HOW ENVIRONMENTAL SCIENCE GRADUATE STUDENTS ACQUIRE STATISTICAL COMPUTING SKILLS

ALLISON THEOBOLD
Montana State University
allisontheobold@montana.edu

STACEY HANCOCK
Montana State University
stacey.hancock@montana.edu

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.

Keywords: *Statistics education research; Data science education; Environmental sciences*

1. 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, Bowles, Willig, & Waide, 2004; Green et al., 2005; Hampton et al., 2017; Hernandez, Mayernik, Murphy-Mariscal, & Allen, 2012; Mislan, Heer, & White, 2016; Teal et al., 2015).

Contrasted with graduate students in the biological sciences, where external structures often exist to support computational knowledge acquisition (Stefan, Gutlerner, Born, & Springer, 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. 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, 2017) 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.

2. 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.

2.1. 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 & Ouellette, 2013; Wing, 2006).

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

2.2. COMPUTATIONAL TRAINING FOR GRADUATE STUDENTS IN THE ENVIRONMENTAL SCIENCES

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, Lortie, Muenchen, Yang, & Ma, 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 & 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).

When considering the issue of curriculum re-evaluation, 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. Even with the inclusion of statistics coursework in the environmental science curriculum, graduate students continue to report that they are not developing these necessary skills.

2.3. 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 & 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 & 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 & 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.

3. 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.

3.1. PARTICIPANTS

At our university, the two-semester graduate-level Applied Statistics course sequence (GLAS I and II) is a service course for graduate students in scientific fields, with GLAS I only assuming 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 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.

Table 1. Academic demographics of participants: GLAS I indicates the academic semester they took the first semester graduate-level Applied Statistics course.

	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

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.

3.2. 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 in Appendix A and the statistical computing tasks are included in Appendix B.

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.

3.3. DATA ANALYSIS

The primary author led a three-stage data analysis process (Miles, Saladaña, & Huberman, 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 labor 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.

4. 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 sub-theme 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.

Table 2. Participants' themes in acquisition of statistical computing knowledge

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

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.

4.1. INDEPENDENT RESEARCH EXPERIENCE

The first theme in acquiring statistical computing knowledge was participation in independent research. Involvement in independent research projects 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 master's student, who attributed nearly all of her computational knowledge as having stemmed from the learning that occurred during 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.

4.2. SINGULAR CONSULTANT

When describing whom 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.

4.3. 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.

5. 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 trouble-shooting 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.

6. 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 & 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.

6.1. 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 & 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.

6.2. 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 from lack of relevant 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.

6.3. 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 & Chor, 2014; Wilson, Alvarado, Campbell, Landau, & Sedgewich, 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.

7. 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 & 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.

REFERENCES

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

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.

Cortina, T. (2007). An introduction to computer science for non-majors using principles of computation. *Proceedings of the 38th SIGCSE technical symposium on computer science education* (pp. 218–222). Covington, KY.

Creswell, J. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Fox, J. A., & 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.

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

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

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., . . . Aukema, J. E. (2017). Skills and knowledge for data-intensive environmental research. *BioScience*, *67*(6), 546–557.

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

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

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, *10*(1).
[Online: https://doi.org/10.1002/ecs2.2567]

Merriam, S. B. (2009). *Qualitative research, a guide to design and implementation*. San Francisco: Jossey-Bass.

Miles, M., Saladaña, J., & Huberman, A. (2014). *Qualitative data analysis: A methods sourcebook*. Thousand Oaks, CA: SAGE Publications.

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

Moustakas, C. (1994). *Phenomenological research methods*. Thousand Oaks, CA: SAGE Publications.

Nolan, D., & Temple Lang, D. (2010). Computing in the statistics curricula. *American Statistician*, *64*(2), 97–107.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
[Online: http://www.R-project.org/]

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

Schram, T. A. (2003). *Conceptualizing qualitative inquiry*. Upper Saddle River, NJ: Merrill Prentice Hall.

Stefan, M. I., Gutlerner, J. L., Born, R. T., & 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., & 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., & Pawlik, A. (2015). Data carpentry: Workshops to increase data literacy for researchers. *International Journal of Digital Curation*, *10*(1), 135–143.

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

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

Wilson, G., Alvarado, C., Campbell, J., Landau, R., & Sedgewich, R. (2008). *CS-1 for scientists. Proceedings of the 39th SIGCSE technical symposium on computer science education* (pp. 36–37). Portland, Oregon.

Wing, J. (2006). Computational thinking. Communications of the *ACM*, *49*(3), 33–35. [Online: https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf]

ALLISON THEOBOLD
Department of Mathematical Sciences
Montana State University
2-214 Wilson Hall
P.O. Box 172400
Bozeman, MT 59717 USA

APPENDIX A: INTERVIEW PROTOCOL

Where and how did you acquire this computing skill?

Why did you use this resource to acquire this skill?

Did you try other resources when learning this skill?

Have you used this resource to acquire other computational skills?

If yes, how often? If no, what other resources have you used to learn computing skills?

Where have you learned the computing skills necessary to implement statistics in your research?

APPENDIX B: STATISTICAL COMPUTING TASKS

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)

## 'data.frame':    18352 obs. of  6 variables:
##  $ trip   : int  1 1 1 1 1 1 1 1 1 ...
##  $ length : num  288 288 285 322 312 363 269 160 213 157 ...
##  $ weight : num  175 190 245 275 300 380 170 40 80 35 ...
##  $ year   : int  1989 1989 1989 1989 1989 1989 1989 1989 1989 1989 ...
##  $ section: chr  "Johnsrud" "Johnsrud" "Johnsrud" "Johnsrud" ...
##  $ species: chr  "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  379
## 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  379  434  391  571  543  407  530  419  325
## 2     RBT  101  141  186  208  244  156  179  320  215  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
```