effective data visualizations, and (4) cleaning and merging data in preparation for analysis and visualization, all using reproducible tools.

The first workshop in this series does not assume that attendees have any previous experience working in `R`, and each workshop builds on knowledge acquired at previous workshop(s), without the expectation that attendees have acquired additional knowledge or skills between workshops. The materials for these workshops were adapted from Data Carpentry's *Data Analysis and Visualization in `R` for Ecologists* lesson (Michonneau et al., 2019), a curriculum maintained by experienced researchers in ecological fields, which uses ecology-specific data contexts [2]. Using the themes which emerged in interviews with environmental science faculty, we were able to tailor this workshop series to suit the needs of this population of graduate students as they prepare for and participate in data-intensive research. The workshop materials developed for this research are available through GitHub [3], with video tutorials recorded and available through our institution's library [4].

[2]This work is a derivative of *Data Analysis and Visualization in `R` for Ecologists* (https://datacarpentry.org/R-ecology-lesson/) by Data Carpentry, used under CC BY (https://creativecommons.org/licenses/by/4.0/).

[3]GitHub repository (link)

[4]MSU Library videos (link)

5.1 Participant-Centered Learning

These workshops are taught in a technology-enhanced active learning (TEAL) classroom that seats up to 35 individuals. The room's monitors make it easy for attendees to watch as the instructor live codes, and the open layout of the room allows helpers to engage with attendees as they work. Each workshop has one lead instructor and two to three workshop assistants. During each workshop, topics are introduced by the instructor, followed by a group discussion of an example. Finally, attendees group up to complete a set of hands-on tasks, applying the concepts covered in that section of the workshop. These tasks allow for attendees to "learn the computational aspects as part of an interesting, challenging, and confidence-building process" (Nolan and Temple Lang, 2010, p. 101).

5.2 Data Context

Emphasized by both the NAS and these faculty, "effective application of data science to a domain requires knowledge of that domain" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 29). Hence, data science instruction ought to be grounded in "substantive contextual examples," to "ensure that data scientists develop the capacity to pose and answer questions" with data relevant to them (p. 30). Therefore, the data used in the creation of these workshops are ecological in nature, originating from Montana Fish, Wildlife, and Parks and the Portal Project Teaching Database (Ernest et al., 2018). These data highlight a variety of aspects that commonly occur in environmental data, including multiple sampling instances, mark-recapture, biological measurements, and meta- and micro-level data.

5.3 Computing Tools for Environmental Science Research

The structure and context of these workshops include a statistical programming language used extensively throughout environmental science research (`R`), environments which facilitate the learning of `R` (RStudio and RStudio Cloud), and tools that promote reproducibility throughout the entire data analysis cycle (R Markdown).

5.3.1 Why R?

The use of R is widespread throughout the environmental science research community, a dramatic change over the last decade (Lai et al., 2019). Presently, R includes over 100 packages frequently used in ecological data analysis, as highlighted in the CRAN Task View: Analysis of Ecological and Environmental Data (https://CRAN.R-project.org/view=Environmetrics). Furthermore, with the invention of the RStudio IDE (RStudio Team, 2015b), this user-ship continues to increase. R is free and open source, so attendees learn a statistical programming language that will be accessible to them throughout their careers. Furthermore, unlike other data analysis software used by environmental scientists, such as program MARK, VORTEX, or RMAS, with R, their results do not depend on remembering the sequence of buttons they clicked. With the shocking realization that large numbers of modern scientific findings cannot be replicated (The Economist Editorial, 2013; Johnson, 2014) and the growing appreciation for reproducible methods of data analysis in ecological research (CASSEY and BLACKBURN, 2006; Ellison, 2010; Morrison et al., 2016; Powers and Hampton, 2019), today's researchers in scientific fields are becoming more aware of the need for a reproducible data analysis workflow.

5.3.2 Why RStudio?

RStudio is a free computer application that allows you access to the resources of R, while providing you with a comfortable working environment (RStudio Team, 2015b). The RStudio IDE "makes [programming] less intimidating than the bare R shell" (Cetinkaya-Rundel and Rundel, 2018, p. 59). Additionally, the RStudio environment is consistent across operating systems, which is not the case for other statistical software packages. Because RStudio is an IDE, it includes integrated help files, intelligent code completion, and syntax highlighting—all of which help to reduce the learning curve. Additionally, RStudio makes reproducibility simple with dynamic R Markdown documents, allowing for a full integration of the data analysis workflow.

