SUPPORTING DATA-INTENSIVE ENVIRONMENTAL SCIENCE RESEARCH:

DATA SCIENCE SKILLS FOR SCIENTIFIC PRACTITIONERS

OF STATISTICS

(PARTIAL DISSERTATION)

by

Allison Shay Theobold

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2020

DEDICATION

I would like to dedicate this dissertation to my partner, Laura Marie Smith. Laughably, inspiration for this research came from an argument about Ornate box turtles, en route to ski. That day and every day after, you've reminded me the power of empathy and have inspired me to be courageous and make a difference. I could not have gotten through this stage in life without you and your support. Even with a global pandemic surrounding us, you've made each and every milestone feel incredibly special. Your thoughtfulness and compassion is an inspiration to me and everyone around you. I love you to the moon and back.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my adviser, Dr. Stacey Hancock, for her unwavering support and belief in me. I cannot thank Stacey enough for the countless hours she spent reviewing my writing, our thoughtful discussions of future directions for my research, and always pushing me to be my best. I also owe gratitude to my committee members: Dr. Jenny Green, Dr. Megan Wickstrom, Dr. Mary Alice Carlson, and Dr. Mark Greenwood. You have each challenged and supported me growing as a researcher and an educator, with seemingly endless encouragement—a role I hope to someday fill with my students.

I want to thank my family for the support and strength they've given me throughout my degree. Reford—you've always pushed me to be my best, to overcome obstacles, and to not be afraid to speak up. Karen—the kindness you give everyone is inspiring, you have shown us how to live the golden rule every day. Kalinda—you are simultaneously a pillar of strength and the most brilliant ray of sunshine, you make me feel like the luckiest sister alive. Marilyn—you have shown me that life has no bounds, that love is unconditional, and that there is no stronger bond than family.

Finally, I would like to thank the strong women that I have had the privilege to learn from. Barbara Milburn was the first educator to show me what it means to teach *every* student. Dr. Tracii Friedman taught me the value of struggling in mathematics, and gave me the the strength to not walk away. In every class, I strive to have the tenacity with which these women approach teaching.

VITA

Allison Shay Theobold was born in Grand Junction, Colorado to Karen and Reford Theobold. After graduating from Grand Junction High School in 2007, Allison attended Colorado Mesa University. In 2014 Allison graduated cum laude with a bachelors in Mathematics, with a concentration in Statistics, and a bachelors in Business Administration, with a concentration in Economics.

In 2014 Allison moved to Bozeman, Montana to pursue a doctoral degree in Statistics from Montana State University. During this time, Allison served as the instructor of record for seven sections of Introductory Statistics and Intermediate Statistics courses, teaching over 225 undergraduate students. Allison also served as a statistical consultant to Montana State University students, faculty, and staff at the Department of Mathematical Science's Statistical Consulting and Research Services (SCRS). As part of her dissertation, Allison created and taught a series of data science literacy workshops. Through partnerships with the Montana State Library and SCRS, these workshops will continue to serve Montana State's broader research community for years to come.

Allison's dedication and success in teaching were acknowledged by the college and department by the receipt of the College of Letters and Sciences Outstanding Graduate Teaching Assistant Award and the Department of Mathematical Sciences Outstanding Graduate Teaching Assistant Award. The importance and impact of Allison's research has been recognized by her receipt of the Kopriva Graduate Student Fellowship, Honorable Mention Speed Session Award from the Statistics and Data Science Education Section of the ASA, and the Department of Mathematical Science's Gary Sackett Research Fellowship.

Upon receiving her degree, Allison moves forward as a statistics educator, excited to create statistics and data science classrooms that promote diversity and inclusion through foundational understandings of how data impact our everyday lives. Allison will begin her next chapter in the fall of 2020 as an Assistant Professor of Statistics at California Polytechnic University in San Luis Obispo, California.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

The importance of data science skills for modern environmental science research cannot be understated, but graduate students in these fields typically lack these integral skills. Yet, over the last 20 years statistics preparation in these fields has grown to be considered vital, and statistics coursework has been readily incorporated into graduate programs. As “data science” is the study of extracting value from data, the field shares a great deal of conceptual overlap with the field of Statistics. Thus, many environmental science degree programs expect students to acquire these data science skills in an applied statistics course. A gap exists, however, between the data science skills required for students’ participation in the entire data analysis cycle as applied to independent research, and those taught in statistics service courses. Over the last ten years, environmental science and statistics educators have outlined the shape of the data science skills specific to research in their respective disciplines. Disappointingly, however, both sides of these conversations have ignored the area at the intersection of these fields, specifically the data science skills necessary for environmental science *practitioners* of statistics.

This research focuses on describing the nature of environmental science graduate students’ need for data science skills when engaging in the data analysis cycle, through the voice of the students. In this work, we present three qualitative studies, each investigating a different aspect of this need. First, we present a study describing environmental science students’ experiences acquiring the computing skills necessary to implement statistics in their research. In-depth interviews revealed three themes in these students’ paths toward computational knowledge acquisition: use of peer support, seeking out a “singular consultant,” and learning through independent research. Motivated by the need for extracurricular opportunities for acquiring data science skills, next we describe research investigating the design and implementation of a suite of data science workshops for environmental science graduate students. These workshops fill a critical hole in the environmental science and statistics curricula, providing students with the skills necessary to retrieve, view, wrangle, visualize, and analyze their data. Finally, we conclude with research that works toward identifying key data science skills necessary for environmental science graduate students as they engage in the data analysis cycle.

DISCLAIMER

This document contains one of three chapters of the completed dissertation. While the later two chapters undergo blinded peer review, they have not been included in this initial document.

INTRODUCTION

Over the last two decades, nearly every scientific field has seen a rapid increase in the volume and variety of available data and a growth in the usage and power of computational tools to model phenomena. This increased focus on data-intensive research has made computationally heavy applications of data science techniques—such as management and coalition of large data sets, high dimensional data visualization, and Bayesian modeling—essential understandings for scientific research. These dramatic changes to scientific practices have created a crucial need to reevaluate how our educational system can better prepare current and future generations of researchers (Green et al., 2005; Hampton et al., 2017). Unfortunately, the gap between the computing included in the education students receive and the computational knowledge required for scientific research has become more pronounced, especially in the environmental and life sciences. When considering the issue of curriculum reevaluation, we note that over the last 20 years, statistics preparation in these fields has become vital.

The History of Computing in Statistics and the Environmental Sciences

The Emergence of Computing in the Statistics Curriculum

In 1962, John Tukey charged the field of Statistics to "seek out novelty in data analysis," reflecting that "in the future [Statistics] can and should contribute much more" to data analysis (Tukey, 1962, p. 3). This charge for innovation in data analysis was echoed in Breiman's organization of the "Conference on the Analysis of Large Complex Data Sets" in 1977 and the symposium on "Modern Interdisciplinary University Statistics Education" by the Committee on Applied and Theoretical

Statistics' (CATS) in 1992. In its August 1992 meeting in Boston, CATS "noted widespread sentiment in the statistical community that upper-level undergraduate and graduate curricula for statistics majors are currently structured in ways that do not provide sufficient exposure to modern statistical analysis and computational and graphical tools." This growth the field of Statistics had experienced "is not reflected in the education that future statisticians receive," and left the need for a more meaningful integration of the "computational and graphical tools that are today so important to many professional statisticians" (National Research Council, 1994, p. vii).

At the close of the century, this call for the need to transform the undergraduate statistics curriculum was reiterated by statistics educators, mapping opportunities for innovation to bring statistics education up to speed with the modern day practice of Statistics. Moore et al. (1995) speculated that "technological advances may at last bring widespread change to college teaching" (p. 250), imagining plausible futures for statistics teaching at the university level. Biehler (1997) added to these musings, elaborating on the need for educators to "critically evaluate existing software and to produce future software more adequate both for learning and doing statistics in introductory courses" (p. 167). Possibilities for modernizing the outdated and overly mathematical undergraduate statistics programs were voiced by Higgins (1999), with Nolan and Speed (1999, 2000) providing recommendations for infusing computing explorations into "traditional" mathematical statistics course(s).

The next century brought the introduction of "data science" (Cleveland, 2001), and continued conversations surrounding how to infuse computing into the statistics curriculum. However, the majority of these conversations revolved around including computing into mathematical statistics course(s) (Reid et al., 2003; Horton et al., 2004). Despite these small changes, some statisticians remained concerned with the trajectory of the field of Statistics. Friedman, reflecting on the absence of Statistics from the development and implementation of data mining methodology, lamented that “the field [of Statistics] should be defined in terms of a set of *problems* rather than a set of tools, namely those that pertain to data” (Friedman, 2001, p. 8, emphasis in original). Furthermore, the field needed to “make peace with computing,” because it had been “one of the most glaring omissions in the set of tools that have so far defined Statistics” (Friedman, 2001, p. 8). That same year, Breiman, the creator of classification and regression trees, published a groundbreaking piece on the two cultures of statistical modeling, data modeling and algorithmic modeling. These cultures, Breiman argued, use two fundamentally different methods to model the relationship between two sets of data: the inputs to a process or mechanism and the outputs from that process. Data modeling assumes that the relationship between these two datasets can be explained by a mathematical model, but relies entirely upon the correct model. In contrast, algorithmic modeling determines a data model solely in terms of correctly predicting an output, provided an input, allowing for the model to potentially have little to no relationship with the underlying data-generating process. Breiman asserted that Statistics’ commitment to data modeling had prevented the field from entering new arenas where “the data being gathered is not suitable for analysis by data models” (p. 200). Hence, Breiman encouraged statisticians to become more familiar with algorithmic modeling, to address this significant change in the data landscape.

Although these calls for an increased focus on computational tools continued to be heard throughout the statistics community, it wasn’t until nearly ten years later that Brown and Kass resumed the discussion around the statistical training of undergraduate and graduate Statistics majors. This work came at a critical time, following Peck and Chance’s detailed description of the assessment of Cal Poly’s undergraduate Statistics program (2005). Brown and Kass argue that "to remain vibrant, the field [of Statistics] must open up by taking a less restrictive view of what constitutes statistical training" (2009, p. 105). The authors acknowledged that "a fear lurks in the heart of many statistics professors" where "statistics as we know it [may] become obsolete" if the field continued to complacently ignore the innovation in data analysis techniques (p. 105). They found that, while programs emphasize the mathematical logic of data analysis, when faced with an actual data analysis, "graduate students in statistics often are reticent to the point of inaction" (p. 106).

At the climax of these discussions, came the publication of "Computing in the Statistics Curriculum" from Deb Nolan and Duncan Temple Lang (2010). In this influential article, Nolan and Temple Lang painted a broad picture of the computing skills that successful statisticians must be facile with, and how these skills had been infused into the Statistics program at the University of California, Berkeley. The authors asserted that, "the skill set needed by a statistician even 20 years ago is very different from what is needed today (p. 98). Moreover, as a statistics education community, we were not preparing students with the computational proficiency, the statistical problem solving, or the "confidence needed to overcome computational challenges" (p. 97). Nolan and Temple Lang reflected that they have "found that Bachelors and Masters students who enter the workforce spend much of their efforts retrieving, filtering, and cleaning data and doing initial exploratory data analysis," (p. 99), but students were not taught these skills in their courses. Instead, students were "told to learn how to program by themselves, from each other, or from their teaching assistant in a two-week 'crash course' in basic syntax at the start of a course" (p. 100). They outlined a series of recommendations for changes the statistics education community should make to bring the statistics curriculum up to date with the tools that modern statisticians use, so that students would leave the statistics curriculum with the “computational understanding, skills, and confidence needed to actively and wholeheartedly participate in the computational arena” (p. 106).

The Emergence of Computing in the Environmental Science Curriculum

Meanwhile, in the 1990s and 2000s, the environmental science community was having similar conversations surrounding the importance of computing for research in biological fields. In 1977, Levin et al. pioneered these conversations by asserting that, with the addition of more powerful computers and new analysis techniques, the “face of the science of computational population biology and ecosystems science will change in the next decade” (p. 341). This conversation went unanswered until in 2001, George Johnson of the New York times published a piece describing the use of computing for research in bioinformatics, concluding that no matter what scientific field one chooses to perform research in, “all science is computer science” (Johnson, 2001).

Finally, by the early 2000s, the conversation around teaching computing in environmental science courses began to flourish. Andelman and colleagues (2004) published a recount of an interdisciplinary seminar they developed for graduate students and what they learned about students’ computational skills—or lack thereof. During their course, the authors learned that students were unprepared with both the statistical and computational skills necessary for data analysis; rather, “ninety-three percent of students did not have skills in the scripted programming languages (e.g., SAS or MATLAB) that are needed for the integration of large data sets” (p. 244). As a consequence, “the greatest limitation [. . .] that the students faced was related to data concatenation, manipulation, and analysis” (p. 245). Environmental science educators continued to press on the issue of integrating computing into environmental science research (Green et al., 2005; Hastings et al., 2005), detailing the formidable computational challenges scientists were facing. To stress the importance of these needs, the NSF sponsored a series of three workshops on "quantitative environmental and integrative biology," to aid in identifying "areas of cutting edge research in ecology and environmental biology that require integration with novel computational, statistical, and informatics tools" (Green et al., 2005, p. 502).

