A Pedagogical Approach to Create and Assess Domain-Specific Data Science Learning Materials in the Biomedical Sciences

Daniel Y. Chen

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Bioinformatics, and Computational Biology

Anne M. Brown, Chair

David M. Higdon

Alexandra L. Hanlon

Stephanie N. Lewis

December 14, 2021

Blacksburg, Virginia

Keywords: data science, data science education, pedagogy, medical education, biomedical sciences

A Pedagogical Approach to Create and Assess Domain-Specific Data Science Learning Materials in the Biomedical Sciences

Daniel Y. Chen

(ABSTRACT)

This dissertation explores creating a set of domain-specific learning materials for the biomedical sciences to meet the the educational gap in biomedical informatics, while also meeting the call for statisticians advocating for process improvements in other disciplines. Data science educational materials are plenty enough to become commodity. This provides the opportunity to create domain-specific learning materials to better motivate learning using real-world examples while also capturing intricacies of working with data in a specific domain. This dissertation shows how the use of persona methodologies can be combined with a backwards design approach of creating domain-specific learning materials. The work is divided into three (3) major steps: (1) create and validate a learner self-assessment survey that can identify learner personas by clustering. (2) combine the information from persona methodology with a backwards design approach using formative and summative assessments to curate, plan, and assess domain-specific data science workshop materials for short term and long term efficacy. (3) pilot and identify at how to manage real-time feedback within a data coding teaching session to drive better learner motivation and engagement. The key findings from this dissertation suggests using a structured framework to plan and curate learning materials is an effective way to identify key concepts in data science. However, just creating and teaching learning materials is not enough for long-term retention of knowledge. More effort for long-term lesson maintenance and long-term strategies for practice will help retain the concepts learned from live instruction. Finally, it is essential that we are careful

and purposeful in our content creation as to not overwhelm learners and to integrate their needs into the materials as a primary focus. Overall, this contributes to the growing need for data science education in the biomedical sciences to train future clinicians use and work with data and improve patient outcomes.

A Pedagogical Approach to Create and Assess Domain-Specific Data Science Learning Materials in the Biomedical Sciences

Daniel Y. Chen

(GENERAL AUDIENCE ABSTRACT)

Regardless of the field and domain you are in, we are all inundated with data. The more agency we can give individuals to work with data, the better equipped they will be to bring their own expertise to complex problems and work in multidisciplinary teams. There already exists a plethora of data science learning materials to help learners work with data; however, many are not domain-focused and can be overwhelming to new learners. By integrating in domain specificity to data science education, we hypothesize that we can help learners learn and retain knowledge by keeping them more engaged and motivated. This dissertation focuses on the domain of the biomedical sciences to use best practices on how to improve data science education and impact the field. Specifically, we explore how to address major gaps in data education in the biomedical field and create a set of domain-specific learning materials (e.g. workshops) for the biomedical sciences. We use best educational practices to curate these learning materials and assess how effective they are. This assessment was performed in three (3) major steps including: (1) identify who the learners are and what they already know in the context of using a programming language to work with data, (2) plan and curate a learning path for the learners and assessing materials created for short and long term effectiveness, and (3) pilot and identify at how to manage real-time feedback within a data coding teaching session to drive better learner motivation and engagement. The key findings from this dissertation suggests using a structured framework to plan and curate learning materials is an effective way to identify key concepts in data science. However, just

creating the materials and teaching them is not enough for long-term retention of knowledge. More effort for long-term lesson maintenance and long-term strategies for practice will help retain the concepts learned from live instruction. Finally, it is essential that we are careful and purposeful in our content creation as to not overwhelm learners and to integrate their needs into the materials as a primary focus. Overall, this contributes to the growing need for data science education in the biomedical sciences to train future clinicians use and work with data and improve patient outcomes.

Dedication

To my family, all the people who helped get me here, and all the great teachers I've had.

Acknowledgments

All chapters included in this dissertation document were written by the candidate. Dr. Anne M. Brown served as primary research advisor and provided editorial comments and suggestions for improvement all manuscripts for publication. The specifics of author contributions are listed below by chapter. Each author's initials are used to specify which author contributed each part. Author names used are as follows:

Daniel Y. Chen (DYC)

Anne M. Brown (AMB)

David M. Higdon (DMH)

Alexandra L. Hanlon (ALH)

Stephanie N. Lewis (SNL)

Chapters 2-4 DYC wrote these papers with input on content and corrections prior to publication from AMB. DYC performed a majority of the survey design, administration, and analysis. DMH, ALH, and SNL provided input during the preliminary phase. ALH and her biostatistics consulting group helped with additional data analysis questions. DYC and AMB collectively conceived the initial project idea and direction. DYC and AMB are responsible for submission of the completed manuscripts and response to reviewers. These chapters has not been published yet.

Contents

List of Figures

List of Tables

List of Abbreviations

AMA American Medical Association

ANA American Nursing Association

BLS Bureau of Labor Statistics

COPSS Committee of Presidents of Statistical Societies

CRAN Comprehensive R Archive Network

CSTA Computer Science Teachers Assoc

CSV Comma Separated Value

DSL Domain Specific Language

EDA Exploratory Data Analysis

FAIR Findability, Accessibility, Interoperability, and Reuse of digital assets

GAISE Guidelines for Assessment and Instruction in Statistics Education

IDE Integrated Development Enviornment

IRB Institutional Review Board

KA Knowledge Area

LO Learning Objective

MOOC Massive Open Online Courses

NIH National Institutes of Health

NNLM National Network of Libraries of Medicine

OHDSI Obserbational Health Data Sciences and Informatics

OMOP Observational Medical Outcomes Partnership

OSEMN Obtain, Scrub, Explore, Model, and i(N)terpret

PBC Public Benefit Corporation

PCK Pedagogical Content Knowledge

REPL The Importance Of Being Earnest

TIOBE User-Centered Design

UCD User-Centered Design

The AMA is a United Staties nationally recognized association that convenes state, speciality medical socites, and critical stakeholders to promote the art and science of medicince and the betterment of public health [2].

The ANA is a United States organization representing the interest of the nation's registered nurses with a goal of improving health care quality for all [3]

The BLS is a US government agency that serves as the principal fact-finding agency for the Federal Government in labor economics and statistics [4].

The COPSS is a prestigious Committee that comprises of the presidents, past presidents, and presidents-elect of several statistics and mathematics societies that are responsible for granting several awards, mainly the COPSS Presidents' Award for "an outstanding contribution to the profession of statistics", which is compared to the "Nobel Prize of Statistics".

CRAN is a network of servers around the world that synchronises and stores versions of code and documentation for the R programming langauge and its extension libraries [5].

The CSTA is a community of K-12 computer science teachers focused on supporting teachers. [6].

CSV files are plain text datasets where each row is a line in the file, and column values are separated by commas (,). This is a specific form of a delimited file, where the comma is used as the delimiter. Other delimiters, such as a tab character, for tab-separated value files are also common.

FAIR is a set of guiding principles for reusability of scholarly data with an emphasis on finding and using data [7].

A DSL is a subsection of a programming that is focused on one primary set of functions (i.e., domain).

EDA uses a combination of summary statistics, visualizations, and basic statistical models to get an understanding of the data to formulate a hypothesis before performing confirmatory data analysis.

The GAISE 2016 report describes a set recommendations to focus on what to teach in introductory statistics courses and how to teach the courses [8].

The IDE is a software tool that makes working with a particular programming language easier to use. It provides an environment where writing and developing code is integrated with useful graphical cools.

The IRB is an administrative body that reviews research proposals that involve human subjects to protect the rights and welfare of research participants.

KAs are topics that are a collection of topics that are related to other sub-domains.

An LO is a short and measureable statement of what a learner will be able to do at the end of a lesson or instructional period. Learning Objectives (plural) are abbreviated as LOs.

MOOCs are online corses that do not have a limitation to the class size and typically have its learning materials freely availiable via open access.

The NIH is a government agency in the United States primarily responsible for biomedical and public health research.

The NNLM is part of the United States Department of Health and Human Services with the goal of improving access to biomedical information to U.S. health professionals and the public [9].

OHDSI (pronounced "Odyssey") is a multi-stakeholder, interdisciplinary, open-science collaborative to bring out the value of health data through large-scale analytics [10].

OMOP is a common data model that allows for the systematic analysis of disparate observational databases [11].

OSEMN (pronounced "awesome") is the acronym in Hiliary Mason and Chris Wiggins's Snice taxonomy that defines the rough order of the data science process: obtain, scrub, explore, model, and i(n)terpret [12].

PBCs is a for-profit company designation that can afford legal protection to prioritize company values over shareholder returns.

Pedagogical content knowledge is the subject matter knowledge for the act of teaching. It includes knowing what are the most commonly taught topics in an area, and knowing what makes specific topics difficult to learn [13].

The REPL is used in interactive programming sessions where users are able to submit code (read) to be interpreted (evaluate) and the results are returned (print). This process then starts over where the programming language waits for the next command (loop). It allows the user to program interactively (as opposed to compiling the code first).

The TIOBE index is a measure of how popular a programming lanuage is. It is named after a play written by Oscar Wilde with the same name [14].

UCD puts effort into thinking about the end user's needs above the needs of the creator in a self-centered design [15, 16].

Chapter 1

Introduction

This dissertation describes the current state of education and pedagogy in the field of data science and explores ways to improve data science education in a specific sub-field, biomedical sciences. By critically appraising the field of data science education and identifying gaps in learning and pedagogical research, this work sought to determine research-backed solutions and approaches in data science education as applied to the domain of biomedical sciences. The main objective is to help future instructors better cater to novice data science learners by providing the tools and methods needed to identify their learner's needs and create more relevant learning materials for better engagement and long-term learning.

This discussion begins by orienting the reader with a broad introduction of the past, current, and future of data science education. This chapter describes the importance of delineating between data science education and computer science education. Additionally, this chapter describes the current gaps in data science education and data literacy in the field of biomedical sciences.

The impact, need, and considerations of creating domain-specific data science materials are included in this body of work. This dissertation is organized in the way you would go about creating a set of domain-specific data science learning materials. First, by identify who the learners are and what are their relevant prior experiences with data and programming; Section 1.3 introduces learner personas as a means to identify your learner audience and the process of creating and using personas are described in Chapter 2. Once our audience is identified, their background, relevant prior knowledge, perception of needs, and special considerations are used to create a set of learning materials. Section 1.4 introduces these learning materials, and assesses how effective the materials are to meet the learning objectives. The details of these learning materials and their effectiveness are described in Chapter 3. In section 1.4.1 we look more closely into how learning works, by looking at the formative assessment questions that are asked throughout the learning materials, and how they play a role in the final summative assessment question that aims to summarize the learning objectives. Details of this experiment are described in Chapter 4. Finally, Chapter 5 summarizes the impact of this work in data science education, and plans for the future in this domain of research and education.

1.1 History of Data Science

Statistics, data science, and computational programming education have co-evolved in the last half-century. While the term "data science" started to grow in popularity in 2014 (according to Google Trends), the ideas behind data science began in the 1960s, if not earlier, and the computing technology behind all the computation needed to create insights evolved in tandem.

1.1.1 Where the Term "Data Science" Originated

In 1962, John Tukey describes data analysis as "intrinsically an empirical science", and points to electronic computers as being vital to data analysis [17, 18]. This scientific approach to data analysis, with hypothesis driven exploration, was later called "Exploratory Data Analysis" (EDA). In 1977, Tukey calls for exploratory data analysis to be used along with confirmatory data analysis by understanding the data first before calculating any confirmatory statistic [19].

EDA becomes a core component of what later becomes "data science", which Peter Naur defines in 1974 as [18]:

> The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.

While data science as a term or field did not take on mainstream adoption as a specific term for the type of skills needed to work with data, statisticians realized that many of the kinds of insights they gather come from using computational tools [20, 21]. In 1997, C. F. Jeff Wu at the University of Michigan called for a rebranding of "statistics" and "statisticians" in favor of "data science" and "data scientists", respectively, since many other "good" names were already taken (e.g., computer science, information science, material science, cognitive science) [18].

The reliance on computers for statistical computing eventually led to John Chambers being awarded the ACM Software System Award in 1998, for creating the S system for graphics and data analysis (Section 1.1.2) [22]. By 2001, the field of "data science" was born with William Cleveland's plan at Bell Labs to expand the technical areas of statistics. Cleveland proposed six (6) technical areas of data science and how university resources should be allocated to expand educational and research offerings into the following fields or categories [23]:

1. Multidisciplinary Investigations (25%): data analysis collaborations in a collection of subject matter areas

2. Models and Methods for Data (20%): statistical models; methods of model building; methods of estimation and distribution based on probabilistic inference.

3. Computing with Data (15%): hardware systems; software systems; computational algorithms

4. Pedagogy (15%): curriculum planning and approaches to teaching for elementary school, secondary school, college, graduate school, continuing education, and corporate training

5. Tool Evaluation (5%): surveys of tools in use in practice, surveys of perceived needs for new tools, and studies of the processes for developing new tools

6. Theory (20%): foundations of data science; general approaches to models and methods, computing with data, teaching, and tool evaluation; mathematical investigations of models and methods, computing with data, teaching, and evaluation

Universities, along with, government research labs, and corporate research organizations have been traditional institutions for innovation. These resource allocations shape what is taught to new graduates to progress data science [23].

Academics typically use journal articles as a means to share knowledge. The push to have more Open Access journals, means research and knowledge no longer needs to be put behind a paywall. In 2002 the "Data Science Journal" was launched, and by 2003, the "Journal of Data Science" was launched. Both journals are open access and publish papers on data science education in addition to academic research. [24, 25].

Hal Varian, Google's Chief Economist, mentions in McKinsey Quarterly in 2009, computer engineers was the "sexy" job in the 1990s, statisticians would be the "sexy" job in the next 10 years [26]. Varian alludes to the need to visualize and learn from the ubiquitous amount of data and many of these skills will need to be transferred to managers who need to understand the data themselves. The "free and ubiquitous" data and the need to communicate findings will go beyond the professional level, and will cascade down to all level of education [26]. As data science becomes a more common term, in 2010, Hilary Mason and Chris Wiggins write "A Taxonomy of Data Science" [12] and Drew Conway creates "The Data Science Venn Diagram" (Figure 1.1) that aims to clarify what is data science and what skills are needed to be a competent data scientist [27]. Mason and Wiggins's "Snice" taxonomy (aka

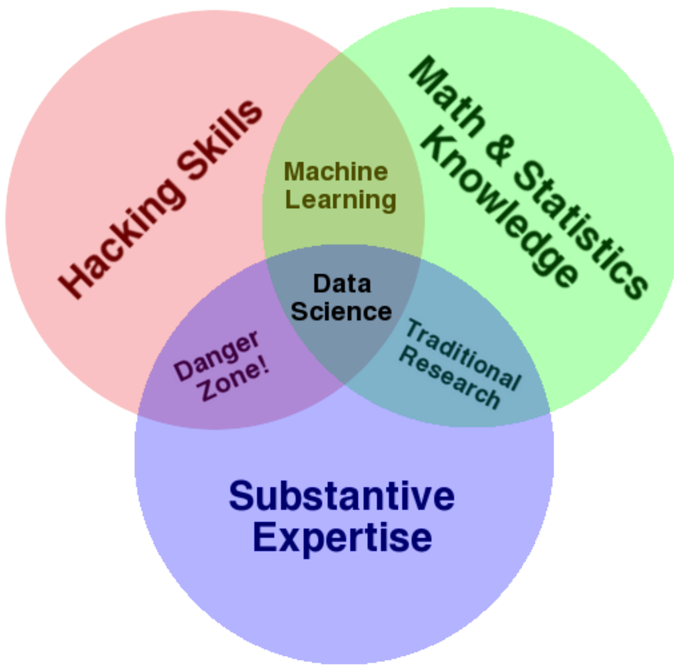


Figure 1.1: Reproduction of Drew Conway's Data Science Venn Diagram [27].

OSEMN, pronounced "awesome") states that a data scientist, in rough chronological order, (o)btains, (s)crubs, (e)xplores, (m)odels, and i(n)terprets data [12]. In 2012, Tom Davenport and D.J. Patil publish "Data Scientist: The Sexiest Job of the 21st Century" in the Harvard Business Review, which talks about the types of insights data science can bring, while also describing the multitudes of skills a data scientist needs to incorporate into analysis, mainly around working with unstructured data to create and present an analysis [28]. Davenport and Patil also talk about the high cost and scarcity of data scientists, how businesses must balance the need to stay competitive with the nonstop flow of data, and waiting for the next wave of talent that is more accessible due to a data scientist's rarefied skills become taught in classes [28]. Patil is later named the First U.S. Chief Data Scientist in 2015 [29]. Meanwhile, while data science has started to impact the growth of open source tools in the academic, scientific, and research domains [30–32] Arfon Smith, Kyle Niemeyer, Dan Katz, Kevin Moerman, and Karthik Ram start the Journal of Open Source Software in 2016 [33] to help academics performing data science research and creating tools for data science get credit for their work and contributions to the open source ecosystem. By 2018, The National Institute of Health (NIH) released its first Strategic Plan for Data Science that provides a road map for the biomedical data science ecosystem [34].

At the time of writing, there is an abundance of data science learning materials [35]. Massive online open courses (MOOCs) have utilized the internet to deliver free data science courses, and "boodcamp" programs filled the immediate training needs (e.g., General Assembly established in 2011). Between EdX, MIT Opencourseware, Khan Academy, Coursera, there are 100s of free data sciences related classes since the early 2000s [36–38]. Publishing tools like Bookdown and JupyterBook also catalogue 100s of free online book resources [39, 40].

While "data science" may be a relatively new term, it may evolve just like the evolution of "statistics", "machine learning", and "artificial intelligence" in mainstream terminology. But, the core skills of obtaining, cleaning, analyzing, and communicating data insights will remain the same. These skills will become more prevalent in fields traditionally not associated with computation, and making the tools and skills more accessible are going to be key ventures moving forward.

1.1.2 Common Tools in Data Science History: R

At the time of writing (November 2021) R ranks #15 on the TIOBE index, a measure of the popularity of a language [14]. R is one of the premier languages used for statistical computing and graphics, that was the successor to the S programming language [41].

John Chambers, Rick Becker, Doug Dunn, Jean McRae, and Judy Schilling implement the S language in the statistics research department at Bell Laboratories in 1976. It was the first implemented statistical computing language [42]. The need for a statistical computing language grew from the necessity of interacting with data using Exploratory Data Analysis (EDA) techniques and graphical output to work with larger data sets and iterate faster [42]. In 1988, a commercial version of S was implemented, S-PLUS, and was succeeded by R, which was in development in 1991 by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and later open sourced in 1995 [42]. The Comprehensive R Archive Network (CRAN) was founded in 1997 by the R Core team as a means to serve as a software repository for users to submit packages that can be used by other R users [43]. R's first stable v1.0 version was released in 2000 [44, 45].

The release of additional data manipulation (reshape v0.4 in 2005 and plyr v0.1.1 in 2008) and visualization (ggplot v0.2.2 in 2006 and its successor, ggplot2 v0.5.1 in 2007) tools in CRAN by Hadley Wickham focused R in exploratory data analysis [19, 46]. Joseph J. Allaire founds RStudio in 2009 [47, 48].

Wickham believed a lot of the work needed to explore and understand data was being overlooked by statisticians [49]:

> ... The fact that data science exists as a field is a colossal failure of statistics. To me, that is what statistics is all about. It is gaining insight from data using modelling [sic] and visualization. Data munging and manipulation is hard and

statistics has just said that's not our domain.

He continued to improve and create a domain-specific language (DSL) in R focused on the data science workflow (reshape2 v1.0 in 2010, dplyr v0.1.1 and tidyr v0.1 2014). During this time, RStudio Inc. works on developing the RStudio integrated development environment (IDE) and is released in 2011 to make interacting and programming in the R programming language much easier for users and also provides tools to make exploratory data analysis more readily available. In 2012, RStudio Inc. releases shiny, a dashboard framework that can be written in R which lowers the barrier of entry to create, interact, analyze, and publish graphics and data on the web. These ideas around "tidy data" principles for exploratory data analysis [50] cumulated in the release of the tidyverse in 2016, which hosts a multitude of libraries for the data science DSL in R. For this work and "influential work in statistical computing, visualization, graphics, and data analysi" and "making statistical thinking and computing accessible to a large audience", Wickham was awarded the international COPSS Presidents' Award in 2019 [51].

Separately, Max Kuhn began to unify the various modeling packages in R with the caret (first release in CRAN in 2007 as v2.27) package and was to be superseded by the tidymodels package (first released in CRAN in 2018 as v0.0.1) for its consideration of Tidyverse principles that where objects share a common philosophy, grammar, and data structure to make things more interoperable in data and analysis pipelines [52, 53]. In 2020, RStudio Inc announced that they have become a Public Benefit Corporation (PBC) [47, 48], and RStudio PBC employees develop, maintain, and advocate for a plethora of R packages. The Tidyverse set of packages are not dependencies of many other R packages in the ecosystem. These packages are also one of the main ways R is taught to new learners [52].

1.1.3 Common Tools in Data Science History: Python and the PyData Ecosystem

At the time of writing (November 2021) Python ranks #1 on the TIOBE index of programming language popularity [14]. It is a general purpose programming language that is being taught to learners of all ages. Along with the R programming language, they are the lingua franca of programming languages in data science.

Python started as a project by Guido van Rossum for developers to write in an interactive high-level programming language that can talk to the underlying operating system but still able to interact with native C libraries [54]. Python was first released in 1991 and v1.0 was released in 1994 [54]. As Python gained in popularity, it started to make its way into the scientific research community. In 1999, the multipack library was released to bring scientific computation to the language by wrapping around C and Fortan libraries. As the scientific community began to grow, The SciPy community and Python package was created in 2001 by Eric jones, Travis Oliphant, and Enthought, Inc. However, the scale of the scipy project was too much to maintain and spawned the creation of smaller scikit projects. During that time, a more interactive Read-Evaluate-Print-Loop (REPL), IPython, similar to other mathematical and programming software like Wolfram Mathematica, was created by Fernando Pérez to make scientific computing and exploration easier [55]. Since much of scientific computation relied on matrix and array objects, the scientific community was slowly being split between the numeric and numarray libraries. In 2005, Travis Oliphant started creating the numpy library to keep the array objects in Python cohesive, since it was used by scipy. In 2006, numpy was released. Since then, many other libraries have been built on top of the NumPy array: theano in 2008, pandas in 2009, and scikit −learn in 2010. These libraries have been paramount to the growth of Python in the data science space [56].

As a means to help sustain and support these critical open source libraries, NumFOCUS was founded in 2012 as a 501(c)(3) public charity status as a nonprofit in the United States along with the first PyData conference series, the educational program of NumFOCUS centered around community organization. Travis Oliphant and Peter Wang also found Anaconda Inc that year with the goal of supporting the open-source Python community and helping the SciPy ecosystem scale to "big data" [56]. In 2013, Anaconda released the blaze package to help scale the SciPy ecosystem. The next iteration of scientific computing interactivity came with Project Jupyter in 2014 with the goal of bringing the interactivity of IPython to other programming languages and with a common user interface, Jupyter Notebooks (Jupyter stands for the Julia, Python, and R programming languages). The dask project spins off of blaze in 2015 with a focused on making multi-core parallelization and distributed computing easier in the SciPy ecosystem.

As "artificial intelligence" gains popularity with the resurgence of neural networks, tools like TensorFlow was released by Google in 2015, and PyTorch was released by Facebook in 2016, both with Python as one of the primary languages. Many of these tools also have bindings into the R programming language as well. As more computation happens on a browser interface, and the success of Jupyter Notebooks and its extensive widget system, Jupyter Lab was released in 2017 as the next iteration of notebook interfaces in the browser. Travis Oliphant leaves Anaconda Inc in 2018 to start QuanSight to help sustain the open-source PyData ecosystem by connecting community members with companies that use the technologies.

1.1.4 Emerging Technologies

The reach of data science goes beyond any single domain and technology stack. The Julia language (Rank #36 on the TIOBE index in November 2021) was released in 2012 with the goal of being a high-level and fast programming language, especially around numerical computing. It was incorporated in the Jupyter ecosystem as a way to unify computational languages into a single development environment for data science tasks. Julia's dataframes.jl library was released in 2012, giving the same data manipulating features as the R dataframe object and Python's pandas library.

Representing a dataframe object across multiple programming languages is a challenge because of how array objects are implement from language to language. The Apache Arrow project aims to create a centralized API for dataframes that can be used across multiple languages, allowing end users to use whatever tool best suits their needs, but using a unified and performant data structure [57]. Groups like Ursa Labs were formed in 2018 (now Voltron Data in 2021) to help provide more support to the Apache Arrow project [58, 59].

The popularity of Javascript over the years (Rank #7 on the TIOBE index in November 2021) stem from it connection with end user interactions with virtually all websites on the internet. It has become a tool for creating interactive figures (e.g., D3.js) for people on the internet. Since essentially every person who connects to the internet uses a web browser, WebAssembly was announced in 2015 as an open standard for any programming language to compile down for web applications to run in a browser [60].

Projects like Apache Arrow and WebAssembly (Wasm) are blurring the lines between tools needed for data science tasks. Apache Arrow is unifying data analysis by providing a single unified memory format to use data for analysis [57]. The project has bindings for many programming languages in an effort to make data programming language independent. These tools are paving the way for end results to not rely on particular programming languages.

1.1.5 Communities in Data Science History: The Carpentries

Programming is not a taught in many academic disciplines. But as research relied more on computational tools, a growing need for teaching and learning these computational tools in academia was needed.

Software Carpentry was founded in 1998 by Greg Wilson and Brent Gorda with the goal of teaching researchers the computing skills to get their work done more efficiently and effectively. These workshops are primarily focused on general programming skills for scientific computation, and were made open source in 2005 with support from the Python Software Foundation (PSF). Software Carpentry continues to grow with the support from the Alfred P. Sloan Foundation and Mozilla Science Lab in 2012 [61, 62].

The following year, 2013, the first workshops geared towards Librarians were run. By 2014, Data Carpentry is founded by Karen Cranston, Hilmar Lapp, Tracy Teal, and Ethan White with support from the National Science Foundation (NSF) with the goal of creating materials focused on data literacy aimed at novices in specific research domains. James Baker is able to expand on Library Carpentry with support from the Software Sustainability Institute (SSI) with the goal of teaching information sciences and best practices in data structures. Software Carpentry Foundation is also founded under NumFOCUS in 2014. In 2015, Data Carpentry gets support from the Gordon and Betty Moore Foundation and in 2018, with the help of Community Initiatives, Software Carpentry, Data Carpentry, and Library Carpentry merge together to form The Carpentries. From 2012 to 2020, The Carpentries have run 2,700 workshops across 71 countries and have touched at least 66,000 novice learners [61, 62].

Today, the Carpentries support over 50 lessons across their 3 Lesson Programs (Data Carpentry, Library Carpentry, and Software Carpentry), with over 150 lesson maintainers [63]. These lessons cover all the basic data literacy, data management, data science, and software programming skills in specific curricula domains (Figure 1.2). Instead of having one-off lesson materials maintained by a small set of authors, The Carpentries lessons are kept up-to-date with a rotating set of maintainers for each lesson, and leverages the broader community to upkeep the teaching infrastructure. This serves as an efficient way to teach workshops, maintain lessons, and train new instructors.

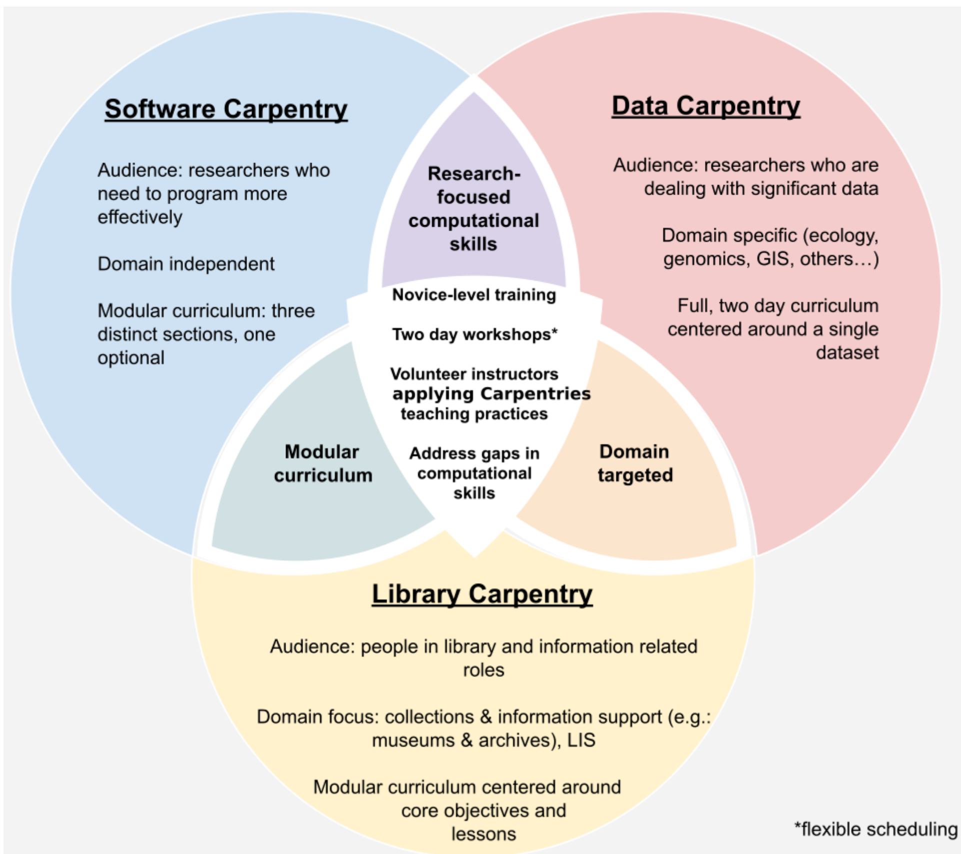


Figure 1.2: The differences between the 3 Carpentries lesson programs: Data Carpentry, Library Carpentry, and Software Carpentry. Data Carpentry focuses more on researchers who work with data in a specific domain, Library Carpentry focuses more on programming tasks in the library and information sciences, and Software Carpentry focuses more on programming concepts. Figure adapted from the Carpentries Trainer Training lesson [64].