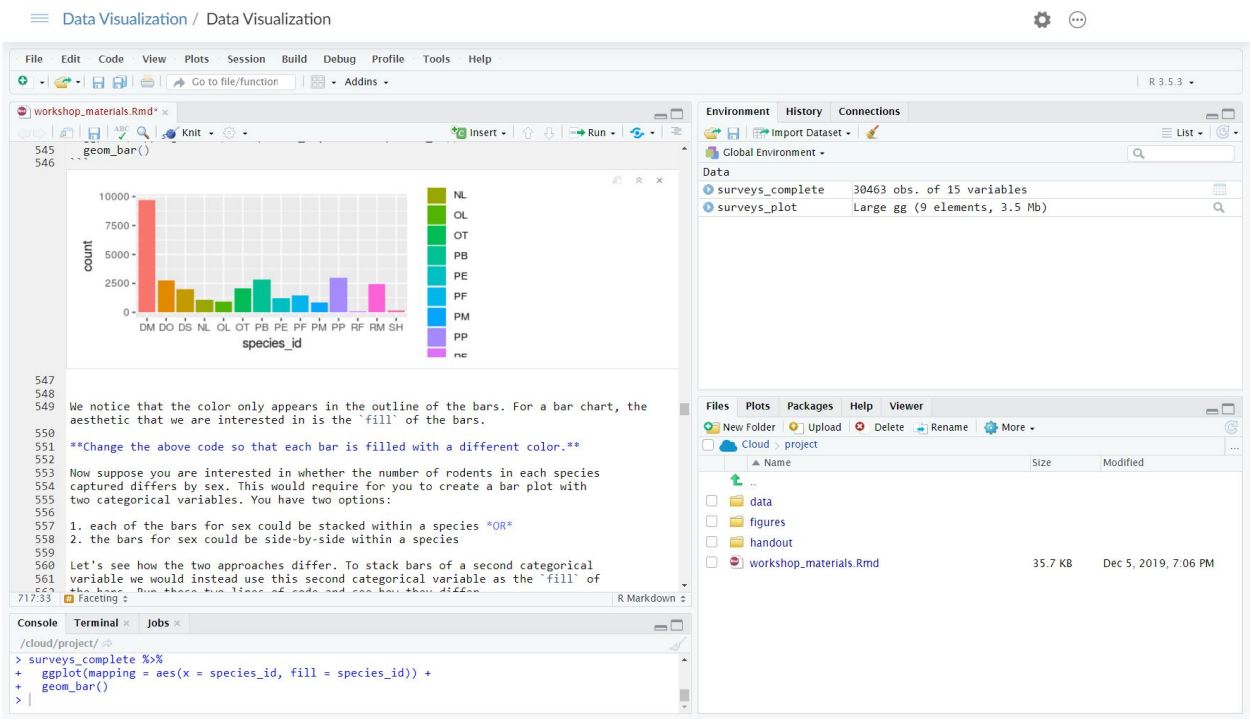

Figure 2: RStudio Cloud workspace environment for *Data Visualization with* `ggplot2` workshop. Every workshop works in an RStudio project, containing a master `R Markdown` file, a data folder containing the data used in the workshop, and the handout produced for attendees.

5.3.3 Why RStudio Cloud?

The RStudio Cloud was created as a platform to make it easy to do, share, teach and learn data science using `R` (RStudio Team, 2015a). Through the Cloud, attendees are able to access publicly available workshop materials, without worrying about software installation, package installation, or transferring data. Each workshop is contained in an organized RStudio project directory, so attendees are exposed to best practices for reproducible project construction. Workshop participants interact with the workshop's materials in the same manner as a locally installed version of RStudio, as seen in Figure 2.

5.3.4 Why R Markdown Documents?

R Markdown documents provide an easy-to-understand framework to combine statistical computing and written analysis in a single document, helping to break the copy-paste paradigm for generating statistical reports (Baumer et al., 2014). During the workshop, R Markdown documents allow for attendees to keep their code organized and their workspace clean, which is unnatural for new learners. Each workshop's master R Markdown document contains blocks of code and descriptions for every topic covered, allowing for participants' exploratory work to be saved within a topic. For additional information on R Markdown documents see Baumer et al..

5.4 Workshop Content

5.4.1 Introduction to `R`

This first workshop in the series covers the basics of learning to program in `R`. The workshop first introduces the RStudio environment and project work flow in RStudio, discussing working directories and relative paths. Next, the workshop progresses through tools for working with vectors and lists of different data types, motivating methods for working with dataframes in `R`. After learning how to import data into `R`, the workshop proceeds through inspecting data, extracting data, and changing data types. Motivated by obtaining unexpected summary statistics when working with missing data, the workshop introduces `R` help files to inspect function arguments and their default values. These help files are called upon as participants use base `R` functions to create summaries of the data, perform basic data cleaning, and produce both univariate and bivariate visualizations of the data.

5.4.2 Intermediate `R`

This second workshop in the series builds off of the content covered in *Introduction to `R`*, without any expectations of attendees having additional knowledge or skills. The workshop begins with a review of creating objects in `R` and working with vectors and dataframes. The workshop then progresses through the use of relational statements in `R` and how to link these statements together using and (`&`), or (`|`), and not (`!`) conjunctions. Next, the workshop

dives into the use of conditional statements, stepping from `if`, to `if else`, to `else if` statements.

The second half of the workshop covers methods to iterate or replicate the same set of instructions many times. For-loops are introduced as a popular way to iterate or replicate the same set of instructions many times. Participants work through examples of a for-loop and a recursive for-loop, in the context of repeated operations on a dataset. These exercises motivate the discussion of why many `R` users recommend instead using vectorization for non-recursive for-loops.

Lastly, functions are presented as an approach to replicate the same set of instructions in multiple locations throughout your code. Persuaded by an `R` script which copies and pastes the same process multiple times, participants understand the difficulty in discerning the underlying process and spotting mistakes. Participants are then tasked with transforming this copy-paste process into a function. By parsing out the function writing process into a set of steps that one walks through once you've copied and pasted your code multiple times, participants have a more intuitive sense for why functions are useful and how they can create them in their own code.

The content in this workshop, excluding relational statements, is not included in Data Carpentry's *Data Analysis and Visualization in `R` for Ecologists* lesson. Instead, many of these concepts are covered in Software Carpentry's *`R` for Reproducible Scientific Analysis* lesson. Yet, creating modularized code—which uses conditional statements, for-loops, and user-defined functions—are skills that many environmental science faculty asserted were necessary for graduate students to possess as they perform independent research.