The importance of every scientific researcher having the ability to reason through computational problems, was further emphasized in 2006 by Jannette Wing in an Association for Computing Machinery (ACM) communication. Wing pioneered the concept of computational thinking in her ACM communication, declaring computational thinking to be a "fundamental skill for everyone" (p. 33). Wing then outlined the variety of mental tools that encompass computational thinking, including but not limited to: thinking recursively, parallel processing, abstraction, and decomposition. A series of articles from the ITiCSE and SIGCSE conferences followed, with computer science educators describing their institution's development of introductory computer science courses for non-computer science majors. These courses were designed for "any student intending to major in science or engineering" (Dodds et al., 2007, p. 23) and were intended to develop students' problem solving and programming skills, paint a compelling picture of the vastness of computer science, and attract students to continue to study computer science (Dodds et al., 2007, 2008; Hambrusch et al., 2009; Wilson et al., 2008).

During these conversations, Greg Wilson and Dr. Brent Gorda at the University of Toronto developed a course named Software Carpentry, to teach "scientists and engineers the 'common core' of modern software development" (Wilson, 2006, p. 66). Software Carpentry addressed a gaping hole in graduate education in the sciences, providing students with the tools to increase their productivity by improving the quality of their code. This course included topics such as version control, scripting, debugging, testing, and continuous integration, all topics which few, if any, students had seen before.

The conversation around scientists' need to be familiar with scripted programming and reproducible documents continued with Stephen Eglen's tutorial piece on how to teach `R` programming to computational biology students (Eglen, 2009). (It is worth noting that Eglen's piece is potentially the first occurrence in which the use of `R` is advocated for research in the biological sciences.) Rounding out the decade, Kelling and colleagues revisited the original argument of "data-intensive" science in 2003, reiterating the importance of computing to biological research, and outlining the big picture "steps in the data-intensive science workflow" (Kelling et al., 2009, p. 614). These researchers paid special attention to the integration of statistics in this data-intensive workflow, with exploratory analysis and confirmatory analysis encompassing the final two stages. The authors concluded with a charge for environmental scientists to "overcome the challenges in organizing and analyzing massive and heterogeneous data" so the field could make headway towards unraveling the complexity of ecological systems" (p. 619).

Data Science in the Environmental Sciences

With the importance of computing to environmental science fields firmly in place, the literature over the next decade focused instead on the important role academic institutions have in preparing undergraduate and graduate students with the skills relevant to research in these fields. Strasser and Hampton (2012) began this conversation, focusing on the importance of data management in undergraduate ecology courses. The authors reported that while ecology instructors rated data management topics, such as workflows, databases, and reproducibility, as very important for their research, less than 20% of instructors included these topics in their courses. These results suggested that—across institutions—"data management education is not currently a priority for ecology instructors" (p. 10).

That same year, Hernandez, an environmental science graduate student, led a large scale study of the "technological and computational experiences of environmental scientists in the formative stages of their career" (Hernandez et al., 2012, p. 1068). These researchers found that over 74% of the students surveyed stated they had no skills in any programming language—including `R`—and only 17% reported basic skill levels in any programming language. These findings suggested that—across institutions—graduate students were not obtaining the knowledge and skills required to navigate the advancing fields of technology, computation, and data management through their coursework or instruction.

Given the poor computational preparation of environmental science graduate students by their curriculum, Hernandez et al. suggested that student-focused workshops could "provide intensive environments" where students could learn "particular methods or technologies" (p. 1075). Furthermore, developing and offering these workshops would be simpler than developing new courses to organize and implement. Over the subsequent years, researchers would reiterate the ability of these external workshops to provide students with on-demand, intensive training to acquire the computing skills necessary for data-intensive environmental science research.

Gutlerner and Van Vactor (2013) led the charge in the development of short-format, skill-building courses, in an article describing tools for evolving the scientific curriculum. These short-courses allowed for students to "take a course on a particular topic or technique at the time when they are most motivated to learn about it" (p. 732). Alternatively, these intensive experiences could be harnessed into a "bootcamp" prior to students' first semester of graduate school. One such "quantitative methods bootcamp" was implemented by instructors at Harvard Medical School, to "enable students to use computational tools to visualize and analyze data" and to "strengthen their computational thinking skills" (Stefan et al., 2015, p. 1). The authors argued introducing graduate students to these concepts before they begin their coursework lowered the computational barrier for students before taking courses, empowered students to learn computational tools on their own, and enabled courses to "build upon this foundation and integrate quantitative methods throughout the curriculum" (p. 2).

Around the same time, Greg Wilson created the Software Carpentry Foundation, transforming the Software Carpentry course into a workshop curriculum, which could be offered to researchers around the world. Software Carpentry found larger success than its predecessors, due in part to the dramatic change in the scientific landscape. Backed by the support of the Mozilla and Sloan Foundations, in 2013, Software Carpentry offered its first Software Carpentry workshops for Librarians in the United States and Canada. Then, in 2014, Data Carpentry was founded to "train researchers in the core data skills for efficient, shareable, and reproducible research practices" (2020), specific to their field of research. These workshops met the need for "good training resources for researchers looking to develop skills that will enable them to be more effective and productive" (Teal et al., 2015, p. 135). Teal and collaborators pressed further into the findings of Hernandez et al. (2012) and Wilson (2016), claiming that "most or all of what [researchers] know about data management, analysis, and sharing has been learned piecemeal, or not learned at all," as "training in data and computing skills is still largely absent from undergraduate and graduate programs" (p. 136). The authors emphasized the importance of developing streamlined training opportunities for researchers in these fields for two main reasons: (1) there is substantial variation in computational training at every institution, and (2) for students not being directly taught computing skills, it is difficult to wade through the plethora of online lessons, MOOCs, and books to find relevant resources. The authors acknowledged 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).

Culminating all of these conversations, in 2017, educators from a variety of environmental science research areas gathered together to write a formative piece on the skills and knowledge necessary for “data-intensive” environmental science research (Hampton et al., 2017). Hampton and colleagues outlined the current state of environmental science education in American universities, stating that “a symptom of the current curriculum’s shortcomings is the recent emergence of a variety of extramural options for acquiring critical technological skills” (p. 547). They then describe the “skillset required by environmental scientists to succeed in the kind of data-intensive scientific collaboration that is increasingly valued” (p. 548). These five classes of skills included: (1) data management and processing, (2) analysis, (3) software skills for science, (4) visualization, and (5) communication methods for collaboration and dissemination. Each of these classes of skills reiterates previous research on the “good enough” (Wilson et al., 2017) computational skills necessary for research these fields.

Over the last 20 years, however, attention has yet to be paid to the substantial role students’ statistics education potentially plays in their attainment of the data science skills necessary for data-intensive scientific research. Andelman and colleagues reflected that, in addition to students’ lack of familiarity with scripted programming languages, students were also “unfamiliar with multivariate statistics and with the range of models for regression and analysis of variance” (p. 244). Almost 10 years later, Strasser and Hampton reported that the most common courses ecology faculty voiced as possibly covering "data-related topics" were "ecology laboratories, advanced ecology courses, or statistics courses" (p. 7-8). The survey administered by Hernandez et al. asked students whether they had taken or planned to take courses related "to the management and analysis of large or complex data" (p. 1070), including courses in spatial or time series analysis. The authors also surveyed students regarding their level of proficiency with programming languages, including `R`. Consistent with the previous discussions, Hampton et al. continued to outline the extensive statistical skills a data-intensive environmental science researcher should possess, but do not admit that today, the majority of students in these fields complete statistics coursework prior to graduation. This discontinuity in these conversations comes as a shock, as, over this time period, "statistical preparation in the environmental sciences has grown to be considered vital" (Hampton et al., 2017, p. 547).

Data Science in Statistics

In statistics education, during the 2010s the research focused on integrating computing throughout the statistics curriculum, revising the program expectations for undergraduate statistics programs, and creating user-friendly tools for streamlined data science workflows. 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). Simultaneously, during the 2011 United States Conference on Teaching Statistics (USCOTS), statistics educators began conversations around how computing could play a larger role in the introductory statistics course through the incorporation of simulation-based methods. Following these conversations, two simulation-based introductory statistics textbooks emerged, both carrying with them a suite of applets for student use (Lock et al., 2013; Rossman and Chance, 2011).

Amidst these conversations, a suite of packages was being created, which would fundamentally change how users interact with `R`. The `ggplot2` `R` package, created by Hadley Wickham in 2005, spearheaded the change toward creating user-friendly `R` tools all "sharing an underlying design philosophy, grammar, and data structures" (Wickham, 2017). The `ggplot2` package was created to produce statistical, or data, graphics; but, unlike most other graphics packages, it had the "deep underlying grammar" (Wickham, 2016, p. 1) of Wilkinson's Grammar of Graphics (2005). Over the next decade, Wickham and his team produced the suite of packages now included in the `tidyverse` package, namely `stringr` (2009), `dplyr` (2014), `RMarkdown` (2014), `tidyr` (2014), `readr` (2015), `purrr` (2015), `tibble` (2016), and `forcats` (2016). The `tidyverse` package houses all of the necessary packages to import, tidy, transform, wrangle, visualize, and model data, and to communicate the results.

With calls for transforming undergraduate statistics education resounding nationally, the 2014 American Statistical Association (ASA) President, Nathaniel Schenker, convened a workgroup to update the association's guidelines for undergraduate programs. The group, with broad representation from academia, industry, and government, put forward guidelines that were endorsed by the ASA Board of Directors in November 2014 (American Statistical Association Undergraduate Guidelines Workgroup, 2014). 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, [and] program in higher-level languages" (p. 7).

Although these changes reflected a growing consensus that computing should be featured throughout statistics programs, much of the statistics education literature up to that point had focused on the introductory statistics and mathematical statistics courses. Hence, 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. The articles in the special issue fell primarily into two themes: the first theme described how computing should be included throughout the statistics curriculum, with articles from Green and Blankenship (2015), Tintle and colleagues (2015), and Hesterberg (2015); the second theme in these articles presented thoughts on how data science topics should be integrated into undergraduate statistics courses, with articles from Nolan and Temple Lang (2015), Grimshaw (2015), Baumer (2015), and 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). In his article, Cobb argued that statistics, like computer science, should be teaching algorithmic thinking at a basic level. But, computing should be mindfully included throughout the statistics curriculum, rather than simply inserting "a new computing course into the existing curriculum" (p. 275).

Despite these technological advances promoting a facile integration of data science in the statistics curriculum and calls for purposeful inclusion of computing in the statistics curriculum, we continue to see students from scientific disciplines leave the statistics classroom without data science skills in hand. A mere 60% of environmental science graduate students reported a basic skill level in `R` (Hernandez et al., 2012, p. 1069), which has become the "primary tool" reported for data analysis in environmental science research (Lai et al., 2019, p. 1). This gap between the importance of data science reiterated by statistics educators and the data science skills environmental science graduate students report leaving their program with demonstrates that data science concepts continue to be absent from many statistics courses. To promote conversations such as this, the *Journal of Statistics Education* will publish a special issue on “Computing in the Statistics Curriculum” in July 2020. To celebrate the 10-year anniversary of Nolan and Temple Lang’s pioneering piece, articles in the special issue will look into what has changed since the publication of “Computing in the Statistics Curriculum,” what still needs to change, and what is needed to implement curricular shifts.

Barriers to Incorporating Data Science in the Curriculum

While calls for incorporating computing throughout the environmental science and statistics curricula have resonated for the last ten years (Jones et al., 2006; Joppa et al., 2013; Laney et al., 2015; Manyika et al., 2011; Mokany et al., 2016; Peters and Okin, 2017; Smith, 2015; Teal et al., 2015), we continue to see researchers reporting the computational ill-preparation of environmental science undergraduate and graduate students by their curriculum (Hampton et al., 2017; Teal et al., 2015). This raises the question, why are these skills still so rare when the need for them is now widely recognized?

Nearly ten years ago, over 70% of ecology instructors reported substantial barriers to incorporating data management topics in their course(s). These barriers include: the instructor’s lack of time or their lack of knowledge of the topics, students’ lack of the necessary quantitative understandings, or a lack of alignment of the data management topics with the content of their course. These obstacles can be distilled into two main components: first we are “attempting to fit more material into already-full courses and curriculum," and second, these courses are potentially "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). Yet, this lack of computational training impedes the progression of scientific research and results in substantial hidden costs.

Instead of acquiring these necessary skills in the coursework required for their programs, these environmental science graduate students "learn much of what they know about programming and data management on their own or the information is passed down within a lab" (Teal et al., 2015, p. 136). Despite the inclusion of statistics courses in these students' programs of study, students continue to be "told to learn how to program by themselves, from each other, or from their teaching assistant in a two-week 'crash course' in basic syntax " at the beginning of their statistics course (Nolan and Temple Lang, 2010, p. 100). This teach-yourself approach sends the signal to students that "computing is not of intellectual importance relative to the material covered in lectures" (Nolan and Temple Lang, 2010, p. 100). Moreover, this structure results in students potentially "picking up bad habits, misunderstandings, and, more importantly, the wrong concepts" (Nolan and Temple Lang, 2010, p. 100). Students' initial knowledge shapes the methods they use to accomplish a task, making some tasks impossible. They may spend weeks or months doing things that could be done in hours or days, unable to abstract what they learned to broader classes of tasks. Furthermore, students may be unaware of the reliability and reproducibility of their results.