1.2 Data Science Education and Pedagogy Research

Computing and statistics education have both created their own sets of higher education curriculum guidelines. Many of these concepts have made their way down from higher education to K-12, suggesting the top-down need for learning data science skills.

1.2.1 Computing Education (Higher Education)

The original Computing Curricula 2005 Overview Report (for undergraduate education) addressed five (5) computing disciplines, which include: [65]: (1) computer engineering, (2) computer science, (3) information systems, (4) information technology, and (5) software engineering. Along with five (5) computational skills [65, 66]: (1) organizational system issues, (2) application technologies, (3) software development, (4) systems infrastructure, and (5) computer hardware and architecture. The term "computing" is generally used to encompass all these various sub-fields that have varying understandings and context [65, 66].

The interaction between computing disciplines and computational skills is reproduced in Figure 1.3, and was highly regarded in computing educational circles for depicting the relationships between the what computation skill is required across the different computing disciplines [65, 66]. The figure (Figure 1.3) suggests that different computing disciplines require varying amount of computing skills, and as computing disciplines grow, new domains were added: cybersecurity in 2017, and data science in 2021 [67].

In addition to computing discipline changes, the learning frameworks on how to teach these disciplines have also changed from knowledge-based in 2005 to competency-based in 2020 [66]. These learning framework changes stemmed from the discrepancy between skills taught in school that focus on individual tasks and skills needed for more complex real-world tasks

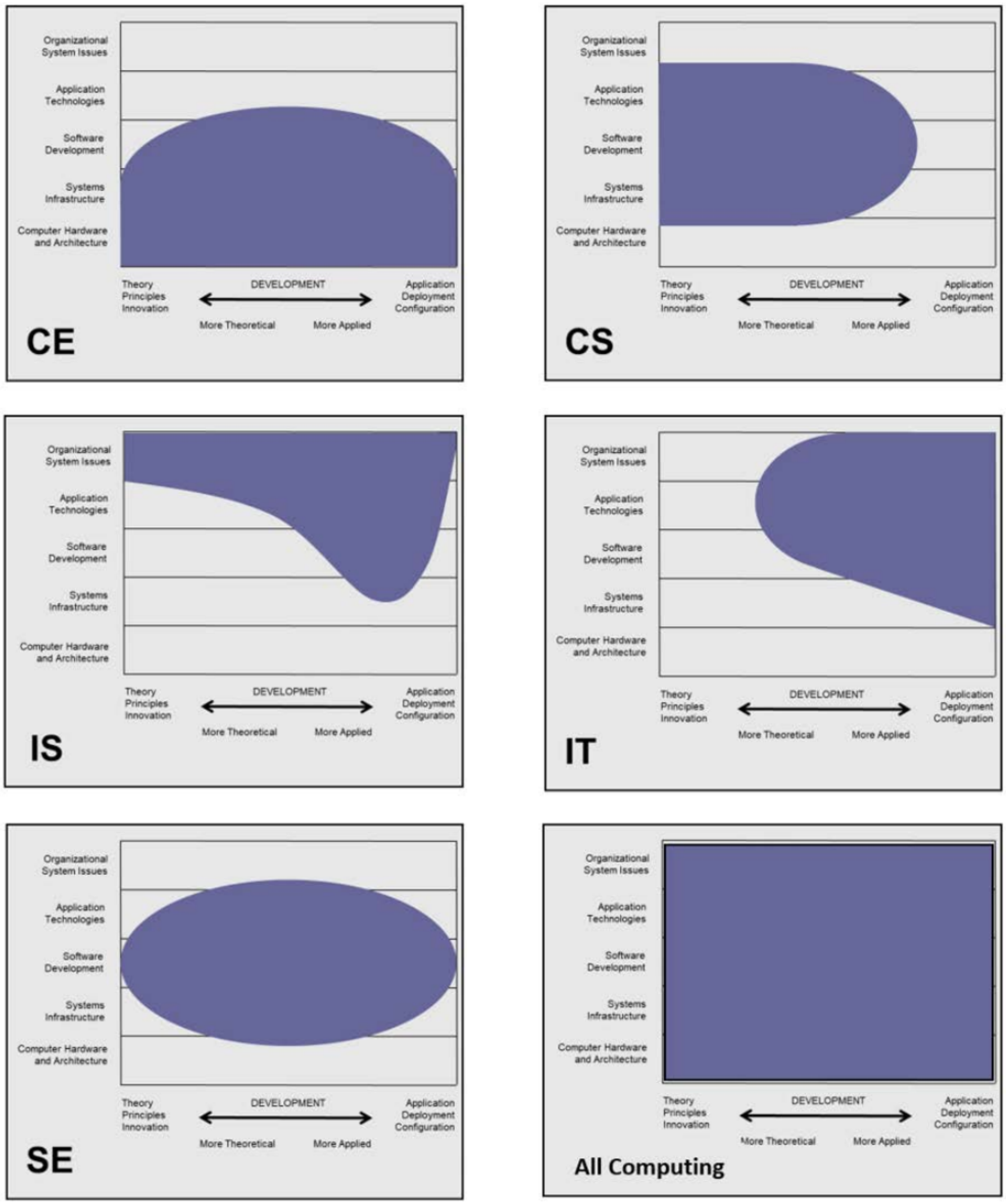


Figure 1.3: The breakdown of how each of the 5 computational skills are utilized across each of the 5 computing disciplines. The 5 computational skills are organizational system issues, application technologies, software development, systems infrastructure, and computer hardware and architecture. The 5 computational disciplines are computer engineering, computer science, information systems, informational technology, and software engineering.

[66]. The new computing curriculum guidelines released in 2020 are now as follows:

1. Focusing on competency
2. Transitioning from knowledge-based learning to competency-based learning
3. Expanding curricular disciplines to include cybersecurity as well as data science
4. Expanding curricular and competency diagrams and visualizations
5. Establishing an interactive website that will bring CC2020 results to public use
6. Charting a framework for future computing curricular activities

This dissertation explores computing skills required for data science, and how to build long-term competency for domain experts. The work also seeks to identify core data science competencies, and how to meet the existing skills gap for individuals who are not enrolled in a course program to learn these competencies. It aims to accomplish these tasks by testing and creating a framework for domain-specific set of materials in data science. Although, domain specific data science materials exist (e.g., The Carpentries), this dissertation will provide validated assessment tools for future educators to identify their learner's needs. Specifically, this dissertation focuses on medical practitioners and biomedical science researchers.

1.2.2 Computing Education (K-12)

The Computer Science Teachers Association (CSTA) is a professional association aimed to support computer science teachers in K-12. Similar to the Computing Curriculum reports, the CSTA also has guidelines and standards for K-12 curriculum and teachers in computer science (Figure 1.4) [68]. The Association and their guidelines and standards suggest the growing need for computing skills from industry down to higher education to the K-12 system.

Figure 1.4: The 5 goals for effective K-12 computer science teachers. This figure is reproduced from the CSTA standards for CS teachers. Since CS teachers come from many different backgrounds, this figure lists the tasks to orient teachers to continuously refine their pedagogical content knowledge (PCK). These goals help the teacher support the students in meeting learning outcomes. Many of these goals apply to teaching computational skills as well.

The CSTA standards are incorporating the need to understand computing literacy in K-12 education. The learning objectives for the standards fall into 5 main categories and have a progression (1A, 1B, 2, 3A and 3B) for students as they get older: (1) Computing Systems, (2) Networks and the Internet, (3) Data and Analysis, (4) Algorithms and Programming, and (5) Impacts of Computing.

The CSTA learning standards align with the Computation Curriculum guidelines and its new focus around complex real-world jobs in undergraduate education with the CSTA guidelines in K-12 education in Data and Analysis, specifically, the CSTA 2-DA-08 standard for data analysis, "collect data using computational tools and transform the data to make it more useful and reliable" [6, 66, 68].

This dissertation builds on the multiple computing reports by acknowledging data science as a separate computing discipline, and requires a different and adjacent set of computational skills from other computing disciplines, and aims to apply pedagogical best practices to effectively teach data science to specific domains (e.g., medical and biomedical sciences). This work focuses on data literacy concepts as a fundamental piece in data science education since many of the tools are built around these data concepts.

1.2.3 Statistics Education (Higher Education)

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) report in 2016 lists recommendations of what and how to teach in statistics [8]. Specifically, the college report recommendations applies to the different variations of introductory statistics courses, while allowing flexibility for courses to implement specific needs. The GAISE 2016 report provides six (6) guidelines:

1. Teach statistical thinking
   (a) Teach statistics as an investigative process of problem-solving and decision-making
   (b) Give students experience with multi-variable thinking
2. Focus on conceptual understanding
3. Integrate real data with a context and a purpose
4. Foster active learning

5. Use technology to explore concepts and analyze data

6. Use assessments to improve and evaluate student learning

Introductory statistics courses differ from one classroom to another. Some courses address statistical literacy, while others focus on statistical methods. The difference stems from the course's focus around consumers of statistical analyses or producers of statistical analyses [8]. This dissertation builds on this differentiation between practitioners and producers of tools by creating learning materials for the former group of learners. It also incorporates the GAISE 2016 recommendations through data literacy concepts, where learners work on processing data to answer their research question though exploratory data analysis.

One of the main challenges with introductory (statistics) courses is catering to a wide audience [8]. Depending on whether a particular course is geared towards a general audience, or a more specific audience (e.g., life sciences, business, engineering, mathematics, etc), different prerequisites are required (e.g., some require calculus, others only need high school algebra). Class sizes and classroom formats taught synchronously and asynchronously also vary (virtual, in-person, Massive Open Online Courses - MOOCs, etc) [8]. This dissertation focuses on a more specific domain (e.g., medical and biomedical sciences), which to aids in learning and motivation [35, 69, 70].

Things To Cut in Statistics Curriculum

One of the more difficult challenges when designing learning materials is fitting the content within a time constraint and balancing adding more content to the learning materials. The GAISE 2016 report has suggestions for topics that could be omitted from introductory statistics courses: (1) Probability theory; (2) Constructing plots by hand; (3) Basic statistics; (4) Drills with z-, t-, 2, and F-tables; (5) Advanced training on a statistical software program (e.g., SAS certification, non-introductory R programming, other more extensive programming topics). The workshop materials created for this dissertation focuses around more practical techniques and skill, over more theoretical ones. It tries to introduce basic data literacy concepts as it pertains to performing statistical analyses. The examples and topics from the workshop materials also serve as a framework on how to prepare data for analysis and other types of statistical tests.

1.2.4 Statistics Education (Pre-K–12)

In 2020, the GAISE II report was released for Pre-K-12 statistics as an updated to the original 2005 report [71]. The new report 2020 report calls for working with data in more non-traditional and multivariable data in many different contexts for Pre-K-12 education, as opposed to working with data in a more static spreadsheet formats [71]. The original problem-solving process (Figure 1.5) of (1) formulating a statistical investigative question, (2) collecting or considering data, (3) analyzing data, and (4) interpreting results sill remain for elementary, middle, and high school (i.e., levels A, B, and C), but the new report adds six (6) more enhancements to account for the growing amount of data and their uses in today's digital age [71]:

1. Questioning through the statistical problem-solving process
2. Different data and variable types
3. Multivariable thinking throughout Levels A, B, and C
4. Probabilistic thinking throughout Levels A, B, and C
5. The role of technology in statistics and how it develops through the Levels
6. Assessment items that measure statistical reasoning

The emphasis around incorporating non-traditional data sources (e.g., GPS coordinates from a fitness application) for analysis and interpretation hopes to give younger learners an earlier explore to the data science process [71].

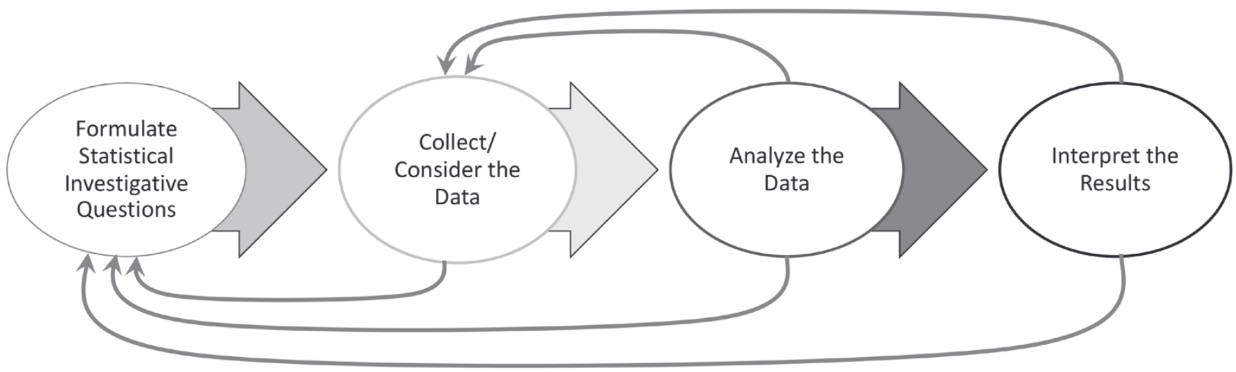


Figure 1.5: Reproduction of the GAISE II Pre-K-12 statistical problem-solving process [71].

1.2.5 Data Science Education (Higher Education)

The permeation of data science into fields other than statistics and into other fields, show how valuable it is for understanding the world [23]. The spike in demand for data science skills in industry led way to the creation of "Bootcamps". These filled the initial void of learning data science skills, while the educational system caught up to train new graduates for the work environment. followed by combining departments and programs in higher education [35], and impacting K-12 education standards [6, 71]. The Bureau of Labor Statistics (BLS) predicts that by 2030 computational employment will increase by 13% in the United States, faster than other occupations [66, 72].

Data science is a set of fundamental concepts that involves principles, processes, and techniques to gain knowledge from data, typically using programming as the tool. [66, 67, 73]. It is a sub-domain of computing education that is adjacent to data analytics and data engineering. There are 11 other topical sub-domains that are incorporated into the core data science knowledge areas (KAs) [66, 67]: (1) Analysis and Presentation (AP), (2) Artificial Intelligence (AI), (3) Big Data Systems (BDS), (4) Computing and Computer Fundamentals (CCF), (5) Data Acquisition, Management, and Governance (DG), (6) Data Mining (DM),

(7) Data Privacy, Security, Integrity, and Analysis for Security (DP), (8) Machine Learning (ML), (9) Professionalism (PR), (10) Programming, Data Structures, and Algorithms (PDA), and (11) Software Development and Maintenance (SDM). Additionally, a full curriculum should be augmented with six (6) fundamental courses in mathematics, statistics, and computer science [66, 67]: (1) Calculus, (2) Discrete structures, (3) Probability theory, (4) Elementary statistics, (5) Advanced topics in statistics, and (6) linear algebra.

These suggestions pertain to data science degree programs in higher education [66, 67], however, this dissertation focuses on introductory materials for working professionals in the biomedical sciences (e.g., clinicians, analysts, academics). There is simply not enough time to cover all the topics as suggested in the Computing Curriculum 2020 report and the Data Science Curricula 2021 report. The goal is not to train working domain experts (e.g., clinicians) to do all the work of a data scientist, rather give them the requisite knowledge and skills to work in multidisciplinary data science teams where they can better utilize their domain expertise. More about domain-specific data science education in adults are described in Section 1.4.2 and 1.5.

1.2.6 Data Science Education (K-12)

K-12 education does not have a separate set of data science curriculum guidelines. The CSTA learning standards and GAISE II Pre-K-12 curriculum guidelines discussed earlier (Sections 1.2.4 and 1.2.2) acknowledge the data science process of working with, collecting, analyzing, and reporting data results.

1.3 Learner Personas and Pedagogical Backed Strategies for Creating Accessible Content in Data Science

Personas come from the field of product design where they are used by designers to help create products for their users. Personas initial conception was created around people succumbing to technology that function, but are clunky to work with (e.g., video cassette records, car alarms, computer software applications). The goal of a persona is to provide a user-centered approach to product design by clearly defining remember-able representations of users. This means trying to identify who the users are, what they want from the product, and how they will use the product. The specifics and understanding that personas are representations of target groups is what separates it from devolving into a stereotype [15].

Persona methodology can be applied to education as well. The “product” is typically a lesson plan or curriculum, and the “users” can be teachers [1] or students. This dissertation explores the use and creation of personas for learning materials (Section 1.3.2). The notion between a persona and a stereotype is somewhat related, i.e., we hypothesize that many domain experts (e.g., medical practitioner) do not work with data regularly, lack data literacy concepts, and mainly interface with data in Excel. What sets this work apart is this hypothesis is tested and incorporated into specific personas that have actionable decisions when putting together learning content.

1.3.1 Personas

We will begin with the basics and overview of persona methodology before applying it to educational contexts

User-Centered Design (UCD)

There are three (3) main difficulties with user-centered design (UCD). First, the natural tendency is to be more self-centered, not user-centered. Design choices typically stem from our own wants and needs, and we even seek out users who are similar to ourselves for product feedback [15, 16]. Self-centered design is a better approach than technology-centered design, where the focus is around capabilities of the technology and less around capabilities of the user, but most people who develop the product are not representative of the audience [15].

Second, Users are varied, and it is not possible to please everyone. Attempt to do so usually have conflicting solutions and are not possible to satisfy everyone's needs [15].

Lastly, the people who collect user and market research are typically not the same people as the ones who design and build a product. If the user and market data are not available at the appropriate time, the people who implement the product will follow a path of what is easiest and cheapest to build [15].

"Users" are Ambiguous

"Users" are an ambiguous, ill-defined, catch-all term. It is more or less meaningless. More details are needed about the "user" to build effective products. What once was a novel approach to product design is now ineffective and ambiguous as it was overused in discourse. We do not refer to cars, bicycles, and chairs as "driver-friendly", "cyclist-friendly", or "bum-friendly", respectively, [15]. So, we should not have to describe teaching and learning materials as "learner-friendly".

Understanding and identifying users are necessary but insufficient for good design. There are multiple steps beyond identifying users: (1) user information needs to be communicated to product designers, (2) product designers need to interpret the information the same way, (3) information about the users need to be incorporated into the product, and (4) the design needs to be evaluated for effectiveness [15].

User Representations

Several important milestones on user representations led to what eventually became known as “personas” [15]. The “market” of people needed to be clearly identified, created, and specified and successful products come from specific definitions of target customers [15, 74, 75]. However, market-segment definitions are usually impersonal and abstract, and focusing on customers using “target customer characterizations” deeply explore individual customers in their work environment [15, 76]. Each characterization incorporates:

1. Personal profile and job description
2. Technical resources
3. A “day in the life” dramatization before the introduction of the proposed product
4. Problem or dilemma that motivates the purchase of the proposed product
5. A “day in the life” after the introduction of the product [15, 76].

Instead of looking at target audiences as a part of a mass population, the notion of “indivisualization” places an emphasis of looking at target audiences as individuals [15, 77]. Descriptive profiles describe the audience from the point of view of others, while individualized profiles contextualize individuals during a purchase decision as they see themselves [15, 77].

The need for a clearer customer "image" begins to bridge the gap between marketing and product design [15, 78]. These images use data to create single sentences that describe some essential characteristic of the customer; desires and suggestions for solutions are not a part of the image statement.

Scenarios are used to aid in product design. These are stories about the target population, not specific users, that provides an overview of everything that is happening to help designers and analysis focus on people and tasks [15, 79]. One of the major limitations with scenario-based designs is how it makes many assumptions about the user to describe the network of interactions, therefore lacking many user-level details in the environment, e.g., relevant details, motivation, preferences, etc. However, scenarios still complement persona methods to promote good UCD [15, 80].

User profiles are detailed representations of users that come from data. These profiles are highly specific, to the point of lacking personality, and provide value to the product design process by consolidating user data from interviews and qualitative research methods [81]. User roles, on the other hand, provide (1) larger context of the user in the environment, (2) user characteristics within the role, and (3) special criteria to support the role. User roles are a precursor to personas and should be explored before personas are created by mapping roles, relationships, and tools together using "contextual design" [82–85].

User archetypes are also a data-driven method created to improve on the concept of user classes and contain user descriptions; attributes; computer skills; concerns and goals; market size and influence; and activities [80]. Personas are imaginary people used to represent a group of users. They are "hypothetical archetypes" that are created with rigor and precision, which provide its usefulness. Personas can be created initially as "assumption personas", and backed by data during the persona lifecycle. Using personas in Goal-Directed Design™are used to inform product design for lasting solutions [15, 86].

Using Personas

There are four (4) main benefits to using personas:

1. Make assumptions about users explicit
2. Place the focus on specific types of users rather than on all possible users
3. Help us make better decisions by limiting choices
4. Engage the product design and development team ([15, 87])

Just creating the personas is not enough. They need to be used in a thoughtful way. Collecting and analyzing user data is the start of creating a report about the users. But these reports also need to be created with the people who implement the product in mind. Many times, a partial understanding of the indented product user gets tinted with personal experiences and biases, and the user report, along with its original message, largely gets underutilized [15].

It's impossible to cater to every persona, nor should all possible types of users be made into a persona. The notion of a primary persona is crucial for managing effort and resources. Even if the personas created were "wrong", the personas were created from user data and the "finished" persona is only the starting point that needs to be validated with more user research. At worst, using a "wrong" persona still provides a consistent product [15]..

This dissertation seeks to identify medical practitioners as its main persona, and use a series of targeted survey questions to distinguish key components of data literacy topics and technical programming skills, they are lacking. The learning materials created will primarily depend on their relevant prior knowledge, and their perception of needs.

1.3.2 Learner Personas in Education

The primary user role for this dissertation are people who work in the biomedical sciences (e.g., students, researchers, and medical clinicians) who are data science novices. This body of work seeks to collect user data (i.e., learners) to create personas (i.e., learner personas) to better engage learners and aim to improve learning outcomes (Table 1.1).

Table 1.1: How learners can be the "user" in product (i.e., lesson) design. The personas column are reproduced from "The Persona Lifecycle" [15] were corresponding list for learner personas were adapted from.

Personas	Learner Personas
• Increase your products' usability, utility, and general appeal	• Increase a lessons' usability, utility, and general appeal
• Streamline your teams' processes and improve your colleagues' ability to work together	• Streamline learning progress and improve learner team Interoperability
• Enable your company to make business decisions that help both your company and your customers	• Enable learners to make data-driven decisions
• Improve your company's bottom line	• Improve learning outcomes

Creating learner personas is not a new concept in education. They have been created for teachers to identify what kind of professional development needs they have (Figure 1.7) [1], and in industry to identify the kinds of learning resources and documentation are needed to better use their products (Figure 1.6) [88]. The personas created by The Carpentries [89] and RStudio, PBC [88], include ones that are looking for data science training. What is lacking are explicit personas for specific domains. This can help identify common knowledge bases to better curate learning materials (Section 1.4), and also serve as marketing and outreach materials to help learners find resources [70].

Persona	In Brief	Domain Knowledge	Statistics Knowledge	Programming Knowledge
Anya Academic	A professor who needs training for her research and to pass on to students.	expert	competent	competent
Celine Certified	A certified RStudio instructor.	competent	competent	competent
Exton Excel	A proficient Excel user working in industry who wants to switch to R.	competent	novice	novice
Jacqui Ofalltrades	A data science generalist at a small consulting company.	expert	expert	expert
Katrin Keener	An R enthusiast.	competent	competent	competent
Larry Legacy	A reluctant learner who would really rather just keep using the tools he knows.	expert	expert	novice
M'shelle Manager	An ex-programmer who now leads a team and needs to make decisions about tool adoption and training.	competent	novice	competent
Nang Newbie	An undergraduate student without statistical knowledge, programming skills, and real-world experience.	novice	novice	novice
Toshi Techsupport	A sys admin who has to support data scientists.	expert	novice	expert

Figure 1.6: RStudio, PBC Learner personas [88]. Reproduction of the RStudio, PBC learner personas with a brief description of each persona, and their competency in domain knowledge, statistics knowledge, and programming knowledge.

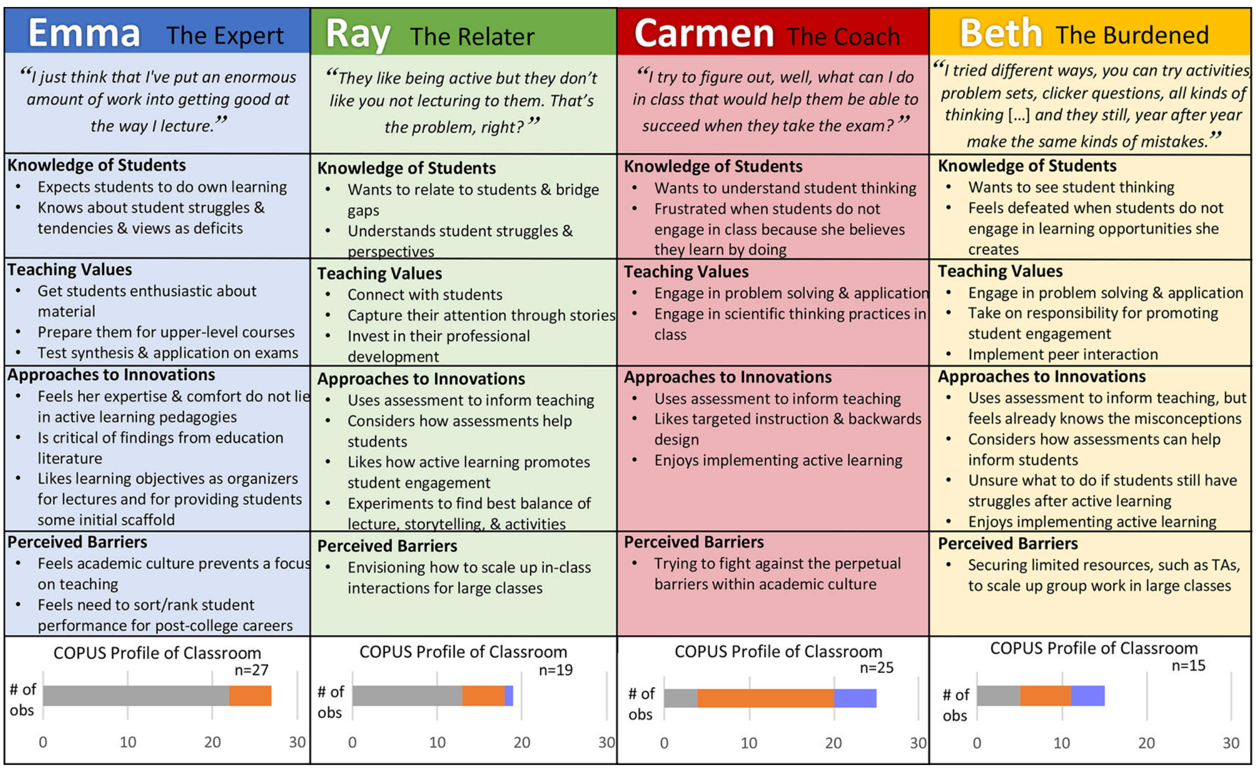


Figure 1.7: Reproduced Teaching personas [1]. These show the 4 personas created looking at professional development needs for teachers.

1.4 Best Practices in Teaching Data Science

Graduates with university computing degrees are having difficulties in workplace settings. This is from the discrepancy between teaching a knowledge-based learning paradigm and an application-based curriculum. The latter incorporates more practical skills that apply knowledge [65, 66]. While modern students have grown up with computers and know how to use the internet for help, more advanced programming skills should not take precedence over data analysis skills or statistical thinking [8].

The GAISE 2016 report mentions that "basic computer skills" can be assumed [8], however "basic" was not defined. For example, modern technologies have made the filesystem more abstract, file sharing and cloud services (e.g., Dropbox, OneDrive, iCloud) does not always make it clear where files may be stored on the filesystem [90, 91]. This obfuscation of the file system does make it confusing when trying to point to datasets around the file system. Additionally, more advanced programming skills would make the course more intimidating for novices [64, 92]. While modern students do have the skills to search for help, they may not know the correct jargon to search for help effectively, if they do manage to find ax solution, many of the solutions novices would find online would be incomprehensible [64, 69, 70, 93, 94]. Parsing and reading error messages is also difficult for novices which would make asking questions online difficult [64, 70, 93, 94]. The workshop materials created for this dissertation address this skills gap by starting with spreadsheets and learning about how to load them into a programming language by understanding a working directory and loading data using absolute and relative paths.

The Computing Competencies for Undergraduate Data Science Curricula 2021 report's Data Cleaning Tier 1 knowledge in data transformation does not talk about tidy data principles [67]. However, that knowledge tier does discuss data standardization and normalization, which is related to tidy data principles as it is the data processing needed to store data for databases and usually comes from tidy data [50].

The concepts around "tidy data" principles are a core component of data science [50, 67]. The work in this dissertation explores how to develop a course around these tidy data principles using persona methodology to identify a domain-specific sub-population and identify gaps in their knowledge and develop lesson materials to meet their needs. This work aims to create a smaller set of learning materials for novices, not creating a full curriculum. So, many of the knowledge areas (KAs) from a data science curriculum guideline does not need to be explicitly taught, and identifying core knowledge and skills to meet our learner's needs and time constraints are a priority in this dissertation work. It is important to set a good foundation to learn and apply additional data science concepts [8, 65, 66, 69, 70, 94]. As a domain-specific course primarily aimed at new learners who are working professionals, It's even more important to meet learners where they are, as they are not going to be primarily in a classroom environment [64, 70, 93].