5.4.3 Data Wrangling with `dplyr` and `tidyr`

Following the *Intermediate `R`* workshop, the *Data Wrangling* workshop continues forward with common data wrangling issues faced by environmental science researchers. Inspired by the difficulty of reading bracket subsetting and how cumbersome it can be to remember the different base `R` functions and formats to wrangle your data, this workshop introduces the `dplyr` (Wickham et al., 2018) and `tidyr` (Wickham, 2014) packages from the `tidyverse` (Wickham, 2017). Much of `R`'s language has not changed over the last 20 years, which leaves the desire for a "smoother, more efficient, and more readable pipeline for modern `R`

workflows" (Ross, Wickham, & Robinson, 2017, p. 19). The `tidyverse` packages share common interfaces and data structures that make it simpler to learn data wrangling tasks and allow for the process to flow naturally from one step to the next.

The workshop begins with a description of the purpose of the `dplyr` package and an outline of six of the common "verbs" that handle common data wrangling challenges. Participants learn how to select columns with `select()`, filter rows with `filter()`, add new columns and modify existing columns with `mutate()`, create a table of summary measures by groups using `summarise()` and `group_by()`, and change the ordering of a tables rows with `arrange()`. Prompted by the need to perform a sequence of multiple data wrangling operations, participants learn how to connect each of these data wrangling verbs using the pipe operator (`%>%`).

Next, the concept of relational data is outlined, impelled by the need to integrate additional data files for analysis. Participants are introduced to the idea of key-value pairs and then use these pairs to map how the data with which they have been working can be joined with additional data files. After discussing the four types of joins (inner, left, right, full), participants use the `left_join()` and `right_join()` functions to join the three datasets used in the workshop.

The final topic of the workshop involves data reorganization, beginning with a discussion of "tidy" data. Up until now, the data were presented to participants in a "tidy" format, where every observation is one row, each variable has a column, and every value has one cell. This idea is then used to describe 'long' and 'wide' data formats, and a discussion around why each format may be useful ensues. The `tidyr` package is introduced to alleviate the burden of these types of data reorganizations, with an introduction to the `gather()` and `spread()` functions. In groups, participants then work through a final exercise applying the skills acquired throughout the entire workshop, starting with creating a data summary for multiple groups, then spreading the values across multiple columns, and finally recombining these multiple columns into a single column.

5.4.4 Data Visualization with `ggplot2`

The final workshop in the series dives into creating data visualizations using the `ggplot2` package (Wickham, 2016). Rather than remembering a list of functions that make different

visualizations, each with it's own unique syntax, arguments, inputs, and outputs, `ggplot2` creates a uniform interface with functions that each solve a particular class of problems. This uniform syntax and set of functions allows participants to create more dynamic visualizations out of the gate. This workshop works with the joined data from the close of the *Data Wrangling* workshop. We begin with participants executing code to generate a scatterplot, using `ggplot()`, which is then used to illuminate the discussion of the `ggplot()` syntax.

Participants learn about the `mapping` argument for specifying aesthetics (`aes`) for the plot and the set of `geom` functions which define the type of plot you produce. By making explicit connections between the addition operator (`+`) and the pipe operator, participants understand addition to be an intuitive metaphor for adding layers to a plot. This discussion directly links to the concepts introduced in the previous workshop, using the pipe operator to motivate why the first argument of `ggplot()` is the data, and the 'long' data format which `ggplot2` requires.

Next, the workshop examines how to modify the `ggplot()` aesthetics and geoms to create violin plots, density plots, bar charts, and line plots. Each of these plots allow for participants to explore different `geom` functions and the aesthetics that pair with each plot. A conversation is had about the importance of plotting raw data rather than simply aggregate measures of the data, by adding a `geom_point()` or `geom_jitter()` layer, further highlighting the layering possibilities in `ggplot()`. Finally, faceting is introduced as an additional tool for creating multivariate visualizations. Participants work with both the `facet_wrap()` and `facet_grid()` functions to create multiple subplots based on categorical variables from the data.

By this point in the workshop, participants have posed many questions on how to modify aspects of a plot that don't depend on the geom. For the final section of the workshop, the group walks through different customizations one can make to each `ggplot` object. Participants learn how to customize a plot's labels, the size of the points, the thickness of lines, the appearance of the plotting window, the color scheme used, and the size, color, and angle of the plot's labels.

Exporting graphics is the final topic of the workshop. At this point, attendees have generated numerous plots, including faceted and arranged plots, which could potentially be used as templates for graphical displays in their research. The workshop closes with a

discussion of the difference between exporting a plot using the "Export" tab and using the `ggsave()` function. As participants may ultimately use the `ggplot` visualizations they create for publications, a discussion of plotting dimension and resolution is included.

5.5 Sustainability of Workshops

To facilitate the sustainability of these workshops, we forged a partnership between our institution's library and the Department of Mathematical Science's Statistical Consulting and Research Services (SCRS). A university's library is an optimal unit for offering these workshops, as it is both department-agnostic and a central hub for the entire university community. Through a partnership with SCRS and a $5,000 faculty excellence grant, a train-the-trainer model was developed and employed to train new graduate students in the fall of 2019 to lead the instruction of these workshops in the spring. Due to the widespread use of `R` across scientific fields, students from a variety of backgrounds hold the potential to be effective instructors, encouraging participation in the workshops throughout their field. Additionally, these workshops require helpers to address issues participants may encounter during the workshop and to circulate during the "hands-on" exercises. Incorporating undergraduate students as workshop helpers provides them with teaching experiences, and allows for the possibility of students becoming lead or co-instructors as they progress through their program.

6 Evaluating Data Science Workshops

The final phase of this research explored the backgrounds and experiences of workshop attendees. For the first iteration of this research, attention was paid to understanding the backgrounds of the workshop attendees, their experiences learning `R`, their motivation for attending the workshop(s), and their experiences in each workshop. Evaluating the learning outcomes of these workshops is left as future research.