Specific Aims for This Research

Clearly, the current situation is unsatisfactory; however, few efforts have been made to better understand the data science skills necessary for environmental science graduate students as they implement statistics in their research. The findings of Hernandez et al. suggest that—by in large—graduate students are not acquiring the data science skills required for participation in data-intensive research in their curriculum. Yet, elements of these skills are necessary for each student as they engage in their research, which surfaces the question: how are environmental science graduate students acquiring the data science skills necessary to implement statistics in the context of their research? Investigating students' experiences navigating the phenomenon of acquiring the data science skills necessary for their research adds a new perspective to the conversation surrounding the acquisition of data science skills for scientific research, and brings to light the pathways through which students successfully acquire these necessary skills.

Multiple environmental science researchers reference extracurricular workshops as a potential solution for researchers acquiring data science skills 'just in time' for their research (Teal et al., 2015; Hampton et al., 2017). Namely, Data Carpentry workshops provide researchers with "high-quality, domain-specific training covering the full lifecycle of data-driven research" (2020). Although Data Carpentry workshops are developed by the community to be tailored to specific areas of research, such as Ecology, there has been no formal investigation on the relevance of the skills taught in these workshops to environmental science graduate students, a population of researchers in critical need for relevant, high quality, and accessible computing instruction. Understanding the data science skills relevant to this population of researchers allows for the tailoring of current workshop resources, by making evidence-based, iterative improvements to the content and structure of the workshops.

These profound changes in the data landscape have also impacted the instructors of graduate courses which are intended to arm environmental science students with the data science skills necessary for their independent research.

Instructors may be experiencing similar barriers as those faced ten years prior (Strasser and Hampton, 2012). Instructors of these courses may not "have not been taught computing formally," so they "have not had the opportunity to learn it well, and feel they cannot teach it effectively" (Nolan and Temple Lang, 2010, p. 106). Educators from both Statistics and the environmental sciences have outlined data science skills of potential relevance to researchers in their respective field, but each of these conversations neglect a critical aspect of data-intensive environmental science research, the data analysis cycle.

While we may see data science concepts integrated into the undergraduate programs in statistics, integrating these topics into graduate-level statistics service courses, often required for environmental science graduate students, has received less attention and poses different issues. These statistics courses that serve a variety of students reflect a snapshot of the statistics curriculum, but often act as students' sole statistics course prior to conducting the research required for their degree. Thus, instructors of these courses are forced to navigate difficult decisions of how they can ensure their students leave the classroom with 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). Regrettably, for instructors unfamiliar with students' scientific disciplines, it can be difficult to "be bold" and infuse data science skills relevant to students' field of research into the classroom (Nolan and Temple Lang, 2010, p. 106).

Each of these issues facing environmental science students and faculty necessitates a better understanding of the specific data science skills relevant to environmental science graduate students as they engage in the data analysis cycle. Understanding the data science skills relevant to this population of researchers allows for the evaluation of the content included in tailored extracurricular workshops and provides statistics and environmental science educators with a set of foundational data science concepts to be included throughout the environmental science graduate curriculum.

With these considerations in mind, the goals of this research are threefold: (1) to outline the experiences of graduate students in the environmental sciences when acquiring the data science skills necessary to apply statistics in the context of their research, (2) to design, implement, and evaluate a suite of data science workshops tailored for graduate-level environmental science researchers, and (3) to describe the data science skills environmental science graduate students employ throughout their research when engaging in the data analysis cycle, and how these skills evolve over time.

For this research, the collection of fields who perform research in the biological and environmental science fields are captured under the term "environmental science." At Montana State University, these are the fields whose students are required or highly-recommended to enroll in the graduate-level Applied Statistics course sequence. In this research, I use the following terms interchangeably: "computing skills necessary to implement statistics," "statistical computing," and "data science skills." Each of these terms are considered to consist of the computing knowledge and skills necessary for the entire data analysis cycle, from data cleaning to data visualization to data analysis to communication. The computing skills necessary throughout the data analysis cycle may include general programming concepts such as loops, user-defined functions, or conditional statements, but the focus of data science skills differ fundamentally from general programming skills. Rather than focusing on computer architecture, design, and application, for data science skills, data are the focus.

Research Journey

I was drawn to research in data science education through my experiences as a second-year graduate student in Statistics. During my second year, I provided statistical consulting for a graduate student in Ecology. This graduate student sought out consulting for assistance to implement a Bayesian framework to Ornate Box Turtle mark-recapture data, having no previous experiences working with statistical software. A component of our consulting collaboration took the form of weekly `R` workshops covering a variety of skills, from importing data to writing for-loops and functions, to fitting models in the `R` package `rjags`. At the close of the semester, I appreciated the computational challenges environmental science researchers face in their attempts to implement statistics in their research, and a realization of the data science skills with which graduate students typically leave their statistics courses. This emboldened me to investigate how environmental science graduate students acquire the data science skills necessary for research in their fields.

Experiences of Environmental Science Graduate Students

With this motivation in hand, I set out to design a study to understand and describe graduate students' transferability of the data science skills learned in the statistics classroom to environmental science applications. The design of this pilot study was comprised of two parts: (1) students' completion of hands-on computational problems, and (2) a survey of students' attitudes and experiences learning and using computing skills.

The computing tasks were included to assess students' abilities to reason through applications of data science skills in an environmental science context. Then, after reasoning through each task, students were asked to detail where and how they had acquired the computational skill(s) they had employed while completing the task. During my initial data analysis, I realized that if a student was unable to reason through a particular task, that did not necessarily capture their ability to reason through that type of data science application in their field. Indeed, it is possible that the student was unable to reason through that type of data science task or, alternatively, the task in question may have been irrelevant to the "typical" data science applications in the student's respective field of research. Therefore, these data science problems were removed from the focus of this study.

Although the statistical computing tasks may not have accurately captured students' data science understandings, the interviews accompanying these tasks shed light on how students acquired the data science skills they were familiar with. In an article which appeared in the *Statistics Education Research Journal* (Theobold and Hancock, 2019), Chapter 2 outlines the results of these in-depth interviews, how these graduate students experienced the phenomenon of acquiring the computational skills necessary to implement statistics in their research. Three themes emerged in students' paths towards computational knowledge acquisition: use of peer support, seeking out a singular "consultant," and learning through independent research experiences. These themes provide descriptions of graduate student experiences absent from the environmental science literature, informing how instruction can be improved, both in and out of the formal classroom.

The findings of this phenomenological study led me to wonder how students' acquisition of the data science skills necessary for their research could be facilitated with extracurricular workshops tailored to research in their specific field. Current ecology focused extracurricular workshops, such as Data Carpentry, aim to provide researchers with the fundamental data skills needed to conduct research in that field. However, the skills included in these workshops may not reflect the key data science skills necessary for the population of environmental science graduate student researchers. Therefore, this research demanded an understanding of the key data science skills necessary for environmental science graduate students to implement statistics in their research. These questions led to the two follow-up studies detailed in Chapters 3 and 4.

Designing Data Science Workshops for Data-Intensive Environmental Science Research

The first follow-up study focused on (1) describing the computing skills environmental science faculty believe are necessary when implementing statistics in graduate-level environmental science research, (2) investigating how these data skills can be infused into currently existing extracurricular data science workshops, and (3) understanding the backgrounds and experiences of attendees of these workshops.

For these investigations, we executed a three-phase design-based implementation research model (Fishman et al., 2013). Phase one encompassed conducting in-depth interviews with faculty members from environmental science fields regarding the computational skills they believed are necessary for graduate students to engage in the data analysis cycle in their research. Phase two then focused on adapting currently existing workshop resources to design a series of data science workshops targeting the key computational skills distilled from these faculty interviews. Phase three consisted of implementing the workshops and collecting survey responses from the workshop attendees regarding their backgrounds prior to the workshop and their experiences participating in each workshop.

For phase one, all university faculty currently overseeing a graduate student from the departments of Ecology, Land Resources and Environmental Sciences, Animal and Range Sciences, and Plant Sciences and Plant Pathology at Montana State University were emailed requesting their participation in this research. Faculty members from these fields were included because of the large degree of overlap in the type of data collected and analyzed in these fields. Therefore, graduate students from these fields would presumably have similar computational skills required of them as they analyze their data. A total of 61 faculty members were invited to participate in the study, and 23 total faculty agreed to participate in an interview.

During these interviews, faculty were asked a series of questions detailing the computational skills they believe are necessary for graduate students in their field to implement statistics in their research. Over the course of transcribing these interviews, it became clear to me that many faculty focused on the statistical skills and understandings necessary for graduate students to succeed in their research, rather than the computing skills necessary to employ these statistical techniques. Upon this discovery, a second round of faculty interviews were conducted. During these interviews, I asked follow-up questions to further explore why each faculty member believed the computational skill(s) in question are necessary for research in their field. If faculty's responses consisted of the statistical understandings necessary for graduate student researchers, I redirected the conversation to understand what computing skills may be required of a student to implement this type of statistical analysis with their data.

Chapter 3 reports on the data science skills outlined in phase one of this research and how they were used to tailor the existing Data Carpentry Ecology curriculum (Michonneau et al., 2019) to design workshops that suit the needs of this population of graduate student researchers. The chapter then reports on the implementation of these workshops during the 2018-2019 academic year, describing the backgrounds and experiences of the workshop attendees. To close, the chapter outlines the next iteration of this design work, reevaluating the content of these workshops using research code produced by environmental science graduate students.

Computing Skills Employed by Environmental Science Graduate Students

With the phenomenon of acquiring the computational skills necessary for graduate-level research in the environmental sciences firmly in place (Andelman et al., 2004; Green et al., 2005; Hampton et al., 2017; Hernandez et al., 2012; Mislan et al., 2016; Teal et al., 2015; Theobold and Hancock, 2019), and an understanding of the skills environmental science faculty believe are necessary for these researchers in hand, I turned my attention to examining the data science skills employed by environmental science graduate students in their research.

Despite the elevated importance of data science to the fields of Statistics and the environmental sciences, research has yet to focus on investigating the data science skills necessary for graduate-level research in the environmental sciences. This final arm of my research focuses on using a qualitative method of investigation to describe and understand the key data science skills necessary for environmental science graduate students as they engage in the data analysis cycle.

For this research, an embedded comparative case study (Yin, 2009) was employed. This comparative case study described the key data science skills used by two environmental science graduate students, Alicia and Ellie, and compared the key skills found for each student, in the context of their educational experiences. Where the phenomenology detailed in Chapter 2 focused on describing the shared experiences of environmental science graduate students when acquiring the data science skills necessary for their research, this case study focused instead on describing the specific data science skills used by two individuals. For this case study, Alicia and Ellie were the cases and the `R` scripts produced for their respective research were the embedded units of analysis.

At the outset of this study, a cohort of eight graduate students from environmental science fields were recruited from first semester Methods of Data Analysis courses, in the spring of 2018. These students were recruited from a variety of environmental science fields to develop an understanding of what key data science skills span across environmental science fields of research. Each of these students participated in at least two interviews, between the fall of 2018 and the fall of 2019. For the first interview, students were asked to submit all of the research code they had produced thus far. For each subsequent interview, students were requested to submit any research code they had produced since the last interview. I produced analytic memos for each of these script files, to synthesize the data science skills used throughout each student's script into higher level analytic meanings (Miles et al., 2014, p. 95). During the interview, students were then asked to describe how they learned the data science skills outlined in these memos.

When the focus came to outlining an analytical framework, however, it became clear that this initial sampling methodology aligned with grounded theory research, with the purpose of generating a substantive theory of the prevalence of specific data science skills used by environmental science graduate students in their research. Regretfully, the sampling logic of a grounded theory methodology did not align with the study's intention to intensively explore *both* the computing skills employed by students when implementing statistics in their research *and* how these skills evolve over time. Instead, an embedded case study aligns with this research goal, by selecting a few individuals and painting a picture of the data science skills they used in their research, and how each individual's skills evolved over time. Furthermore, a comparative case study allows for the comparison of the data science skills used by each student, in the context of their personal experiences.

The rationale for selecting Ellie and Alicia were two-fold. Their experiences represent two ends of the spectrum in the computational preparation and support of environmental science graduate students as they perform data-intensive research in their field. These experiences differed in four primary ways: (1) their programming backgrounds, (2) the statistics coursework they completed for their degree, (3) the field-specific quantitative methods coursework they completed for their degree, and (4) the computing and statistical support of their adviser. Second, the research code produced by Ellie and Alicia also represents two substantially different types of computational tasks environmental science graduate students might face as they engage in the data analysis process.

Chapter 4 describes the design, analysis, and findings of this embedded case study research. Finally, Chapter 5 concludes our work and presents directions for future research outlining a learning trajectory for how students build understandings of data science concepts.

HOW ENVIRONMENTAL SCIENCE GRADUATE STUDENTS ACQUIRE STATISTICAL COMPUTING SKILLS

Contribution of Authors and Co-Authors

Author: Allison Theobold

Contributions: Designed study, collected data, performed analyses, interpreted results, and wrote manuscript.

Co-Author: Stacey Hancock

Contributions: Discussed results and implications and edited earlier manuscripts.

