Dissertation Proposal

Allison Theobold

Introduction & Motivation

With society's increased computational power and the volume and variety of available data, the practice of environmental science is changing. These changes have reshaped the vital computational requirements for researchers to successfully manage their data, create data visualizations, and analyze their data. However, in this changing climate, the curricula in environmental science remains stagnant. In 2012 a survey of graduate students in environmental sciences found that "over 80% of students reported that they had received no formal training in computing, even at the most basic level, and 74% stated that they had no skills in any programming language" (Hernandez, Meyernik, Murphy-Mariscal, & Allen, 2012).

Select undergraduate programs across the nation have begun to incorporate CS courses for non-majors into their life science curriculum. Many chemistry, biochemistry, and bioinformatic graduate programs have begun to incorporate computational training into their programs, however, a similar revolution, affirming the importance of computational proficiency, has yet to be experienced in the environmental sciences. However, in the environmental sciences, preparation in applied statistics is considered fundamental. Hence, statistics courses have been readily incorporated into graduate environmental science programs, with the implicit assumption that these courses would also introduce students to the computational skills necessary for their research.

Statistics educators across the nation have called for a re-evaluation of the statistics curriculum to incorporate computing, stating that the curriculum as it stands does not prepare students for the modern practice of statistics (Cobb, 2015; Lumley, 2001; Hampton et al., 2017; Nolan & Temple Lang, 2010). All arguments considered, there are barriers to incorporating computing that many institutions fall victim to (Hampton et al., 2017). The question is thus: if these students are not being taught these vital computational skills needed for them to successfully implement statistics within their research, where are they acquiring them?

Most, if not all, of "what environmental science researchers know about data management, visualization, and analysis has been learned piecemeal, through their lab, or not learned at all" (T. Teal et al., 2015). This communal process of acquiring the computation knowledge necessary for environmental science research lends itself well to the theoretical lens of distributed cognition. Distributed cognition, like any cognitive theory, investigates the organization of cognitive systems. However, unlike traditional cognitive theory, distributed cognition (1) considers the cognitive system beyond the individual, (2) includes interactions between individuals, and (3) incorporates resources and materials into an individual's environment (Hollan, Hutchins, & Kirsh, 2000). While social constructivism attends to the knowledge created through interactions between individuals, it does not attend to the tools and resources available to the individual in their acquisition of knowledge.

The theory of distributed cognition allows for us to seek to understand the relationships between elements in an individual's cognitive process. The literature on computational knowledge acquisition in the environmental sciences notes how an individual distributes the cognitive processes of learning computing to their social environment. Additionally, the literature acknowledges the variety of materials and resources available to students when seeking to acquire computational knowledge (Eglen, 2009; Hampton et al., 2017; Sedgewich & Wayne, 2008, 2015). Yet, the literature does not embrace the coordination required of an individual when balancing these internal (social) and external (resource) materials, in the process of acquiring the computational skills necessary for their research.

This study intends to understand how these individuals coordinate their environmental influences, such as their backgrounds, coursework, peers, and online resources, in order to acquire the computational skills necessary to implement statistics within their research. Additionally, I plan to investigate how an additional resource, workshops targeted towards key computational skills necessary for research in environmental science, can influence individuals' attainment of these necessary skills.

Research Questions

To address the aims outlined above, we propose three research questions:

1. What computing skills are necessary for environmental science graduate students to successfully implement applications of statistics in their research?
2. How are students filling the gap between the computing skills they know and the computing skills they need to know, in order to perform applications of statistics in their research?
3. How can workshops help to alleviate this gap between statistical computing preparation and expectations?

Methodology

In this section I outline the methodologies I propose to frame both aspects of my study. The first section addresses the methodology I plan to frame the study to develop workshops for key computational skills necessry for research in the environmental sciences. In the second section I frame my proposed theoretical lens from which I will interpret the findings of the study of computational knowledge acquision.

Workshops

This research proposes to situate itself within a classroom design study methodology. This design study aims to systematically study the development of computational skills, in the context of environmental science research, while also designing a means by which to support the development of these computational skills. In a design study I am able to both investigate how to support students' learning of computational skills, while also developing, testing, and revising my conjectures about this learning process and how to best support it.

This methodology is appropriate as it would be difficult to authentically observe these students acquiring the computational skills necessary for them to implement statistics in their research, as these skills are not necessarily acquired daily, monthly, or even yearly. Additionally, small literature base on the computational skills necessary for environmental science research establishes broad areas of importance, but does not adequately specify what specific computational skills are relevant for these disciplines. Additionally, there is a dearth of literature on the instructional methods that are best suited for teaching these key computational skills, in the context of research in environmental science.