6.1 Data Collection

In the week prior to the workshop, a survey is sent out to those registered for the workshop through a public Google Form. The pre-workshop survey details individuals' areas of study, current occupations, statistics and computer science experiences, participation in independent research, and their chosen method of storing any data they've collected. Following each workshop, attendees are asked to complete another survey. This post-workshop survey details the workshop participants' experiences in the workshop environment, their familiarity with the content covered in the workshop, their ability to implement the skills they learned, the best thing about the workshop, and what could use improvement. The content of these surveys was informed from the pre- and post-workshop surveys disseminated for Data and Software Carpentry workshops [5], with revisions to the disciplines and occupations provided, and removal of questions regarding the degree of agreement with statements provided. The full pre- and post-workshop surveys are included as supplementary materials.

6.2 Backgrounds of Workshop Participants

The workshops tailored to environmental science graduate students, during phase two, were offered each semester of the 2018-2019 academic year. A workshop schedule was created prior to the start of each semester, with care taken to not overlap with any required graduate-level environmental science seminars. The schedule of workshops was then distributed to department administrative assistants in these environmental science disciplines, to be advertised to each department's students, faculty, and staff. Additionally, these workshops were advertised through the library and our institution's weekly news and events.

A total of 150 unique students, faculty, and staff attended at least one of the workshops. Many participants elected to return for the subsequent workshops, but nearly 70% only attended the *Introduction to* `R` workshop. A total of 65 individuals attended the *Introduction to* `R` workshop, 40 attended *Intermediate* `R`, and 20 attended each of the *Data Wrangling* and *Data Visualization* workshops.

[5] This work is a derivative of the Carpentries pre- and post-workshop survey materials (https://github.com/carpentries/assessment/), used under CC BY (https://creativecommons.org/licenses/by/4.0/).

The majority of the workshop attendees were in biological fields—from departments such as Land Resources and Environmental Sciences (LRES), Ecology, Plant Sciences, Animal and Range Sciences, Earth Sciences, and Biochemistry or Microbiology. The majority of workshop attendees were masters and doctoral students in these fields. It is worth noting that 12 faculty, staff, and postdocs also attended these workshops. Figure 3 displays the department affiliations of the workshop attendees and their current occupation.

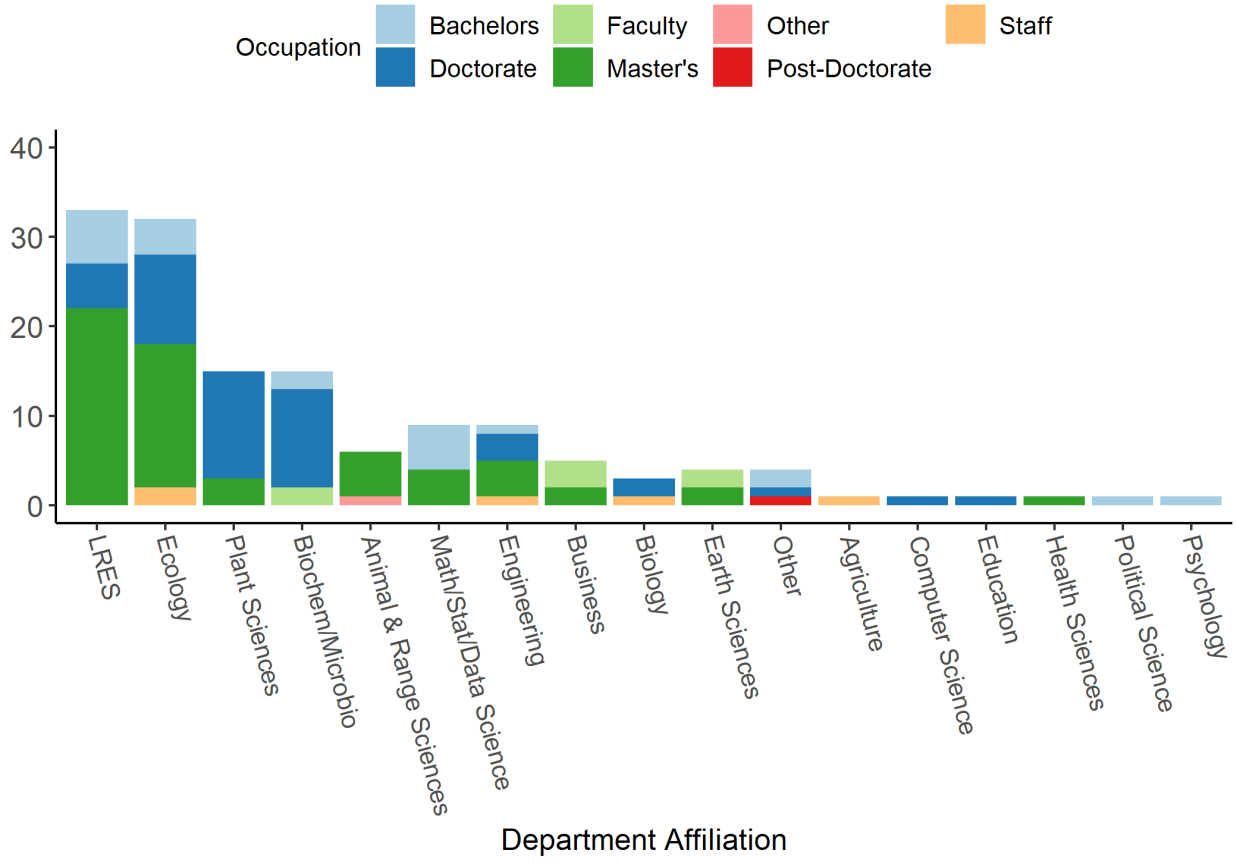

Figure 3: Number of attendees by department and current occupation, selected from an itemized list of campus departments and positions.