Manuscript Information Page

Allison Theobold & Stacey Hancock

Statistics Education Research Journal

Status of Manuscript:
_____ Prepared for submission to a peer-reviewed journal
_____ Officially submitted to a peer-reviewed journal
_____ Accepted by a peer-reviewed journal
__X__ Published in a peer-reviewed journal

Publisher: *International Association for Statistical Education & International Statistical Institute*

Submitted: January 7, 2019; Published Online: November 20, 2019; `https://iase-web.org/documents/SERJ/SERJ18(2)_Theobold.pdf?1575083627`

Theobold, A. and Hancock, S. (2019). How environmental science graduate students acquire statistical computing skills. *Statistics Education Research Journal.* `https://iase-web.org/documents/SERJ/SERJ18(2)_Theobold.pdf?1575083627`

Abstract

Modern environmental science research increasingly requires computational ability to apply statistics to environmental science problems, but graduate students in these scientific fields typically lack these integral skills. Many scientific graduate degree programs expect students to acquire these computational skills in an applied statistics course. A gap remains, however, between the computational skills required for the implementation of statistics in scientific research and those taught in statistics courses. This qualitative study examines how five environmental science graduate students at one institution experience the phenomenon of acquiring the computational skills necessary to implement statistics in their research and the factors that foster or inhibit learning. In-depth interviews revealed three themes in these students' paths towards computational knowledge acquisition: use of peer support, seeking out a singular "consultant," and learning through independent research experiences. These themes provide rich descriptions of graduate student experiences and strategies used while developing computational skills to apply statistics in their own research, thus informing how to improve instruction, both in and out of the formal classroom.

Introduction

With the increased focus on data-intensive research, statistical computing has become essential in many scientific fields. Yet, the gap between science education and students' computational knowledge has become more evident, particularly in the environmental and life sciences. The growth in computational power and the volume and variety of available data has multiplied the computational and statistical expectations of scientific researchers' abilities. Yet an abundance of literature in the environmental sciences suggests graduate students are not acquiring the computational skills necessary for their research (Andelman et al., 2004; Green et al., 2005; Hampton et al., 2017; Hernandez et al., 2012; Lai et al., 2019; Teal et al., 2015).

Contrasted with graduate students in the biological sciences, where external structures often exist to support computational knowledge acquisition (Stefan et al., 2015), environmental science graduate students are often assumed to acquire computational skills in graduate-level statistics courses. The requirement of graduate-level statistics coursework is intended to help these students acquire the statistical knowledge necessary for their research along with any essential computational skills, but little is known about the paths graduate students actually rely upon when faced with statistical computing problems in their research. The intention of this study is to describe the experiences of graduate students in the environmental sciences to illuminate the phenomenon of acquiring the computing skills necessary to apply statistics in the context of their research. We consider the following research question: Through what paths do graduate students in the environmental sciences gain the computational knowledge necessary to implement statistics for research applications in their disciplines?

The subjects of this study were graduate students enrolled in a second semester graduate-level Applied Statistics course at a mid-size university in the Western United States. The target audience of this course is non-statistics graduate students, and, at this institution, this two-semester Applied Statistics sequence is either required or highly recommended for the completion of a master's degree in departments such as Ecology, Land Resources and Environmental Sciences (LRES), Animal and Range Sciences (ARS), and Plant Sciences & Plant Pathology. This sequence of two one-semester courses covers the foundations of statistical inference, including a wide variety of statistical methods, starting from two sample inferences and moving through regression and generalized linear models to mixed models. Taught using an R (R Core Team, 2020) programming environment, students are typically given code to modify, covering base R graphics, data and model summaries, and built-in functions, while also being exposed to a few computational concepts such as loops, and conditional and relational statements.

The majority of graduate students in Ecology, LRES, ARS, and Plant Sciences departments enroll in the graduate-level Applied Statistics course sequence or solely in the first course in this sequence. Thus, this terminal statistics sequence often serves as graduate students' sole statistical computing course, and consequently, their only formal preparation for the computational problems they may face when implementing statistics as researchers. In examining the experiences these environmental science graduate students face when acquiring the computational skills necessary to use statistics in their research, we seek to capture an in-depth understanding of the successes and shortfalls these students encounter in their computational journey.

Though the term "Environmental Science" refers to a specific discipline in the literature, in this paper we will refer to the collection of fields that perform research in the biological and environmental sciences as "environmental science." At our institution, these are the fields whose students are required or highly-recommended to enroll in the graduate-level Applied Statistics course sequence described above. For this study, "statistical computing" is considered to consist of the computing knowledge and skills necessary for the entire process of statistical analyses, from data cleaning to data visualization to data analysis. These computing skills may include programming concepts such as loops, user-defined functions, or conditional statements, and methods for importing, cleaning, and subsetting data.

We begin by describing areas of the research literature that address the computational and statistical training of graduate students in the environmental and biological sciences. We then outline the qualitative study we implemented to explore the experiences of graduate environmental science students in acquiring the statistical computing skills necessary for their research. The results presented reveal the prevailing experiences of these students when faced with computational problems beyond their understanding, and articulate the paths students employed to gain the required computational skills for carrying out statistics in their research.

Computing and the Environmental Sciences

Research in the computational abilities of environmental science students is in its infancy, with only a handful of institutions performing research that specifically addresses the computational training necessary to prepare students for careers post undergraduate or graduate degree. Literature related to this area has primarily focused on resources that students could potentially use to increase their computational abilities, with no studies focusing on the resources graduate students actually employ when wrestling with the computing problems necessary to apply statistics in the context of their research.

In this section, we discuss briefly three broad areas of the literature that informed this study. First, we review the literature on the foundational role computation has in the sciences. We then discuss research efforts detailing computational training in the environmental sciences, as compared with the computational training of graduate students in other biological fields. Finally, we detail research in statistics education declaring the importance of computing in the statistics curriculum.

Computing and Statistics in the Sciences

Over the last two decades, nearly every scientific field has seen a rapid increase in the use of computation and analytical tools to model phenomena across many disciplines of inquiry. In some scientific fields, such as biology and chemistry, the recent ability to collect multitudes of data easily and quickly have made computational abilities vital to researchers and practitioners. Fields previously thought to be niche disciplines, such as computational biology, are now “becoming an integral part of the practice of biology across all fields” (Stefan et al., 2015, p. 2). Across a large sector of scientific domains, computationally heavy applications of mathematical and statistical techniques, such as management of large data sets, dynamic data visualization, and computationally intensive modeling and prediction, have become essential computational understandings for field applications (Weintrop et al., 2016). With these advances in computational power, analytical methods, and detailed computational and statistical models, scientific fields are undergoing a renaissance. These advances have, however, created a growing need for scientists to receive an appropriate education in computational methods and techniques (Fox and Ouellette, 2013; Wing, 2006).

Many chemistry, biochemistry, and bioinformatics programs have begun to incorporate computational training into their programs. A similar revolution affirming the importance of computational proficiency has yet to be experienced in environmental science fields.

Computational Training for Graduate Students in Environmental Science

The volume and variety of data collected by environmental science researchers for statistical analysis continues to increase at a rapid pace due to the availability of data from “long-term ecological research, environmental sensors, remote-sensing platforms, and genome sequencing” (Hampton et al., 2017, p. 546). These technological advances have created a crucial need to reevaluate how our system of training can better prepare current and future generations of environmental researchers (Green et al., 2005; Hampton et al., 2017).

Facing the new frontiers of “big data,” programming skills to manipulate, analyze, and visualize data are becoming necessary for many ecologists. Moreover, most environmental science graduate students are required to write their own code as part of their research (Mislan et al., 2016), with the use of `R` as the “primary tool reported in data analysis increasing from 11.4% in 2008 to 58% in 2017” (Lai et al., 2019, p. 1). In a survey of a seminar course for graduate students in ecology across 11 American universities, however, Andelman and colleagues (2004) found that “ninety-three percent of students did not have skills in the scripted programming languages (e.g., SAS or MATLAB) that are needed for the integration of large datasets” (p. 244), and that one of the greatest limitations students experienced was related to data concatenation, manipulation, and analysis. Furthermore, in a recent survey of graduate students in the environmental sciences, “74% of students reported they had not completed any coursework related to the management and analysis of complex data” and only 56% of students “claimed a basic skill level in statistical applications, including `R`” (Hernandez et al., 2012, p. 1069).

This lack of computational training required for data analysis inhibits the progress of research and is laden with hidden costs. Teal and colleagues (2015) suggest that “researchers learn most of what they know about programming and data management on their own or the information is passed down within a lab” (p. 136). The costs associated with this process are substantial. Graduate students “can spend weeks or months doing things that could be done in hours or days,” they may be unaware of the reliability of their results, and they are often unable to reproduce their work.

Not all biological graduate students, however, are experiencing a lack of computational training. For example, researchers in the Department of Biological and Biomedical Sciences at Harvard have developed an intensive workshop that introduces graduate bioinformatics students to the “fundamentals of programming, statistics, and image and data analysis through the use of MATLAB” (Stefan et al., 2015, p. 2). This course is framed not only with the goals of developing programming skills and statistical understandings, but also emphasizing how to algorithmically reason through computational problems. The structure of the two-week intensive “bootcamp” consists of five full, mandatory days. The workshop dedicates the first two days to an introduction to programming using MATLAB, where students learn a variety of topics, including creating variables, performing basic variable operations, indexing, logicals, functions, conditionals, and loops. Day 3 is dedicated to developing statistical understandings, including probability distributions, hypothesis testing, p-values, bootstrapping methods, and multiple testing. Day 4 covers topics in image analysis, and Day 5 assists students in working with their own data. These workshops are given twice a year, once prior to the start of the school year as new graduate students are attending orientation, and a second time for upper-level graduate students and post-doctoral fellows (Gutlerner and Van Vactor, 2013). In introducing beginning graduate students to these concepts, researchers hoped to lower the computational barrier for students taking courses, empower students to learn computational tools on their own, and allow for other courses to “build upon this foundation and integrate quantitative methods throughout the curriculum” (Stefan et al., 2015, p. 2).

Providing effective training in data-intensive computational skills for researchers is wrought with challenges. Strasser and Hampton (2012) reported that ecology instructors indicated eight barriers to covering data-intensive computational skills. These barriers included limited time, students did not have the necessary level of quantitative or statistical skills to cover the topics, lack of resources, the instructor was not knowledgeable in these topics, topics should be included in a lab, and the topics should be covered in other courses. These obstacles 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).

When considering the issue of curriculum reevaluation, however, we note that, for many environmental science fields, statistics preparation is considered vital, and statistics courses have readily been incorporated into undergraduate and graduate programs across the country.

Computing in the Statistics Curriculum

The digital age is also having an overwhelming impact on the practice of statistics and the nature of data analysis, which necessitates a "reevaluation of the training and education practices in statistics" (Nolan and Temple Lang, 2010, p. 97). The skills needed by today's statistics practitioners differ profoundly from what was needed 20 years ago. For scientific research today, computing skills are vital, especially for scientific research requiring statistical analysis (Hardin et al., 2015, p. 344).

Nearly 20 years ago, Friedman (2001) noted that "computing has been one of the most glaring omissions in the set of tools that have so far defined statistics" (p. 8). This statement is echoed in the calls from statisticians advocating for changes in the statistics curriculum (Cobb, 2015 [Discussions from Gelman, Gould, Duncan Lang, Kass, Nolan]; Nolan & Temple-Lang, 2010), as "what we teach lags decades behind what we practice" (Cobb, 2015, p. 268). Furthermore, computing has become more necessary to implementing statistical methods than even ten years ago such that "a 'just enough' level of understanding of computing is not adequate" (Nolan and Temple Lang, 2010, p. 106).

Many statisticians would agree that more computing should be included in the statistics curriculum so that students leave the classroom more computationally capable and literate. However, many statistics students are "told to learn how to program by themselves, from each other, or from their teaching assistant in a two-week 'crash course' in basic syntax at the start of a course" (Nolan and Temple Lang, 2010, p. 100). This do-it-yourself approach signals to students that statistical computing is not of intellectual importance compared to materials covered in lectures. Additionally, this structure inherits additional hidden costs, where students may pick up bad habits, misunderstandings, or the wrong concepts. Students may learn "just enough to get what they need done, but they do not learn the simple ways to do things," and the knowledge they possess when approaching a problem limits the tasks they are able to accomplish (p. 100). This brings us to question whether students in our statistics courses acquire the confidence necessary to overcome computational challenges they may face in their scientific research.

Due to the historical importance of statistics in environmental science fields, graduate students are often required or highly recommended to enroll in statistics courses for completion of their degree. As evidenced by literature in the environmental sciences, however, graduate students are not being prepared by their current curricula with the computational skills necessary to perform data-intensive environmental science research. Indeed, these commentaries by statistics educators also illuminate the lack of computational preparation with which students often leave the statistics classroom.

Methodology

In this study, we examined experiences of environmental science graduate students in gaining the computational knowledge necessary to implement statistics in their research, and the paths that impacted these experiences. Implementation of statistics is necessary for many of these graduate students to succeed in their master's and doctoral research. Across these fields, however, students may not be acquiring these necessary skills within their graduate curriculum.

Phenomenology is a study of "people's conscious experience of their life-world" (Schram, 2003, p. 71) or their "lived experiences" (Van Manen, 1990, p. 9). As compared to case study research, which stresses the "unit of analysis, not the topic of investigation" (Merriam, 2009, p. 41, emphasis in original), a phenomenology aims to depict the essence or the structure of a shared experience through analyzing and comparing the experiences of different people (Patton, 2002).