For this study I will use the broad, foundational knowledge skills suggested by Hampton et al. (2017) to frame my workshops' learning goals. These foundational skills state that for data-intensive research in the environmental sciences, researchers should be able to

- work with messy or organized datasets, stored in varied data formats, in a reproducible workflow,
- employ a variety of statistical methods and simulation,
- make use of basic coding skills, and
- create effective data visualizations (Hampton et al., 2017, p. 547).

Each arm of these computational abilities coincides with a workshop currently implemented or in development to be administered through partnership with the Library. The current Introduction to `R` workshop gets researchers started on working with `R`, exploring the basics of: structure and organization of datasets, data summary techniques, and data visualization techniques. The current Intermediate `R` workshop guides researchers through the use of: conditional and relational statements for working with data, and looping and user-defined functions for repeated processes.

Based on my pilot study in spring 2017 and faculty interviews I performed in spring 2018, I was able to create starting points for what computational skills to begin teaching in these workshops, and how to structure the instruction of these computing skills. In the interviews with graduate environmental science students in Stat 512, where they reasoned through applications of statistical computing, it became clear that these students had misconceptions of fundamental concepts for working with data in `R`. Select participants stated that the computational knowledge they left their statistics courses with, was hardly the foundational knowledge necessary for understanding basic code. These participants abilities to explain the higher-level computational concepts they used in their independent research and inability to explain foundational concepts in `R`, lead me to designing my workshops to teach computing skills from foundational concepts.

In my interviews with environmental science faculty, I questioned them about what computational skills they believed are necessary for students to successfully perform research in the environmental sciences. As expected, answers varied between fields of research, however, many faculty emphasized students' ability to write their own functions, use conditional statements to process their data, looping and/or vectorization to automate their processes, database storage, and data manipulation. These "key skills" outlined by the environmental science faculty serve as the initial computing skills to integrate into these workshops.

Longitudinal Interviews

The cohort of "control" and "intervention" environmental science graduate students, recruited in spring 2018 and spring 2019 will be interviewed, as outlined in Section 3. These interviews, will be situated within the theoretical lens of distributed cognition. As I am seeking to understand what computational skills are necessary for research in environmental science fields, I am assuming students are acquiring these skills through use of their social environment (as seen in the pilot study) and utilization of resources, such as coursework and online or written tools. This study does not attend to the structure of these resources and how students interact with them, instead focusing on what the critical components of these environments are and under what situations students use them.

3 Data Collection

In this section I outline the data I propose to collect to address the research questions I have proposed. Each section contains the data I am planning to collect for each component of my study. The description of how the proposed data address my research goals is included in Section 4 (Data Analysis).

Workshops

The day prior to each workshop, the students, faculty, and staff that are registered are sent an email. This email requests that they complete the workshop pre-survey (Appendix), so that I am able to gauge everyone's area of study, computational background, why they are attending the workshop, and what they hope to get out of the workshop. During the workshop, applications of the computational concepts are given to participants for them to work through. Participants are requested to submit their answer to each of these questions during the workshop (through a Google Form). The day after each workshop, I send an email with any necessary information (e.g. `R` code from applications we did not get to), and a request for participants to complete a workshop post-survey (Appendix), which gauges their experiences with the workshop.

Workshop materials have been kept since their first implementation in fall 2017, however no recordings of the workshops were made during the 2017-2018 academic year. The Introduction to `R` and Intermediate `R` workshops were audio recorded, with one video recording of each workshop. I propose to create audio recordings for every workshop in spring and fall 2019. Video recordings could be considered, if the committee believes they would allow for a larger perspective on the workshop environment.

Control Cohort

The “control” cohort was recruited in spring of 2018 from the Statistics 511 courses. These students agreed to allow for me to follow them through their program of study, participating in interviews twice a semester. Approximately two weeks prior to each interview, participants will be requested to submit the most recent code they have generated for their research. During each interview, participants will be questioned about where they acquired the computational methods identified in their code.

Intervention Cohort

I propose to recruit an “intervention” cohort in spring 2019 to follow one year through graduate their program. I intend to recruit these individuals from the first year environmental science graduate students enrolled in Statistics 511, and through contacts with the Ecology, LRES, Plant Sciences, and Animal Range Sciences Departments. These students will be required to participate in all four workshops (Introduction to `R`, Intermediate `R`, Data Visualization, Data Wrangling), agree to an interview prior to the first workshop, and interviews following each workshop. These students will then be followed through the fall of 2019, with the same protocol as the “control” cohort (code submissions & computational knowledge acquisition interviews).