Consistent with the environmental science literature (Andelman et al., 2004; Hampton et al., 2017; Hernandez et al., 2012; Teal et al., 2015), a large number of workshop participants were either unfamiliar with the concept of a programming language or had no experience with any programming languages. As seen in Table 2, nearly 75% of the attendees reported having never formally taken a course in computer programming.

Programming Languages	Participants
What is a programming language?	30
None	35
`R`	22
SQL	12
Java or Javascript	11
C or C++	7
Fortran	4

Table 2: Workshop attendees' responses to the question of "What programming languages do you have experience with? Select all that apply."

Many attendees, however, stated that they had taken courses in statistics. The majority of participants reported having some undergraduate of graduate experiences with introductory level statistics courses. Notably, over 15% of attendees reported having no formal statistical training. Most graduate students had enrolled in discipline-specific introductory statistics courses in their own department or a graduate-level applied statistics course offered by the Department of Mathematical Sciences. Table 3 consolidates themes of workshop participants' previous statistical experiences, when asked to report the statistics courses they have taken over the course of their education.

Stat Courses	Participants
Introductory Statistics	46
Applied Statistics	42
None	24
Discipline Specific Introductory Statistics	20
Intermediate Statistics	10
Experimental Design	8
Probability Theory	6
Statistical Computing	3
Sampling	3
Biostatistics	2
Spatial Analysis	2
Econometrics	1
Time Series Analysis	1

Table 3: Workshop attendees' responses to the question "What are your previous statistical experience(s)? List course names.," thematically organized based on content of the course.

6.3 Motivation for Attending

In Table 4 we outline what workshop participants reported as their "most important reason for attending this workshop." As expected from the prevalence of the use of `R` in environmental science research (Lai et al., 2019; Mislan et al., 2016), over half of the master's and doctoral workshop participants attended for assistance with their research. Others were seeking additional assistance for learning the `R` skills necessary for their coursework, refreshing or updating their `R` skills to include new tools they were unfamiliar with (e.g. `ggplot`, `dplyr`), or undergraduates preparing for graduate school.

Reason Attended	Participants
Research assistance	58
Coursework assistance	35
Refresh or update skills	16
Department/Professor recommended	13
Preparation for graduate school	12
Professional Development	7
Adviser recommended	6
Expand Skills	6

Table 4: Workshop attendees' responses to the question "What is your most important reason for attending this workshop? Select all that apply."

As echoed by previous studies of environmental science graduate students (Teal et al., 2015; Theobold and Hancock, 2019), attendees overwhelmingly stated that they primarily use the internet, their peers, or their lab mates when learning `R`. Based on the statistical backgrounds of these participants and the statistics education literature on computing in the statistics classroom, it is not surprising that nearly two-thirds of these individuals reported using resources other than course materials as their main resource for learning `R`. Table 5 details the resources workshop participants selected having used when learning to program in `R`.

The three themes which emerged from attendees' responses to "what are you hoping to learn in this workshop?" were requests specific to the content taught in the workshop, requests for content not applicable to the workshop. and general interest in learning about `R`. The number of participant responses to ‘ regarding the desire to learn more about `R` are indicative of the widespread use of `R` for data-intensive research. Furthermore, regardless

Resources Used to Learn R	Participants
Internet Resources	55
Peers	43
Course Materials	35
Lab Mates	29
Adviser	20
These Workshops	15
Books	3
Professor	1

Table 5: Workshop attendees' responses to the question "What resources have you used while learning to program in R? Select all that apply."

of statistical background, the bulk of these workshop attendees voiced that they were hoping to learn R for some aspect of analyzing their data. On occasion these requests from attendees were out of focus of the broader workshop content, with requests to understand "how R can be used to analyze microbiom data," "geospatial analysis in R," "a refresher on stats," and "how to integrate data into the `Ternary` package I'm learning." Given the backgrounds of workshop attendees, their reasons for attending, and the resources they've used to learn R, these workshop learning requests can be viewed as attendees cry for help. While other attendees' requests focused largely around the desire to learn R more fluently for their research—these specific requests for data analysis help could be attendees' final resort in understanding the analyses involved in their research.

6.4 Reflections of Workshop Participants

Every participant attending the workshops reported that they felt the workshop environment to be welcoming, with many participants voicing that the "enthusiasm of the instructor" was the best part of the workshop. The percentage of individuals reporting that all of the information presented was new to them differed by workshop, with 40% of *Introduction to R* participants, 30% of *Intermediate R* participants, 80% in *Data Wrangling*, and 50% in *Data Visualization*.

The themes which emerged from these attendees' reflections to what they enjoyed most about the workshop were hands-on learning, workshop atmosphere, instructor attributes, and confidence. Many individuals commented on how walking through the code step-by-step

made the information more clear, and how this process left them feeling more "confident figuring things out on my own, now that I understand the general lay out and 'way' commands or functions are set up." Furthermore, these participants voiced that the workshop left them with a more substantial feeling of independence, because "I have a better understanding of how to read code, what certain symbols/terms/etc mean and how they work" and "I feel like I can better interpret [`ggplot`] code to work through it better individually now." Participants expressed that the hands-on exercises used throughout the workshops also contributed to "fostering a much greater level of understanding." This deeper level of understanding was facilitated by providing participants with an adequate amount of time to "explore `R` on our own" and then spending time, as a group, talking through a variety of ways to address the task at hand. Some individuals who reported using the internet as a resource to learn `R` stated that "it's easy to walk away from `R` workshops wondering if anything was learned, however the exercises were a clear tool which allow me to see what I gained."