A phenomenology was appropriate for this study, as it focuses on the experiences of graduate environmental science students as they acquire the computational skills necessary to apply statistics in their research. Participants for this study were not chosen to illustrate different aspects of a shared experience. Rather, these participants act as a cohort to illuminate and understand the phenomenon of acquiring the computational skills necessary to implement statistics through participants' lived experiences. Aspects of the backgrounds from each of the study's participants may characterize a "typical" graduate student in the environmental sciences, however, it is not the intention of these participant characterizations to focus on how backgrounds impact the experience of this phenomenon.

Participants

At our university, the two-semester graduate-level Applied Statistics course sequence (GLAS I and II) serves as a service course for graduate students in scientific fields, and only assume prerequisite knowledge of Introductory Statistics. Additionally, GLAS I serves as the required prerequisite course for other statistics courses in the department.

Students were recruited from GLAS II in the spring of 2017. These students were interviewed following their spring break, nearly halfway through the course. Only graduate students taking the course for their respective master's or doctoral programs in environmental science fields were considered.

We requested all eight environmental science graduate students enrolled in GLAS II in the spring of 2017 to complete a survey detailing their previous statistics and computer science courses, the computer languages with which they had experiences, and their independent research experience. All eight of these students completed the survey and were then asked to participate in an in-depth interview, of which five agreed. Names of participants used in this paper are pseudonyms.

Details of the five interview participants are summarized in Table 2.1. All five identified as women, and all had taken GLAS I within the last two years. Four of the interview participants had begun or were nearly finished with their master's thesis, while Robin had just begun to work on the projects associated with her dissertation.

Of the five interview participants, Catherine's only prior statistics course had been GLAS I, Beth, Kelly, and Robin had all taken another statistics course outside of GLAS I and II, and Stephanie was completing a Graduate Certificate in Applied Statistics. The Graduate Certificate in Applied Statistics requires the completion of GLAS I and II, as well as Sampling or Experimental Design, and one additional upper-level statistics course. The Experimental Design course covers the foundations of design and analysis of experiments, including a large variety of experimental methods, starting from matrix forms and moving through factorial, balanced complete and incomplete blocking, and split plot designs. The Sampling course covers the cornerstones of sampling methodology, including a wide variety of probability samples, from simple random sampling to systematic sampling and cluster sampling. Both courses are taught using a SAS programming environment, where students are typically given code to modify. Other courses often taken for completion of this certificate include Time Series Analysis, Multivariate Analysis, Mixed Effects Models, and Generalized Linear Models.

	Beth	**Catherine**	**Kelly**	**Robin**	**Stephanie**
Degree Seeking	MS	MS	MS	PhD	MS
Department	ARS	LRES	Ecology	LRES	LRES
GLAS I	Fall 2015	Fall 2015	Spring 2016	Fall 2015	Fall 2015
Additional Statistics Courses	Experimental Design	None	Sampling	Time Series	Time Series, Experimental Design
Languages Introduced in Coursework	`R`	`R`, SQL	`R`, SQL	`R`, SQL, Python	`R`, SQL, Python, Java
Languages Employed in Research	`R`, SQL	`R`	`R`	`R`, SQL, Python	`R`, SQL, Python
Independent Research	Thesis	Thesis	Thesis	Thesis	A few projects

Table 2.1: Academic demographics of participants: GLAS I indicates the academic semester they took the first semester graduate-level Applied Statistics course.

Over the last five years, this first semester graduate-level Applied Statistics course sequence has serviced 101 students from the departments of Ecology, LRES, ARS, and Plant Sciences. Of those 101 graduate environmental science students, 63% have gone on to complete the second semester graduate-level Applied Statistics course sequence, and only 5% have completed the Graduate Certificate in Applied Statistics.

Every interview participant from the Ecology and LRES departments voiced that they had taken a required course for their graduate coursework which introduced Access databases, providing them with experiences working with a structured query language (SQL). Robin and Stephanie continued to use SQL during their independent research and Beth learned SQL independently at the recommendation of her adviser. Unlike many environmental science graduate students, Stephanie had experience with Java from her undergraduate coursework and gained knowledge for working in Python and R from a year's work as a research assistant prior to enrolling in graduate school.

Data Collection

Following the preliminary survey, students who agreed to be interviewed were audio recorded while working through a set of ecological applications of statistical computing. These tasks assessed students' abilities to reason through applications of statistical computing, covering a broad range of problems that may be necessary for research in environmental science. These tasks were not intended to determine what statistical computing knowledge each participant did or did not possess, but rather as an entry point to capture the experiences of these participants in acquiring the statistical computing skills with which they were familiar.

After reasoning through each task, students were asked where and how they had acquired the computational skill they had employed. Based on participants' responses, the interviewer asked a follow-up question to gain additional information regarding why the participant used this resource to acquire the computational skill in question. For instance, if a participant voiced acquiring the statistical computing skill in a course, further information was sought out regarding why she enrolled in that particular course. Alternatively, participants who voiced the Internet as their resource in acquiring the statistical computing skill were asked for additional information regarding what Internet resources they had employed. All participants were asked whether they attempted to use other resources when acquiring each skill, as well as how often they had used each resource when acquiring computational skills. Finally, every participant was asked to summarize where they have learned the computational skills necessary for implementing statistics in their research. The full interview protocol is included as an Appendix and the statistical computing tasks are included as Supplementary Materials.

The analysis in this paper is based on participant responses to questions regarding their experiences acquiring the computational skills they employed while reasoning through these statistical computing tasks.

Data Analysis

The primary author led a three-stage data analysis process (Miles et al., 2014). In the first stage, the interviews for each participant were transcribed verbatim, with participants' names removed and pseudonyms given. Subsequently, the primary author read the transcripts independently and created descriptive codes for the paths through which the participants voiced having acquired the computational skills they employed when reasoning through the statistical computing tasks. Concluding this stage, the author looked for specific references to how the courses taken by the participants had influenced their acquisition of statistical computing skills.

After working through each transcript in this manner, the primary author began the second stage of analytical coding. In this process, every path was given equal value and "nonrepetitive constituents of experiences" were linked thematically

(Moustakas, 1994, p. 96). Categories of experiences that held across multiple interviews were retained. For example, every participant voiced specific individuals they sought out as paths for knowledge acquisition. These activities were initially coded to belong to the category of "learning from others." Based on these groupings, initial categories of course work, research experience, and learning from others were constructed. Next, the primary author searched through the data to identify successes and limitations voiced by the participants when acquiring statistical computing skills within the initially identified categories. Through this step we learned that certain categories were instead subcategories, whereas others were independent of one another. For example, some participants voiced exposure to computational skills in the statistics classroom but emphasized that their understanding of these skills instead came through interactions with their peers or when using the methods in their own research. Additionally, participants who learned from others found great success in acquiring statistical computing skills from a single person in their lab or department, as compared to the limited success select participants had when using their peers to acquire statistical computing skills.

In the final analysis stage, the primary author identified emerging themes arising from these categories to describe the phenomena of acquiring statistical computing skills. The author searched for instances which reiterated the themes, as well as negative cases, with attention paid to the transcripts throughout the validation process. Following the validation process, both authors met to discuss the rationale for coding, scrutinizing the situation of each participant's description of their paths of knowledge acquisition in the context of the emergent themes. Ultimately, we reached consensus regarding the categories in which each participant's response was placed.

Although the frequency of use varied across participants, every participant voiced experiences acquiring statistical computing skills across every path, supporting the themes that emerged. The final themes were exhaustive, mutually exclusive, and sensitizing, so that the name of the theme authentically represented the data (Merriam, 2009). These final themes present the "essence of the phenomenon" (Creswell, 2007, p. 62) of acquiring the computational skills necessary to implement statistics in environmental science fields.

Following this process, we provided the participants with the table outlining the computational skills they employed when completing the statistical computing tasks and the paths from which they voiced acquiring each skill. The participants recommended no change to be made to the table they were provided. This inclusion of member checking allows participants to check for accuracy of their statements. The ability of this study to authentically capture the experiences of students is enhanced with the lack of researcher engagement with students prior to their participation in the study. This helped to ensure that no student felt more comfortable in the interview environment, articulating their experiences, than any other student.

Results

When investigating the phenomenon of acquiring the computational knowledge necessary to implement statistics in environmental science research, we expected themes of coursework and support structure to emerge. The experiences that emerged from every participant's interview, however, related primarily to the support structures they employed, rather than the coursework that helped them to acquire the computational knowledge necessary for applying statistics in their research. In this section, we present themes describing the phenomenon of statistical computing knowledge acquisition that developed throughout the participants' interviews: (1) independent research, (2) singular consultant, and (3) peer support. A subtheme of coursework appeared within peer support and independent research, where participants voiced the importance of their coursework on their knowledge of statistical computing. Participants consistently voiced this sub-theme to depend on either peer assistance or independent research in its impact on participants' understanding of statistical computing. The themes and sub-themes are summarized in Table 2.2.

Theme	Sub Theme	Description
Independent Research	Coursework	Research experiences that allowed students to take their course knowledge and transfer it to statistical computing applications
Singular Consultant		All-knowing past or current graduate student whom students sought out for computational assistance
Peer Support	Coursework	Assistance from peers with statistical computing tasks

Table 2.2: Participants' themes in acquisition of statistical computing knowledge.

In the sections that follow, we provide a detailed description of each theme, supplemented with quotations from participants to ensure authentic descriptions of their experiences.

Independent Research Experience

The first theme in acquiring statistical computing knowledge was participation in independent research. Involvement in independent research helped students transfer their course knowledge to statistical computing applications. This environment helped students to see the messiness of non-classroom applications and feel the unease that comes when attempting to perform statistical computing tasks beyond one's knowledge. These experiences came predominantly in the form of working as a research assistant prior to entering graduate school, collaborating on a project in the first year of graduate school, or performing research for a master's thesis, or ultimately, a doctoral dissertation.

Catherine, a master's student in Environmental Science, who still faced everyday computational struggles, attributed the majority of her application-specific computational knowledge to her experiences in independent research. She emphasized the importance of understanding how to work in a statistical computing environment, such as `R`, which she learned by performing research, before she was able to begin to transfer the statistical knowledge she had learned in the classroom to her research:

> What I struggled with is [GLAS I] covers theory really well, but since I was new, I spent most of my time trying to figure out how to apply that theory in [`R`]. And even now I struggle transferring from R into actual statistical theory, when I'm writing my thesis. The way I had to approach it was I had to learn the `R` first, then I was able to look back on what I had actually done, in order to learn the statistics.

Kelly, an Ecology master's student, described her experiences with data management for her master's thesis as having produced the most substantial contributions to her computational abilities. Often, she attributed her intuition for solving statistical computing problems to experiences she had "merging data sets" and learning to use conditional statements for her research project. She emphasized the importance of her statistical knowledge gained in both graduate-level statistics courses in understanding "what statistical method to use," whereas she attributed becoming more fluent in statistical computing to her research experiences: "The data management stuff comes from independent research, trial and error, getting myself through." In this context, Kelly seemed to be reflecting on the computational skills she acquired when applying the statistical methods from the classroom in the context of her research, not the skills she acquired from the "trial and error" process involved with performing research. Similar sentiments were voiced by Beth, an Animal & Range Science master's student, who attributed nearly all of her computational knowledge as having stemmed from her independent research. With the recommendation of her adviser, she taught herself how to create an Access database to store her data. In storing her project data in this manner, she was able to learn important concepts about data structures, subsetting data "using qualifiers and criteria," and sorting data, skills which were then easily transferred into `R` to manage data for analysis.

Singular Consultant

When describing who they seek out for computational help, every participant described first seeking out an "all-knowing" past or current graduate student. These individuals served as "singular consultants," with whom these students had the "best," most productive experiences in finding solutions to computational problems that had arisen in their implementation of statistics to their research. For Beth, this singular consultant came in the form of a past graduate student from Animal & Range Sciences who was hired to help faculty complete projects:

> We have a guy who used to be a student in our department and then he was hired on again to help finish some projects, after he got his master's in Statistics. He is very helpful with [pointing out what's wrong with your code]. He's very good with code and if I have a quick question he can always answer it.

For Kelly, another graduate student on her same project served as this consultant. Kelly described turning to this particular graduate student for help with computational problems she had encountered in her thesis; she added that other graduate students in their department also used this person as a consultant for their computational problems:

> The other grad student on this project is so well versed in R that he's unofficially become the person that people go to with questions.

Throughout her computational struggles, Catherine found assistance from previous graduate students from the department, but she found the most assistance from a previous graduate student "who had left the department and was off professionally somewhere else, but he still took the time to help walk me through [my code]."

One participant, Stephanie, an Environmental Science master's student, served as this singular computational consultant for many members of the Environmental Science department. With her experiences teaching herself R, she was able to "explain code in a way that makes sense," says Robin, a fellow Environmental Science doctoral student who has often sought out help from Stephanie. With an adviser from a computational background and a project which required sophisticated statistical modeling, Stephanie "had to learn to code." Additionally, her laboratory often worked in collaboration with Computer Science faculty, where she and her lab-mates were taught computer science coding practices and jargon. "Stephanie has gotten good at teaching it, because everyone on our floor is like 'I can't do this, Stephanie help me'," said Robin. Stephanie stated that graduate students have sought her assistance "daily" or "at minimum two to three times a week." In contrast, when Stephanie experiences difficulty in performing computational tasks, she has found solace in her lab-mates and ultimately, when necessary, with her adviser:

> My entire lab works in the same room and my adviser's door is always open. So, if someone is having a major issue, whoever is in the room can hear that. If [my adviser] hears me ask [a lab-mate] how to do something and he knows how, he just shouts how to do it. So, it's a very group oriented dynamic. I've never had to go beyond the people in my lab.