1.4.1 Designing Lesson Materials

Tidy data principles are important because it feeds into the rest of the data science process [50, 52, 67]. This dissertation creates the learning materials using a backward design to keep the content focused on its objective: teaching practitioners how to do data science in the biomedical sciences. The backward design process begins with identifying learner personas (Section 1.3). In its simplest form, the seven (7) steps for a backward design, adapted from "Teaching Tech Together" are [70]:

L.1 Create or recycle learner personas (Section 1.3) to figure out who you are trying to help and what will appeal to them.

L.2 Brainstorm to get a rough idea of what you want to cover, how you're goign to do it, what problems or misconceptions you expect to encounter, what's *not* going to to be included, and so on (Section 1.4.1). Drawing concept maps can help a lot at this stage.

L.3 Create a summative assessment to define your overall goal. This can be the final exam for a course or the capstone project for a one-day workshop; regardless of its form or size, it shows how far you hope to get more clearly than a point-form list of objectives.

L.4 Create formative assessments that will give people a chance to practice the things you're [sic] learning. These will also tell you (and them) whether they're [sic] making progress and where they need to focus their attention. The best way to do this is to itemize the knowledge and skills used in the summative assessment you developed in the previous step and then create at least one formative assessment for each.

L.5 Order the formative assessments to create a course outline based on their complexity, their dependencies, and how well topics will motivate your learners.

L.6 Write material to get learners from one formative assessment to the next. Each hour of instruction should consist of three (3) to five (5) such episodes.

L.7 Write a summary description of the course to help its intended audience find it and figure out whether it's right for them.

Mental Models, Concept Maps, and Cognitive Load

Going through the order of planning a lesson using a backward design approach, identifying who our learners are by using personas (**L.1**) is the first step. The persona methodology (Section 1.3) requires collecting information from potential learners. This process requires some planning between what needs to be covered (**L.2**).

Concept maps are one way to visualize all the relationships between concepts. A similar activity to creating concept maps is task deconstruction [93]. Concept maps can be used to help the instructor plan out learning materials by identifying: (1) concepts to be covered in a lesson, (2) ordering of lesson content, (3) concepts that can be cut from a lesson.

From the learner's perspective, concept maps can help in identifying their mental model of how topics are related to one another. The number and density of connections are one way to distinguish the competency of a topic. The progression from novice to expert was first described in the Dreyfus model of skill acquisition [95, 96]. The same cognitive transitions were described in looking at skill acquisition and clinical judgment in nurses [96].

The original Computational Curriculum guidelines for knowledge-based learning is similar to the differences between concepts from the learner's perspective, and the lesson's concept map by Starting from the learner's knowledge base, and adding new nodes of knowledge and connections to build up a knowledge base. The competency-based approach is similar, but since it focuses more on integrated skills, these guidelines point to using a backward design approach to best teach computational skills.

The concept maps also represent the amount of information being taught at any moment during a lesson. For any given point during the learning process, there are three (3) main areas where confusion can occur: knowledge, information, and processing power [94]. Each of these three areas relate to differet areas of memory: long-term memory (LTM) stores knowledge, short-term memory (STM) stores information, and working memory (WM) is the processing power. These concepts are not just specific to learning but apply to all cognitive activities [94]. For novices, because they lack the necessary density of connections and knowledge (LTM), each new bit of information (STM) requires more processing power (WM). Concept maps help plan how much working memory and short-term memory learners are using during a lesson, and lessons should follow George Miller's 7 ± 2 rule of how many items people can store in their STM [97]. More recent research suggests that the STM is even smaller, two (2) to six (6) items [94], or 4 ± 1 [98], which means lessons need to be curated and presented to lower learner's cognitive load. Concept maps are just one tool that can be used to plan how much cognitive load a learner may be experiencing.

The backward design approach used in this dissertation ended up with meeting learner's where they are, spreadsheet applications like Excel. This because the starting point for the learning materials by exploring common issues people encounter while working with spreadsheets, and how to avoid them [99]. The lessons here start with a common starting point to introduce tidy data concepts so by the time the actual lesson on tidy data comes up, enough of the rationale, need, and concepts of tidy data have already been covered, just not explicitly until that point.

Learning Objectives, Formative Assessment, and Summative Assessments

After planning out the general concepts with a concept map (or similar task deconstruction process), operationalizing and creating measures for whether or not concepts are being learned are the next steps in a backward design process for curriculum building [70]. The notion of whether or not a set of learning materials being "effective" is vague. Learning objectives are explicit, measurable, actions that are student-centered [69]. They articulate the lesson intentions to learners for them to direct learning efforts, and provide a framework to organize lesson content and assessments to the instructor.

The measurable verbs that are used for learning objectives typically come from Bloom's Taxonomy [69, 70, 100]. The taxonomy is a popular framework that was originally published in 1956 and updated in 2001 as a framework for understanding that is often depicted as a pyramid [100–102]. The pyramid representation makes each part of the taxonomy seem as discrete steps part of a hierarchy and cause issues with how to classify specific objectives [103]. Daniel Willingham's diagram represents Bloom's taxonomy with knowledge as its foundational base, with the other terms on top of "knowledge" without any hierarchy [102, 104]. Fink's Taxonomy of Significant learning is another set of terms that looks at learning objectives as a complementary evolution of change and is typically represented in a non-hierarchical manner [105].

Regardless of how the learning frameworks are depicted, they are still useful for creating explicit, measurable, student-centered actions [69, 70]. Because learning objectives are measurable, they lead to creating assessments that can gauge how much of the material has been retained by the learner. There are two (2) main forms of assessments, formative and summative. Formative assessments are relatively quick and low-stakes questions that are presented to the learner during the time of instruction. These are typically represented as "clicker" questions, homework assignments, or questions asked in the middle of a lecture, and are typically focused on a few concepts. The goal of formative assessment questions is to provide both the instructor and learner feedback about what needs to to review and focused on. Summative assessments, on the other hand, are relatively longer and higher-stakes questions that are presented to the learner at the end of an instruction period. These are typically mid-term examinations, final examinations, or final projects that encompass multiple topics and require the learner to consolidate what they have learned to answer the question. Formative and summative assessments provide a small delay after learning, which improves comprehension [106].

In creating domain-specific data science learning materials for the biomedical sciences, after identifying learner personas and core concepts that need to be taught, this dissertation creates a set of learning objectives and a summative assessment question and uses a backward design to plan out the learning materials with three (3) to five (5) formative assessment questions roughly every instructional hour [70].

1.4.2 Engaging Learners for Active Learning

Keeping learners engaged and motivated guide their behaviors and energy to what they spend time learning [69]. To keep learners motivated, the learning materials need to be interesting or relevant. The learners also need to have a perception that they will be successful [69]. Domain-specific learning materials add to the perceived value of the learning content. Spending the effort with persona methodologies can help create authentic, real-world examples learners are most likely to encounter, and incorporate them into the lesson materials. This dissertation work used a backward design to trace back to spreadsheet programs as a base of knowledge for most of the potential learners. Combining their familiarity with spreadsheets to show why data is sometimes difficult to work with as a means to implicitly introduce tidy data concepts slowly introduces the concepts to reduce their cognitive load when tidy data concepts are explicitly described.

The starting point of a backward design approach is the complex real-world example that aims to show how all the skills relate to one another and can be applied broadly. Tidy data principles stand at the core of these ideas because data science tools rely on on tidy data for more complex data tasks, e.g., plotting, and model fitting [50, 52]. Having formative assessment targets to pace the lesson content helps identify what topics are needed for "quick wins" [93]. Having the appropriate level of formative assessments also provide a means for timely feedback, where learners can iterate and focus their efforts.

Teaching Live Coding

When teaching programming-related tasks in a more formal setting, live-coding is an effective way to teach students as it allows for the flexibility to follow learner's interests in real-time. Instead of showing the correct solutions in a slide deck, live coding fosters more active teaching and learning and also promotes unintended knowledge transfer from the instructor by showing learners how things are done. Learners are able to see how problems are diagnosed when the instructor makes mistakes in front of the students (either on purpose or by accident) and forces the instructor to work and think through the error in front of the students. The process of live coding also slows down the instructor and gives students a way to follow along. This way multiple sensory inputs are working together to encode the same bit of information, and helps retain knowledge [70, 94].

Pair-Programming

Pair programming is the process of "pairing" 2 people together on a task where one person does the actual programming, and the other person talks them through the process. Usually, the more experienced person is guiding the other person what to program, however, it can work if both people have the same experience or learning something together. This delegation of tasks allows the programmer to deal with the nuances of programming syntax, while the other member can think about the overall program flow. Separating these tasks is useful for new learners as it reduces the cognitive load of managing a workflow and data science tasks with the syntax and errors from programming. It also gives the opportunity for both members to learn together and from one another. What makes pair programming different from a traditional "group project" is that 2 people are working on the same part of the project at the same time. Only the cognitive load of accomplishing the task is delegated.

In a group project, the entire project is delegated into separate tasks, so members in the group do not necessarily work on the same task. However, the main downside with pair programming is that it uses a lot of resources, two people need to be assigned to work on the same exact problem [70, 94].

1.5 The Need for Pedagogically-Backed Data Science Curriculum in Medicine

Many data science students go on to teach, but are rarely trained in pedagogy [23]. What is also lacking are rigorous evaluations of data science tools that statisticians advocate for process improvements in other disciplines [23]. Industry and government can use professional advisory boards, work-study programs, and internships to generate the demand for modern computational skills [66]. Academic institutions can benefit their graduates by proactively supporting strong, contemporary computing programs [66]. The discrepancy between student need for clinical informatics courses, and the availability and opportunity to take a clinical informatics course in the biomedical sciences [107, 108], show that academic institutions, industry, and government can play a role in meeting training needs for the biomedical sciences. This dissertation aims to fill the gaps of improving the data process statisticians advocate for, while also providing a set of learning materials that meet the need for the biomedical sciences.

Healthcare providers (e.g., medical doctors, veterinarians, nurses, physician assistants, other clinicians, and administrators) play a key role in the interdisciplinary needs in health care [109–111]. Being able to work with open collaborative data science platforms, the healthcare domain experts are able to better understand, utilize, and contribute to the “wisdom of crowds" [109–111]. A coordinated effort between the biomedical communities and data science communities is a priority in order to create effective curricular frameworks for mutual understanding of each respective domain [111].

In the United States, the National Institute of Health (NIH) is responsible for biomedical and public health research. The NIH has made a strategic data science plan to improve the storage, management, standardization and publication of biomedical research. Professional organizations such as the American Medical Association (AMA), American Nurses Association (ANA), American Academy of Physician Associates (AAPA), and American Veterinary Medical Association (AVMA) serve as national organizations for medical doctors, nurses, physician associates, and veterinarians. All of these organizations have made calls for the importance of data science in their respective professions [3, 34, 111–114]. The NIH's strategic data science plan (Figure 1.8) has five (5) main goals and objectives: (1) data infrastructure, (2) modernized data ecosystem, (3) data management, analysis, and tools, (4) workforce development, and (5) stewardship and sustainability.

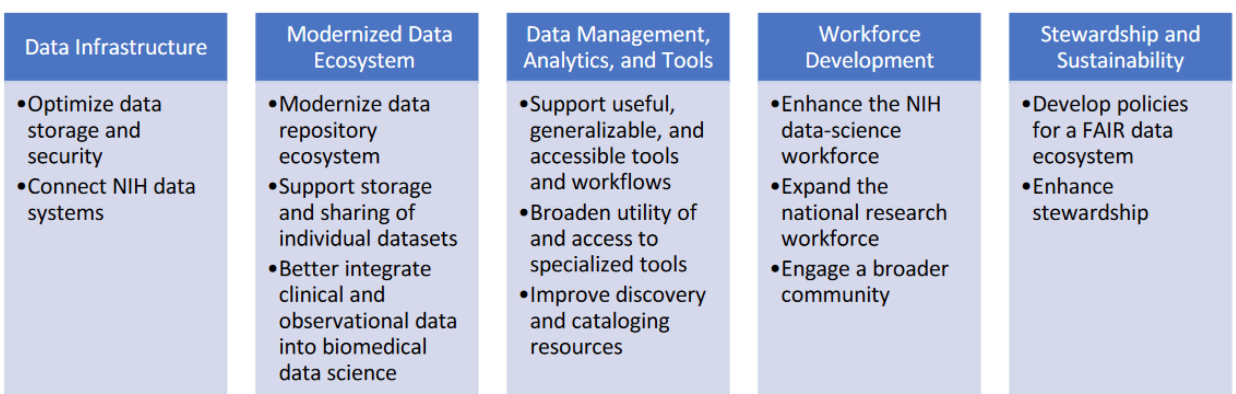


Figure 1.8: Reproduction of the NIH strategic plan for data science [34]. The plan consists of five (5) main goals and objectives: (1) data infrastructure, (2) modernized data ecosystem, (3) data management, analysis, and tools, (4) workforce development, and (5) stewardship and sustainability. The plan's goal is to address the quantity and highly distributed nature of biomedical data and metadata that lead to a variety of data formats which lead to data cleaning complications.

The objectives in data management, analytics, and tools; workforce development; and stewardship and sustainability relate to the more individual data science skills that integrate with larger data infrastructure [34].

Incorporating computing skills and domain knowledge can be thought of computing + x and x + computing (where x is a knowledge domain) [66]. In computing + x, computing systems extend to non-computing disciplines. These fields usually have "informatics" in the term e.g., medical infomatics, bioinformatics, health informatics, legal informatics, etc [66]. In x + computing, computing systems are extensions to an already existing and established field of study. One prominent example is computational biology, where established laboratory methods expanded to computing. Both mechanisms of combining computing and a discipline allow for the discovery of transformational relationships, only the starting point is different [66].

Medicine is inherently an information-management task [115]. This dissertation focus on the core and fundamental skills of data science computing, and how these skills can better areas in the biomedical sciences (computing + x and x + computing) by focusing on core data literacy concepts. By combining domain, computing, and integrative knowledge and skills, non-computing individuals can make the connections from their domain to the transformative opportunities created by using computing [66].

1.6 Building a Community (of Practice)

This dissertation does not explicitly explore the process of online community building, but this is an important idea to keep in mind when creating educational materials, as connecting with other educators to form a community of practice where (typically geographically co-located) people with a common set of goals, interests, and concerns can learn from one another [70]. Organizations like The Carpentries can provide online teaching communities of practice where other instructors can learn from one another around builing the pedagogical content knowledge for teaching, not just around data science [13, 61].

Community building is a slow process, and there are four (4) main components on building and sustaining a community. The first step is onboarding and recruiting new members. A Code of Conduct should be prominent during this phase of community building and ensure members are in a safe environment. Onboarding should also include resources to bring new members up to date on current community events. The second step is around retention. Community members should have some sense of agency and be able to contribute to the group, and these contributions should be acknowledged. As communities grow, eventually a governance model will need to be created to help provide top-down guidance for major decisions. These governance models do not need to be strictly hierarchical, and can adapt with the size of the community. Finally, retention and onboarding is a never-ending process. Healthy communities plan for the efflux of members by onboarding more members and helping to retain them.

This dissertation primarily focuses on the learners, and finding ways to improve their learning. Teaching is an art and skill on its own [70, 116]. Knowing what needs to be taught (content knowledge), how the materials are taught (pedagogical content knowledge), and why topics are taught in context (curricular knowledge) are all aspects of teaching knowledge that instructors can benefit from joining a teaching community of practice [13, 70, 116].

1.7 Ethics

One of the more prominent data science workflow figures comes from the R4DS book reproduced in Figure 1.9 [52]. Many of these figures that discuss the data science workflow miss the opportunity to discuss how the data products that come out of data science and the decisions that come from communicating findings, affects the world ((e.g., Figures 1.1 and 1.5)).

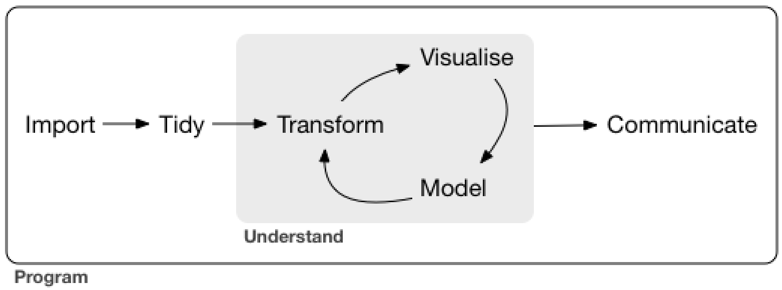


Figure 1.9: Reproduction of the data science workflow from R4DS [52].

In practice, there many cycles within the data science process; it is typically not a linear process from data collection to communicating results for a decision. Each of the steps within the data science process has the potential for errors and biases. Among one of the cycles within the data science process is the cycle between how decisions made from the data science process affect the world (Figure 1.10). Many of the data ethics issues can be accounted for by being more mindful of how data science have real-world consequences.

Reproducibility failures (Figure 1.11 adapted from [118]) are one way where data science can have consequences in the real-world [118–124].

Aside from technical issues that can cause adverse real-world impacts, data science and healthcare have numerous ethical questions as well [125–129]. In healthcare, there are primiarly 4 ethical challenges: (1) informed consent to use, (2) safety and transparency, (3) algorithmic fairness and biases, and (4) data privacy [125]. While data science has the potential to revolutionize medical research, a high level of scrutiny must be employed to adequately and safely work with data that will have real impacts on patients.

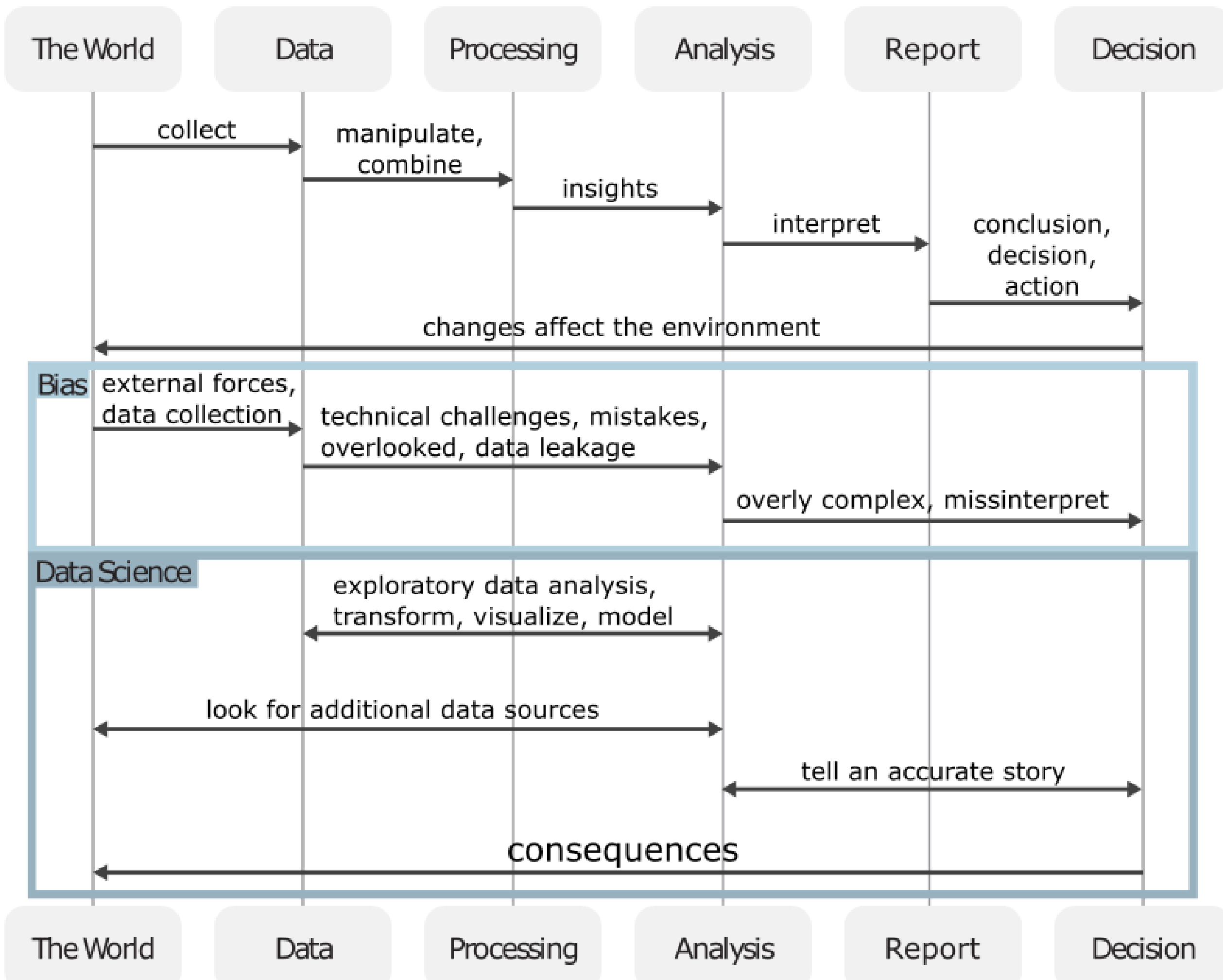


Figure 1.10: The data science process showing the same cycles as previous data science figures. This figure puts an emphasis on how the decisions from data products have real-world consequences that feedback into the data science process [117].

Reproducibility error	Consequence	Source(s)
Limitations in Excel data formats	Loss of 16,000 COVID case records in the UK	(Kelion 2020)
Automatic formatting in Excel	Important genes disregarded in scientific studies	(Zeeberg et al. 2004; Ziemann, Eren, and El-Osta 2016)
Deletion of a cell caused rows to shift	Mix-up of which patient group received the treatment	(Wallensteen et al. 2018)
Using binary instead of explanatory labels	Mix-up of the intervention with the control group	(Aboumatar and Wise 2019)
Using the same notation for missing data and zero values	Paper retraction	(Whitehouse et al. 2021)
Incorrectly copying data in a spreadsheet	Delay in the opening of a hospital	(Picken 2020)

Figure 1.11: Reproduction of the Reproducibility Failures table by Ostblom and Timebers [118] showing the real-world consequences of data reproducibbility errors.

Chapter 2

Identification of Biomedical Data Science Learner Persons: Implications and Lessons Learned for Domain-Specific Data Science Curriculum

Abstract

Many data science learning resources are geared towards general audiences. This provides an opportunity to fill the need for more domain-specific learning materials which can provide more context, motivation, and reliability to the content to help engage learners. One of the first steps in creating new learning materials is identifying the audience who will be using the materials. Persona methodology provides a user-centered framework to empirically identify the learning audience (i.e., learner personas), how they will interact with the lesson, what their prior relevant knowledge are, and their perception of needs. This study demonstrates that personas can be empirically created used in education to create learner personas. Specifically, this study looks into learner personas in the biomedical sciences to

create more relevant and domain-specific data science learning materials.

A survey was distributed asking participants about their prior programming experience; data cleaning and processing experience; project management satisfaction; and statistics. The survey items were validated using Confirmatory Factor Analysis (CFA), and the learner personas were identified with hierarchical clustering. Survey responses were combined with the clusters to create 3 learner personas: Ash Academic, Clare Clinician, and Samir Student. The primary persona was Clare Clinician and was used to create a domain-specific set of learning materials for the biomedical sciences. The survey can also be used for future educators to identify their learners and can be used to create learner personas for other domains.

2.1 Introduction

The abundance of data science learning materials have made it a commodity [35]. Massive online open courses (MOOCs) providers like Coursera, Udacity, and EdX offer dozens of free data sciences courses (57, 7, and 68, respectively) [36–38]. Publishing tools like Bookdown and JupyterBook also catalogue free online book resources (1075 and 63, respectively) [39, 40]. However, many of these resources are geared towards general populations or general topics, and examples are not always relevant to new learners [35]. For working practitioners we also need to consider their needs and time constraints.

There are community projects that try to create compilations of more resources, e.g., "The Big Book of R" has 267 free books on R programming with 21 books under its life sciences section, and only 4 resources with a medical or epidemiology focus, one of them is the work the authors are developing [130]). Instead of creating data science materials for general audiences, there is the opportunity to create more domain-specific lesson materials with a direct focus on pedagogy, identifying more data science concepts, and relevant examples [35]. In the medicine domain, the demand for informatics courses outpaces and exceeds the training opportunities [107, 108, 112, 131]. Increasing the quantity, quality, and publicity of medically-focused data science materials can alleviate the training and opportunity gap [107, 112]. More relevant and more content-accessible lesson materials often leads to better engagement and motivation from learners [69, 70, 93]. Online courses, including massive open online courses (MOOCs), do have the flexibility to create domain-specific and focused learning materials. But the lack of clinical informatics training and mentoring opportunities available to medical students suggest that there is still a need to increase the quantity, quality, and publicity of data science materials catered towards the biomedical sciences [107].

Healthcare providers (e.g., medical doctors, veterinarians, nurses, physician assistants, other clinicians, and administrators) play a key role in the interdisciplinary needs in health care [109–111]. Being able to work with open collaborative data science platforms, the healthcare domain-experts are able to better understand, utilize, and contribute to the "wisdom of crowds" [109–111]. A coordinated effort between the biomedical communities and data science communities is a priority in order to create effective curricular frameworks for mutual understanding of each respective domain [111].

In the United states, the National Institute of Health (NIH) is responsible for biomedical and public health research. They have made a strategic data science plan to improve the storage, management, standardization and publication of biomedical research [34]. Professional organizations such as the American Medical Association (AMA), American Nurses Association (ANA), American Academy of Physician Associates (AAPA), and American Veterinary Medical Association (AVMA) serve as national organizations for medical doctors, nurses, physician associates, and veterinarians. All of these organizations have made calls for the importance of data science in their respective professions [3, 34, 111–114]. The NIH's strategic data science plan has five (5) main goals and objectives: (1) data infrastructure, (2) modernized data ecosystem, (3) data management, analysis, and tools, (4) workforce development, and (5) stewardship and sustainability. The objectives in data management, analytics, and tools; workforce development; and stewardship and sustainability relate to the more individual data science skills that integrate with larger data infrastructure [34].

Persona methodology is a systematic way to collect data on our potential learners, and identify key components of our target audience grounded in empirical data. These "learner personas" are fictional characters based on empirical data that represent key characteristics of our learners. This is a technique is commonly used by the design industry to bridge the gap between product designers and product users [1, 15, 87]. Here, we apply these methods to the educational space, identifying what our learners know and how to build on their existing knowledge.

Persona methodologies have been used to identify college instructors and teacher professional development needs [1]. While learner personas have been created in the past to help focus educational materials [88, 89], the process, data, and survey instruments, were not published. This study aims to use a data and analysis driven approach on creating learner personas for the biomedical sciences in developing relevant data science materials to people who work in the biomedical space. These personas provide a more memorable character and also serve to reduce the cognitive load of instructors when preparing and tailoring lesson content since all the considerations are incorporated into specific characters, and not disparate lists [15, 86, 87]. There are for (4) main benefits of using the persona methodology, they: (1) make assumptions about users explicit, (2) place the focus on specific types of users rather than on all possible users, (3) help us make better decisions by limiting choices, and (4) engage the product design and development team [15, 87]. We create personas to understand what our learners know and how to get them competent and confident in performing data literacy tasks and basic data analysis techniques. The personas can also help with the shortage of clinical informatics training and mentoring opportunities by using them as a marketing and outreach tool to promote learning resources.

One of the secondary goals of this study is to validate the survey used in persona identification, and have it be a useful tool for other educators to identify how much programming, data literacy, and technical background their potential learners will have, to create a better learning experience for the learners. We can use the learner persona information with concept maps along with a backwards design to create new lesson content.

2.2 Methods

Results from a pre-workshop student self-assessment survey (i.e., persona survey) were clustered to create and identify biomedical data science learner personas. Only complete survey responses were used for the analysis. There was no data imputation for missing or incomplete responses. The learner persona survey was validated with factor analysis and calculating Cronbach's alpha. Clustering used all the questions in the survey, regardless of factor analysis results, for persona identification. The factor analysis results were not used to shorten the survey questions. All duplicate response IDs were identified as coming from the same person. These responses were coalesced and only the first set of responses were kept for analysis.Survey design, validation, and analysis are discussed below. The R programming language was used for all data processing and analysis [132–181]

2.2.1 Learner Self-Assessment Survey (Persona Survey)

Survey questions were adapted from “How Learning Works”, “Teaching Tech Together”, and The Carpentries pre-, post-, and long-term workshop survey questions [69, 70, 182–185]. Additional demographic questions were also added. A total of 33 questions across eight (8) topics were included in the final survey: (1) Demographics, 5 (2) Programs used in the past, 1 (3) Programming experience, 6 (4) Data cleaning and processing experience, 4 (5) Project and data management, 2 (6) Statistics, 4 (7) Data and programming Likert table, 7 (8) Workshop framing and motivation, 3.

Two (2) of the workshop framing questions were used to design follow-up biomedical data workshops. These free-response questions asked: (1) what learners hoped to learn in a biomedical data science workshop, and (2) what they would they want to be able to do in working with data after a biomedical data science workshop or training event that they cannot do right now.

Survey Questions

Of the 33 questions in the survey, 16 were on an ordinal Likert scale. All survey questions included in the survey can be seen in (Supplemental 2.5.1). In order to better relate to learners and aid in their assessment of skills, Likert table questions were not asked in the more common “Disagree”, “Neither”, or “Agree”, etc, format, but rather asked, “I wouldn’t know where to start”, “I could struggle through, but not confident I could do it”, “I could struggle through by trial and error with a lot of web searches”, and “I could do it quickly with little or no use of external help”. These Likert responses were framed to aid in the clarity of selection of the survey taker, and potentially make the responses more consistent between participants.