Across every workshop, nearly all participants stated that they "strongly agreed" that they "learned skills that [they] will be able to use in [their] research/work." Over 75% of the workshop participants reported that they would use the skills they learned in their research immediately or in the next 30 days. Numerous participants expressed that "[in their field] the value of learning `R` cannot be underestimated," because "[`R`] is an essential tool for researchers." While, a number of graduate students reflected on the importance of these workshops "filling a critical hole in the curriculum of many college programs." Other students stated that they attended the workshops because they were taking a class which uses `R`, but had received no formal training in `R` at the start of the course. Therefore, as new `R` users, they felt that "[they] were missing some of the most basic understanding," and the workshop left them "feeling much more prepared to address the content in class now."

Lastly, when attendees were asked what in the workshop needed to most improvement, themes of time and content appeared. The largest volume of constructive feedback from workshop participants focused on the amount of time allocated for the workshop. Many individuals stated that "it could be better to have more time," suggesting an additional hour, with a "break in between." However, others reflected that it could be more beneficial to have "shorter, but more workshops," since "it's easy for the brain to get tired after an hour or so." Due to the large attendance at many *Introduction to `R`* workshops, often times a large

number of questions would arise over the course of the afternoon. This lead to an inability to cover some of the workshop content in as much depth as hoped, which some participants remarked on. However, others felt that "with so many people, [the workshop] had better discussions." There exists a balancing act between the amount of workshop content, the time allocated to addressing questions that may arise during the workshop, and the duration of the workshop. However, the three hours allocated to each of these workshops far exceeds the 60 minutes typically given to this content during a standard Data or Software Carpentry workshop.

7 Limitations, Dissemination, & Future Research

Currently, this design research is focusing on incorporating the content of these workshops into the *Data Analysis and Visualization in* `R` lesson within Data Carpentry's Ecology curriculum. Infusing the skills outlined in this research into the *Data Analysis and Visualization in* `R` *for Ecologists* lesson helps to create a Carpentries curriculum that best reflects the "core data skills" necessary for data-intensive environmental science research. For skills outlined by this research where there is no room in the current *Data Analysis and Visualization in* `R` *for Ecologists* lesson, the Carpentries Incubator and Carpentries Lab provide potential avenues to produce additional lesson materials that are broadly available to The Carpentries community. These avenues allow for the continued discussion of the importance of integrating user-defined functions, conditional statements, and loops into the broader Data Carpentry Ecology curriculum.

The next iteration of this design work will be informed by research concentrating on the computational skills employed by environmental science graduate students in their research code. Collecting the research (`R`) code produced by graduate students in the environmental sciences provides insight into the key computational skills students are using to implement statistics in their research. The skills outlined by this research aid in re-evaluating the content of these workshops, to ensure they cover the skills necessary for graduate-level environmental science research.

The attendance of these workshops by students, faculty, and staff from disciplines outside of the environmental sciences brings to question whether this type of tailored design

work is necessary. While these workshops were originally designed to facilitate environmental science graduate students' acquisition of computing skills, over a third of the workshop attendees came from disciplines outside of this focus. Strikingly, these attendees reported similar workshop experiences to attendees from these targeted disciplines. This brings to question if there are common computational understandings necessary for research in *any* scientific field, which should be infused into *every* statistics and data science course.

Alternatively, in this research we saw a larger number of attendees from environmental science fields persist across workshops, rather than solely attending *Introduction to* `R`. What are the drivers behind these individuals' continued attendance? Future research focusing on the learning outcomes of the workshop attendees could provide fruitful insight on the necessity of these discipline-specific learning opportunities.

8 Conclusion

Ten years ago, Nolan and Temple Lang declared that "modernizing the statistics curricula to include computing [...] is an issue that deserves widespread attention and action" (p. 106). In 2014, the American Statistical Association endorsed a new set of curriculum guidelines for undergraduate programs in statistical science. These new guidelines include an increased emphasis on data science skills and real applications, specifically students' ability to "access and manipulate data in various ways, use a variety of computational approaches to extract meaning from data, program in higher-level languages" (American Statistical Association Undergraduate Guidelines Workgroup, 2014, p. 7). While we may see these changes reflected in undergraduate and graduate programs in statistics, integrating these topics into graduate-level statistics service courses has received less attention and poses different issues.

Statistics courses that serve a variety of students (undergraduate, graduate, statistics major, non-major) reflect a snapshot of the statistics curriculum, but often act as many students' sole statistics course prior to conducting scientific research. Instructors of these courses thus grapple with difficult decisions of how they can ensure their students have both the statistical and "computational understanding, skills, and confidence needed to actively and wholeheartedly participate" in the scientific research arena (Nolan and Temple Lang, 2010, p. 106). For instructors unfamiliar with students' scientific disciplines, it can be

difficult to "be bold and design curricula from scratch" (Nolan and Temple Lang, 2010, p. 106).

The topics suggested by Nolan and Temple Lang (2010), represent a starting point towards building a taxonomy for computing in statistics for undergraduate and graduate statistics programs. These topics, however, may not be relevant to or emphasized by other scientific disciplines whose students enroll in graduate-level statistics service courses. In our research, we found that environmental science faculty stressed the importance of graduate students developing skills surrounding the fundamentals of working with data in `R`, software skills for data processing and preparation, creation of data visualizations, and usage of reproducible work flows. This research equips statistics service course instructors with knowledge of the computing skills necessary for data-intensive environmental science research, and facilitates a meaningful integration of computing topics into these types of courses.