Peer Support

The third theme in acquiring computational knowledge that all participants spoke of was the support they had received from fellow graduate students when performing computational tasks related to applications of statistics. The students described how, when they are unsure of how to complete a computational task for their research and their singular consultant is not available to them, they turn to fellow graduate students for help. Participants described instances when the computational tasks required of them were beyond their current knowledge or occasions when they had been unsuccessful at attempting to complete a problem and sought out help from a fellow graduate student. For example, Kelly, an Animal & Range Science master's student, shared that when she reached a point in coding when she didn't know how to do something, she turned to one of her lab-mates:

> I've been to a point where I didn't know how to do something with my knowledge or what I can find online, and then I'll go to one of my lab-mates.

Catherine, a master's student in Environmental Science, spoke of the expectations of her advisers that the computational problems she was being asked to perform were "easy, since she had all the information." Catherine has had numerous experiences, however, where she did not have the knowledge necessary to perform the task or she was missing "little caveats" that kept her from fully being able to perform the tasks. When faced with these problems, she "reached out to previous students that had taken the course."

Robin, a doctoral student in Environmental Science, reiterated Catherine's experiences, describing how she reached out to other graduate students in other labs for help with computational problems. Alternatively, Stephanie, as a singular consultant, voiced that when she was faced with computational problems beyond her knowledge, she had never been forced to "go beyond talking to her lab-mates" for assistance.

Unfortunately, peer support did not always provide an optimal solution. This may be a potential reason that participants sought help from peers only when their singular consultant was unavailable. For example, Kelly described negative experiences when seeking computational assistance from graduate students not of close proximity to her:

> When I'm struggling with something and I go to other grad students, they'll say "I did this the other day. I'll send you my code." I've found most of the time I don't understand what they've done enough to plug in what I want and make it work. There have been a few times when making tables and plots and someone sends me their code and I can just plug in my data and it works just fine. I've had less success with that.

Discussion

The present study, although exploratory in nature, outlines the experiences of environmental science graduate students to shed light on the phenomenon of obtaining the computational skills necessary to apply statistics in the context of environmental science research. The themes identified, and their corresponding examples, illustrate the essence of the structure of the shared experience of these participants. These results help to illuminate the gaps that exist between the statistical computing skills these students acquire through their curriculum and the computing skills required for them to successfully implement statistics in their research.

Our expectation of coursework to be a primary source of statistical computing knowledge was not found for these participants. When these graduate students encountered a statistical computing problem, they would pull upon the knowledge they had acquired through their graduate coursework, but this knowledge was often insufficient. Rather, the computational understandings that these students attributed to their statistics coursework were primarily low-level concepts, such as using built-in R functions, adding comments to their code, and limited troubleshooting of error messages. Additionally, these low-level concepts were said to only be fully understood through participants' peer interactions, or as they were being implemented independently within their own research.

Instead, participants voiced that having experiences performing independent research substantially influenced their abilities to reason through and perform the computational tasks required for various statistical analyses. Through independent research, the participants were able to play with real-world data and applications more complex than what they had encountered in the classroom. The programming skills developed during a student's independent research, in conjunction with peer collaboration, were described largely as high-level concepts, such as conditional statements, loop implementation, and user-defined functions. Students described their independent research as having opened the door to experiencing the unease that comes when one is asked to perform statistical computing tasks beyond one's knowledge, a feeling they had not encountered in their courses. In these circumstances, students stated that they would ask for help from the people with whom they had the most prior success or felt the most comfortable.

In a direct connection to the participants' discomfort in asking for help from an adviser, the theme of a singular consultant emerged. These singular consultants served as an "all-knowing" individual, from whom the participants had either had the "best" experiences with, where the individual spent the necessary time to explain the concepts, or the consultant had always been capable of providing the participant with a solution to their problem. These individuals served as the first line of defense when statistical computing problems arose, where participants were both able to seek computational help and acquire new computational skills and understandings through their interactions. If this consultant was unavailable to the graduate student due to time or physical constraints, these students then turned to their peers.

Peer support was initially discussed by the participants in their interviews as a mechanism they used when their "code doesn't run" or when they were asked (or needed) to do something beyond their current computational understandings. However, this theme continued to emerge as the participants worked through computational problems, often attributing their knowledge of a computational procedure to a friend or fellow graduate student helping them "do it with their data." These peers offered a path for students to seek help, often voiced to be more comfortable than asking an adviser, where participants described both the fear of asking and "feeling dumb" or being "brushed off" because their adviser thought they should "be able to figure out how to do it." As opposed to the help participants received from their singular consultant, these students also voiced negative experiences they had encountered when seeking help from their peers, such as a peer sending them "helpful" code that they did not understand.

Lastly, the adviser played an important role in students acquiring the computational knowledge necessary to perform applications. Despite students' reluctance to seek out computational assistance from their adviser, advisers did often emphasize the importance of statistical computing skills, as well as introduce (or recommend) students to store their data using a relational database. The participants' ability to understand both data structures and sorting or filtering data was largely attributed to their experiences working with these types of databases. Although these interviews found that advisers were often considered as the last line of defense, they were, however, viewed as an accessible way for students to better understand the statistical computing necessary for their independent research projects, which overall contributed to better computational understanding and skills for these students.

Implications

The implications for statistics and environmental science education focus on identifying and understanding the importance of the computational knowledge necessary to apply statistical methods in environmental science research, and the paths graduate students employ to acquire these essential skills. Environmental science fields have long understood the importance of statistics education for their students, so a preponderance of programs recommend or require at least one graduate-level statistics course. Conversely, many of these graduate programs are not actively incorporating computational courses into their degree, instead assuming that students are acquiring these skills in their recommended statistics courses. Unfortunately, computational skills necessary for research are not typically included in these statistics courses (Friedman, 2001; Hardin et al., 2015; Nolan and Temple Lang, 2010). As evidenced in the research on computational preparation of environmental science students (Andelman et al., 2004; Green et al., 2005; Hampton et al., 2017; Hernandez et al., 2012; Mislan et al., 2016; Teal et al., 2015), the experience of poor computational preparation is not unique to students at this institution. A restructuring dilemma is faced by both fields—statistics education and the environmental sciences—with intractable differences between the curricula of statistics service courses and the expectations of environmental science research.

Implications for Statistics Educators

Statistics educators should consider the power an applied statistics course sequence has to provide graduate students with a year-long introduction to statistical computing. As seen by Stephanie, who entered graduate school after completing a year's work as a research assistant working in R, these learning experiences can help to alleviate the power differential students feel when asking their advisers or peers for assistance. However, the content covered by graduate applied statistics sequences is expected to paint a vast picture of the field of Statistics, with topics ranging from a difference in means to mixed-models. Consequently, many educators feel they do not have the time to incorporate statistical computing into the classroom, and some feel that they have limited computational expertise to teach these concepts (Hampton et al., 2017; Nolan and Temple Lang, 2010). The inflexibility of graduate programs further complicates this issue, as many graduate students are unable to enroll directly in a statistical computing course due to an already full and demanding course load. Thus, questions should be raised about how to best bridge this gap between coursework and research expectations for statistical computing skills.

The importance of playing with statistical applications on real-world data, as voiced by these participants, should also be considered by statistics educators at all levels. This transition to incorporating authentic, research-like tasks, which engage students in statistical computing, can be supported by online resources, data-discovery tools, example datasets and code, and instructional tools, along with collaborative course designs and the sharing of instructional materials.

Implications for Environmental Science Educators

Due to the extensive research on computational preparation of environmental science graduate students, faculty in these fields have a growing awareness of these issues of computational ill preparation. Yet, most of this research has focused on a vast array of computational skills students do not possess, rather than focusing on the computational skills necessary to implement statistics in their research. Environmental science faculty should thus have an increased awareness of the statistical computing preparation with which graduate students leave the statistics classroom. As echoed by the participants in this study, the implementation of statistics in the context of environmental science research is not always as tidy as is presented in the classroom. Hence, to better support these students' acquisition of the computational skills necessary for implementing statistics in their research, additional preparation focusing specifically on statistical computing should be considered by faculty in these fields.

The impact of an undergraduate education on students' experiences as graduate researchers should be considered by all statistics and environmental science faculty in higher education when recognizing the importance of developing data-intensive statistical computing skills early on in undergraduate statistics courses. In this study, none of the participants voiced having any experience working with R in their undergraduate coursework. Instead, these students encountered R for the first time in their first semester of graduate school during the first graduate-level Applied Statistics course. The participants who had computing experiences in their undergraduate coursework or post baccalaureate research work or experience with Access databases were able to navigate learning R with greater ease than students with no computing experiences. This lack of computing experience was further compounded when students began their independent research, where students with fewer computational skills and understandings had substantially different independent research experiences than their counterparts with more. The frustrations of simple tasks, such as subsetting data or removing NA's, were felt by the participants who had completed a bachelor's without any computational elements to their coursework, whereas those who were exposed to even small amounts of computing in their undergraduate coursework were able to begin computational tasks in their research walking and not crawling.

Limitations and Future Research

Although the methodology we used to describe the phenomenon of acquiring the computing knowledge necessary to implement statistics for graduate students in the environmental sciences provided important themes of knowledge acquisition, it is not without its limitations. Eliciting descriptions of computational knowledge acquisition yielded varied experiences with each of the main themes, but richer data could be gathered in a future longitudinal study. Following graduate students throughout their program of study could further identify where students are acquiring statistical computing knowledge, as well as instructional methods that best assist students in obtaining these understandings. To better inform environmental science and statistics faculty, a thorough investigation of both the coursework and structure of courses completed by these participants could be performed. This would allow for a discussion of how to best integrate these computational concepts into current coursework requirements, so that students leave the classroom with understandings they can implement immediately in their own research.

The focus of this study of environmental science graduate students' experiences acquiring the statistical computing skills necessary for their research should not be generalized to experiences acquiring general computing or programming skills. Whereas general programming skills may overlap with statistical computing skills, the foundation of study of each set of skills differs. Rather than focusing on computer architecture, design, and application, statistical computing skills center around the study of data. Select universities have, however, begun to require general computing courses for undergraduates majoring in environmental science fields (Cortina, 2007; Rubinstein and Chor, 2014; Wilson et al., 2008). The doors to future research will open as these students begin to enroll in graduate programs in environmental sciences. This future research can instead focus on understanding how students transfer their general programming knowledge to acquiring statistical computing knowledge, and which skills possess the greatest overlap.

Conclusion

Statistical computing has become a foundational aspect of research in the environmental sciences. This small-scale exploratory study brings forward the experiences of graduate environmental science students in acquiring the computational understandings necessary to successfully perform statistical applications for independent research. Participants found the greatest success in acquiring the computational skills required for their research through independent research, a singular consultant, and peers. Whereas others have noted the importance of integrating computing into the statistics curriculum (Friedman, 2001; Hardin et al., 2015; Nolan and Temple Lang, 2010) or the lack of computational preparation for environmental science graduate students (Andelman et al., 2004; Green et al., 2005; Hampton et al., 2017; Hernandez et al., 2012; Mislan et al., 2016; Teal et al., 2015), we instead explored the phenomenon of acquiring the computational knowledge necessary to implement statistics in graduate environmental science research. The computational burdens experienced by these participants when implementing statistics in the context of their research and the computational understanding with which they left the statistics classroom suggest the need for integration of formal computational training into these programs. The present study helps to emphasize the importance of computing skills necessary for data-intensive environmental science research.

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 Mary Alice Carlson, Jennifer Green, Megan Higgs, Megan Wickstrom, co-editor Jennifer Kaplan, assistant editor Beth Chance, and reviewers for their insightful comments on this paper.

CONCLUSION

This body of research is intended to provoke thought and discussion surrounding the computational preparation of graduate students in the environmental sciences with the data science skills necessary to engage in the entire cycle of data analysis. We began by outlining the evolution of the fields of statistics and the environmental sciences, catalyzed by the rapid increase in the volume and variety of data, and the computational tools available for analysis. These dramatic changes to the data landscape created a crucial need to re-evaluate the preparation of scientific researchers. Calls for this revitalization were echoed across both statistics and the environmental sciences, yet environmental science researchers continue to report that students are are not learning these critical skills in their curriculum (Hampton et al., 2017; Hernandez et al., 2012; Teal et al., 2015).

While Hernandez et al. outlined the shape of this ill preparation by describing the coursework and topics that students reported never encountering in their graduate program, no study had focused on the experiences of graduate students acquiring the computing skills necessary to analyze their data. This gap in knowledge inspired me to investigate how environmental science graduate students experience the phenomenon of acquiring the computing skills necessary to implement statistics in the context of their research. Through in-depth interviews with five environmental science graduate students, we uncovered three themes in students' paths to computational knowledge acquisition: use of peer support, seeking out a singular "consultant," and learning through independent research experiences. Furthermore, we described the statistical computing skills students reported leaving the statistics classroom with, and how their backgrounds affected their experiences acquiring these necessary skills. By in large, these students reported learning the computing skills necessary to analyze their data on their own or through information that was passed down within their social network. As suggested by statistics and environmental science educators alike, this “do it yourself” system results in substantial hidden costs and impedes the progress of scientific research (Nolan and Temple Lang, 2010; Teal et al., 2015). Because the computing included in the environmental science curriculum continues to lag behind, environmental science educators have repeatedly recommended extracurricular workshops as a bridge for students to acquire the foundational data science skills needed to conduct research (Hampton et al., 2017; Hernandez et al., 2012; Teal et al., 2015; Wilson, 2006).