This survey is also a part of a larger workshop longitudinal study. In order to track participant responses longitudinally and protect their privacy by not collect identifying information (e.g., names, email addresses, phone numbers, etc), participants created a unique identifier that was generated based on their results to demographic questions. The IRB approved surveys can be found here in supplemental section 2.5.1.

Survey Dissemination

Surveys were created in the Qualtrics platform and emailed out in two (2) rounds [186]. The first round was only sent to biomedical relevant university listservs or to departmental administrators to post on our behalf at Virginia Tech. The second round included the same listservs and contacts from the first round, in addition to slack groups (The Carpentries, R/Medicine, Nursing & Data Science Collaboration), emails collected from teaching two (2) Carpentries instructor workshops, one of them for the National Network of Libraries of Medicine (NNLM), and Claude Moore Health Science Library at the University of Virginia.

Survey Validation

The learner self-assessment survey (i.e., persona survey) was created with the goal of becoming a tool for future instructors to help identify the learning audience. The only biomedical domain-specific questions involve the statistics related questions where the example uses a health-related research question. This was so participants had a better sense of what kind of analysis they would do given a particular scenario. All of the other questions do not assume any particular domain, and the statistics questions are framed in a way where domain knowledge is not needed.

Face Validity The survey was structured around many data literacy concepts and asked questions around programming, data processes, and statistics experience. It was assumed survey takers had some familiarity with spreadsheet programs (i.e., Excel), and focused the questions around spreadsheet proficiency before asking questions around "tidy data" principles and more specific statistics analysis questions. These were all topics we were hoping to cover in preparing workshops materials, and the questions from the survey provided a starting point for the workshop material content. These questions and their use case provided the face validity for the survey.

Factor Analysis The psych R package was used for factor analysis [165]. The demographic, free-response, and prior programming languages used questions were not used and the rest of the responses were scaled prior to running the analysis. Items for factor analysis were picked based on how many other items they were significantly correlated with. The scree plot from the R nFactors package suggested trying 1-factor to 5-factor models [163]. Different rotations and factoring methods were tested, and the best model was picked based on TLI, RMSEA, and BIC scores, in addition to model interoperability.

Internal Consistency The psych R package was also used to calculate Cronbach's alpha for internal consistency. A separate Cronbach's alpha measure was calculated for each set of questions that loaded into each factor. Kendall's kappa was not calculated because survey results were not paired.

2.2.2 Identification of Biomedical-Specific Data Science Learner Personas

All survey questions were used for clustering; results from factor analysis did not select which questions would be used for clustering. Hierarchical clustering with Euclidean distance and Ward's clustering method was used to identify respondent groups [164]. These groups were then combined with demographic information and survey responses to create learner personas.

Results from hierarchical clustering and survey responses were used to create the fictional learner persona characters. To make the character more complete, we combined the empirical survey results for the background, relevant prior knowledge and experience, perception of needs, and special considerations for each persona. The background and special consideration sections were not backed by empirical data, but they were needed to create a complete persona.

2.3 Results

There were a total of 68 responses to the persona survey. 57 complete observations were used for the factor analysis, Cronbach's alpha, and clustering analysis. The persona survey had 1 set of responses that shared a duplicate ID. Looking at the individual responses, it was determined that this particular set of IDs came from the same person. Duplicate responses were back filled for any missing values (i.e., coalesced), and only the initial set of responses were kept for analysis. 57 complete observations were used for clustering.

Items from exploratory factor analysis (EFA) grouped respondents using hierarchical clustering with Euclidean distance and Ward's clustering method to identify learner personas.

Overall, we used the 3-factor model and the clustering results yielded 3 clusters for our personas.

2.3.1 Survey Study Participants

67 survey respondents self-reported their current occupation and career stage. These responses were further grouped into 3 overall occupations: (1) student, (2) researcher, and (3) clinician. Clinical students were placed in the student category, and the "researcher" role was a catch-all for all the other responses. Figure 2.1 shows the grouped occupation counts. The original ungrouped demographic counts are shown in Supplemental Figure 2.7.

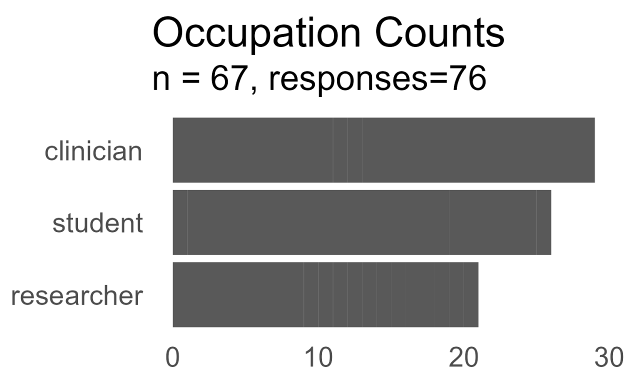


Figure 2.1: Total number of survey respondents from the persona study and their self-reported occupation. This particular survey question was a "select all that applies" question with the ability to type in an "other" response. The responses were then grouped together into the 3 major occupation groups as shown.

Of the 67 respondents, 29 reported as being a student, 21 reported as being a clinician, and 26 were classified as being a researcher.

Overall (Figure 2.2), we count that the respondents agree and strongly agree with the statement that having access to raw data is important to be able to repeat an analysis. However, they also strongly disagree with the notion that they can write small programs to work with data or address a problem with their own work. They either have neutral or strong agree-ments towards programming languages (e.g., R, Python, etc) making analysis more efficient and easier to reproduce. Most respondents were also leaning towards an agreement about being able to search for technical help online.

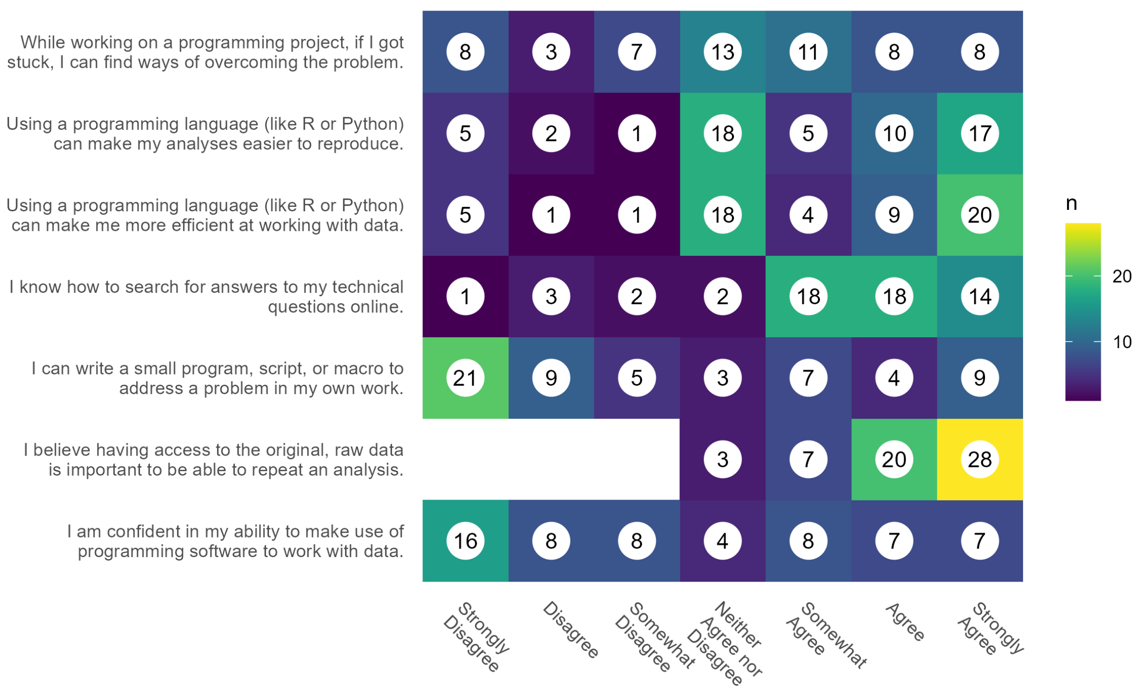


Figure 2.2: Number of responses for each summary Likert scale question. Likert questions were initially created as a summary table to gauge respondents attitudes programming in data analysis. Results show that most respondents believe having raw access to data is important to repeat an analysis, but are less likely able to use a programming language to perform an analysis. There is a bimodal distribution between being indifferent and strongly agreeing about programming languages making analysis easier and more efficient.

2.3.2 Survey Validation

Factor Analysis and Cronbah's Alpha was performed on 14 of the the 23 survey items. The results were used to validate the survey. The clustering for persona identification was performed in a separate step.

Factor Analysis

Survey items were selected for factor analysis based on its correlation with other variables and their significance ($|\rho| \geq 0.5$ and $p < 0.05$). Figure 2.3 shows the correlation matrix between all items. (only significant correlations are shown). Items with a high correlation coefficient ($|\rho| \geq 0.5$) were candidates to be selected for factor analysis. Item candidates that were both significant and highly correlated with at least 20% of the other items were selected for factor analysis. These parameters reduced the number of items for factor analysis, while still keeping at least 1 item from each section of the survey.

14 items met the criteria for factor analysis. The scree plot (Figure 2.4) suggested testing factor models between two (2) and four (4) factors (Supplemental Figure 2.9 shows the scree plot for all the survey items).

A 3-factor model was chosen based on varimax rotation and "ten Berge" scores. The type of rotation did not affect the results. The biggest differences occurred with the number of factors and factoring method.

The maximum likelihood (ML) factoring method was ruled out because data was not normally distributed as confirmed by Q-Q plots and Shapiro Wilk's tests for teach item. The 3-factor model had the most interpretability, and was used for the final model. Multiple factoring methods were tested using the 3-factor model, outside of the ML factoring method, the minimum chi-square (minchi) method had the lowest BIC score, and also had good cut-off values (TIL > 0.9, RMSEA < 0.08). However, a slightly lower fit model (TIL $= 0.93$, RMSEA $= 0.09$) using principal axis factoring (PA) was chosen because it better suited our data (Table 2.1, 2.2) [187, 188].

The results in Table 2.3 show the item loadings for the 3-factor model. The factors can be interpreted as programming experience (PA1), statistics confidence (PA2), and programming

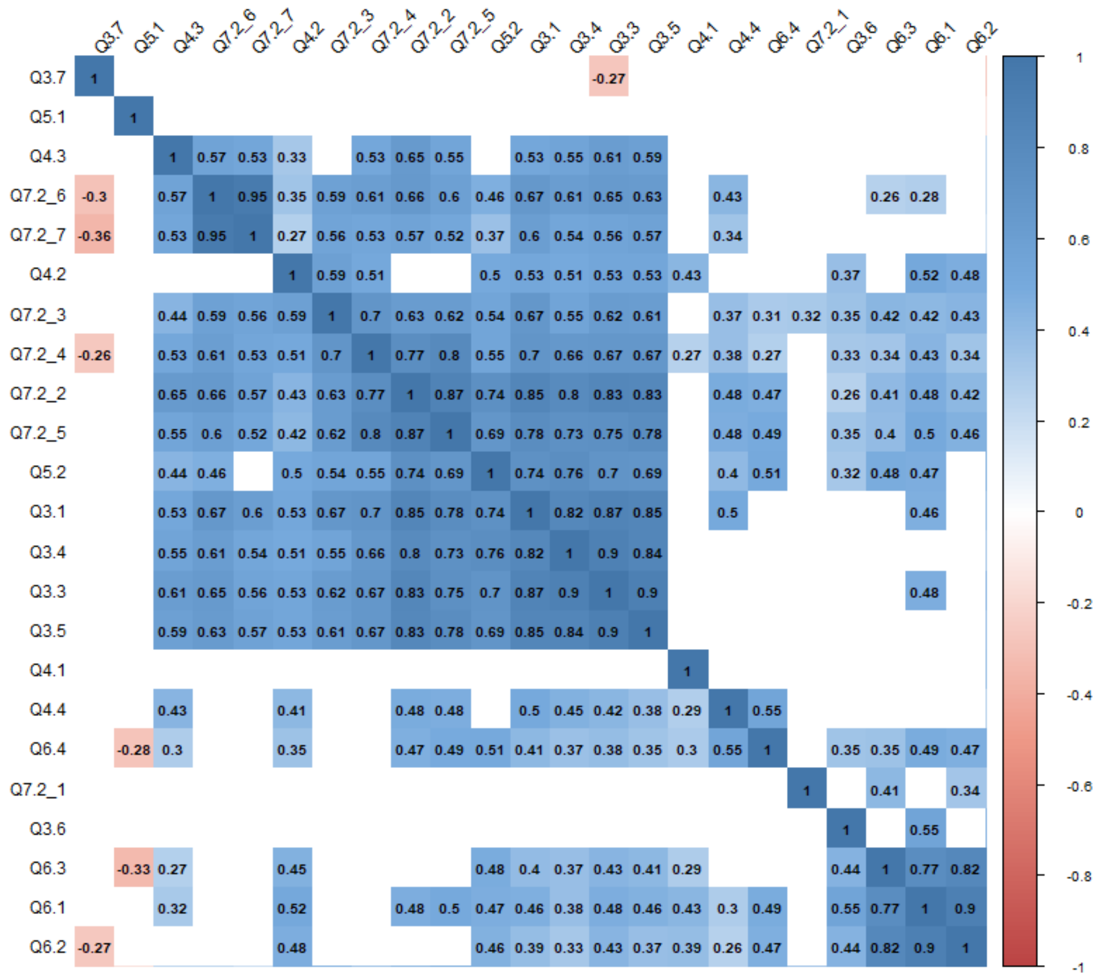


Figure 2.3: Correlation plot of all the items in the persona survey. Plot uses the same variable ordering as the clustering analysis, hierarchical clustering with Euclidean distance using Ward's (ward.D2) clustering method. Only significant correlations are shown ($p < 0.05$). Items that had a high correlation coefficient ($|\rho| \geq 0.5$) were candidates to be used for factor analysis.

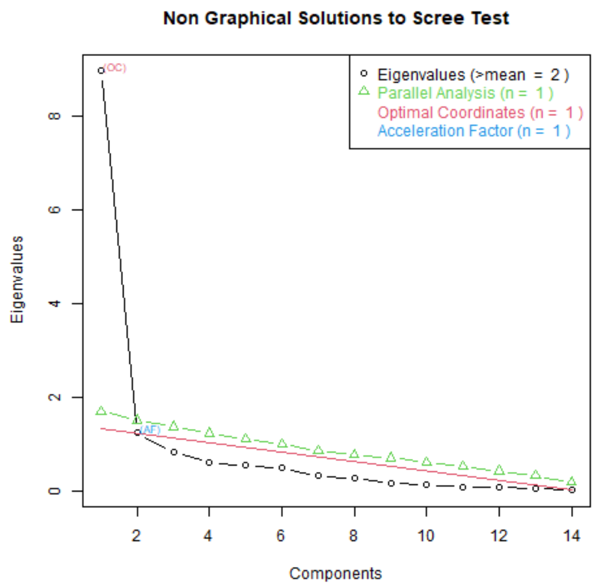


Figure 2.4: Scree plot for factor analysis showing the number of components (x) and the eigenvalues (y). The figure suggests exploring 2 to 4-factor models.

Table 2.1: Cuttoff scores sorted by BIC for all 3-factor varimax rotated models. The maximum likelihood (ML) and minimum chi-square (minchi) factoring methods had the best cutoff values, but both had to be ruled out because the methods were not suitable for the data. Principal axis factoring (PA) was choosen because it does not require normally distributed data [187, 188]. Only varimax rotations are shown, since the rotation did not change any of the cutoff scores, even when comparing orthogonal rotations to oblique transformations.

	nfactors	rotation	fm	tli	rmsea	bic
1	3	varimax	ml	0.98	0.05	-151.18
2	3	varimax	minchi	0.97	0.06	-145.35
3	3	varimax	old.min	0.94	0.09	-134.97
4	3	varimax	uls	0.93	0.09	-130.59
5	3	varimax	ols	0.93	0.09	-130.59
6	3	varimax	minres	0.93	0.09	-130.59
7	3	varimax	pa	0.93	0.09	-130.58
8	3	varimax	minrank	0.92	0.10	-126.12
9	3	varimax	alpha	0.91	0.11	-124.42
10	3	varimax	wls	0.84	0.14	-96.67
11	3	varimax	gls	0.81	0.16	-82.37

Table 2.2: All varimax rotation factor analysis models that met cutoff values of TLI $\geq$ 0.9 and RMSEA < 0.08 sorted by BIC. The 4-factor model had less interpretability than the 3-factor model. However, the maximum likelihood and minimum chi-squarefactoring methods of the 3-factor model did not suit the data [187, 188].

	nfactors	rotation	fm	tli	rmsea	bic
1	3	varimax	ml	0.98	0.05	-151.18
2	3	varimax	minchi	0.97	0.06	-145.35
3	4	varimax	ml	1.02	0.00	-130.86
4	4	varimax	minchi	1.02	0.00	-130.85
5	4	varimax	pa	1.01	0.00	-128.45
6	4	varimax	minres	1.01	0.00	-128.33
7	4	varimax	uls	1.01	0.00	-128.33
8	4	varimax	ols	1.01	0.00	-128.26
9	4	varimax	alpha	1.00	0.00	-125.00
10	4	varimax	old.min	0.99	0.03	-122.20
11	4	varimax	minrank	0.99	0.04	-120.50

for data analysis (PA3).

Table 2.3: Factor loadings, communality, uniqueness, and complexity based on minimum sample size weighted chi square factoring method with varimax rotation and tenBerge scores. The three factors are: programming experience (PA1), programming for data analysis (PA2), and solving technical problems (PA3). Loadings < 0.5 are suppressed. Loadings $\geq$ to 0.6 were used for Cronbah's α.

	item	PA1	PA2	PA3	Communality	Uniqueness	Complexity
1	Q3.4	0.811			10.34	0.17	1.54
2	Q3.3	0.799			10.34	0.13	1.75
3	Q7.2_2	0.793			10.34	0.13	1.78
4	Q3.5	0.774			10.34	0.16	1.82
5	Q3.1	0.729			10.34	0.17	2.12
6	Q7.2_5	0.695			10.34	0.25	2.08
7	Q5.2	0.682			10.34	0.35	1.73
8	Q4.3				10.34	0.56	2.36
9	Q7.2_7		0.944		10.34	0.02	1.19
10	Q7.2_6		0.87		10.34	0.07	1.45
11	Q4.2			0.718	10.34	0.40	1.32
12	Q7.2_3			0.64	10.34	0.31	2.26
13	Q6.1			0.538	10.34	0.60	1.68
14	Q7.2_4	0.506		0.509	10.34	0.33	2.85

Cronbah's Alpha

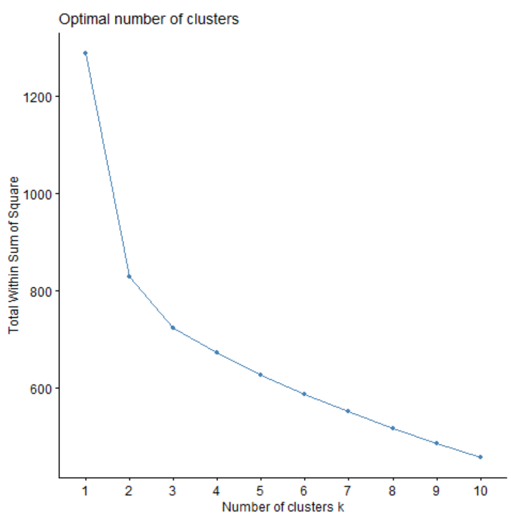
Internal consistency was measured with Cronbah's alpha. A separate Alpha was calculated from each of the factor analysis results. Questions that had a loading above 0.6 was used for the alpha calculation. The 0.6 cutoff was mainly selected to make sure all the major question groups in the survey was represented and each factor had more than 1 question. The programming experience subscale (PA1) consisted of 7 items ($\alpha = 0.96$), the programming for data analysis (PA2) consisted of 2 items ($\alpha = 0.98$), and the solving technical problems subscale (PA3) consisted of 2 items ($\alpha = 0.75$).

2.3.3 Clustering and Persona Identification

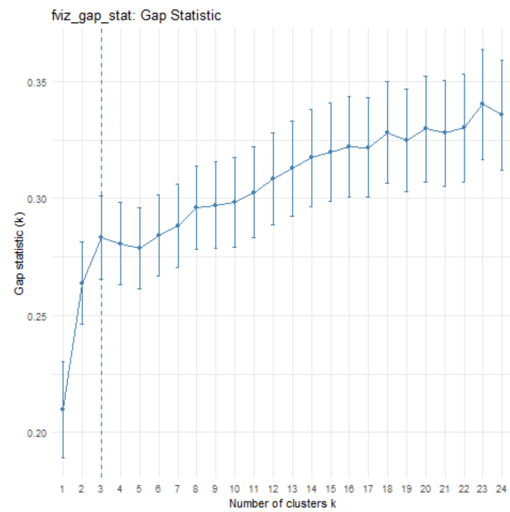
Hierarchical clustering was used on all 23 survey items to identify personas. Clusters were combined with survey demographic and item responses to create the learner personas. Questions with high factor loadings along with demographic results were used to guide the clustering interpretations (Figure 2.6).

Clustering

The agglomerative coefficient was used to select Ward's clustering method (0.88) over average (0.54), single (0.37), and complete (0.72). The total within-cluster sum of squares was used to generate an elbow plot (Figure 2.5a) and gap statistic for different numbers of clusters (Figure 2.5b). These results were were used to determine that three (3) clusters were the most optimal number of clusters. Three (3) clusters were and also the most interpretable (Figure 2.5). The dendrogram of the three (3) clusters are shown in Supplemental Figure 2.10.



(a) Elbow plot showing the total within-cluster sum of squares (y-axis) for increasing number of clusters, k (x-axis).



(b) Gap statistic comparing intra-cluster variation as compared to a distribution with no clustering (y-axis) for different number of clusters, k (x-axis).

Figure 2.5: Determining the number of optimal clusters for the persona survey data. Each cluster would become a learner persona. (2.5a) shows the elbow plot for the data and suggests that the optimal number of clusters is between $k = 2$ and $k = 4$. (2.5b) shows the gap statistic for the data. The dotted line at $k = 3$ shows the optimal number of k clusters. The three (3) cluster model had the best interpretation from the survey data. Each cluster became its own learner persona.

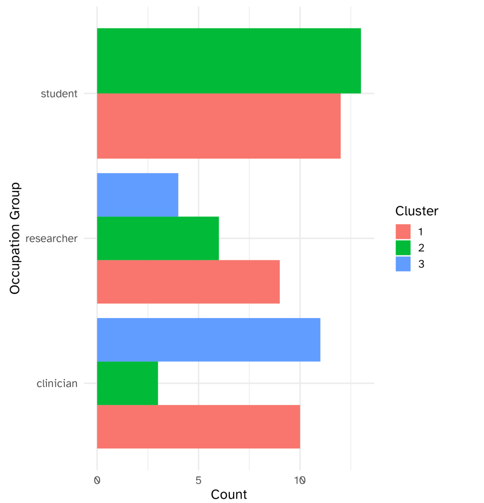
Persona Identification

Results from clustering were incorporated into the survey results in order to identify learner personas (Figure 2.6). The highest loaded item in each factor (Figures 2.6b, 2.6c, 2.6d) and the occupations demographic breakdowns (Figure 2.6a) give an overview of the clusters.

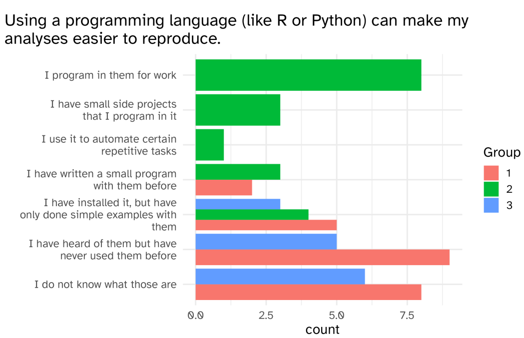
Clinicians (the target audience) were primarily in Groups 1 and 3. There were no students in Group 3. Groups 1 and 2 were spread out across the 3 occupations, however, Group 2 had the fewest number of clinicians (Figure 2.6a). Group 3 were either clinicians or academics, with most of the Group 3 occupations being a clinician. Original ungrouped occupation breakdown are shown in Supplemental Figure 2.8.

Looking at programming experience (Q3.4, Figure 2.6b), Group 2 had the most amount of experience and Groups 1 and 3 were predominately "never" programmers. Group 2 had the most number of respondents strongly agreeing to the statement that using a programming language can make an analysis easier to reproduce (Q 7.2-7, Figure 2.6c). Groups 1 and 3 are mostly neutral towards the statement and are the only respondents who disagree and strongly disagree with the statement. Looking at solving technical questions (Q4.2, Figure 2.6d).

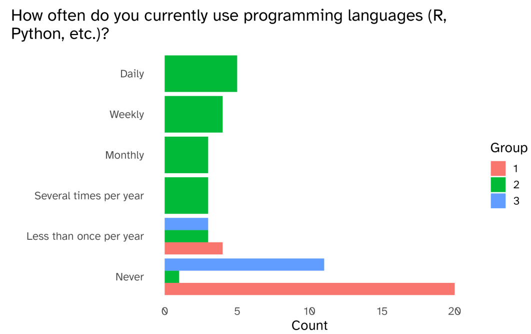
The 2-cluster results did not provide enough details between our participants. The major splits were the student group (Group 2 in the 3 cluster results) and everyone else. In the 4-cluster results (Supplemental Figure 2.11), the additional split occurred in the student group. This split was difficult to interpret and the new cluster was unable to be distinguished from the other split group.



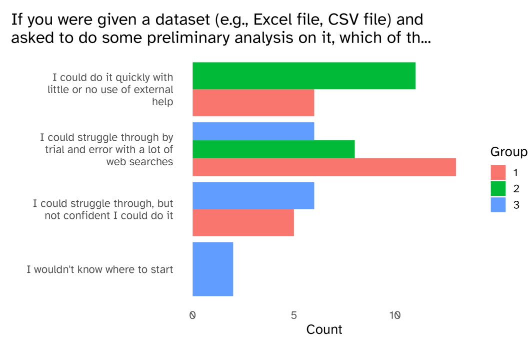
(a) Grouped demographics for 3 clusters.



(c) Clustered responses for Q7.2-7: Programming for analysis factor. "Using a programming language (like R or Python) can make my analyses easier to reproduce."



(b) Clustered responses for Q3.4: Programming experience factor. "How often do you currently use programming languages (R Python, etc.)?".



(d) Clustered responses for Q4.2: Solving technical problems factor. "If you were given a dataset (e.g., Excel file, CSV file) and asked to do some preliminary analysis on it, which of these best describe how easily you can accomplish the task?"

Figure 2.6: Survey responses for each of the 3 clusters. (2.6a) Group 1 and 3 had common demographics, However Group 1 also included students. Group 3 had the most number of clinicians. Group 2 were predominantly students. (2.6b) Group 2 are more consistent users of programming languages than Group 1 and 3. (2.6c) Results show the likert table responses for each of the 3 groups. Group 2 feel programming will make analyses easier to reproduce than Group 1 and 3. (2.6d) Group 3 were the only group of people who did not how to start an analysis. Group 2 had the most number of responses for being able to do the analysis with minimal help. These results were similar to the question that asked a more specific statistics related question (Supplemental Figure 2.12)

Persona Creation

Using the clustering data (Supplemental Figure 2.10) with the survey results, three (3) learner personas were created: Ash Academic (Group 1), Samir Student (Group 2), Claire Clinician (Group 3). Each persona had 4 components: (1) background, (2) relevant prior knowledge or experience, (3) perception of needs, and (4) special considerations [70, 88, 89]. The results from the survey were used to write out the relevant prior knowledge or experience and perception of needs sections. The background and special considerations were written to make each learner persona more complete [15]. Full write-ups with each persona's background, relevant prior knowledge or experience, perception of needs, and special considerations can be found in Supplemental 2.5.3.

2.4 Discussion

There were a few existing parameters heading into this study. The target audience are: (1) adult working professionals, (2) work as a clinician and/or do research in the biomedical field, and (3) use Excel as the main, if not only, program for data work. These parameters put in constraints into the lesson development phase: (1) The lesson needs to be modular and sections need to be taught within 1 hour blocks. This can account for busy work schedules and can possibly be given during break times (e.g., lunch). (2) Written lesson materials need to be provided online for asynchronous learning and can be used as reference material later on. (3) Classes will be recorded and posted online for the same reason written materials are posted. (4) Data examples must be medically or health-related. (5) A separate lesson needs to be given about working with data in Excel and setting up spreadsheet data. (6) An example of loading an Excel spreadsheet into the programming language so learners can import any existing datasets. A separate study will build on these concepts to create concrete learning objectives (LOs) for the lesson materials.

These studies are learner-focused and does not focus on the skills needed to properly present the materials as an instructor. Teaching knowledge does play a factor in how materials are presented and how the instructor handles questions from the class. Assessing content knowledge and pedagogical content knowledge of the instructor was not performed [13], but since all the workshops will be taught by the same instructor, those components were consistent throughout out the future workshops that assess the learning materials. The Carpentries Instructor Training materials is a good reference to learn more about evidence-based teaching practices [93]. This study will provide a foundation for future instructional material creators identify core data literacy components that need to be taught.

2.4.1 Identification of Biomedical Data Science Learner Personas Informs Curriculum Design

There was a relatively even balance between the 3 main occupation groups: students, researchers, and clinicians. The student group mainly clustered into their own persona, which also happened to be the most experienced persona. The programming experience can be because higher education is beginning to incorporate data science courses and into existing programs, from students learning skills doing their own research. If the more experienced students come from a self-taught programming background, then a more formalized class can solidify their existing mental model, and build the way for more connections to progress their learning.