Workshop participants' attribution of the internet, their peers, and lab mates as the resources they've used to learn `R` suggests that statistics courses continue to under prepare students with the computing understandings and skills necessary to "engage in and succeed at statistical inquiry" (Nolan and Temple Lang, 2010, p. 97). Furthermore, the content covered by these workshops has been voiced by workshop participants to be extremely important and impactful, stating that the content covered in these workshops "fills a critical hole in the curriculum of many college programs."

As we work toward a more thorough integration of computing into the statistics curriculum, this research offers a model for facilitating external workshops—promoting university-wide data science literacy. External workshops create opportunities every semester for a university's students, faculty, and staff to acquire computational skills in a hands-on, interactive environment. Moreover, these workshops provide avenues for faculty to acquire computing knowledge and skills "they have not had the opportunity to learn well" (Nolan and Temple Lang, 2010, p. 106), and provide resources and tools for instructors to integrate computing into their classroom.

9 Acknowledgements

We would like to specially thank the participants from this study, without whom this research would not have been possible. We would also like to thank the workshop helpers for their time and assistance, helping to grow the data literacy across our campus.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2019). *rmarkdown: Dynamic Documents for R*. R package version 1.16. Available at https://CRAN.R-project.org/package=rmarkdown.

American Statistical Association Undergraduate Guidelines Workgroup (2014). *2014 curriculum guidelines for undergraduate programs in statistical science*. American Statistical Association, Alexandria, VA.

Andelman, S. J., Bowles, C. M., Willig, M. R., and Waide, R. B. (2004). Understanding environmental complexity through a distributed knowledge network. *BioScience*, 54(3):240–246.

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician*, 69(4):334–342.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., and Horton, N. J. (2014). R markdown: Integrating a reproducible analysis tool into introductory statistics. *Technology Innovations in Statistics Education*, 8(1):1–22.

Baumer, B. S., Horton, N. J., and Wickham, H. (2015). Setting the stage for data science: Integration of data management skills in introductory and second courses in statistics. *CHANCE*, 28(2):40–50.

CASSEY, P. and BLACKBURN, T. M. (2006). Reproducibility and repeatability in ecology. *BioScience*, 56(12):98.

Cetinkaya-Rundel, M. (2018). Intro stats, intro data science: Do we need both? Presented at the 2018 Joint Statistical Meetings.

Cetinkaya-Rundel, M. and Rundel, C. (2018). Infrastructure and tools for teaching computing throughout the statistical curriculum. *The American Statistician*, 72(1):58–65.

Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, 69(4):266–282.

Cobb, P. A., Confrey, J., diSessa, A. A., Lehrer, R., and Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1):9–13.

Data Carpentry (2020). https://datacarpentry.org/.

Dodds, Z., Alvarado, C., Kuenning, G., and Libeskind-Hadas, R. (2007). Breadth-first CS 1 for scientists. In *Proceedings of the 2007 ITiCSE*. ACM.

Dodds, Z., Libeskind-Hadas, R., Alvarado, C., and Kuenning, G. (2008). Evaluating a breadth-first CS 1 for scientists. In *Proceedings of the 2008 SIGCSE*. ACM.

Eglen, S. J. (2009). A quick guide to teaching R programming to computational biology students. *PLOS Computational Biology*, 5(8):1–4.

Ellison, A. M. (2010). Repeatability and transparency in ecological research. *Ecology*, 91(9):2536–2539.

Ernest, M., Brown, J., Valone, T., and White, E. P. (2018). Portal project teaching database.

Fishman, B. J., Penuel, W. R., Allen, A.-R., Cheng, B. H., and Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. In Fishman, B. J. and Penuel, W. R., editors, *Design Based Implementation Research*, volume 112, pages 136–156. National Society for the Study of Education.

Gould, R. (2010). Statistics and the modern student. *International Statistics Reveiw*, 78(2):297–315.

Green, J. L. and Blankenship, E. E. (2015). Fostering conceptual understanding in mathematical statistics. *The American Statistician*, 69(4):315–325.

Green, J. L., Hastings, A., Arzberger, P., Ayala, F. J., Cottingham, K. L., Cuddington, K., Davis, F., Dunne, J. A., Fortin, M.-J., Gerber, L., and Neubert, M. (2005). Complexity in ecology and conservation: Mathematical, statistical, and computational challenges. *BioScience*, 55(6):501–510.

Grimshaw, S. D. (2015). A framework for infusing authentic data experiences within statistics courses. *The American Statistician*, 69(4):307–314.

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., Hernandez, R. R., Boettiger, C., Collins, S. L., Gross, L. J., Fernndez, D. S., Budden, A., White, E. P., Teal, T. K., Labou, S. G., and Aukema, J. E. (2017). Skills and knowledge for data-intensive environmental research. *BioScience*, 67(6):546–557.

Hardin, J. (2018). Dynamic data in the statistics classroom. *Technology Innovations in Statistics Education*, 11(1):1–22.

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

Hastings, A., Arzberger, P., Bolker, B., Collins, S., Ives, Anthony, R., Johnson, N. A., and Palmer, M. A. (2005). Quantitative bioscience for the 21st century. *BioScience*, 55(6):511–517.

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

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386.

Horton, N. J. and Hardin, J. S. (2015). Teaching the next generation of statistics students to "think with data": Special issue on statistics and the undergraduate curriculum. *The American Statistician*, 69(4):259–265.

Johnson, G. (2014). "new truths that only one can see". *The New York Times*.