The computational ill preparation of environmental science graduate students by their curriculum leaves a need for high-quality, relevant, and accessible trainings, equipping students with the data science skills needed to conduct research in their field. Although Data Carpentry workshops offer domain-specific training intended to provide researchers with the foundational skills necessary for data-driven research (Data Carpentry, 2020), no attention had been paid to the relevance of the content of these workshops to specific populations of researchers. This need motivated me to investigate how these discipline-specific workshops could be tailored to meet the needs of environmental science graduate students. However, this investigation required a more comprehensive understanding of the computing skills necessary for environmental science graduate students throughout their data analysis cycle.

Through interviews with environmental science faculty, we learned that faculty believed students need extensive experiences working with data and visualizing data, both using reproducible tools. Additionally, these faculty reiterated sentiments heard throughout the environmental science literature, that these students are not learning the data skills necessary for their research in the coursework required for their degree. The foundational data skills outlined by these faculty were then infused into Data Carpentry's *Data Analysis and Visualization in* `R` *for Ecologists* lesson (Michonneau et al., 2019). In addition, faculty also outlined data skills, such as conditional statements and repeated operations, that were not currently included in this Data Carpentry lesson. Software Carpentry, however, offers a lesson for learning to program in `R`, which teaches many of these additional programming skills (Wright and Zimmerman, 2016). Thus, these additional skills were integrated into the `R` *for Reproducible Scientific Analysis* lesson, tailored to have the same environmental science context as the other workshops. In the end, a suite of four workshops were developed: *Introduction to* `R`, *Intermediate* `R`, *Data Wrangling with* `dplyr` *and* `tidyr`, and *Data Visualization with* `ggplot2`.

During the 2018-2019 academic year, this suite of workshops was offered through a partnership with the Montana State University library. Advertised across campus, a total of 202 students, faculty, and staff attended at least one of the workshops. Although, many of the attendees solely attended the *Introduction to* `R` workshop, many attendees still selected to return for subsequent workshops. The attendees' pre-workshop surveys were consistent with what had been heard in the literature, as over 75% of workshop attendees had completed no formal courses in computer programming (Andelman et al., 2004; Hampton et al., 2017; Hernandez et al., 2012; Teal et al., 2015). Additionally, we discovered that the preponderance of these attendees had only taken a single statistics course, covering introductory concepts. As might be expected from the prevalent use of `R` in environmental science research (Lai et al., 2019; Mislan et al., 2016) and the current state of computing in the environmental science curriculum, over half of the master's and doctoral workshop participants attended the workshops for assistance with their research. Furthermore, the majority of these attendees reported using the internet and their peer networks as the main resource for learning `R`, consonant with the suspicions of environmental science educators (Teal et al., 2015). Finally, numerous participants reported that, at the workshop, they were hoping to learn something related to analyzing data, a desire which reiterates the importance of data science skills throughout the data analysis cycle.

The iterative nature of design-based implementation research demands the researcher revisit the content of their teaching innovation, to reassess its alignment with the desired learning outcomes. As the goal of these workshops is to equip environmental science graduate students with the data science skills necessary to conduct their research, this reevaluation of the workshop content requires an understanding of the data science skills these students are actually using in their research. Paired with the need to distill the broad classes of computing skills outlined by statistics and environmental science educators (Hampton et al., 2017; Nolan and Temple Lang, 2010), I embarked on research which would illuminate these foundational skills.

Case study research allows for the investigation of "a contemporary phenomenon within its real-life context," (Yin, 2009, p. 18) when the boundaries between the phenomenon and the context are not easily deciphered. Where the prevalence of the phenomenon of environmental science graduate students acquiring the computing skills necessary for their research has been outlined by environmental science educators (Hampton et al., 2017; Hernandez et al., 2012; Strasser and Hampton, 2012; Teal et al., 2015) and the first arm of this research (Theobold and Hancock, 2019), none of these studies have sought to understand this phenomenon in the context of students and their research. To illuminate the data science skills environmental science students use throughout the data analysis cycle, we conducted an embedded, comparative case study. By analyzing the research code generated by two environmental science graduate students, Alicia and Ellie, we identified themes of data science skills each student used throughout their code and created concept maps outlining the interwoven nature of these skills. This longitudinal exploration of the data science skills used by each of these women allowed us to map how each student's skills evolved over time. Furthermore, interviews with these women regarding their experiences acquiring the data science skills they made use of adds new perspectives to the discussions surrounding the computational preparation of graduate students in the environmental sciences by contrasting the computational preparation and support Ellie experienced with that of Alicia.

Rather than generating extensive evidence of the need for training for environmental science graduate students in the computing skills related to data, our research has focused on describing and understanding the nature of this need, through the voices of the graduate students. Our research has described these students' experiences acquiring the computing skills necessary to implement statistics in their research, explored how extracurricular workshops can be tailored to meet the needs of this population of researchers, and began the work toward identifying key data science skills necessary for these students as they engage in the data analysis cycle. Moreover, this research attests to the inseparable nature of statistics and data science, consistently focusing on the data science skills necessary for environmental science graduate students as they endeavor to implement the statistical analyses dictated by their research.

Directions for Future Research

Data science is here to stay—giving a name to the computing skills necessary for researchers to engage in the entire data analysis cycle. Potentially reflecting the growing awareness that statistical analyses are but one piece in the puzzle, the Google searches for "data science" now greatly overshadow that of "statistics" (Figure 3.1).

However, this dramatic shift has left statistics and environmental science educators grappling with what data science topics belong in the curriculum, when to teach them, and how they should be taught.

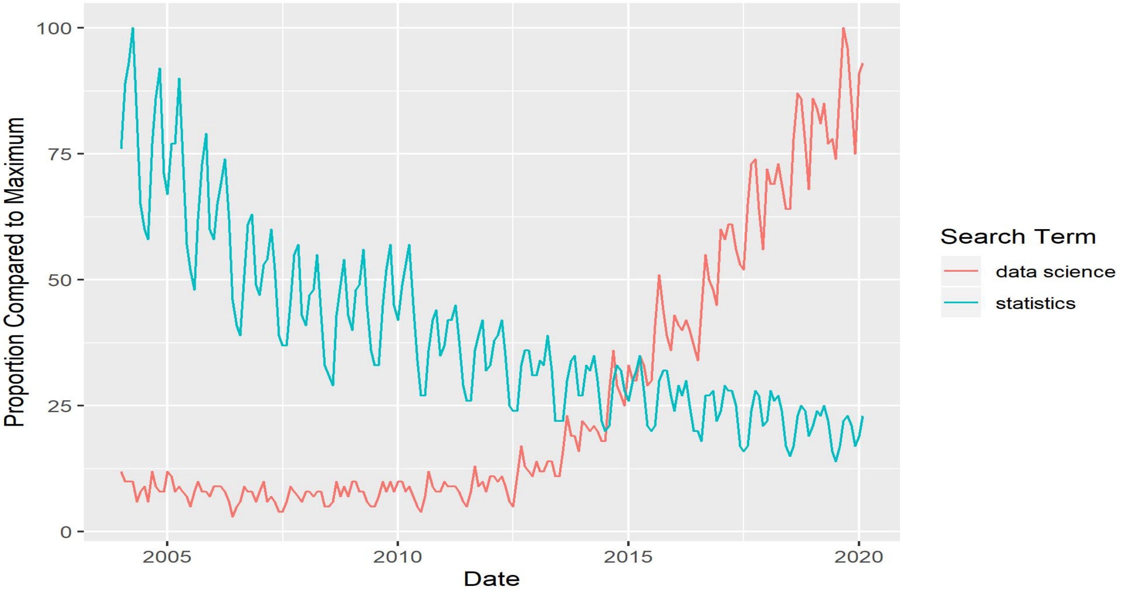

Figure 3.1: Google trends for search terms "data science" and "statistics" as of February 23, 2020. The y-axis represents search interest relative to the highest point on the chart between 2004 and 2020, where 100 is the peak popularity for the term.

While these disciplines have extensively outlined data science topics of potential relevance to researchers in their respective field, there currently is no understanding of how students learn concepts in data science, how these concepts build on each other, and what understandings foster or inhibit new learning. Research outlining a learning trajectory for data science concepts should be of utmost concern to the discipline of statistics education. The beginnings of this research can be seen as Alicia acquired the ability to filter rows of her data using a variety of tools, but lacked the ability to synthesize how each tool could solve a broader array of data tasks. Would Alicia's understanding of selecting columns have fostered her understanding of how to filter her data? Or could this understanding have been built from a fluency with data structures?

Statistics educators are in a position of great responsibility to communicate how to appropriately teach data science concepts. With data science growing in popularity, we need to remind researchers that statistics is more than data analysis. Rather, statistics, like data science, encompasses the entire process of “extracting value from data” (Wing, 2019). We hope to have provided environmental science researchers with an understanding and appreciation for the computing skills necessary for graduate students to implement statistics, while also emphasizing to statistics educators the importance of incorporating data science concepts into *every* statistics course.

REFERENCES CITED

Bibliography

Altadmri, A. and Brown, N. C. (2015). 37 million compilations: Investigating novice programming mistakes in large-scale student data. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, pages 522–527. ACM.

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). Use of R as a toolbox for mathematical statistics exploration. *Technology Innovations in Statistics Education*, 8(1):1–30.

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.

Bernard, H. R. (1988). *Research Methods in Cultural Anthropology*. Sage Publications, Inc., Newbury Park, California.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review*, 62(2):167–189.

Bilkstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, pages 110–116. ACM.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

Brown, E. N. and Kass, R. E. (2009). What is statistics? *The American Statistician*, 63:105–110.

Brown, N. C. C. and Altadmri, A. (2014). Investigating novice programming mistakes: Educator beliefs vs. student data. In *Proceedings of the 10th Annual Conference on International Computing Education Research*, pages 43–50. ACM.

Bryce, G. R., Gould, R., Notz, W. I., and Peck, R. L. (2001). Curriculum guidelines for bachelor of science degrees in statistical science. *The American Statistician*, 55(1):7–13.

Bulmer, J., Pinchbeck, A., and Hui, B. (2018). Visualizing code patterns in novice programmers. In *23rd Western Canadian Conference on Computing Education*. ACM.

Caceffo, R., Wolfman, S., Booth, K. S., and Azevedo, R. (2016). Developing a computer science concept inventory for introductory programming. In *Proceedings of the 47th ACM Technical Symposium on Computer Science Education*, pages 364–369. ACM.

Cannon, A., Hartlaub, B., Lock, R., Notz, W., and Parker, M. (2002). Guidelines for undergraduate minors and concentrations in statistical science. *Journal of Statistics Education*, 10(2).

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.

Chang, W. (2019). *R6: Encapsulated Classes with Reference Semantics*. R package version 2.4.0.

Cherenkova, Y., Zingaro, D., and Petersen, A. (2014). Identifying challenging CS1 concepts in a large problem dataset. In *Proceedings of the 47th ACM Technical Symposium on Computer Science Education*, pages 695–700. ACM.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1):21–26.

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.

Cortina, T. J. (2007). An introduction to computer science for non-majors using principles of computation. *SIGCSE Bull.*, 39(1):218–222.

Creswell, J. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications, Inc., 2nd edition.

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

Denzin, N. (1978). *Sociological Methods*. McGraw-Hill, New York.

Dodds, Z., Alvarado, C., Kuenning, G., and Libeskind-Hadas, R. (2007). Breadth-first CS 1 for scientists. *ACM SIGCSE Bulletin*, 39(3):23–27.

Dodds, Z., Libeskind-Hadas, R., Alvarado, C., and Kuenning, G. (2008). Evaluating a breadth-first CS 1 for scientists. *ACM SIGCSE Bulletin*, (1):266–270.

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.

Fox, J. A. and Ouellette, B. F. (2013). Education in computational biology today and tomorrow. *PLOS Computational Biology*, 9(12):1–2.

Friedman, J. (2001). The role of statistics in the data revolution. *International Statistics Review*, 69:5–10.

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.

Gutlerner, J. L. and Van Vactor, D. (2013). Catalyzing curriculum evolution in graduate science education. *Cell*, 153(4):731–736.

Hambrusch, S., Hoffman, C., Korb, J. T., Haugan, M., and Hosking, A. L. (2009). A multidisciplinary approach towards computational thinking for science majors. In *Proceedings of the 2009 SIGCSE*, pages 183–187. ACM.

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., Fernandez, 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., Temple Lang, D., and Ward, M. D. (2015). Data science in statistics curriculua: Preparing stuents 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.

He, X., Madigan, D., Yu, B., and Wellner, J. (2019). Statistics at a crossroads: Who is for the challenge. Technical report, The National Science Foundation.

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.

Higgins, J. J. (1999). Nonmathematical statistics: A new direction for the undergraduate discipline. *The American Statistician*, 53(1):1–6.

Hodder, I. (1994). The interpretation of documents and material culture. In Denzin, N. K. and Lincoln, Y. S., editors, *Handbook of qualitative research*, pages 393–402. Sage Publications, Inc., Thousand Oaks, California.

Horton, N. J., Brown, E. R., and Qian, L. (2004). Use of R as a toolbox for mathematical statistics exploration. *The American Statistician*, 58(4):343–357.

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.