Gender, race, and ethnicity demographic information was collected for these surveys. However, these results were not used for the persona clustering. Future iterations of this survey should not include these demographic questions since they are not relevant to creating the personas.

The survey had 23 items with 57 responses for analysis. We performed a correlation analysis to reduce the number of items down to 14, for factor analysis but these values may change with more respondents. The main motivation to drop items from the factor analysis was due to the low number of responses. A general rule of thumb is to have 10 observations per item [189–191], and the correlation analysis to select items was one way to mitigate the effects of having a low number of observations. The survey questions, data, and results of this study are published for future reseachers and studies to expand the responses have more power in validating the survey.

One of the assumptions for the ML factoring method is the items need to be normally distributed. Our analysis showed that our data was not normal, and many of the other factoring methods did not seem relevant for our data. The PA factoring method was selected because it does not assume the normality of data. This method better suited the data distribution and while it did not meet every cutoff value, this was enough to show some validity for the survey and can be used in a future study to collect more data. The Cronbah's alpha results showed that the survey responses were consistent and can be a useful tool for researchers moving forward.

2.4.2 Creation of Biomedical Data Science Learner Personas

The demographic information was a main contributor on how the personas were named (Figure 2.6a). The target audience for creating learning materials were clinicians. This group also clustered into their own group from the survey responses. It is possible that the names and number of personas will change with more responses. The benefits of creating these explicit learner personas is so future educators can have a more concrete understanding of their audience. These personas can be used to cater and focus learning materials and better serve learners' needs. The overall differences between the personas are listed in Table 2.4. In theory, each combination of differences in Table 2.4 would be its own learner persona, but not every combination would require a different set of learner needs. Also trying to create a persona for every possible learner would be impractical [15].

Table 2.4: How each persona differs from one another. In terms of data science and data literacy skills. Results come from persona survey results and are based off the representatice question from factor analysis. The "Programming", "Data Programming", and "Technical" columns represent the relative results from the programming experience (Q3.4 Figure 2.6b), programming for analysis (Q7.2-7 Figure 2.6c), and solving technical problems (Q4.2 2.6d) questions that had the highest loadings in each factor. The "Statistics" and "Excel" columns represesnt the relative responses from a statistics question and Excel proficency question that the survey also asked in the survey (Q6.2 and Q4.1 in Supplemental Figures 2.12 and 2.13, respectively).

Group	Programming	Data Programming	Technical	Statistics	Excel
1	Low	Medium	Medium	Medium	Medium
2	High	High	High	High	Medium
3	Low	Low	Low	Low	Medium

Samir Student (Group 2)

Samir Student (Group 2) is indicative of many introductory programming classes. They serve as the more experienced individuals who have previously seen the materials or have experience with data programming (Table 2.4). These students may find some of the materials too simple for their needs, and can cause a distraction in the classroom by asking questions that are too complicated for the intended audience to understand. There are a few ways to manage these students in the classroom. Overly technical questions outside the scope of the lesson materials can be answered during a break or after the class as to not confuse the rest of the learners. These more experienced learners can also provide help to other learners around them. This will also reinforce their own understanding of the materials being covered, while keeping them engaged in the class. In a virtual setting, these students can provide help and answer questions in the text chat [70, 93].

This bimodal learner experience is also confirmed when comparing the Where the clusters are split. Samir Student was the first clustering split with survey responses showing they had the most experience with data science concepts. The 3-cluster model partitioned the less experienced programmers (Supplemental Figure 2.10).

Making the lessons modular and publicly available gives these types of learners the chance to judge if the materials are suited for their needs and experience levels. More advanded learners can self-learn at their own pace and use the learning materials as a reference.

Clare Clinician (Group 3)

Groups 1 and 3 are mainly "never" programmers. What distinguishes group 3 from the rest, is they are also not confident in their ability to perform a statistical analysis, and may not agree that programming makes analyses easier to reproduce (Table 2.4). This group was also predominantly clinicians.

Ash Academic (Group 1)

Group 1 was also mainly in the "never" programmers group. This group was also a catch-all for all the other occupations. Group 1 is more confident in their abilities to do statistical analysis. They also lean slightly toward programming languages being helpful for reproducible analysis.

2.4.3 More Data for Personas

The persona survey was given in 2 waves. The results in the first wave had 4 clusters: Expert, Clinician, Academic, and Student. The expert and student clusters combined into a single group with the addition of wave 2 data. As more data gets collected and the persona survey gets distributed to more instructors, a more accurate stratification of the responses can be created and more accurately identify learner personas and needs. One of the main reasons the data from this study is published is for future research to add to the results and perform the same analysis again with more observations. This would help create a more generalizable set of personas and potentially identify more specific learning needs.

2.4.4 Domain-Specific Learner Personal Survey Validation

There are four (4) main benefits to using personas: (1) make assumptions about users explicit, (2) place the focus on specific types of users rather than on all possible users, (3) Help make better decisions by limiting choices (4) Engage the product design and development team [15, 87]. As educators, these learner personas provide a starting point to curate, create, and design domain-specific learning materials for the biomedical sciences. The primary user group of Clare Clinician can be used to focus on content and examples. The personas created in this study are still in their initial conception phase, but they should have enough detail to begin planning lesson materials. Even if the personas are “wrong”, the lesson materials will still be consistent for learners [15]. The survey questions from this study are also published along with the responses as a means to pool together future research attempts to validate this survey. Responses from the survey can not only help with creating more targeted learning materials, but also help balance the prior knowledge experience of learners in the classroom.

2.4.5 Propagating Results to Other Domains

The results from this study are highly focused for the medical and biomedical sciences. When designing domain-specific learning curriculum, it is possible to systematically and objectively collect information on the learner's prior relevant knowledge to begin the process of creating a targeted learning curriculum. The personas from this study will help future instructors in the biomedical sciences focus on what their learners need. Several suggestions where the survey, data, results, and personas can be hosed are The National Library of Medicine (NLM), The National Institutes of Health (NIH), Patient Centered Outcomes Research Institute (PCORI), Agency for Healthcare Research and Quality (AHRQ), or the The National Science Foundation (NSF) [111].

The process and methods used from this study can also be expanded to to help improve data science education in other domains. The structure of writing a learner persona in this study was adapted from existing personas used by RStudio, PBC and The Carpentries [88, 89]. This shows that even though domains may differ, there are many commonalities between data science learners across many disciplines. The personas, data, and results from this study can used in other disciplines and hosed in their respective foundation or research agency to share with other potential instructors in their domain.

2.5 Supplemental

Supplemental materials for the "Identification of Biomedical Data Science Learner Persons: Implications and Lessons Learned for Domain-Specific Data Science Curriculum".

2.5.1 Pre-Workshop Student Self-Assessment Survey (Persona Survey) Questions

The surveys can be downloaded from the GitHub URL that holds the IRB proposal for the study: https://github.com/chendaniely/dissertation-irb/tree/master/irb-20-537-data_science_workshops/survey. The numbers in parenthesis correspond to the numeric code in the original downloaded results.

Demographics

Q2.2 Please create a unique identifier. This unique identifier will be used for long-term assessment but keep your personal information anonymous.

To create an identifier type in: Number of siblings (as numeric) + First two letters of the city you were born in (lowercase) + First three letters of your current street (lowercase).

E.g., (Sherlock Homes has **1** brother, was born in **Po**rsmouth, and lives on **Bac**ker Street - **1pobac**)

Q2.5 What is your current occupation/career stage (select all that apply).

- DO/MD (1)
- RN/PA (2)
- Academic (3)
- Analyst (4)
- Student (Masters e.g., MPH) (5)

- Student (MD/DO) (6)
- Student (Nurse, PA) (7)
- Student (Graduate) (8)
- Student (Undergraduate) (9)
- iTHRIV Scholar (11)
- Other, please describe (10)

Q2.6 What operating system will be on the computer you are using at the workshop or to participate in the online materials?

- Windows (1)
- macOS (2)
- Linux (3)
- Not sure (4)

Programming Experience

Q3.1 In general, which of these best describes your experience with programming?

- I have none (1)
- I took some programming related class in the past but have not used it since (5)
- I have written a few lines now and again (2)

- I have written programs for my own use that are a couple of pages long (3)
- I have written and maintained larger pieces of software (4)

Q3.2 What programming languages have you used in the past? Select all that apply.

- VBA (Visual Basic for Applications) (1)
- Python (2)
- R (3)
- Perl (4)
- Matlab (5)
- Javascript (6)
- C (7)
- C++ (8)
- Fortran (9)
- Other, please list (10)

Q3.3 How familiar are you with interactive programming languages like Python or R?

- I do not know what those are (1)
- I have heard of them but have never used them before (2)
- I have installed it, but have only done simple examples with them (3)

- I have written a small program with them before (4)
- I use it to automate certain repetitive tasks (5)
- I have small side projects that I program in it (6)
- I program in them for work (7)

Q3.4 How often do you currently use programming languages (R, Python, etc.)?

- Never (1)
- Less than once per year (2)
- Several times per year (3)
- Monthly (4)
- Weekly (5)
- Daily (6)

Q3.5 Which of these best describes how easily you could write a program (in any language) to find the largest number in a list?

- I wouldn't know where to start (1)
- I could struggle through, but not confident I could do it (4)
- I could struggle through by trial and error with a lot of web searches (2)
- I could do it quickly with little or no use of external help (3)

Q3.6 How often do you currently use a specialized software with a point-and-click graphical user interface on your own (e.g., for statistical analysis: SPSS, SAS, …; for Geospatial analysis: ArcGIS, QGIS, … ; for Genomics analysis: Geneious, …)?

- Never (1)
- Less than once per year (2)
- Several times per year (3)
- Monthly (4)
- Weekly (5)
- Daily (6)

3.7 How often do you currently use Databases (SQL, Access, etc.)

- Never (1)
- Less than once per year (2)
- Several times per year (3)
- Monthly (4)
- Weekly (5)
- Daily (6)

Data Cleaning and Processing Experience

4.1 How familiar are you with Microsoft Excel?

- I have never used it, or I have tried it but can't really do anything with it. (1)
- I have used it as an electronic todo list and planner putting schedules and task deadlines in a single place (2)
- I've used it to store datasets and able to calculate basic aggregate values, such as mean and sums (3)
- I've used data aggregation, pivot tables, formulas, and plotting feature to understand how my data breaks down. (4)
- I've coded up VBA macros and made VLOOKUP calls integrating multiple sheets for a simulation task (5)

4.2 If you were given a dataset (e.g., Excel file, CSV file) and asked to do some preliminary analysis on it, which of these best describe how easily you can accomplish the task?

- I wouldn't know where to start (1)
- I could struggle through, but not confident I could do it (4)
- I could struggle through by trial and error with a lot of web searches (2)
- I could do it quickly with little or no use of external help (3)

Q4.3 Are you familiar with the term “tidy data”?

- I have never heard of the term (1)
- I have heard of it but don’t remember what it is. (2)
- I have some idea of what it is, but am not too clear (3)
- I know what it is and could explain what it pertains to (4)

Q4.4 Do you know what “long” and “wide” data are?

- I have never heard of the term (1)
- I have heard of it but don’t remember what it is. (2)
- I have some idea of what it is, but am not too clear (3)
- I know what it is and could explain what it pertains to (4)

Project and Data Management

Q5.1 Please rate your level of satisfaction with your current data management and analysis workflow (e.g. how you collect, organize, store and analyze your data).

- Very unsatisfied (1)
- Unsatisfied (2)
- Neutral (3)
- Satisfied (4)

- Very satisfied (5)
- Not sure (6)
- Not applicable (7)
- Never thought about this (8)

Q5.2 Which of the following best describes how do you manage your data and analysis?

- I don't do data and/or analysis work (1)
- My data and analysis are all in excel files, possibly with multiple sheets. (2)
- I work on carefully time-stamped excel files for my version control and analysis (3)
- I use some programming language to load in my data sets for analysis, but sometimes modify my original data files when cleaning the data (4)
- I hold my original data sacred, and only work on it from another program and save out intermediate and final data projects as separate files (5)
- I have a very specific project structure where data and analysis are kept in separate areas and have a version control system (e.g., Git, SVN) (6)
- I have version controlled project templates along with build scripts (e.g., Makefile) to reproduce various aspects of the analysis (7)

Statistics

Q6.1 If you were given a dataset containing 2 cholesterol treatment options (drug and placebo), patients' baseline cholesterol values, and cholesterol values 4 weeks after treatment has started, would you know how to conduct a statistical analysis to see if there is a difference between the 2 groups? Any type of model will suffice.

- I wouldn't know where to start (1)
- I could struggle through, but not confident I could do it (4)
- I could struggle through by trial and error with a lot of web searches (2)
- I could do it quickly with little or no use of external help (3)

Q6.2 If you were given a dataset containing an individual's smoking status (binary variable) and whether or not they have hypertension (binary variable), would you know how to conduct a statistical analysis to see if smoking has an increased relative risk or odds of hypertension? Any type of model will suffice.

- I wouldn't know where to start (1)
- I could struggle through, but not confident I could do it (4)
- I could struggle through by trial and error with a lot of web searches (2)
- I could do it quickly with little or no use of external help (3)

Q6.3 If you were given a dataset comparing different treatment methods for cancer patients. Would you know how to conduct an analysis to see which treatment had a higher survival rate of patients?

- I wouldn't know where to start (1)
- I could struggle through, but not confident I could do it (4)

- I could struggle through by trial and error with a lot of web searches (2)
- I could do it quickly with little or no use of external help (3)

Q6.4 Are you familiar with the term ”dummy variable”? It is sometimes also called ”one-hot encoding”.

- I have never heard of the term (1)
- I have heard of it but don’t remember what it is (2)
- I have some idea of what it is, but am not too clear (3)
- I know what it is and could explain what it pertains to (4)

Workshop Framing and Motivation

Q7.1 Why are you participating in this workshop? Please check all that apply.

- To learn new skills (1)
- To refresh or review my skills (2)
- To learn skills that I can apply to my current work (3)
- To learn skills that I can apply to my work in the future (4)
- To learn skills that will help me get a job or a promotion (5)
- As a requirement for my program or current position (6)

7.2 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- I believe having access to the original, raw data is important to be able to repeat an analysis. (1)
- I can write a small program, script, or macro to address a problem in my own work. (2)
- I know how to search for answers to my technical questions online. (3)
- While working on a programming project, if I got stuck, I can find ways of overcoming the problem. (4)
- I am confident in my ability to make use of programming software to work with data. (5)
- Using a programming language (like R or Python) can make my analyses easier to reproduce. (6)
- Using a programming language (like R or Python) can make me more efficient at working with data. (7)

7.3 Please share what you most hope to learn from participating in this workshop and/or workshop series.

7.4 What do you want to know or be able to do after this workshop (or series of sessions) that you don't know or can't do right now?

2.5.2 Pre-Workshop Student Self-Assessment Survey (Persona Survey) Supplemental Factor Analysis Results

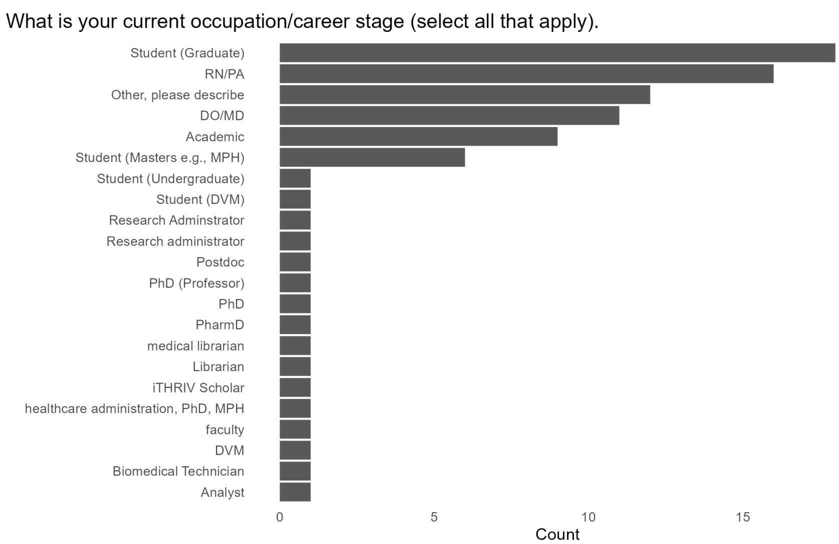


Figure 2.7: Self reported current occupation and career stage. Respondents are able to select multiple options and also write in their own choices.

2.5.3 Learner Personas

Three (3) learner personas were identified int he analysis. The initial wave of results had four (4) learner personas, but as more data was collected, Patricia Programmer (already competent programmers) was was dropped as the results were combined with Samir Student.

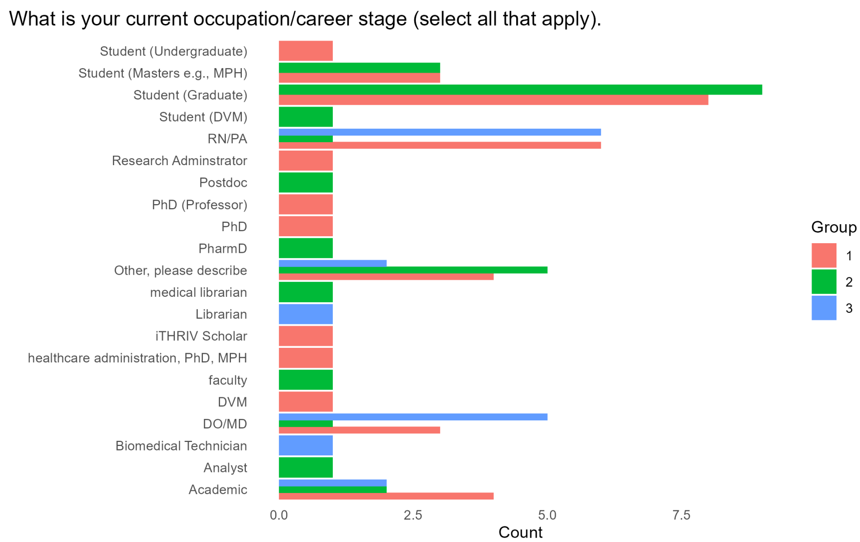


Figure 2.8: Self reported current occupation and career stage by clustering group. Respondents are able to select multiple options and also write in their own choices.

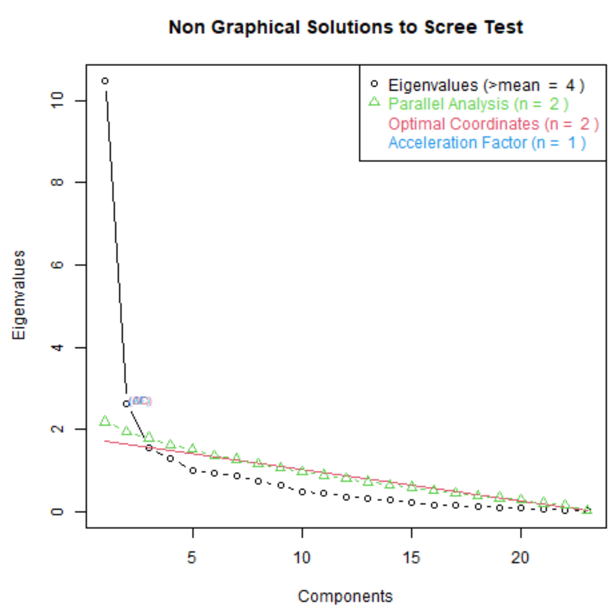


Figure 2.9: i
Scree plot for factor analysis showing the number of components (x) and the eigenvalues (y) for all 23 survey items. The figure suggests exploring 2 to 4 factor models.

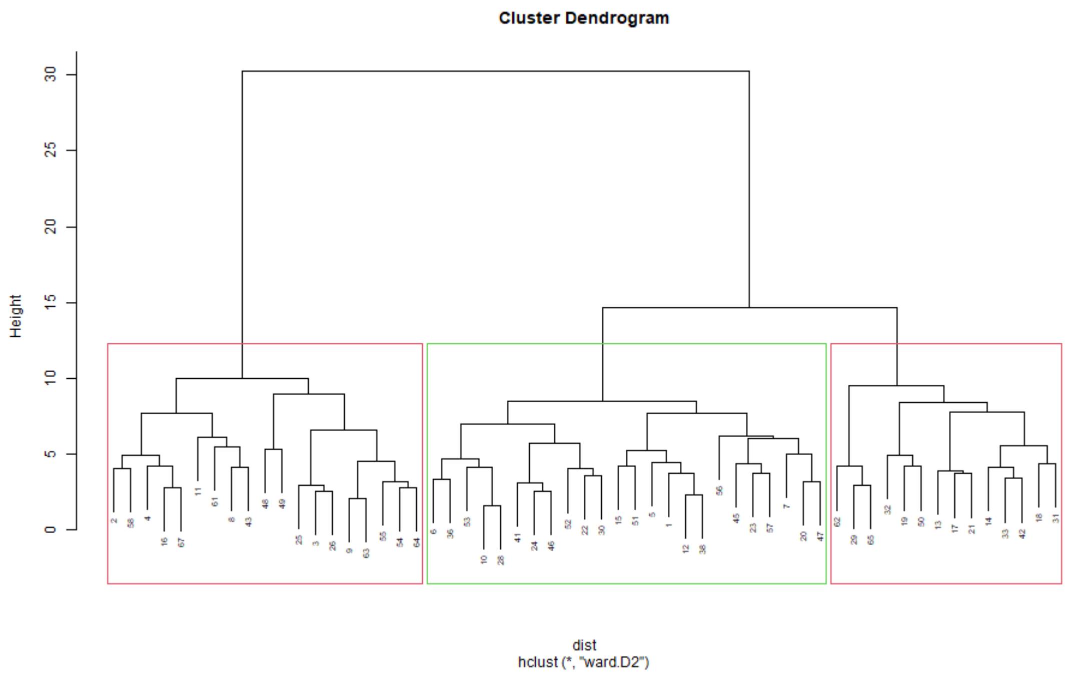


Figure 2.10: Dendrogram of the three (3) learner persona clusters. The clusters were combined with the survey responses to identify the learner personas. From left to right, the clusters are: Samir Student (Group 2), Ash Academic (Group 1), and Clare Clinician (Group 3).

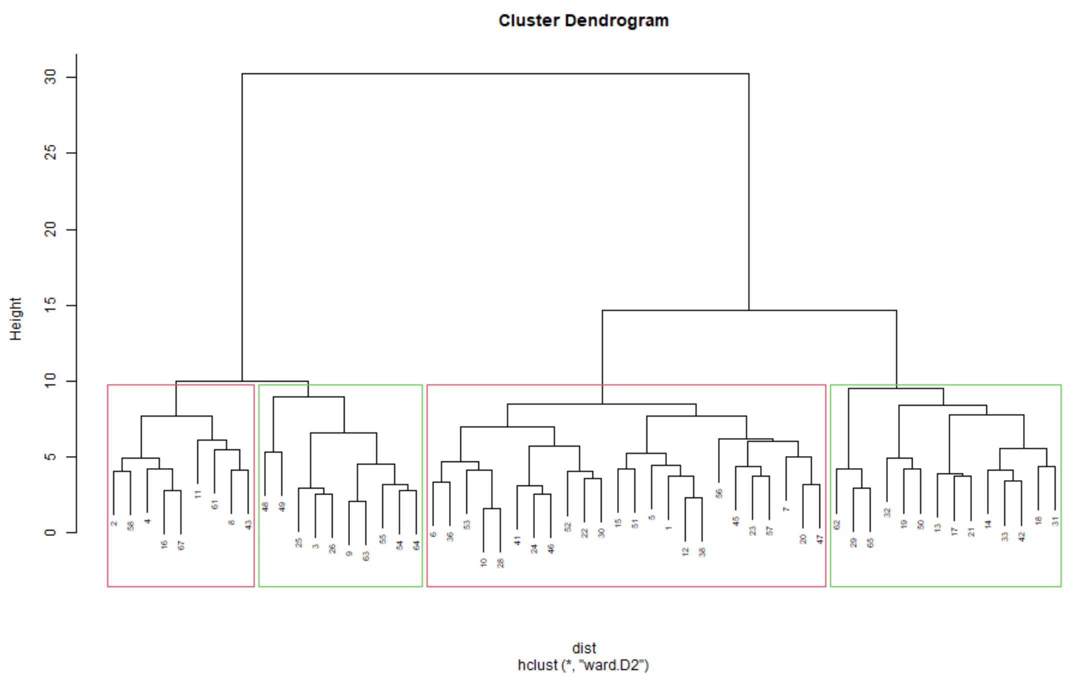


Figure 2.11: Dendrogram showing 4 clusters. Looking at the differences in how the clusters were split from the 3-cluster model (Figure 2.10), the same group that eventually was named Samir Student (Group 2) was split into another cluster. This split was hard to interpret, so the 3-cluster model was used for creating the learner personas.

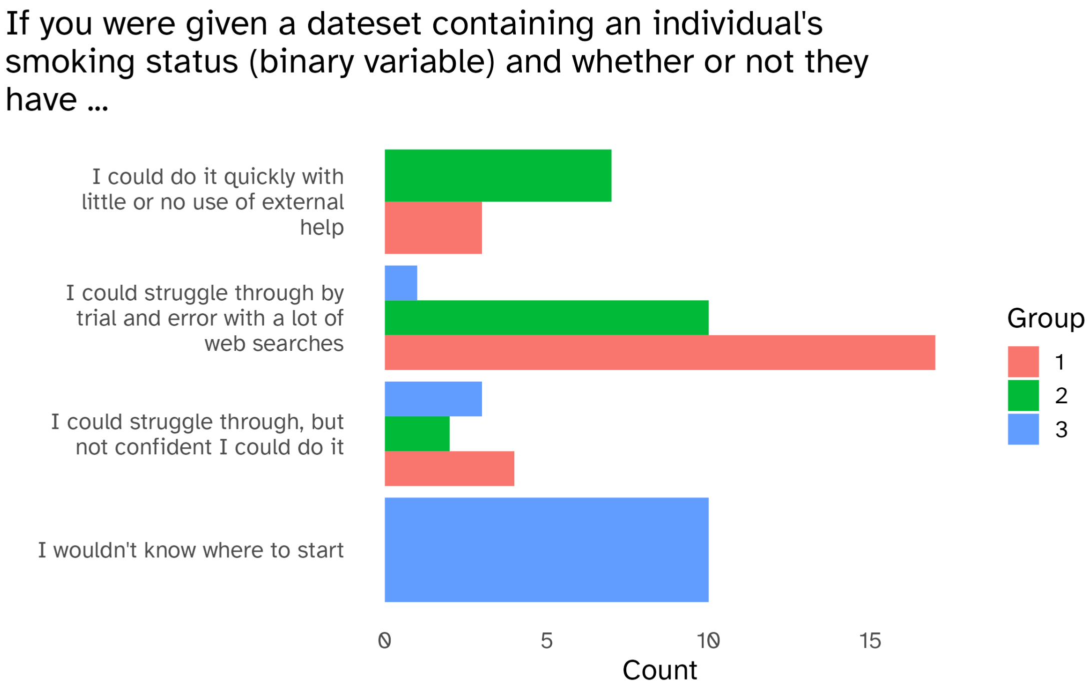


Figure 2.12: Participants were asked if they were able to perform an analysis with a binary outcome. The full question asked "If you were given a dataset containing an individual's smoking status (binary variable) and whether or not they have hypertension (binary variable), would you know how to conduct a statistical analysis to see if smoking has an increased relative risk or odds of hypertension? Any type of model will suffice." Typically, logistic regression would be performed for this kind of problem, but the authors were less concerned with identifying a particular analysis, and more focused if participants could perform any analysis. Group 3 were the only group that would not know how to start this kind of analysis, and also leaned towards struggling through this kind of analysis. Groups 1 and 2 leaned towards being able to perform this analysis task, with Group 2 having the most responses for performing the task with minimal external help. Group 2 also had a high number of responses for completing the task with some struggle, but Group 1 had the most frequent number of responses.

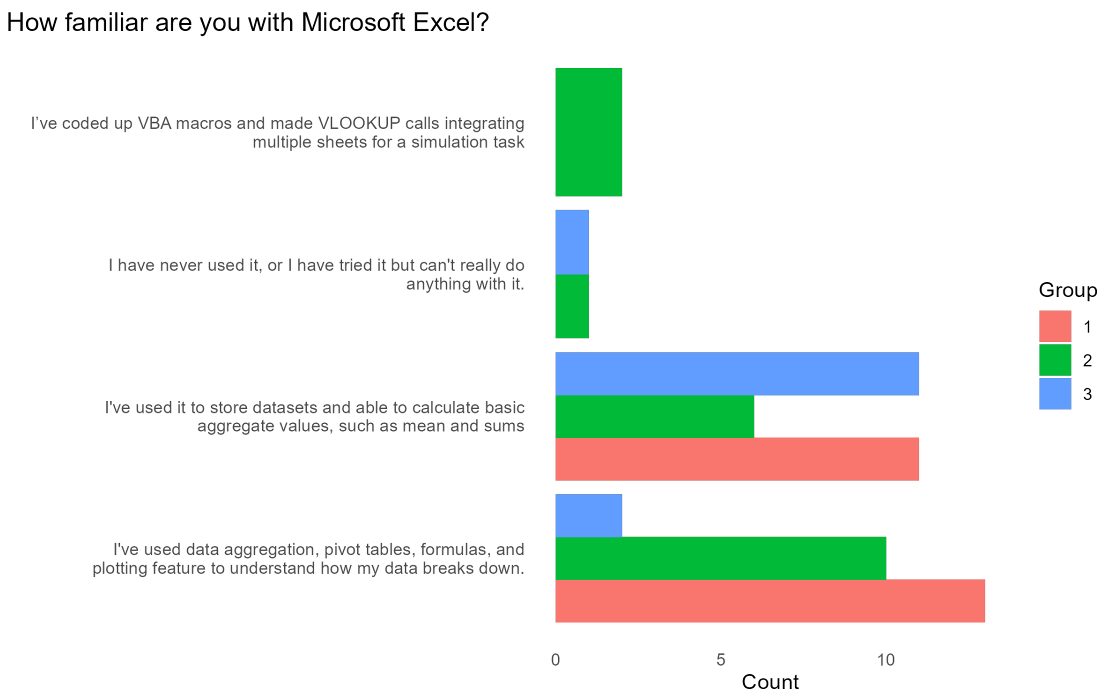


Figure 2.13: Participants were asked about their familiarity with Microsoft Excel. The full question asked "How familiar are you with Microsoft Excel"? The responses showed that many of the respondents how how to use Excel for basic data tasks and calculations, and very few of the respondents were highly proficient in using Excel for data tasks.