Kaplan, D. (2018). Teaching stats for data science. *The American Statistician*, 72(1):89–96.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, (59):613–620.

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., and Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, 10(1).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute.

McNamara, A. and Horton, N. (2018). Wrangling categorical data in R. *The American Statistician*, 72(1):97–104.

Michonneau, F., Teal, T., Fournier, A., Seok, B., Obeng, A., Pawlik, A. N., Conrado, A. C., Woo, K., Lijnzaad, P., Hart, T., White, E. P., Marwick, B., Bolker, B., Jordan, K. L., Ashander, J., Dashnow, H., Hertweck, K., Cuesta, S. M., Becker, E. A., Guillou, S., Shiklomanov, A., Klinges, D., Odom, G. J., Jean, M., Mislan, K. A. S., Johnson, K., Jahn, N., Mannheimer, S., Pederson, S., Pletzer, A., Fouilloux, A., Switzer, C., Bahlai, C., Li, D., Kerchner, D., Rodriguez-Sanchez, F., Rajeg, G. P. W., Ye, H., Tavares, H., Leinweber, K., Peck, K., Lepore, M. L., Hancock, S., Sandmann, T., Hodges, T., Tirok, K., Jean, M., Bailey, A., von Hardenberg, A., Theobold, A., Wright, A., Basu, A., Johnson, C., Voter, C., Hulshof, C., Bouquin, D., Quinn, D., Vanichkina, D., Wilson, E., Strauss, E., Bledsoe, E., Gan, E., Fishman, D., Boehm, F., Daskalova, G., Tavares, H., Kaupp, J., Dunic, J., Keane, J., Stachelek, J., Herr, J. R., Millar, J., Lotterhos, K., Cranston, K., Direk, K., Tyln, K., Chatzidimitriou, K., Deer, L., Tarkowski, L., Chiapello, M., Burle, M.-H., Ankenbrand, M., Czapanskiy, M., Moreno, M., Culshaw-Maurer, M., Koontz, M., Weisner, M., Johnston, M., Carchedi, N., Burge, O. R., Harrison, P., Humburg, P., Pauloo, R., Peek, R., Elahi, R., Cortijo, S., Umashankar, S., Goswami, S., Yanco, S., Webster, T., Reiter, T., Pearse, W., and Li, Y. (2019). datacarpentry/r-ecology-lesson: Data carpentry: Data analysis and visualization in R for ecologists, July 2019.

Miles, M. B., Huberman, A. M., and Saldana, J. (2014). *Qualitative Data Analysis, A Methods Sourcebook*. SAGE, Thousand Oaks, CA, 3rd edition.

Mislan, K., Heer, J. M., and White, E. P. (2016). Elevating the status of code in ecology. *Trends in Ecology & Evolution*, 31(1):4–7.

Morrison, C., Wardle, C., and Castley, J. (2016). Repeatability and reproducibility of population viability analysis (pva) and the implications for threatened species management. *Frontiers in Ecology and Evolution*, 4:98.

National Academies of Sciences, Engineering, and Medicine (2018). *Data Science for Undergraduates: Opportunities and Options*. The National Academies Press, Washington, DC.

Nolan, D. and Temple Lang, D. (2010). Computing in the statistics curriculum. *The American Statistician*, 64(2):97–107.

Nolan, D. and Temple Lang, D. (2015). Explorations in statistics research: An approach to expose undergraduates to authentic data analysis. *The American Statistician*, 69(4):292–299.

O'Neill, D. K. (2012). Designs that fly: What the history of aeronautics tells us about the future of design-based research in education. *International Journal of Research and Method in Education*, 35(2):119–140.

Powers, S. M. and Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, 29(1).

Revision Committee, A. (2014). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. American Statistical Association, Alexandria, VA.

Ross, Z., Wickham, H., and Robinson, D. (2017). Declutter your R workflow with tidy tools. Technical report, PeerJ Preprints.

RStudio Team (2015a). *RStudio Cloud*. RStudio, Inc., Boston, MA.

RStudio Team (2015b). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Software Carpentry (2020). https://software-carpentry.org/.

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., and Pawlik, A. (2015). Data carpentry: Workshops to increase data literacy for researchers. *International Journal of Digital Curation*, 10(1):343–353.

The Carpentries (2019). https://carpentries.org/.

The Economist Editorial (2013). *"Trouble at the lab. (Cover story)"*. .

Theobold, A. and Hancock, S. (2019). How environmental science graduate students acquire statistical computing skills. *Statistics Education Research Journal*, 18(2):68–85.

Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T., and VanderStoep, J. (2015). Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. *The American Statistician*, 69(4):362–370.

Wang, X., Rush, C., and Horton, N. J. (2017). Data visualization on day one: Bringing big ideas into intro stats early and often. *Technology Innovations in Statistics Education*, 10(1):1–22.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Wickham, H., Franois, R., Henry, L., and Mller, K. (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6.

Wilson, G. (2006). Software carpentry: Getting scientists to write better code by making them more productive. *Computing in Science & Engineering*, 8(6):66–69.

Wilson, G., Alvarado, C., Campbell, J., Landau, R., and Sedgewich, R. (2008). CS-1 for scientists. In *Technical Symposium with Computer Science Education*, pages 36–37. ACM.

Wing, J. (2006). Computational thinking. *Communications of ACM*, 49(3):33–35.

Word, K. R., Jordan, K., Becker, E., Williams, J., Reynolds, P., Hodge, A., Belkin, M., Marwick, B., and Teal, T. (2017). When do workshops work? a response to the 'null effects' paper from Feldon et al. Technical report, Software Carpentry.