Computational Courses

Observations and content analysis of courses taken by a large proportion of environmental science graduate students, which teach aspects of computing, will be carried out (e.g. Stat 511 & 512, PSPP 516, WILD 401 & 501). Course materials, such as syllabi, labs, and online resources, from these courses will be solicited from the instructor or downloaded from course websites.

Data Collection Matrix

Research Question	**“Control“ Cohort**	**“Intervention“ Cohort**	**Computational Courses**
RQ 1: Skills	Code for Research	Code for Research	• Classroom Observations • Course Materials
RQ 2: Pathways	Research Interview	Research Interview	• Classroom Observations • Course Materials
RQ 3: Workshop Impact	NA	• Pre-interview • Post-workshop interviews • Research interviews	NA

Table 1: Data collected to address research questions, for control and intervention cohorts and common computational courses.

Research Question	***All* Workshop Participants**	**Workshop Design**
RQ 1: Skills	“What do you hope to learn from this workshop?"	• Demographics • Research Experience • Backgrounds • Post-workshop Survey (Material, Instruction, Improvement, Strengths)
RQ 2: Pathways	“Why did you choose to come to this workshop?"	• Workshop Materials • Audio & Video Recordings • Post-workshop Survey (Environment, Strengths, Improvement, Recommendation)
RQ 3: Workshop Impact	"I feel prepared to implement the concepts from this workshop in my own research."	• Post-workshop Survey (Strengths, Improvement, Recommendation)

Table 2: Data collected to address research questions, for every workshop participant and every workshop facilitation.

4 Data Analysis

In this section I outline how I plan to analyze the data I will collect, and how these data will answer my proposed research questions. The aim of this section is to explicitly outline how Tables 1 and 2 relate to my proposed study.

Workshops

The pre-workshop surveys can be used to assess the contingent aspects of each workshop. These surveys contain important background information that will aid in identifying what components of the workshop learning environment were potentially due to participants' backgrounds. Additionally, responses to "What do you hope to learn from this workshop?" can be used to inform research question 1, detailing the computational skills these participants feel are necessary for performing their (environmental science) research. Responses to "Why did you choose to come to this workshop?" help to delineate participants seeking computational skills for their coursework from those seeking skills for their research, as well as outline participants who are using the workshops as a resource for their research. Lastly, "What resources have you used while learning to program in R?" can help in identifying and reinforcing the dominant themes of pathways for computational knowledge acquisition, found in the control and intervention cohort interviews.

The submissions for each application of computational concepts help to outline any misunderstandings students are having when independently coding. These misunderstandings can be used to inform restructuring of the workshop, and to revise my proposed learning trajectory of computational skills. Finally, the post-workshop survey will provide guidance on both the environment of the workshop (number of facilitators, structure and location of workshop) and the structure of the workshop (material presented in an understandable way, confidence in implementation in research). Directly asking participant "What did you get out of this workshop?" highlights the aspects of the workshop that were successful for each participant, which can be used to note any aspects that could use improvement. Lastly, asking participants "What changes would you recommend for this workshop?", helps to revise workshop structure, content, and learning environment.

The changes to the workshops suggested by the assessments and post-workshop survey can be reinforced by the audio (and video) recordings of the sessions. These recordings can help to document the evolving workshop learning environment.

Control Cohort

This aspect of the study seeks to investigate the evolving nature of graduate students' computational tasks, the history of their computational tasks, and their relationship with other information-sharing resources (peers, consultants, advisers, courses, online resources, etc.). The historical comparison of computational abilities over time provides rich information about the evolving nature of these skills, and the resources students employ throughout their research experiences.

For the "control" cohort, I will have the ability to track each student's research code from each interview to the next. The first submission of research code will be inspected to identify the computational skills demonstrated by each student. These will be identified as skills pertaining to each of the four broad areas of learning goals outlined in the workshop learning goals (working with data of a variety of formats, employing a variety of statistical methods, basic software skills, and effective data visualizations). Each subsequent submission of research code will be inspected to identify any new computational skills employed.

During the interview, participants will be asked where they learned each of the computational skills identified in their research code, and their experiences employing these resources. The interviews will be open-coded to feature the dominant resources these students are utilizing in their acquisition of the computational skills necessary for their research, along with their justifications for employing these resources.