2.5.4 Ash Academic

Alex, a professor of bioinformatics, studies molecular dynamics of proteins and protein-protein interactions. They are also responsible for teaching an introductory research class to 100 freshman and sophomore students every year. They also run a data consulting service at their university, providing support for data-related challenges to research and instructors from any discipline.

Their students complain that intro courses in programming are too theoretical and require more programming knowledge than they have. Many students in the department also cannot register for similar classes. Alex, has 10s of students working for them in computational molecular dynamics simulations and other data analytics projects.

Relevant prior knowledge or experience

Alex performs their research using a combination of Excel spreadsheets and specialized software. But the are starting to move to using R or Python (which they taught themselves during a sabbatical). They have never taken a formal programming course, and suffers from impostor syndrome in discussions about programming. Alex would like to learn more about how programming can help their research and keep up with the tools their students are learning in class. They also need resources to give to their student researches to help work on projects in the lab.

Perception of needs

Alex needs workshops and how-to guides so they can better block off time and apply new skills to their research. They would like ready-to-use lesson material that they can learn from and also be be remixed for their students. The materials should be at in introductory level demystify jargon (e.g., what is “tidy data”?). Alex also wants to be able to teach the same materials they use for their own research to amortize learning costs and stay in practice.

Special considerations

Alex wants to provide technical training to students, but does not have the actual time to teach all the relevant skills. As a person in STEM, they typically find themselves isolated and alone when taking formal technical classes and is scared to appear ignorant, and are reluctant to speak up and ask questions.

2.5.5 Clare Clinician

Clare has spent the last 6 years working in the Cardiothroasic ICU in a large medical hospital system. They see the impacts of data science in their day-to-day job and want to learn these skills to better understand their patients. However, nothing makes sense when trying to learn it on their own, and there are few formal opportunities to learn that fit into their work schedule. Clare has always been a good student and always excelled at things they tried to learn; they are hard on themselves when struggling to learn a new skill and would rather place blame on the long hours at work than having their peers know they could use assistance.

2.5.6 Relevant prior knowledge or experience

Clare keeps up with medical research, but has little to no experience in doing medical research. They use Excel for non-data related tasks (e.g., making lists), or manually inputting patient data into spreadsheets for chart reviews. Clare wants to be able to collect and manage data as well as learn about the process behind data analysis to perform their own analysis and study one day.

2.5.7 Perception of needs

Clare wants self-paced tutorials with practice exercises that uses health related data they can directly relate to. Since programming and data analysis are relatively new to them, Since they are new to programming, a class that provides overviews to orient them to programming and ask questions along with a community forum where they can ask for help asynchronously is needed. The introductory tutorials should contain detailed visualizations on how to install, setup, and use the tools they will learn on datasets that they can relate to. They do not mind having references to other materials, but do want a single learning path.

2.5.8 Special considerations

Clare is a single parent who juggle their time at work and at home who are strapped for time to learn a new skill. They are typically only able to take additional classes early in the morning or late at night, outside of "normal" work hours.

2.5.9 Samir Student

Samir is a graduate student in a bioinformatics program. They worked in a wet lab during their undergraduate days studying neuroscience. These days, Samir is doing more computational work and starting to use programming based tools to look at protein structures with Ash Academic. They've taken a few classes that had had programming based homework assignments and projects, but the lectures themselves were mostly around theory, and many of the programming skills were self-taught.

Relevant prior knowledge or experience

Samir is fairly proficient in Excel, does works with spreadsheets regularly, and knows how to load up Excel spreadsheets into R and do basic data processing and analysis. However, they do not have much programming practice outside of classroom homeworks and projects. They spend a lot of their time on StackOverflow copying and pasting code so they don't consider themselves a "real programmer". They have no problem getting their work done, but usually involves a lot of googling to eventually get the solution.

Perception of Needs

Samir wants a formal workshop and reference materials that can be used to build a good foundation of the programming skills they were never taught. They want a better understanding of the terminology and jargon used in data science so they have the vocabulary to search for and understand solutions posted online. They are also looking for a community to help in their growth as a student in this domain, and also materials they can pass on to the more junior researchers in the lab.

Special considerations

Samir has a disability (vision, hearing, attention, etc) that make it difficult to learn in "traditional" classroom settings. They typically need a recording of the lectures so they can rewatch lectures at their own pace.

Chapter 3

Assessing the Efficacy of Domain-Specific Data Science Curriculum in the Biomedical Sciences: How Learner Personas Can Guide Educational Needs in the Short-Term and Long-Term

Abstract

The demand for clinical informatics training is outpacing the supply and opportunities for training. Even though there is an abundance of data science training materials, there are not enough domain-specific data science materials in the biomedical sciences. In an effort to plan on developing these domain-specific materials, a previous first identified learner personas which included background information, relevant prior knowledge, perception of needs, and special considerations of learners. These personas were used to create a set of learning materials using backward design to create a set of open access relevant domain-

specific learning materials for the biomedical sciences.

This study looks at the efficacy of these learning materials using a set of cross-sectional longitudinal surveys that track confidence in meeting learning objectives and answering a summative assessment question. 200 total workshop participants participated in 67 pre-workshop surveys, 43 post-workshop surveys, and 11 long-term workshop surveys. The study sees an improvement in learner's confidence in meeting learning objectives post-workshop, but in the long-term survey (at least 4 months out), confidence in meeting learning objectives and confidence to complete a summative assessment question that covered data loading, subsetting, saving, tidying, and model fitting were back to pre-workshop levels. This suggests that lesson materials may have an effect to learners in the short term, but more resources need to be created and provided in the long-term for learners.

The authors recommend that time and effort into creating domain-specific data science materials from scratch can be better served by creating more case-study and data examples to serve learners in the long term. Introductory materials can be created by remixing existing bodies of work, and efforts into creating relevant real-world examples can be used for formative assessment questions during a lesson. This curation of domain-specific exercises and case studies will be easier to maintain for the original authors as well as levering a broader community of practice for lesson content maintenance.

3.1 Introduction

There is a growing need for clinical informatics training, but the demand is exceeding the opportunities to learn the relevant skills to improve patient care [107, 108]. Some of the demand can be met with more marketing and publicly promoting existing resources [107]. However, increasing quantity, quality, and publicity would all need to be incorporated in meeting the growing demand [107, 111].

There are a plethora of introductory data science materials [35], as the field continues to grow and curriculum are adapted from K-12 to higher education [8, 67, 68, 71], the base of knowledge in the general population will increase, however, what is currently lacking is providing domain-specific data science curriculum for working professionals who have more immediate data science needs, e.g., those who work in medicine [34, 35, 111, 192, 193].

"How Learning Works" provides seven (7) principles of learning [69]: (1) Students' prior knowledge can help or hinder learning. (2) How students organize knowledge influences how they learn and apply what they know. (3) Students' motivation determines, directs, and sustains what they do to learn. (4) To develop mastery, students must acquire components skills, practice integrating them, and know when to apply what they have learned. (5) Goal-directed practice coupled with targeted feedback enhances the quality of students' learning. (6) Students' current level of development interacts with the social, emotional, and intellectual climate of the course to impact learning. (7) To become self-directed learners, students must learn to monitor and adjust their approaches to learning. Getting a sense of what learners know (principles 1 and 2) is crucial in targeting the correct learning content to them [69, 70, 93]. if we narrow the scope of potential learners to a specific domain, it will be easier to provide better and more applicable teaching examples to learners to help motivate them to learn (principle 3) [15, 87]. Using persona methodologies to create learner personas help give more concrete examples of what learners know and what their needs are. This can also help identify any special needs to make the learning environment more conducive to learning (principle 6) [15, 87, 88].

Since courses take place over a fixed period, spurring internal motivation to continue practicing and learning (principle 5) is a challenge when putting together a curriculum [69]. Applying and practicing component skills is a core component of developing mastery (principle 4). Metacognition (principle 7) is a higher-order skill and requires of self-reflection on the learner's end to adapt how they think and approach a problem. This is a skill that is often overlooked and neglected in many courses [69]. Since many other steps in the learning process can happen after the course time, a way to scale education is to identify and leverage learning communities and build a community of practice.

Creating domain-specific materials helps learners by showing more relevant examples [35, 69, 70, 93]. This helps with internal factors for motivation and aids in creating learning feedback loops. These are all components of creating self-directed learners [35, 69, 70, 93]. Previous work used a learner self-assessment survey (i.e., persona survey) to create learner personas (Chapter 2): Ash Academic, Clare Clinician, and Samir Student were the main groups of learners. Each group had varying amounts of programming experience, data programming experience, and confidence to search for and understand technical help on the internet. When we looked at overall responses to Excel usage and data literacy questions, we found that the vast majority of responses are familiar with basic data analysis features to understand their data, e.g., able to calculate aggregate descriptive statistics, pivot tables, formulas, and plotting (Supplemental Figure 2.13). Most of the respondents also have an Excel-centric data workflow, and only a fraction interact with data programmatically (Figure 2.6). These responses feed into how familiar respondents were to data literacy jargon, i.e., "long" and "wide" data, and "dummy variable" and "one-hot encoding" of variables.

The effectiveness of a curriculum is operationalized by creating a concrete set of measurable learning objectives (LOs). Bloom's taxonomy serves as a useful model to create LOs. However, creating the learning objectives is not enough. The Computing Curricula guidelines have moved away from knowledge-based teaching to competency-based teaching [65–67]. Knowledge-based teaching aims to identify learners' current amount of knowledge, and starts to expand on the existing knowledge base. However, this approach has led to gaps in applicable skills between the classroom and the workforce. Since, a more competency-based approach has been adapted, combining knowledge, skill, and disposition (i.e., what, how, and why) in order to have a deeper understanding and applicable set of skills [65–67].

The learner personas helped identify "tidy data" as a core component in the data science process, and all of the LOs stemmed from tidy data principles. Since our learners primarily use Excel, diving directly into loading data into a programming language would be too jarring of a transition. The curriculum was modeled off Data Carpentry lessons that put a spreadsheet lesson before going into programming with data. The spreadsheet lesson introduces tidy data principles without explicitly naming them, rather it talks about why some datasets are more difficult to work with, and has learners curate hypothetical pharmacokinetics (PK) study where different ways of entering data are discussed. The data is then used to import into a programming language to help with example continuity.

The learning objectives were used to assess learning material and workshop presentation efficacy, and provided longitudinal data with a pre-workshop, post-workshop, and long-term survey. These results could be compared to look at baseline data science and data literacy competencies, how they change after the workshop, and how much information is retained long-term. One key component after the workshop was to provide other learning communities for learners to join that can help continue the learning process. In order to motivate motivate learners to continue learning and work towards being self-directed learners, resources were provided to ask questions, other communities of practice, and other learners in the same biomedical domain.

3.2 Methods

A backward design approach was used to create the lesson materials. The overall steps of backward design are: (1) Identify your learners, (2) Plan out your lesson content, (3) Define overall goals, (4) Outline the course, and (5) Write a summary of the course [70]. The first step of identifying learners was done in a previous study where leaner personas were identified. This study uses the rest of the backward design approach to create and assess the efficacy of learning materials. The R programming language [164] along with serveral R packages were used to extract Qualtrics survey data for analysis [136, 186, 194–219].

3.2.1 Creating Learning Objectives

The initial persona survey showed that many of the potential learners who do not have existing data programming experience use Excel for basic calculations (Supplemental Figure 2.13). We used the information from the persona survey to create a roadmap of topics to get learners from no programming experience to being able to fit a logistic regression model. These learning objectives were used to write the lesson materials and used as a Likert scale question in the pre-workshop and post-workshop surveys.

3.2.2 Workshop Materials: ds4biomed

Results from the survey pre-workshop learner self-assessment survey (i.e., persona survey) were used to create the workshop materials. The survey results from the first round of respondents were used to create the initial learning materials. Prioritizing data literacy concepts was the main goal of the learning materials. We used the framing from Data Carpentry lessons to frame the first spreadsheet chapter. The authors felt that this provided the most tangible connection from existing knowledge working in spreadsheets, mainly Excel, where common "tidy data" issues are discussed and why learners may encounter difficulties doing data analysis after data collection.

3.2.3 Workshop Surveys

Workshop effectiveness was assessed using a series of pre-workshop, post-workshop, and long-term workshop surveys. Some of the questions in the surveys were paired and repeated to perform a longitudinal analysis.

Pre-Workshop Survey

The pre-workshop survey had 4 main sections: (1) demographics, (2) persona, (3) prior and background knowledge, and (4) workshop framing and motivation

The "Persona" questions presented participants with 4 learner personas: (1) Ash Academic, (2) Clare Clinician, (3) Patricia Programmer, and (4) Samir Student These personas were created using only data from the first of two waves of survey responses from the learner persona survey. Participants were asked to pick which one of the personas most resonated with them. Subsequent persona analysis clustered into 3 groups, where Patricia Programmer was removed.

The "prior and background knowledge" questions were taken from a 3-factor model from the first wave of the persona survey. These questions were the highest loaded items in each factor. Later results from the persona factor analysis had a different set of item loadings, but the new items were still captured in the "Workshop Framing and Motivation" questions. The full survey can be found in Supplemental 3.5.1

Post-Workshop Survey

The post-workshop survey had 6 main sections: (1) demographics, (2) workshop environment, (3) workshop framing and motivation, (4) summative assessment, (5) workshop content, and (6) open feedback.

Questions around workshop environment, content, and open feedback were not used to assess workshop efficiency, rather, they served as feedback to the instructor for things they may need to change while teaching. The "workshop framing and motivation" questions were the same from the pre-workshop survey. These questions were used to determine if there were changes in self-reported confidence in completing tasks specified in the learning objectives. The summative assessment question provided a small data task that also asked the participants self-reported ability to complete data tasks. The full survey can be found in Supplemental 3.5.2.

Long-term Workshop Survey

The long-term survey had 4 main sections:

(1) demographics, (2) workshop framing and motivation, (3) summative assessment, and (4) impact

The "workshop framing and motivation" and "summative assessment" questions were the same ones from the post-workshop survey. This provided a way to longitudinally track long-term results. The full survey can be found in Supplemental 3.5.3.

3.2.4 Workshop Survey Analysis Questions

Participants provided a unique identifier that would be used to track their results across the 3 longitudinal surveys. Since participants were not forced to take any of the surveys, 2 sets of analyses can be performed: (1) a longitudinal study looking at survey responses between paired responses across all the surveys, or (2) a series of cross-sectional studies that does incorporate participants that took multiple surveys. The longitudinal analysis would be highly dependant on the number of participants taking at least one of the surveys. Composite scores for each type of analysis were created by summing up survey Likert scale responses; None of the survey questions analysed were reverse-coded.

Workshop Framing and Motivation There was 2 Likert tables of questions for workshop framing and motivation. Both sets of questions were on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree" and included a "Neither Agree nor Disagree" neutral term in the center.

The first set of questions asked respondents on their level agreement with the following 7 statements: (1) I believe having access to the original, raw data is important to be able to repeat analysis, (2) I can write a small program, script, or macro to address a problem in my own work, (3) I know how to search for answers to my technical questions online, (4) While working on a programming project, if I got stuck, I can find ways of overcoming the problem, (5) I am confident in my ability to make use of programming software to work with data, (6) Using a programming language (like R or Python) can make my analyses easier to reproduce, and (7) Using a programming language (like R or Python) can make me more efficient at working with data.

The second set of questions used the same 7-point scale, and asked respondents to rate their agreement about their ability to perform the following tasks: (1) Name the features of a tidy/clean dataset, (2) Transform data for analysis, (3) Identify when spreadsheets are useful, (4) Assess when a task should not be done in spreadsheet software, (5) Break down data processing into smaller individual (and more manageable) steps, (6) Construct a plot and table for exploratory data analysis, (7) Build a data processing pipeline that can be used in multiple programs, and (8) Calculate, interpret, and communicate an appropriate statistical analysis of the data

These 2 sets of likert scale questions were asked in both the pre-workshop and post-workshop survey. The second set of questions asked each participant on how they would self-rate their ability to meet each of the learning objectives the learning materials sought after. These self-reported confidence reports served as a proxy for learning and meeting learning objectives.

Summative Assessment The post-workshop survey also had a summative assessment question where an "untidy" dataset is presented along with a "tidy" version of the same dataset and asked participants if they were able to (1) load the "untidy" dataset into R/Python, (2) Filter the data for a particular set of observations, (3) Save the filtered dataset to send to a colleague, (4) Tidy the dataset into the format presented, (5) Plot a histogram of one of the variables, and (6) Fit a logistic regression model with the dataset. Each of the tasks were presented as a Likert scale question, and respondents were asked to rate their ability to accomplish each task: (1) I wouldn't know where to start, (2) I could struggle through, but not confident I could do it, (3) I could struggle through by trial and error with a lot of web searchers, and (4) I could do it quickly with little or no use of external help.

3.3 Results

The learning materials used for the workshops can be found at https://ds4biomed.tech/. Workshop recordings are published on "Brown Lab YouTube channel [1] with the "ds4biomed" tag. There were 8 workshop sessions with a total of 279 registrants, with a total of 200 learners across all workshop days, and 124 just counting day 1 participants (Table 3.1).

Table 3.1: Number of registrants and workshop attendees. Table shows when each of the 8 workshops were given, the programming language taught during the workshop (type), Number of registrants of the workshop, and the attendee count virtually (V) and in-person (IP). There were a total of 200 non-unique participants across all the 2 workshop days, and 124 unique attendees on the first day. The workshop on 2020-12-09 was taught in a single session. More participants need to be enrolled into the study for better statistical power, however, that requires many more workshop iterations.

	Date 1	Date 2	Type	Registrants	D1-V	D1-IP	D2-V	D2-IP
1	2020-10-20	2020-10-21	R	27	20	-	11	-
2	2020-12-09	-	R	11	5	-	-	-
3	2021-02-02	2021-02-03	R	45	19	-	16	-
4	2021-05-17	2021-05-18	R	40	18	-	15	-
5	2021-06-29	2021-06-30	R	43	10	18	11	10
6	2021-09-20	2021-09-21	R	33	14	1	4	1
7	2021-09-22	2021-09-23	Python	36	9	2	4	0
8	2021-09-27	2021-09-28	R	44	8	0	4	0
Total	-	-	-	279	103	21	65	11

67 responses were collected for the pre-workshop survey, 43 responses were collected for the post-workshop survey, and 11 responses were collected for the long-term survey. Across all the survey responses, 28 respondents took a combination of the surveys. 2 respondents took all 3 surveys, 25 respondents took only the pre-workshop survey and post-workshop survey, and 1 respondent took only the pre-workshop and long-term survey (Figure 3.1).

[1]Brown Lab YouTube channel URL: https://www.youtube.com/channel/UCStv5ui5yBc6_kH91h2o_qA

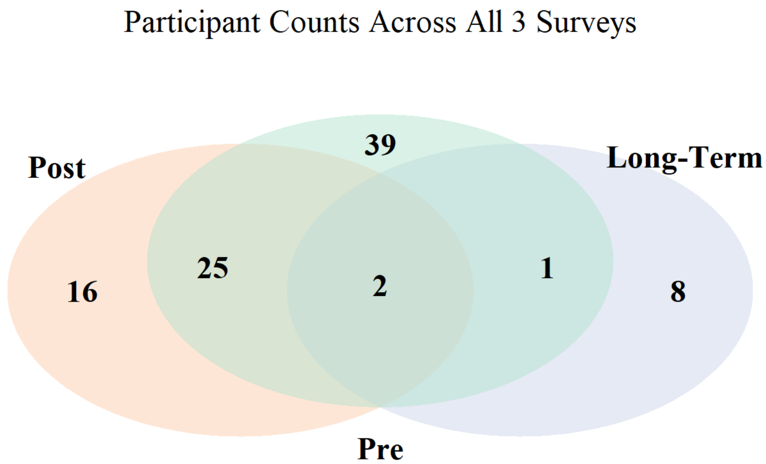


Figure 3.1: Number of participants who responded to each of the workshop surveys: pre-workshop, post-workshop, and long-term workshop. 67 responses were collected for the pre-workshop survey, 43 responses were collected for the post-workshop survey, and 11 responses were collected for the long-term survey.

3.3.1 Create Learning Objectives

The learning materials were developed around 8 learning objectives: (1) Name the features of a tidy/clean dataset, (2) Transform data for analysis, (3) Identify when spreadsheets are useful, (4) Assess when a task should not be done in spreadsheet software, (5) Break down data processing into smaller individual (and more manageable) steps, (6) Construct a plot and table for exploratory data analysis, (7) Build a data processing pipeline that can be used in multiple programs, and (8) Calculate, interpret, and communicate an appropriate statistical analysis of the data.

3.3.2 Workshop Materials: ds4biomed

An open set of workshop materials were created, “Data Science for the Biomedical Sciences” (ds4biomed), and published with a CC0 1.0 Universal (CC0 1.0) license. At the time of writing, the book has 15 programming language agnostic chapters. The full table of contents is listed in Supplemental 3.5.4. The workshop teaching materials begin with basic spreadsheet concepts and best practices where it introduces "tidy data" concepts without using the jargon term.

The lesson then shows how to interact with a programming language (in R or Python) using its respective Integrated Development Environment (IDE) and how to use the read-evaluate-print-loop (REPL) to submit code and commands to be evaluated and results printed in the console. The authors followed the mantra from The Carpentries of teaching the most useful tasks as early as possible to motivate learners. The first programming lesson is around loading and viewing different subsets of a comma-separated value (CSV) and Excel file. The first analysis task learners encounter is performing grouped (i.e., aggregate) statistics. This decision was made to also fit a conference workshop time block; the introduction, spreadsheets, programming, loading data, and descriptive calculation lessons take about 3 hours to teach.

The next few lessons focus on data cleaning. The first lesson explicitly covers "tidy data" principles, how to identify common data problems, and how to fix them. This lesson walks through the 2014 "Tidy Data" paper [50], but uses the more recent definition of "Tidy Data" from the "R for Data Science" book [52]. and illustrations by Allison Horst [220]. With an understanding of what a clean and "tidy" data set is, the next 2 lessons go through plotting and model fitting, which take tidy data as inputs. Teaching tidy data, plotting, model fitting, and conclusion also take 3 hours.

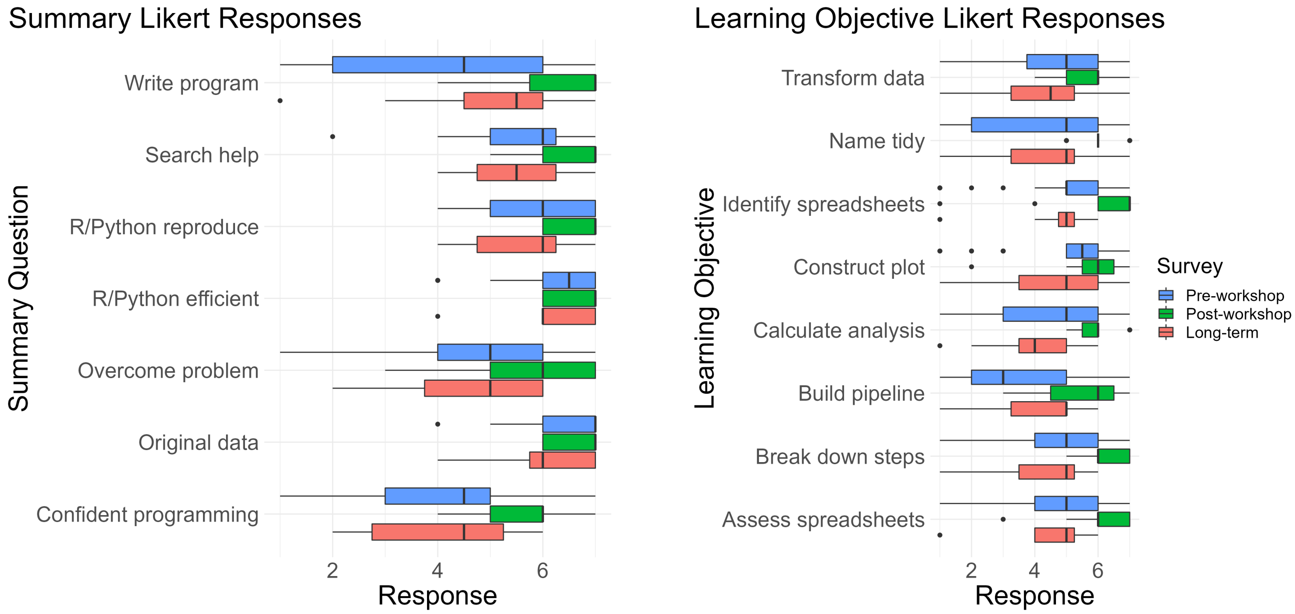
3.3.3 Longitudinal and Cross-Sectional Study

Since there were only 2 participants who took all 3 surveys, a proper longitudinal study could not be performed. Instead, a series of cross-sectional observations were analyzed. To make all the results comparable across each of the cross-sections (i.e., surveys), only the participants who took that particular survey were used for the analysis. Any participant who took more than 1 survey was not included in these results. This meant that only 39 observations were used in the pre-workshop analysis, 16 in the post-workshop analysis, and 8 in the long-term analysis (Figure 3.1).

There were 3 main sets of cross-sectional observations: (1) summary and learning objective Likert responses across all 3 surveys (Figure 3.2) , (2) summative assessment Likert questions across just the post-workshop and long-term workshop surveys (Figure 3.3), and (3) a sub-analysis looking at just summary and learning objective results from the pre-workshop and post-workshop surveys (Figure 3.4 and Supplemental Figure 3.6).

Figure 3.2 looks at each of the Likert questions in the summary and learning objective Likert questions across all 3 surveys. The overall trend is that learners have an increase in confidence across all tasks after the workshop, but confidence wanes off in the long-term survey. The same post-workshop to long-term findings applied for the summative assessment question, where participants were asked about their confidence to accomplish a particular data task (Figure 3.3)

In order to make the pre-workshop and post-workshop results more comparable, due to varying sample sizes in each group, proportions of all the Likert results across the summary and learning objective questions were compared. The proportions of responses from the pre-workshop and post-workshop were then filled to a 100% scale, to create a proportion of proportions analysis (Figure 3.4). These results show that there was a general trend to ”Agree” and ”Strongly Agree” to each of the Likert questions between the 2 longitudinal points.



(a) Summary likert table questions

(b) Learning objective likert table questions

Figure 3.2: Cross-sectional results for 2 Likert table questions in the pre-workshop, post-workshop, and long-term workshop surveys. Results show a general increase in confidence across all questions between pre-workshop and post-workshop responses, and a drop in confidence across all questions between post-workshop and long-term workshop responses.

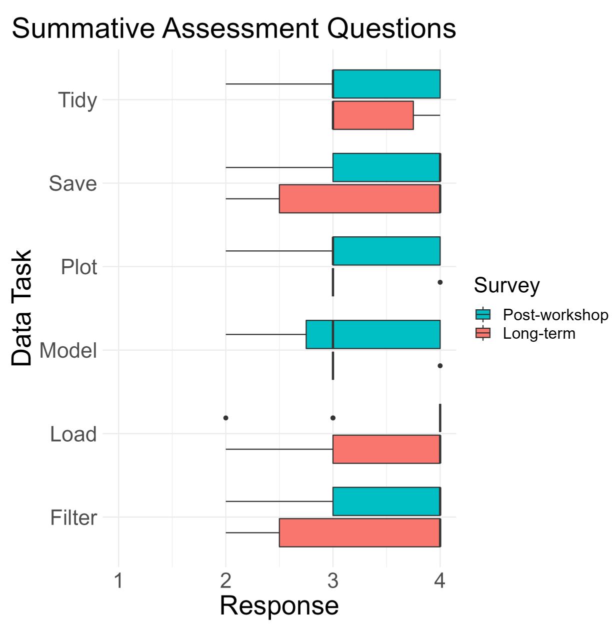
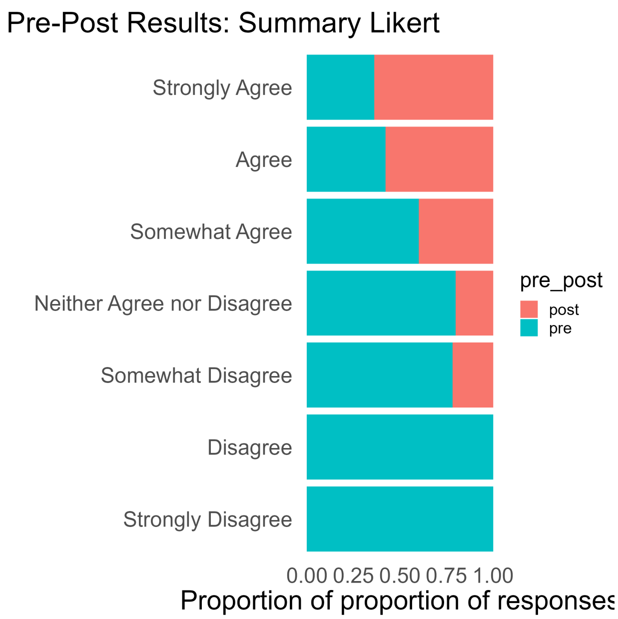
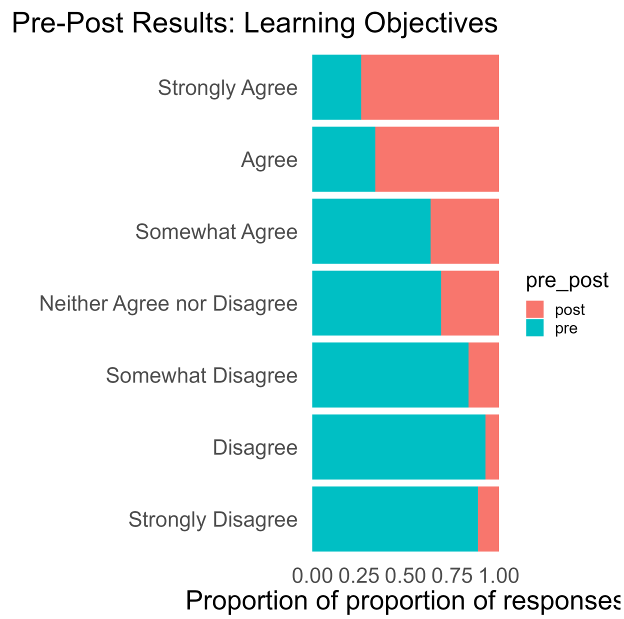


Figure 3.3: Cross-sectional results for summative assessment questions in the post-workshop and long-term survey. Results show a decline in confidence to complete a particular data task.