Hristova, M., Misra, A., Rutter, M., and Mercuri, R. (2003). Identifying and correcting Java programming errors for introductory computer science students. In *Proceedings of the 34th ACM Technical Symposium on Computer Science Education*, pages 153–156. ACM.

Johnson, G. (2001). The world: In silica fertilization; all science is computer science. *New York Times*.

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

Jones, M. B., Schildhauer, M. P., Reichman, O., and Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):519–544.

Joppa, L. N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D., and Emmott, S. (2013). Troubling trends in scientific software use. *Science*, 340(6134):814–815.

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.

Keon-Woong, M. (2019). *ggiraphExtra: Make Interactive 'ggplot2'. Extension to 'ggplot2' and 'ggiraph'.* R package version 0.2.9.1.

Kitzes, J., Turek, D., and Deniz, F. (2018). *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*. University of California Press, Okland, CA. Available from: https://www.practicereproducibleresearch.org/.

Kross, S., Carchedi, N., Bauer, B., Grdina, G., Schouwenaars, F., and Wu, W. (2018). *swirl: Learn R, in R.* R package version 2.4.3.

Lahtinen, E., Ala-Mutka, K., and Jarvinen, H. M. (2005). A study of the difficulties of novice programmers. *ACM SIGCSE Bulletin*, 37(3):14–18.

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

Laney, C. M., Pennington, D. D., and Tweedie, C. E. (2015). Filling the gaps: sensor network use and data-sharing practices in ecological research. *Frontiers in Ecology and the Environment*, 13(7):363–368.

Levin, S. A., Grenfell, B., Hastings, A., and Perelson, A. S. (1977). Mathematical and computational challenges in popluation biology and ecosystems science. *Science*, 275(5298):334–343.

Lock, R., Lock, P. F., Lock Morgan, K., Lock, E. F., and Lock, D. F. (2013). *Unlocking the Power of Data*. Wiley, Hoboken, New Jersey.

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.

Merriam, S. B. (2009). *Qualitative research, a guide to design and implementation*. Jossey-Bass, 3rd edition.

Michonneau, F., Teal, T., Fournier, A. M., Seok, B., and Conrado, A. C. (2019). Data carpentry: Data analysis and visualization in R for ecologists.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative Data Analysis, An Expanded Sourcebook*. Sage Publications, Inc., Thousand Oaks, California.

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

Milne, I. and Rowe, G. (2002). Difficulties in learning and teaching programming views of students and tutors. *Education and Information Technologies*, 7(1):55–66.

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

Mokany, K., Ferrier, Simon amd Connolly, S. R., Dunstan, P. K., Fulton, E. A., Harfoot, M. B., Harwood, T. D., Richardson, A. J., Roxburgh, S. H., Scharlemann, J. P. W., Tittensor, D. P., Westcott, D. A., and Wintle, B. A. (2016). Integrating modelling of biodiversity composition and ecosystem function. *Oikos*, 125(1):10–19.

Moore, D. S., Cobb, G. W., Garfield, J., and Meeker, W. Q. (1995). Statistics education fin de siecle. *The American Statistician*, 49(3):250–260.

Moreno, J. L. (2002). Toward a statistically literate citizen: What statistics everyone should know. In *Proceedings of the 6th International Conference on Teaching Statistics*. IASE.

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.

Moustakas, C. (1994). *Phenomenological research methods*. Sage Publications, Inc.

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

National Research Council (1994). *Modern Interdisciplinary University Statistics Education: Proceedings of a Symposium*. The National Academies Press, Washington, DC.

Newman, H. B., Ellisman, M. H., and A., O. J. (2003). Data-intensive e-science frontier research. *Communications of the ACM*, 46(11):68–77.

Nolan, D. and Perrett, J. (2016). Teaching and learning data visualization: Ideas and assignments. *The American Statistician*, 70(3):260–269.

Nolan, D. and Speed, T. (2000). *Stat Labs: Mathematical Statistics Through Applications*. Springer, New York.

Nolan, D. and Speed, T. P. (1999). Teaching statistics theory through applications. *The American Statistician*, 53(4):370–375.

Nolan, D. and Temple Lang, D. (2010). Computing in the statistics curricula. *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.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Sage Publications, Inc., 3rd edition.

Peck, R. and Chance, B. (2005). Assessing effectiveness and the program level: Undergraduate statistics program evaluation. In *Proceedings of the 2005 Joint Statistics Meetings*. American Statistical Association.

Peters, D. and Okin, G. (2017). A toolkit for ecosystem ecologists in the time of big science. *Ecosystems*, 20:259–266.

Petre, M. and van der Hoek, A. (2016). *Software Design Decoded: 66 Ways Experts Think*. MIT Press.

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. R version 4.0.0.

Ramsey, F., modifications by Daniel W. Schafer, D. S., Sifneos, J., vignettes contributed by Nicholas Horton, B. A. T., Loi, L., Aloisio, K., Zhang, R., and with corrections by Randall Pruim (2019). *Sleuth3: Data Sets from Ramsey and Schafer's "Statistical Sleuth (3rd Ed)".* R package version 1.0-3.

Reid, N., Efron, B., and Morris, C. (2003). Is the math stat course obsolete?

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.

Rossman, A. J. and Chance, B. L. (2011). *Workshop Statistics.* Wiley, Hoboken, New Jersey.

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

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

Rubinstein, A. and Chor, B. (2014). Computational thinking in life science education. *PLOS Computational Biology*, 10(11):1–5.

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLOS Computational Biology*, 9(10):1–4.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R.* Springer, New York.

Schram, T. A. (2003). *Conceptualizing qualitative inquiry.* Merrill Prentice Hall, 3rd edition.

Smith, D. (2015). Vision and change in undergraduate biology education: Chronicling change, inspiring the future. Technical report, American Association for the Advancement of Science.

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

Stake, R. E. (2006). *Multiple Case Study Analysis.* The Guilford Press, New York.

Stefan, M. I., Gutlerner, J. L., Born, R. T., and Springer, M. (2015). The quantitative methods boot camp: Teaching quantitative thinking and computing skills to graduate students in the life sciences. *PLOS Computational Biology*, 11(4):1–12.

Strasser, C. A. and Hampton, S. E. (2012). The fractured lab notebook: Undergraduates and ecological data management training in the united states. *Ecosphere*, 3(12):1–18.

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):135–143.

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.

Tukey, J. (1962). The future of data analysis. *Annals of Statistics*, 33(1):1–67.

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, 57(2):74–79.

Van Manen, M. (1990). *Researching lived experience: Human science for an action sensitive pedagogy*. State University of New York.

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.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., and Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, 25(1):127–147.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59(10).

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. (2019). *Advanced R.* Chapman & Hall, Boca Raton, Florida, 2nd edition.

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

Wickham, H. and Grolemund, G. (2017). *R for Data Science.* O'Reilly, Sebastopol, California.

Wickham, H., Hester, J., and Chang, W. (2019). *devtools: Tools to Make Developing R Packages Easier.* R package version 2.2.1.

Wilkinson, L. (2005). *The Grammar of Graphics.* Springer, Hoboken, New Jersey, 2nd edition.

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. (2016). Software carpentry: lessons learned. *F1000 Research*, 3(62).

Wilson, G. (2019). *Teaching Tech Together: How to Make your lessons work and build a teaching community around them.* Chapman and Hall, Boca Raton, Florida.

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.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6):1–20.

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

Wing, J. M. (2019). The data life cycle. *Harvard Data Science Review*, 1(1).

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.

Worsley, M. and Blikstein, P. (2013). Programming pathways: A technique for analyzing novice programmers learning trajectories. In Lane, H., Yacef, K., Mostow, J., and Pavlik, P., editors, *Artificial Intelligence in Education.* AIED 2013.

Wright, T. and Zimmerman, N. (2016). Software carpentry: R for reproducible scientific analysis.

Yin, R. K. (2009). *Case Study Research: design and methods.* Sage Publications, Inc.

APPENDICES

APPENDIX A

STATISTICAL COMPUTING TASKS FROM CHAPTER TWO

We have data on fish caught in the Blackfoot River by Fish, Wildlife, & Parks personnel over a number of years. They used electrofishing equipment to attract the fish to the boat, then dipped them out of the water with nets, measured length in cm and weight in grams. They are often working in cold conditions in late autumn or early spring, so some measurement error is expected.

These data are not from a random sample. The goal is to catch all fish within a reach or section of the Blackfoot River every few years to assess the health of the population. Changes over years are important to the biologists.

The data were collected by making two trips per section (Johnsrud or Scotty Brown) each sampling year. The fish caught each trip of a given year, had their weight, length, and species recorded.

```
head(blackfoot)

##   trip length weight year  section species
## 1    1    288    175 1989 Johnsrud     RBT
## 2    1    288    190 1989 Johnsrud     RBT
## 3    1    285    245 1989 Johnsrud     RBT
## 4    1    322    275 1989 Johnsrud     RBT
## 5    1    312    300 1989 Johnsrud     RBT
## 6    1    363    380 1989 Johnsrud     RBT

summary(blackfoot)

##       trip         length          weight            year     
##  Min.   :1.0   Min.   : 16   Min.   :   0   Min.   :1989  
##  1st Qu.:1.0   1st Qu.:186   1st Qu.:  65   1st Qu.:1991  
##  Median :2.0   Median :250   Median : 150   Median :1996  
##  Mean   :1.5   Mean   :262   Mean   : 246   Mean   :1997  
##  3rd Qu.:2.0   3rd Qu.:330   3rd Qu.: 330   3rd Qu.:2002  
##  Max.   :2.0   Max.   :986   Max.   :4677   Max.   :2006  
##                              NA's   :1796                 
##    section            species         
##  Length:18352       Length:18352      
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
##                                       
##                                       
```

```
str(blackfoot)

## Observations: 18,352
## Variables: 6
## $ trip    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ length  <dbl> 288, 288, 285, 322, 312, 363, 269, 160, 213, ...
## $ weight  <dbl> 175, 190, 245, 275, 300, 380, 170, 40, 80, ...
## $ year    <dbl> 1989, 1989, 1989, 1989, 1989, 1989, 1989, ...
## $ section <chr> "Johnsrud", "Johnsrud", "Johnsrud", ...
## $ species <chr> "RBT", "RBT", "RBT", "RBT", "RBT", "RBT", ...
```

- What type of variable did `R` store `species` and `section` as? How would you change species and section to categorical variables?
- If the researchers were only interested in Rainbow trout and Brown trout, how would you remove Bull trout and WCT (whitefish) from the data set?
- Sometimes when sampling the fish, a technician fails to record one of the variables. How would you remove all the fish with missing values? How would this change if you instead removed the fish with only missing weight?
- The sampling methods used by Fish, Wildlife, & Parks on the Blackfoot River has changed over the years. In the years 1989 - 1996 they used gill nets and since 1996 they have used electrofishing. How would you create a new variable named `method` to reflect these different sampling methods used over the years?
- The researchers are interested in how many fish are caught each year that weigh over 1500 grams. How would you find these numbers to report?
- Which pairs of (weight, length) combinations seem difficult to believe? One way to look for unusual pairs is to use what fisheries biologists call a "condition index": $\frac{w^{1/3}}{l} \times 50$, where w = weight and l = length of the fish. If fish are highly unusual in this scale, it would be best to remove them, but you might need to compare only within species.
- How would you calculate each trout's condition number?
- How would you summarize these condition numbers for each of the two species of trout (Rainbow and Brown)?
- How would you plot the condition numbers of each trout, making sure to differentiate between Rainbow and Brown trout?

- The researchers are interested in trends in fish size over the sampling period (1989-2006). How would you create a visualization of fish lengths over the sampling period?
- Researchers are also interested in the number of fish from each species caught each year. How would you create a visualization of the number of fish caught from each species over the sampling period?

Lastly, the researchers are interested in trends in average fish weight over the sampling period. They want you to create a visualization of the average fish weight across years, differentiated by species of trout.

- First, you need to create a data frame of the mean weight of fish caught each year for the two species of trout. The end product should look something like the data frame below. How would you create this data frame of mean weights?

```
##   year species mean
## 1 1989   Brown  297
## 2 1989    Bull  429
## 3 1989     RBT  101
## 4 1989     WCT  120
## 5 1990   Brown  380
## 6 1990    Bull  422
```

- Next, to plot these mean weights for each year you need to transform the data from the current long format to wide format. This process is done by spreading the year variable across 10 different columns, one for each year (1989, 1990, etc.). The end product should look something like the data frame below. How would you transform these data from long format to wide format?

```
##   species 1989 1990 1991 1993 1996 1998 2000 2002 2004 2006
## 1   Brown  297  380  435  391  571  543  408  530  420  326
## 2     RBT  101  142  187  209  245  156  179  321  216  173
```

- There are additional data about the sections of the Blackfoot river for the sampling days each year. Researchers wish to merge these data (shown below) with the data on the fish caught during the sampling period. The `year`, `trip`, and `section` variables are keys that connect the two data sets. How would you merge these two data sets together?

```
head(water)

##   trip year      section temp water_level
## 1    1 1989 Scotty Brown 48.9        3.74
## 2    2 1989     Johnsrud 64.2        3.69
## 3    1 1990 Scotty Brown 53.9        3.37
## 4    2 1990     Johnsrud 65.3        3.69
## 5    1 1991 Scotty Brown 40.1        3.67
## 6    2 1991     Johnsrud 52.0        3.53
```