Intervention Cohort

The pre-workshop interview for the intervention cohort will act as a "baseline" to outline their backgrounds, computational experiences, and current working computational knowledge. Similar to the workshop design study, this will allow for me to both separate the contingent aspects of a student's learning from the necessary aspects, and track the development of each students' reasoning during the study. Following each workshop, I propose to interview each participant to detail the aspects of the workshop environment that helped promote learning, and outline how each successive form of their reasoning emerged as a reorganization of prior reasoning (e.g. how their understanding of material learned in the workshop build off of itself and/or other workshops). These interviews will allow me to test and improve my envisioned learning trajectory of how these computational understandings relate and build off of one another.

These students will then be followed through the fall of 2019, with the same protocol as the "control" cohort (code submissions & computational knowledge acquisition). The collection of research code and interviews, similar to the "control" cohort, allows for a comparison of the dominant resources these students are utilizing in their acquisition of the computational skills necessary for their research with that of their "control" counterparts. This comparison addresses the third proposed research question.

Computational Courses

Course observations will take place at the consent of the instructor, and are intended to identify the culture of the classroom and the classroom discourse surrounding the learning of computational skills. The analysis of each course's materials will be used to compare the computational understandings taught across different course instructors (Stat 511), and to outline what computational concepts are taught, and the context in which they are taught.

5 Study Aims

Outlining the key computational skills necessary for successful implementation of statistics to research in environmental science fields and the paths students employ when faced with computational challenges for their research, directly benefits the field of environmental science. These findings help to emphasize the importance of core skills for data-intensive environmental science research, helping to "facilitate the integration of training into the university" (Hampton et al., 2017, p. 555).

The workshop materials developed through this research will be publicly available through the Montana State University Library. The creation of these materials adds resources into these graduate students' environment, which help to enhance their learning of these core computational skills necessary for environmental science research. Importantly, these workshops have no barriers to entry, unlike the large amount of online resources that require a basic understanding of how to program in `R`. Additionally, each workshop builds off of the prior workshop(s), allowing researchers to have a guided progression through the computational skills necessary for environmental science research.

Finally, these workshops can help to guide Statistics and environmental science faculty in better integrating these skills into the curriculum. Materials for the workshops, which address these key computational skills, can be used by university faculty to pool resources for teaching computing. These materials are available in formats that can be quickly adapted and customized to the context of each instructor's course.

6 Timeline of Research Study

Semester	Project(s)
Spring 2017	• Environmental science graduate students in Statistics 512 were interviewed
Spring 2018	• Introduction to R & Intermediate R were advertised broadly to graduate environmental science students, through partnership with Library, with pre- and post-surveys administered • Environmental science faculty were interviewed about computational expectations • Environmental science graduate students in Statistics 511 were recruited for longitudinal study
Fall 2018	• Introduction to R & Intermediate R revised and advertised broadly to graduate environmental science students, through partnership with Library • First interviews with longitudinal environmental science graduate cohort • Recruit first year environmental science graduate students through workshop participation • Perform PSPP 516 classroom observations and course material analysis • Submit manuscript on computation knowledge acquisition strategies to SERJ
Spring 2019	• Revise and administer Introduction to R & Intermediate R workshops • Create and administer Data Visualization and Data Wrangling workshops, through partnership with Library • Continuing interviews with longitudinal environmental science graduate cohort • Perform WILD 401 and 501 observations and course material analysis • Begin draft of manuscript documenting gaps between computational abilities and expectations of graduate students in environmental science
Summer 2019	• Complete manuscript of gaps between computational abilities and expectations • Analyze research code provided by longitudinal cohort • Analyze Stat 511 and 512 materials from previous instructors
Fall 2019	• Revise and administer Introduction to R, Intermediate R, Data Visualization, and Data Wrangling workshops, through partnership with Library • Continuing interviews with longitudinal environmental science graduate cohort • Perform Stat 511 and 512 observations and course materials analysis • Draft paper on computational skills necessary for research in the environmental sciences
Spring 2020	• Finalize paper on computational skills necessary • Draft manuscript on impact of computing workshops • Defend dissertation

References

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.

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

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.

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

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196.

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

Sedgewich, R., & Wayne, K. (2008). *Introduction to programming in java.* Addison Wesley.

Sedgewich, R., & Wayne, K. (2015). *Introduction to programming in python.* Addison Wesley.

Teal, T., Cranston, K., 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.