(a) Summary Likert table proportion of proportions



(b) Learning objective Likert table proportion of proportions

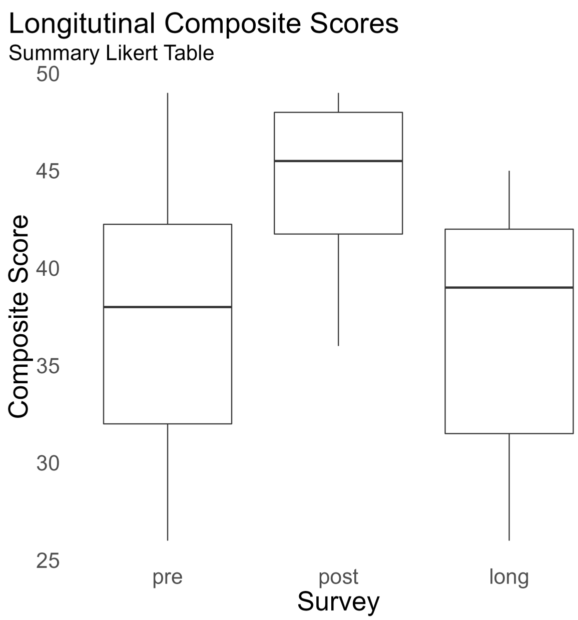
Figure 3.4: Cross-sectional results for 2 Likert table questions in the pre-workshop and post-workshop surveys.

Composite Scores

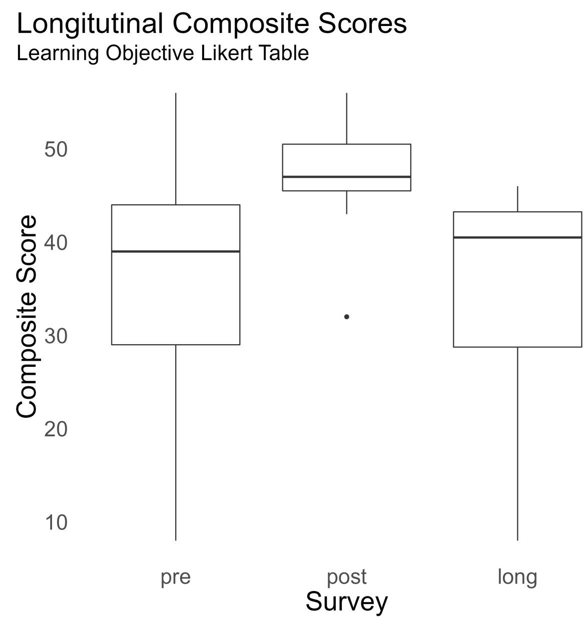
The responses from the Likert scale were summed together for each participant to create a composite score. These results confirmed the trends found in Figures 3.2 and 3.3: learners felt more confident in the post-workshop results than the pre-workshop, and confidence waned off in the long-term survey (Figure 3.5).

3.3.4 Learning Environment

The environment plays a critical role in a learner's ability to learn. There were a few instructor feedback questions in the post-workshop survey to make sure the presentations and learning environment was welcoming. Due to COVID-19 quarantine guidelines, the vast majority of workshop classes were held virtually. The classes that were able to be held in-



(a) Summary composite likert table questions



(b) Learning objective composite likert table questions

Figure 3.5: Taking the sum of the Likert scale results were used to create composite scores for each respondent across each of the surveys. In general, there was an increase in the scores from pre-workshop to post-workshop, and a decline in scores from post-workshop to long-term.

person were also given virtually in a hybrid setting. The same instructor was used across all the workshops, to ensure consistency between teaching sessions.

3.4 Discussion

Survey response rates were generally not high enough to make definitive conclusions. The work we presented in this study serves as a baseline set of values for future research. The surveys and data are published so future researchers can expand and combine their data with this study.

3.4.1 Limitations

One of the main limitations with the survey responses is there are no graded summative assessment question. All the responses to the learning objectives are self-reported confidence in task completion. The actual summative assessment question was not graded, or were participants asked to provide code to solve the data challenge.

In addition, due to the low response rates across all the surveys, a longitudinal analysis looking at the same participant across multiple time points was unable to be conducted. A cross-sectional observational study was used instead that only looked at participants who only took one (1) of the three (3) surveys.

3.4.2 Learner Personas and Concept Maps Help Curate Lesson Content

The lesson materials were created using learner personas. While, not directly measured in the initial persona survey, thinking about learner's special considerations helped make the materials more accessible, and has helped with improving the greater open-access set of materials.

The learner personas guided the creation of the lesson content. Tidy data concepts were a core part of the lesson materials. Using a backward design approach, The spreadsheet lesson was created to indirectly introduce tidy data concepts. The concepts in this lesson were used through all of the lesson materials. By starting from spreadsheets, the learners are presented with a data workflow that they are most familiar with. This transitioned into eventually loading an Excel dataset as one of the initial programming commands.

However, before learners are able to load data into a programming language, they needed to understand where files are on their computers. These concepts even pertain to younger audiences, since operating systems have conflated cloud storage with local storage, and younger generations typically do not have an understanding of how files are stored on their computers when compared those who learned computer skills in the 90s and 00s [221]. A separate lesson was created on project structures, working directories, and file paths so all learners can load up a dataset from the lesson, and potentially their own work.

3.4.3 Language-Agnostic Lessons Guide Presentation Order

The main focus was to have modular lesson content but focused in a workshop format. Since learners were able to opt-in for the classes, the workshop presented basic dataset filtering, workflow management, and calculating descriptive and grouped statistics first. This catered to a more Excel-based workflow, and the opt-in allowed us to postpone plotting a bit later into the workshop.

Not all programming libraries will have the ability to load up the same datasets from installing external packages. This also guided the lessons to show manually loading data as one of the first programming lessons. The lesson then goes into inspecting data, rather than going directly into plotting. This gives us the foundation of understanding how to make summary statistics which can be visualized later.

Since the individual lessons are modular, these lessons can be re-arranged as needed. The current arrangement of (1) introduction, (2) spreadsheets, (3) programming language setup, (4) load data, (5) descriptive statistics takes about 2.5 hours to complete teaching. Each of the remaining topics in: (6) tidy data, (7) visualization, and (8) logistic regression takes about an hour to complete. The currently presented order can be used for conference workshop blocks as is or with minor tweaks (usually 3 to 4 hours minimum). This leaves statistics to be the last topic covered, and first topic to be cut out if more time is needed.

Aspects of Our Work Confirm Existing Bodies of Work.

The R for Data Science book is commonly referenced for new R learners. This book places data visualization as the first main topic as a means to attract learners and keep them motivated. From there, the book slowly introduces workflow steps, and basic data transformation and exploratory data analysis steps using visualization as the guiding aid. The first part of the book ends with project workflow, where they talk about files and paths to load data. The ds4biomed materials are catered towards working practitioners that may be self-learning or participating in a workshop, the visualization steps are deferred in lieu of getting existing data loaded into the programming language.

The Data Carpentry materials typically begin with lessons on spreadsheet and workflow management. Since Data Carpentry lessons are organized by curriculum domains, more domain-specific methods and techniques are presented. The ds4biomed follows the Data Carpentry ordering by providing a foundation on structuring data.

3.4.4 Data Science Lessons Differ from Computer Science Lessons

The initial use case of loading and working with data makes teaching data science different from computer science classes. Computer science is primarily focused on the implementation of algorithms around basic data structures, data science typically work around the higher-order "dataframe" object. The implementation and internals of dataframes are typically not needed by novices, and focusing on more data literacy concepts is a more direct need for data scientists.

Many concepts that are taught in introductory computer science classes will eventually be learned by novice data scientists, but framing many data manipulation steps around tidy data principles can circumvent teaching topics like loops until much later when pragmatically curating data is needed. Even knowing how to write functions can be delayed in teaching data science topics. The ds4biomed materials does not cover functions until much later in the learning materials.

3.4.5 Intermediate Materials Will Be Difficult to Plan

Many of the core learning objectives related to tidy data principles can be relatively mapped out. However, intermediate level materials may be harder to cater to broad and domain-specific audiences. Even specific domains have sub-fields with their separate analysis needs. In the biomedical sciences, informatics needs will differ from hospital billing needs. The tools, methods, and analysis techniques will begin to diverge, which make planning cohesive materials more difficult.

The Carpentries have approached this problem by creating separate Data Carpentry curriculum. This allows common materials to be taught while proving lessons for specific techniques. After the core data programming skills are learned, learning and applying a different method should become much easier for learners. The more modular lessons are, the more flexibility learners can have after the fundamentals are taught.

3.4.6 Long-Term Practice is important

The longitudinal survey results show that there is an increase in confidence to complete certain data science tasks, but the confidence wanes down to pre-workshop levels several months after the workshop. Future adaptations of the long-term survey should ask more about programming usage to get a better understanding of why long-term confidence in skills decreased. One hypothesis is the skills taught at the workshop were not being used. Since our target audience are working professionals, this can be explained as not having the time to use the data science skills they were taught. Another possibility is that the skills taught were not relevant for the kind of work they wanted to do.

Another hypothesis stems from unable to acquire data to explore. Health data is generally highly protected and regulated. Researchers and clinicians may not have the means to immediately work on a research project with medical data.

Expanding ds4biomed Content

Having example problems to work on is the best way to retain information and learn new skills. The ds4biomed materials expanded beyond the content used in this study. Later chapters went into more case-study based lessons with a main topic: (1) 30-day re-admittance; (2) working with multiple datasets; (3) application programming interfaces (APIs); (4) functions; (5) survival analysis; (6) machine learning.

The 30-day re-admittance mostly deals with the same selecting, filtering, and mutating columns from the first programming chapter. It builds on these same skills by introducing date columns and how to perform basic date arithmetic to find patients in the SynThea dataset which had a 30-day re-admittance to the emergency department for a heart attack. It asked the learners to explore more of the data to discover how to filter down the dataset for the problem.

The following lesson takes the results from 30-day re-admittance and combines the encounters dataset with patient level information to look at age of heart-attack admittance to the emergency room. This required knowledge on how to join and combine datasets on "primary keys", a term used in databases that points to an ID column or columns that point to a unique observation. Since hospitals store data in databases, and not flat Excel or CSV files, we use this opportunity to show how concepts from the previous lessons translate into database querying, and how to query data into a dataframe object.

We then go into APIs by using the US Census API to download census data and combine it with leading causes of death data in the United States to calculate death rates. This builds on the joining lesson from earlier, but also teaches how rates always have a time factor involved (e.g., deaths over the last year), how rates are calculated with a reference population, and how rates with different reference populations cannot be adequately compared.

The functions chapter introduces string methods and how users can re-code values. This reinforces tidy data principles by making sure only a variables are stored in one column. The example shows how to prototype a representative example from the data, how to test (i.e., unit test) the function's behavior to make it more robust to accidental changes, and how to apply a function to a column of variables. If the dataset was not tidy, apply a function to process data in a column would be more difficult.

The final two (2) examples are more statistics related. The lesson materials shifts to understanding how to interpret statistical models. Survival analysis is a common tool used to look at the efficacy between a treatment group and control group. The lesson mainly works by loading a dataset that has been processed to perform this analysis, and was more concerned about how to create Kaplan-Meier curves, and how to interpret survival models. We finish off the lesson materials with a discussion about machine learning, and defined machine learning as the process primarily focused on creating predictive models, and not inferential models. The lesson uses the same logistic regression model, but introduces more machine learning concepts, such as training and testing data splits. This is the process where users can simulate new data and compare how different models are able to predict data it was not trained (i.e., fitted) on. Here we emphasize the importance of having a workflow that works on training and testing data separately, and how data leakage can artificially boost the performance of a model. We introduce many of the concepts that would be covered in a machine learning class, but only focus on a single case-study. This allows the materials to be more focused, while being able to cover the main points.

Work on Relevant Problems Solidify skills

The results from the long-term study suggest that what may be more important is to have more case-study problems that reinforces concepts from the introductory materials. The second set of ds4biomed materials cover many of the core data science components. What can be beneficial is to have more worked out examples that can be posed as problems. Future workshops would spend more time working through these problems and can even be presented where all participants work asynchronously. Having relevant examples will motivate learning.

3.4.7 Communities of Practice

The Synthea project used in ds4biomed provides one mechanism to explore and practice working with health datasets. Another resource is the the Observational Health Data Sciences and Informatics (OHDSI). Leveraging these data sources can provide new examples for long-term learning. The other challenge is finding a cohort of learners who are at roughly the same learning level to learn together. For example, r4ds has an online slack learning community, there are other communities that use slack where users can ask for help. The Carpentries provide a mechanism through the curriculum development program to host similar learning materials together. The organization also follows many of the best teaching practices that can help with scaling the maintenance and teaching of domain-specific data science materials. They already do so for ecology, genomics, social sciences, and geospatial data. The medical and biomedical sciences could also be a curriculum program by using the ds4biomed materials as the introduction for other lessons. This feeds into the transdisciplinary approach of looking at health, One Health. Another growing community in this space is the “r/pharma” and “r/medicine” groups where core learning materials can be centralized.

One of the main benefits of The Carpentries is they are already a globally and funding agency recognized organization. Lesson materials can be maintained by multiple people, and there are mechanisms to recruit new maintainers. The lessons are also hosted openly and not bound to proprietary services. This prevents the lesson from going stale after its initial conception, which is usually the cause to re-create the same materials over.

3.4.8 Conclusion

The ds4biomed materials serve as a strong foundation for a data science curriculum in the biomedical sciences. The methodologies used to create the learning materials can be applied to other domains. Data sets can be swapped out to make the examples more relevant, but the learning objectives would remain the same. This study does not indent to create a completely self-contained lesson curriculum, rather it serves as a foundation and hopes to link to already exiting materials for continuing education. However, what may be more important for long-term learning is to link to external learning resources but provide a series of case-study questions where learners can continuously practice and improve on their data science skills. This can work as a spaced repetition model where case studies can drive long-term engagement while practicing foundational skills.

3.5 Supplemental

Supplemental materials for the "Assessing the Efficacy of Domain-Specific Data Science Curriculum in the Biomedical Sciences: How Learner Personas Can Guide Educational Needs in the Short-Term and Long-Term".

3.5.1 Pre-Workshop Survey Questions

The surveys can be downloaded from the GitHub URL that holds the IRB proposal for the study: https://github.com/chendaniely/dissertation-irb/tree/master/irb-20-537-data_science_workshops/survey

Demographics

Q2.2 Please create a unique identifier. This unique identifier will be used for long-term assessment but keep your personal information anonymous.

To create an identifier type in: Number of siblings (as numeric) + First two letters of the city you were born in (lowercase) + First three letters of your current street (lowercase).

E.g., (Sherlock Homes has **1** brother, was born in **Po**rsmouth, and lives on **Bac**ker Street - **1pobac**)

Q2.3 Please select the first date of your workshop

- Monday, September 20, 2021: Virtual (9)
- Monday, September 20, 2021: In-Person (8)
- Wednesday, September 22, 2021: Virtual (10)
- Wednesday, September 22, 2021: In-Person (11)
- Tuesday, June 29, 2021 (7)
- Monday, May 17, 2021 (6)
- Tuesday, February 2, 2021 (5)

- Wednesday, December 9, 2020 (4)
- Tuesday, October 20, 2020 (1)
- I went through the online materials on my own (2)

Q2.5 What is your current occupation/career stage (select all that apply).

- DO/MD (1)
- DVM (12)
- RN/PA (2)
- PhD (13)
- Academic (3)
- Analyst (4)
- Student (Masters e.g., MPH) (5)
- Student (MD/DO) (6)
- Student (Nurse, PA) (7)
- Student (Graduate) (8)
- Student (Undergraduate) (9)
- iTHRIV Scholar (11)
- Other, please describe (10)

Q2.6 What operating system will be on the computer you are using at the workshop or to participate in the online materials?

- Windows (1)
- macOS (2)
- Linux (3)
- Not sure (4)

Persona

Q3.1 Which of the below personas do you most identify with? Be less concerned about the actual occupation, and more with what relates to your skill and workshop needs.

More detailed descriptions of Ash Academic, Clare Clinician, Patricia Programmer, and Samir Student can be found here: https://ds4biomed.tech/who-is-this-book-for.html

Ash Academic

Alex performs their research using a combination of Excel spreadsheets and specialized software, but is switching to R or Python (which they taught themself during a sabbatical). They have never taken a formal programming course, and suffers from impostor syndrome in discussions about programming. Alex would like to learn more about how programming can help their research and keep up with the tools their students are learning in class.

Alex needs workshops (so they can allocate focused time) and how-to guides (for research). They would like ready-to-use lesson material that could be remixed for their students and some orientation material to demystify jargon (what is ”tidy data”?). Alex also wants to be able to use the same tools in their research as in their teaching to amortize learning costs and stay in practice.

Clare Clinician

Clare keeps up with medical research, but has little to no experience in doing medical research. They use Excel for non-data related tasks (e.g., making lists), or manually inputting patient data into spreadsheets for chart reviews. Wants to be able to collect and manage data as well as learn about the process behind data analysis to perform their own analysis and study one day.

Clare wants self-paced tutorials with practice exercises, plus forums where they can ask for help. They also need short overviews to orient them and introductory tutorials that include videos or animated GIFs showing exactly how to drive the tools, and that use datasets they can relate to. Clare wishes they had a community of other people in the medical field who are interested in learning how to do data work so they can learn and ask questions.

Patricia Programmer

Patricia regularly connects to a remote server to do their work. They write SQL statements to pull data out of Epic and processes the data in both Python and R to generate reports and dashboards for their team and management. Patricia writes data pipelines for all their work either by combining shell scripts or build scripts.

Patricia wants how-to guides and reference material for their day-to-day work and short, intensive online training for very specific topics. Because they often jump around between various tools, Patricia wants a way to quickly review topics before starting a new project.

Samir Student

Samir is fairly proficient in Excel and does works with spreadsheets regularly and knows how to load up Excel spreadsheets into R and do basic data processing and analysis. However, they do not have that much practice outside of a classroom homework and project setting, and spends a lot of their time on StackOverflow copying and pasting code so they don't consider themselves a "real programmer". They have no problem getting their work done, but usually involves a lot of googling to eventually get the solution.

Samir wants a formal workshop and reference materials that can be used to build a good foundation of the programming skills they were never taught. They want a better understanding of the terminology and jargon used in data science so they have the vocabulary to search for and understand solutions posted online. They are also looking for a community to help in their growth as a student in this domain.

- Ash Academic (1)
- Clare Clinician (2)
- Patricia Programmer (3)
- Samir Student (4)

Prior and background knowledge

Q4.1 How familiar are you with interactive programming languages like Python or R?

- I do not know what those are (1)
- I have heard of them but have never used them before (2)
- I have installed it, but have only done simple examples with them (3)
- I have written a small program with them before (4)

- I use it to automate certain repetitive tasks (5)
- I have small side projects that I program in it (6)
- I program in them for work (7)

Q4.2 Are you familiar with the term “tidy data”?

- I have never heard of the term (1)
- I have heard of it but don’t remember what it is. (2)
- I have some idea of what it is, but am not too clear (3)
- I know what it is and could explain what it pertains to (4)

Q4.3 If you were given a dataset containing an individual’s smoking status (binary variable) and whether or not they have hypertension (binary variable), would you know how to conduct a statistical analysis to see if smoking has an increased relative risk or odds of hypertension? Any type of model will suffice.

- I wouldn’t know where to start (1)
- I could struggle through, but not confident I could do it (4)
- I could struggle through by trial and error with a lot of web searches (2)
- I could do it quickly with little or no use of external help (3)

Workshop Framing and Motivation

Q5.1 Why are you participating in this workshop? Please check all that apply.

- To learn new skills (1)
- To refresh or review my skills (2)
- To learn skills that I can apply to my current work (3)
- To learn skills that I can apply to my work in the future (4)
- To learn skills that will help me get a job or a promotion (5)
- As a requirement for my program or current position (6)

Q5.2 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- I believe having access to the original, raw data is important to be able to repeat an analysis. (1)

- I can write a small program, script, or macro to address a problem in my own work. (2)
- I know how to search for answers to my technical questions online. (3)
- While working on a programming project, if I got stuck, I can find ways of overcoming the problem. (4)
- I am confident in my ability to make use of programming software to work with data. (5)
- Using a programming language (like R or Python) can make my analyses easier to reproduce. (6)
- Using a programming language (like R or Python) can make me more efficient at working with data. (7)

Q5.3 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- Name the features of a tidy/clean dataset (1)
- Transform data for analysis (2)
- Identify when spreadsheets are useful (3)
- Assess when a task should not be done in a spreadsheet software (4)
- Break down data processing into smaller individual (and more manageable) steps (5)
- Construct a plot and table for exploratory data analysis (6)
- Build a data processing pipeline that can be used in multiple programs (7)
- Calculate, interpret, and communicate an appropriate statistical analysis of the data (8)

Q5.4 Please share what you most hope to learn from participating in this workshop and/or workshop series.

Q5.5 What do you want to know or be able to do after this workshop (or series of sessions) that you don't know or can't do right now?

3.5.2 Post-Workshop Survey Questions

Demographics

Q2.2 Please create a unique identifier. This unique identifier will be used for long-term assessment but keep your personal information anonymous.

To create an identifier type in: Number of siblings (as numeric) + First two letters of the city you were born in (lowercase) + First three letters of your current street (lowercase).

E.g., (Sherlock Homes has **1** brother, was born in **Po**rsmouth, and lives on **Bac**ker Street - **1pobac**)

Q2.3 Please select the first date of your workshop

- Tuesday, June 29, 2021 (7)
- Tuesday, May 18, 2021 (6)
- Tuesday, February 2, 2021 (5)
- Wednesday, December 9, 2020 (4)
- Tuesday, October 20, 2020 (1)
- I went through the online materials on my own (2)
- Other (3) _____

Q2.4 What is your current occupation/career stage (select all that apply).

- DO/MD (1)
- DVM (12)
- RN/PA (2)
- PhD (13)
- Academic (3)

- Analyst (4)
- Student (Masters e.g., MPH) (5)
- Student (MD/DO) (6)
- Student (Nurse, PA) (7)
- Student (Graduate) (8)
- Student (Undergraduate) (9)
- iTHRIV Scholar (11)
- Other, please describe (10)

Workshop Environment

Q3.1 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- I felt comfortable learning in this environment (1)

- I can immediately apply what I learned (2)
- I was able to get clear answers to my questions from the instructors (3)
- The instructors were enthusiastic about the workshop (4)
- I felt comfortable interacting with the instructors (5)
- The instructors were knowledgeable about the material being taught (6)

Q3.2 Do you have accessibility requirements?

- No (1)
- Yes (2)

Q3.3 Where there any accessibility issues that affected your ability to participate in this workshop?

- No (1)
- Yes (2)
- Not applicable (3)

Q3.4 Please describe what the accessibility issues were.

Workshop Framing and Motivation

Q4.1 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- I believe having access to the original, raw data is important to be able to repeat an analysis. (1)
- I can write a small program, script, or macro to address a problem in my own work. (2)
- I know how to search for answers to my technical questions online. (3)
- While working on a programming project, if I got stuck, I can find ways of overcoming the problem. (4)
- I am confident in my ability to make use of programming software to work with data. (5)
- Using a programming language (like R or Python) can make my analyses easier to reproduce. (6)
- Using a programming language (like R or Python) can make me more efficient at working with data. (7)

Q4.2 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- Name the features of a tidy/clean dataset (1)
- Transform data for analysis (2)
- Identify when spreadsheets are useful (3)
- Assess when a task should not be done in a spreadsheet software (4)
- Break down data processing into smaller individual (and more manageable) steps (5)
- Construct a plot and table for exploratory data analysis (6)
- Build a data processing pipeline that can be used in multiple programs (7)
- Calculate, interpret, and communicate an appropriate statistical analysis of the data (8)

Summative assessment

Q5.1 Cytomegalovirus (CMV) is a common virus that normally does not cause any problems in the body. However, it can be of concern for those who are pregnant or immunocompromised.

Suppose you have the following Cytomegalovirus dataset [1] of CMV reactivation among patients after Allogenetic Hematopoietic Stem Cell Transplant (HSCT) in an excel sheet (first 10 rows shown below):

It contains a patient's: ID age priror.radiation: whether or not patient had prior radiation treatment (0 = no, 1 = yes) aKIRs: Number of donor activating killer immunoglobulin-link receptors donor_negative: the recipient's CMV status when the donor was CMV negative donor_positive: the recipient's CMV status when the donor was CMV positive

It is believed that the donor activating KIR genotype is a contributing factor for CMV reactivation after myeloablastive allogenetic HSCT. You want to do some data analysis to see what variables are associated with CMV reactivation.

Assuming this is the version of the data you need for the tidying, plotting, and modeling:

How would you rate your ability to accomplish the following tasks:

[1]: Sobecks et al. "Cytomegalovirus Reactivation After Matched Sibling DonorReduced-Intensity Conditioning Allogeneic HematopoieticStem Cell Transplant Correlates With Donor KillerImmunoglobulin-like Receptor Genotype". Exp Clin Transplant2011; 1: 7-13.

- I wouldn't know where to start (4)
- I could struggle through, but not confident I could do it (5)
- I could struggle through by trial and error with a lot of web searches (6)

- I could do it quickly with little or no use of external help (7)

- Load the excel sheet into R (1)
- Filter the data for individuals over the age of 65 (in R) (2)
- Save filtered dataset (in R) as an Excel file to send to a colleague (6)
- Tidy the dataset (in R) so we have a donor CMV status and a patient CMV status in separate columns (3)
- Plot a histogram (in R) of the age distribution of our data (4)
- Fit a model (e.g., logistic regression) to see which variables are associated with patient CMV reactivation (in R) (5)

Workshop Content

Q6.1 Were there any topics you wish were covered?

Q6.2 What topic would you take out of the workshop to make room for the topics mentioned above?

Q6.3 In general, how would you prefer to have the workshop content (4 - 5 hours) taught?

- 1 day 4-5 hour workshop on the weekday (1)
- 1 day 4-5 hour workshop on the weekend (2)
- 2 days about 2-3 hours each on the weekday (3)

- 2 days about 2-3 hours each on the weekday (4)
- Multiple days in a row about 1 hour each day (5)
- Multiple days spread across multiple weeks on the weekdays (6)
- Multiple days spread across multiple weeks on the weekends (7)
- Not applicable (8)

Open Feedback

Q7.1 Please provide an example of how an instructor or helper affected your learning experience.

Q7.2 What is something you liked about the workshop?

Q7.3 What is something you **did not** like about the workshop?

3.5.3 Long-Term Workshop Survey Questions

Demographics

Q2.2 Please create a unique identifier. This unique identifier will be used for long-term assessment but keep your personal information anonymous.

To create an identifier type in: Number of siblings (as numeric) + First two letters of the city you were born in (lowercase) + First three letters of your current street (lowercase).

E.g., (Sherlock Homes has **1** brother, was born in **Po**rsmouth, and lives on **Bac**ker Street - **1pobac**)

Q2.3 Please select the first date of your workshop

- Tuesday, June 29, 2021 (7)
- Tuesday, May 18, 2021 (6)
- Tuesday, February 2, 2021 (5)
- Wednesday, December 9, 2020 (4)
- Tuesday, October 20, 2020 (1)
- I went through the online materials on my own (2)
- Other (3) _____

Q2.4 What is your current occupation/career stage (select all that apply).

- DO/MD (1)
- DVM (12)
- RN/PA (2)
- PhD (13)
- Academic (3)
- Analyst (4)
- Student (Masters e.g., MPH) (5)
- Student (MD/DO) (6)
- Student (Nurse, PA) (7)

- Student (Graduate) (8)
- Student (Undergraduate) (9)
- iTHRIV Scholar (11)
- Other, please describe (10)

Behaviors and Confidence

Q3.1 Which of the following behaviors have you adopted as a result of completing the workshop / going through the materials.

- Improving data management and project organization (1)
- Developing a data management and analysis plan (2)
- Transforming step-by-step workflows into scripts (3)
- Using programming languages like R or Python to automate repetitive tasks (4)
- Reusing code (5)
- Sharing code or data publicly (6)
- None (12)
- Other (13) ___

Q3.2 Before the workshop, how often did you use programming languages?

- I had not been using tools like these (1)

- Less than once a per half-year (2)
- Several times per half-year (3)
- Monthly (4)
- Weekly (5)
- Daily (6)

Q3.3 Since taking the workshop, how often did you use programming languages?

- I had not been using tools like these (1)
- Less than once a in the last 6 months (2)
- Several times in the last 6 months (3)
- Monthly (4)
- Weekly (5)
- Daily (6)

Q3.4 How would you rate your change in confidence in the tools that were covered during your workshop compared to before the workshop?

- I'm more confident now (1)
- I'm equally confident now (2)
- I'm less confident now (3)

Workshop Framing and Motivation

Q4.1 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- I believe having access to the original, raw data is important to be able to repeat an analysis. (1)
- I can write a small program, script, or macro to address a problem in my own work. (2)
- I know how to search for answers to my technical questions online. (3)
- While working on a programming project, if I got stuck, I can find ways of overcoming the problem. (4)
- I am confident in my ability to make use of programming software to work with data. (5)
- Using a programming language (like R or Python) can make my analyses easier to reproduce. (6)

- Using a programming language (like R or Python) can make me more efficient at working with data. (7)

Q4.2 Please rate your level of agreement with the following statements:

- Strongly Disagree (1)
- Disagree (2)
- Somewhat Disagree (3)
- Neither Agree nor Disagree (4)
- Somewhat Agree (5)
- Agree (6)
- Strongly Agree (7)

- Name the features of a tidy/clean dataset (1)
- Transform data for analysis (2)
- Identify when spreadsheets are useful (3)
- Assess when a task should not be done in a spreadsheet software (4)
- Break down data processing into smaller individual (and more manageable) steps (5)
- Construct a plot and table for exploratory data analysis (6)
- Build a data processing pipeline that can be used in multiple programs (7)
- Calculate, interpret, and communicate an appropriate statistical analysis of the data (8)

Summative assessment

Q5.1 Cytomegalovirus (CMV) is a common virus that normally does not cause any problems in the body. However, it can be of concern for those who are pregnant or immunocompromised.

Suppose you have the following Cytomegalovirus dataset [1] of CMV reactivation among patients after Allogenetic Hematopoietic Stem Cell Transplant (HSCT) in an excel sheet (first 10 rows shown below):

It contains a patient's: ID age priror.radiation: whether or not patient had prior radiation treatment (0 = no, 1 = yes) aKIRs: Number of donor activating killer immunoglobulin-link receptors donor_negative: the recipient's CMV status when the donor was CMV negative donor_positive: the recipient's CMV status when the donor was CMV positive

It is believed that the donor activating KIR genotype is a contributing factor for CMV reactivation after myeloablastive allogenetic HSCT. You want to do some data analysis to see what variables are associated with CMV reactivation.

Assuming this is the version of the data you need for the tidying, plotting, and modeling:

How would you rate your ability to accomplish the following tasks:

[1]: Sobecks et al. "Cytomegalovirus Reactivation After Matched Sibling DonorReduced-Intensity Conditioning Allogeneic HematopoieticStem Cell Transplant Correlates With Donor KillerImmunoglobulin-like Receptor Genotype". Exp Clin Transplant2011; 1: 7-13.

- I wouldn't know where to start (4)
- I could struggle through, but not confident I could do it (5)
- I could struggle through by trial and error with a lot of web searches (6)

- I could do it quickly with little or no use of external help (7)

- Load the excel sheet into R (1)
- Filter the data for individuals over the age of 65 (in R) (2)
- Save filtered dataset (in R) as an Excel file to send to a colleague (6)
- Tidy the dataset (in R) so we have a donor CMV status and a patient CMV status in separate columns (3)
- Plot a histogram (in R) of the age distribution of our data (4)
- Fit a model (e.g., logistic regression) to see which variables are associated with patient CMV reactivation (in R) (5)

Impact

Q6.1 The statements below reflect ways in which completing the workshop may have impacted you. Please indicate your level of agreement with these statements.

- Strongly disagree (1)
- Disagree (2)
- Neutral (3)
- Agree (4)
- Strongly agree (5)

- I have used skills I learned at the workshop to advance my career. (1)

- I have been motivated to seek more knowledge about the tools I learned at the workshop. (2)
- I have made my analysis more reproducible as a result of completing the workshop. (3)
- I have improved my coding practices as a result of completing the Workshop (4)
- My research productivity has improved as a result of completing the workshop (5)
- I have gained confidence in working with data as a result of completing the workshop. (6)

Q6.2 Did you go back to the online materials after the workshop?

- I went back to the code I wrote for reference (1)
- I went back to the code the instructor posted for reference (2)
- I went back to the video recording for the workshop (3)
- I went back to the online written materials for reference (4)
- I did not go back to any of the workshop materials (5)
- Other (6) ___

Q6.3 Why did you not go back to a particular workshop resource?

Q6.4 Please tell us the most important way you were impacted as a result of the workshop.

Q6.5 Please provide any outcomes as a result of attending this workshop.

Q6.6 If you would like to make additional comments about the workshop experience, or ways you've used the tools you learned in the workshop please comment below.

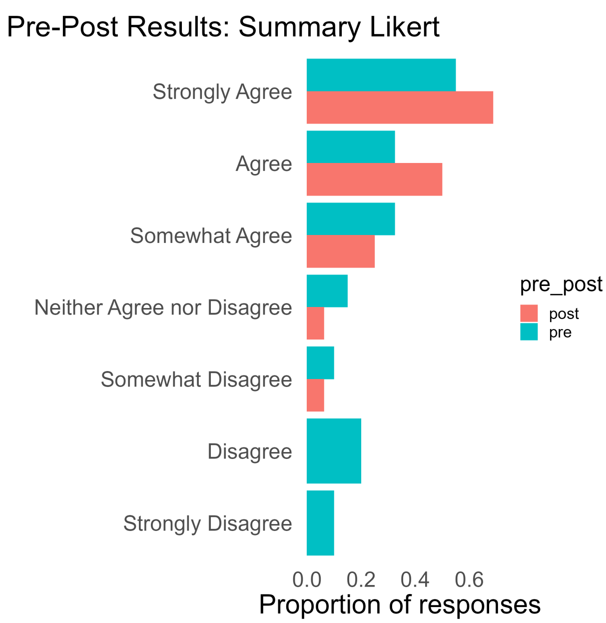
3.5.4 ds4biomed Table of Contents

The materials for "Data Science for the Biomedical Sciences" can be found at the following URL: https://ds4biomed.tech/

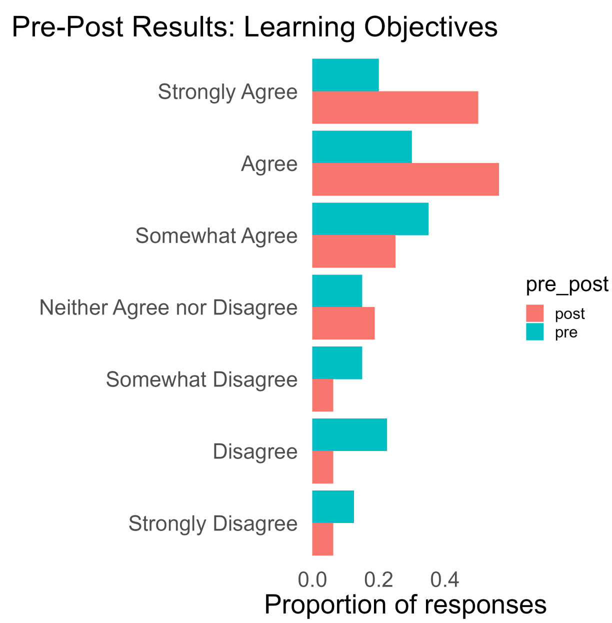
1. Welcome
2. Preface
3. Who is this book for
4. Code of Conduct
5. Setup
6. Workshop logistics
7. Introduction
8. spreadsheets R + RStudio
9. Load Data
10. Descriptive Calculations
11. Clean Data
12. Visualization
13. Analysis Intro

14. 30-Day Readmittance
15. Working with multiple datasets
16. Application Programming Interfaces (APIs)
17. Functions
18. Survival Analysis
19. Machine Learning (tidymodels)
20. Additional Resources

3.5.5 Longitudinal Study



(a) Summary likert table proportions



(b) Learning objective likert table proportion

Figure 3.6: Cross-sectional results for 2 Likert table questions in the pre-workshop and post-workshop surveys. Figure shows the unadjusted proportions of responses between pre-workshop and post-workshop across the two (2) survey questions. Since the number of participants used to calculate the proportions differ between the pre-workshop and post-workshop results, Figure 3.4 compares the proportions as a "proportion of proportions".

Chapter 4

Refining Feedback and Guidance in Data Science Workshops: Making Time for Formative and Summative Assessments Engages Students and Refines Lesson Content

Abstract

A backward approach to lesson development start with identifying learner personas, planning out what content needs to be covered and assessment questions. These assessment questions can be used to outline the overall lesson and used to guide learners from one assessment question to another. Designing assessment questions rely on the amount of information that is being taught at any given point of a lesson. Learner's are only able to keep about 4 ± 1 new bits of information in working memory. This has ramifications about what kind of formative assessment questions are asked during the teaching portion of a class. Different types of exercise types can be used to reduce cognitive load in an assessment question.

This pilot study looks primarily at how faded examples help with learner engagement during a workshop with and without an auto code grader. Faded examples are questions that have the solution partially removed (i.e., faded out), and learners need to “fill in the blanks” to solve the solution. This allows the cognitive load of the question to be reduced by filling in parts of the solution that are not necessary for the conceptual point of the lesson. Faded examples are then compared to regular assessment questions that only have the question, and an empty space for the code solution.

Formative assessment questions are beneficial in the classroom, regardless of the two exercise types used. In the online setting where the lesson was conducted, a high percentage of responses were collected for the exercises compared to the expected amount of attrition. This suggests that learners will engage with formative assessment questions even when there is little or no interaction during the online chat system. Instructors are encouraged to provide ample time to complete these formative assessment questions during active class instruction, and provide additional learning resources for more details after the core concepts are taught.

4.1 Introduction

Putting together learning materials for learners typically usually leads to some kind of assessment as to whether or not the materials created are effective [69, 70]. However, the term “effective” is vague and ill-defined. This leads to the creation of using learning objectives, concrete tasks and goals learners are expected to meet at the end of teaching instruction. The benefit of creating learning objectives is they are able to be measured and assessed. These assessments can be used to gauge the efficacy of a lesson.

4.1.1 Mental Models and Cognitive Load

The Dreyfus model of skill acquisition describes how competency is acquired from novies to expert practitioner [95, 96]. Mental models are one way the bits of knowledge are connected together, with novices having a lower number of nodes and connections compared to the density of connections in expert practitioners. Mental models can be represented physically as concept map diagrams [70, 93]. Learner's existing mental models represent existing knowledge in their long-term memory (LTM). When teaching new concepts, more nodes are added to the model model of the learner. Before these new nodes can be solidified, they are stored in working memory (WM) before transitioning into short-term memory (STM) and LTM. [70, 93, 94].

Understanding how memory works in the context of teaching helps determine the amount of information presented. For novices, because they lack the necessary density of connections and knowledge (LTM), each new bit of information (STM) requires more processing power (WM). Concept maps help plan how much working memory and short-term memory learners are using during a lesson, and lessons should follow George Miller's 7 ± 2 rule of how many items people can store in their STM [97]. More recent research suggests that the STM is even smaller, two (2) to six (6) items [94], or 4 ± 1 [98].

4.1.2 Assessments and Learning Objectives

One of the steps in planning out a lesson using a backward design is creating the assessment questions. Assessments come in 2 main forms: formative and summative. Formative assessments are the exercises students do during the course of instruction in order to monitor student learning [70, 222]. They can interweave within a single instructional period (e.g., clicker type questions), or between instructional periods (e.g., quizzes, homework assignments). The main goal of formative assessments is to keep the learner engaged with the learning materials, and for both the instructor and learner to gauge learning by identifying areas that have not been grasped by the learner or areas that need more review. Formative assessments are typically low-stakes and given out frequently so the instructor can get an accurate gauge of how well the learners are doing [70, 222].

In contrast, summative assessments are given at the end of an instructional period to evaluate student learning [70, 222]. These types of assessments are the "summation" of multiple topics and can be given during a course of instruction (e.g., midterm exam) or towards the end of a course of instruction (e.g., final exam, thesis paper). Summative assessments are typically high-stakes and are used to gauge whether or not learners met learning objectives or prerequisite knowledge for subsequent courses or lessons [70, 222].

When designing assessment questions, there are many different ways questions can be formatted. Two that we explore the use of faded examples (i.e., fill in the blanks) and their performance to a "blank" problem. Faded examples are code questions with the solution partially removed (i.e., components faded out) and require the user to fill in the missing components to solve the question. This type of question focuses the learner on what is important about the question, and allows the instructor to lower the cognitive load of a question by filling in extraneous parts of the solution (e.g., function input parameters). This study then looks at time-to-completion and solution correctness when faded examples are compared to a regular question without pre-populated solutions, with and without an auto-code grader that can parse out the code submission and tell the student where exactly in their submission does not match the solution, instead of a binary correct-incorrect response.

4.2 Methods

Two (2) separate workshops were run. All data and access to the exercises were conducted through Qualtrics. Participants consented at the start of the study and provided a unique user identifier (Supplemental 4.5.1). The identifier was used to randomize participants into one (1) of four (4) treatment arms. All exercise questions were created using R learnr documents, depending on the treatment arm, each exercise question was paired with or without the gradethis auto code grader.

The workshop began with a pre-workshop exercise. During the workshop, three (3) 3 topics were covered and a short exercise (i.e., formative assessment) was presented at the end of each topic. The exercises fall into one of the four treatment arms for how the exercise is presented. The end of the main workshop content was followed up with a final summative assessment question. The amount of time to access the exercise and submit solution code was collected along with user identifier and code solution for all of the coding questions. Time to completion and code solutions were analysed and graded with a rubric.

4.2.1 Treatment Arms

The study created 4 treatment arms to look at exercise type and whether or not an auto-grader for real-time feedback helps with student learning: (1) blank exercise + no auto grader, (2) faded example + no auto grader, (3) blank exercise + auto grader, and (4) faded example + auto grader.

Blank exercises only contained the programming question and a space for participants to type and execute R code. Faded exercises contained the same programming question, but the space participants would type and execute R code would be pre-populated with the solution with function calls and arguments blanked out with a ___ (e.g., if the solution was read_csv("mydata.csv"), the faded example would look like ___(___)).

4.2.2 Randomization

Block randomization was used to randomize participants. A randomization list was pre-generated with a seed of 42, 4 treatment groups, block size of 8, and no stratification factors [223]. The second workshop re-generated the randomization list to balance out the group responses. Allocation ratios were adjusted for the second workshop due to varying amounts of participation and attrition after initial randomization from the workshop sign-in from the first workshop. Group 3 had 0 responses from the first workshop. The second randomization doubled the weight for Group 3, effectively creating a 5th treatment group, and used a block size of 10 for randomization. Participants were asked to sign-in at the start of the workshop, and their unique identifier was used for randomization.

4.2.3 Workshop Content

The workshop delivered was the "Tidy Data" portion of the ds4biomed materials It starts with the understanding that learners know about (1) spreadsheets, (2) loading data into R, (3) subsetting columns and rows of data, (4) calculating grouped aggregate summary statistics. The workshop for this study started off with a pre-workshop survey that served as an assessment of prerequisite knowledge. The content workshop remained exactly the same as previous workshops and studies that covered tidy data principles. The only major change was forcing time for the formative and summative assessment questions. Participants were given 5 minutes to work on the pre-workshop and formative assessment questions and the solution was reviewed before continuing to the next topic. 15 minutes were allocated for the summative assessment question at the very end of the workshop. The summative assessment question solution was provided after the workshop to fit within the 90-minute time limit for the workshop.

4.2.4 Exercise Questions

There was a total of 5 exercises presented to participants for this study: (1) one pre-workshop exercise that asks participants to perform a small data pipeline task that they would have been taught by now in the full 6-hour workshop version, (2) three formative exercise questions, and (3) one final summative assessment question. Each exercise question had a space for the participant to enter and run the existing R code. The exercise questions were written in R using the learnr package to create the documents, and the gradethis package to grade the solutions and provide feedback results. Workshop questions were deployed to shinyapps.io[1]. The links provided to the participants led them to the corresponding exercise across each of the treatment arms.

Pre-workshop Exercise

The pre-workshop exercise question is used as a baseline example since participants should be able to accomplish these tasks by this point of the workshop. The summative assessment example also uses concepts here during its data processing example. Participants were asked the following question:

> Please write the code for the following pipeline steps:
>
> 1. Load the tidyverse and readxl libraries.

[1] https://www.shinyapps.io/

2. Read in the Excel file located in: "data/medicaldata_tumorgrowth.xlsx" into a variable tumor.
3. Select the all the columns except Grp, and filter the rows such that Day is 0 or 20. Save this data subset into a variable tumor_subset.
4. We want to compare baseline tumor sizes (Day 0) with tumor sizes at Day 20 between each of the groups. Using tumor_subset, calculate the average tumor Size for each Grp and Day.
5. Save tumor_subset into a CSV file located in "data/tumorsubset.csv".

Formative Assessment Exercise 1

Take a look at the ebola dataset.

1. Tidy the dataset such that you get the dataset below.
2. You can use the last_col() to select the last column of the dataset.
3. Remember to drop missing values as the last step.

Formative Assessment Exercise 2

This is a different version of the ebola dataset.

1. Tidy the dataset such that you get the dataset below
2. Remember to drop missing values as the last step

Formative Assessment Exercise 3

This is a different version of the ebola dataset.

1. Tidy the dataset such that you get the dataset below
2. Remember to drop missing values as the last step

Summative Assessment

The summative assessment question is the same question from the post-workshop and long-term survey question given to participants who participated in the full workshop sessions. The question given in the previous studies did not ask participants to code up the solution, rather it asked how confident participants are in their ability to complete a set of data tasks. This study asks the participants to actually code up the results that will be graded.

This is the cmv dataset you will load:

1. Use the `readxl` library to load the ”data/cmv.xlsx” into a variable, `cmv`
2. Filter the `cmv` dataset such that only age > 65 are remaining. Save this to a variable, `cmv_subset`.
3. Save the `cmv_subset` variable to a csv file in ”data/cmv_subset.csv”.

1. Tidy the `cmv` dataset such that it looks like the clean dataset below. Save the tidy dataset into a variable, `cmv_tidy`.

1. In the `cmv_tidy` dataset, calculate the average age for each value of cmv.

4.2.5 Grading Rubric

A grading rubric was created to score the participant-submitted code solutions. A composite score for each exercise solution was created based on multiple factors. 1 point was awarded for each correct function call or similar function that achieves the same results. Grading was done with the actual treatment group blinded to the grader, scores were then combined with the rest of the full data for analysis.

Pre-Workshop Exercise 7 points total. Submission was graded on: (1) loading library packages using the library function, (2) loading data using the read_excel function, (3) subsetting columns using the select function, (4) subsetting rows based on a condition using the filter function, (5) aggregating summary statistics with the group_by function, (6) calculating the mean on grouped variables with the summarize function, and (7) writing out results to an external file with the read_csv function.

Exercise 1 4 points total. Submission was graded on: (1) reshaping the dataset with the the pivot_longer function, (2) correctly selecting the columns for the pivot_longer function, (3) correctly specifying the new columns after the pivot_longer function call, and (4) dropping missing values with drop_na function.

Exercise 2 6 points total. Submission was graded on: (1) reshaping the dataset with the the pivot_longer function, (2) correctly selecting the columns for the pivot_longer function, (3) splitting column values with the the separate function, (4) selecting the correct column to separate, (5) using the correct separating delimiter, and (6) ping missing values with drop_na function.

Exercise 3 3 points total. Submission was graded on: (1) reshaping the dataset with the the pivot_longer function, (2) reshaping the dataset with the the pivot_wider function, and (3) dropping missing values with drop_na function.

Summative Assessment Exercise 7 points total. Submission was graded on: (1) loading data using the read_excel function, (2) subsetting rows based on a condition using the filter function, (3) saving the filtered dataset to a file with the read_csv function, (4) reshaping and tidying the dataset with the the pivot_longer function, (5) aggregating summary statistics with the group_by function, (6) calculating the mean on grouped variables with the summarize function, and (7) writing out results to an external file with the read_csv function.

4.2.6 Analysis

Three (3) different sets of values were compared across each of the 4 treatment arms: (1) continuous variable of exercise solution score based on the grading rubric, (2) continuous variable of time to exercise completion, and (3) binary variable of whether the solution submitted is correct. The final results did not include the binary variable analysis due to sample size.

Non-code solutions were dropped (e.g., pasted in a URL instead of solution code) and not graded, as opposed to a score of 0. Participants who attended both workshop sessions had their responses removed from the repeat session. Participants who attempted the exercise questions more than once had the higher score kept for analysis, in the event of a score tie, the first submission was used for analysis, and other responses were dropped.

The R programming language and various supporting packages were used for analysis [164, 224–250].

4.3 Results

The interpretations of the results from this study are limited due to attrition and low response rates. The results are presented as preliminary data. Due to the low sample size, treating the final summative assessment question as a binary variable of correctness was not performed. In general, none of the results had any statistical significance.

4.3.1 Participants and Randomization

Two (2) workshop sessions that covered the same materials and exercise content were given. There were a total of 92 registrants (43 for session 1, and 49 for session 2) and 44 workshop attendees (25 for session 1, and 19 for session 2). 30 participants were randomized across 4 arms (Group 1: 8, Group 2: 7, Group 3: 8, Group 4: 7).

One (1) participant attended both workshops; Any data from the second workshop was dropped from analysis. One (1) participant attempted a question twice, since the graded scores were the same for both attempts, the first submission was kept for analysis. One (1) code solution appeared to be code for a different question and was not scored and counted in the analysis. A total of 16 randomized participants submitted at least 1 of the 5 exercises for analysis: 1 participant only took 1 exercise, 5 participants took 2 of the exercises, 2 participants took 3 of the exercises, 4 participants took 4 of the exercises, and 4 participants took all 5 exercises. Table 4.1 shows the number of responses for each exercise and treatment group.

Table 4.1: Number of code submissions for each group and exercise. There were a total of 16 randomized participants submitted at least one (1) of the five (5) exercises for analysis. The values listed in the Total column are the exercise response counts, and the values listed in the sum column are the group counts. The columns represent the treatment group, exercise 1 (ex1), exercise 2 (ex2), exercise 3 (ex3), pre-workshop exercise (pre), and post-workshop summative assessment (sum).

	treatment	ex1	ex2	ex3	pre	sum
1	Group 1	1	3	3	4	4
2	Group 2	3	4	2	3	2
3	Group 3	1	1	1	2	1
4	Group 4	5	4	3	3	3
5	Total	10	12	9	12	10

4.3.2 Exercise Scores

The pre-workshop exercise was given out as a means to cover pre-requisite knowledge for the summative assessment question. Otherwise, the summative assessment would be no different from any of the other exercise questions. Since the summative assessment exercise relied on knowledge that was covered in the pre-workshop exercise, the 5 participants who received a full score of 7 in the pre-workshop exercise were analysed first. Participants who scored well on the pre-workshop exercise also did well in all the other exercises (Figure 4.1). The 1 participant in Group 4 scored less than 50% in the summative assessment question. Figure 4.2 shows the percentage scores across each of the exercises and treatment group.

Some of the code solutions suggested that not every participant used the feature to execute their code before submission. Since we were unable to confirm if the autograder was used in these exercise treatment groups, a separate analysis was performed that ignored the auto-grader, so treatment groups 1 and 3 were combined together and treatment groups 2 and 4 were combined together (Figure 4.3). These preliminary results seem to suggest that faded examples do not affect code results during the formative assessments, but hinder results in the summative assessment question. These differences disappear when pre-workshop exercise

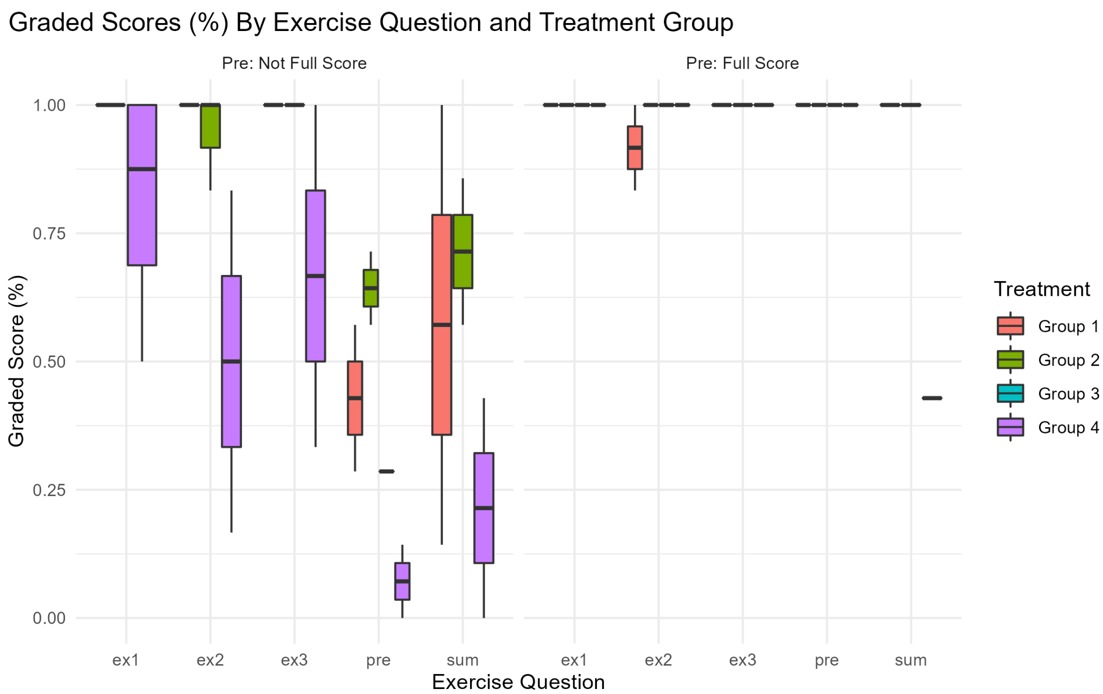


Figure 4.1: Distribution of scores for each treatment group across each exercise of respondents who received and did not receive a full score in the pre-workshop exercise. Participants who scored well on the pre-workshop exercise also did well in all the other exercises. The 1 participant in Group 4 scored less than 50% in the summative assessment question.

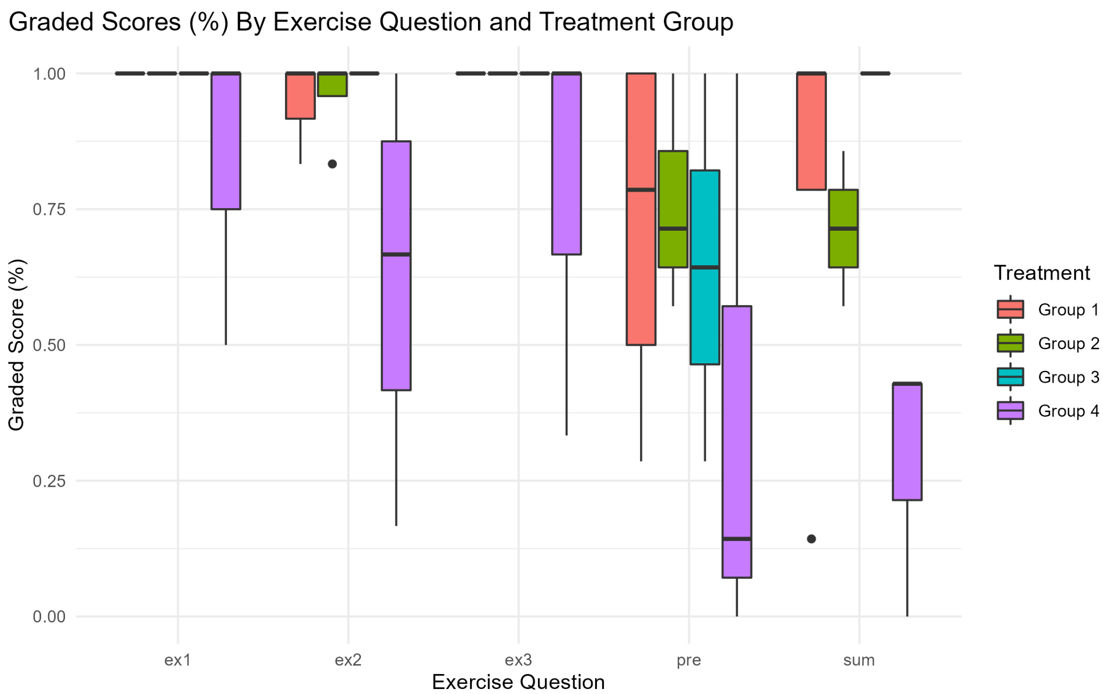


Figure 4.2: Distribution of exercise scores. Each participant's code submission was graded on a rubric. Scores are reported as a percentage because each exercise has a different total score. Score distributions were separate by whether or not a participant had a full score in the pre-workshop exercise (pre), since components of that are also used in the summative assessment question (sum). The other 3 exercises (ex1, ex2, ex3) were giving during the workshop as formative assessment questions. Results are also compared across 4 treatment groups: (1) Group 1: blank example with no auto code grader, (2) Group 2: faded example with no auto code grader, (3) Group 3: blank example with an auto code grader, and (4) Group 4: faded example with an auto code grader. Group 1 is the control group, and Group 4 is the main treatment of interest.

The low sample size in each of the groups makes it difficult to make definitive conclusions. The data currently shows that participants who were able to get a full score in the pre-workshop exercise tend to also do well on the remaining exercise questions (Figure 4.1). More variation between exercises exists with participants who did not receive a full score in the pre-workshop exercise.

performance is taken into account (Figure 4.4).

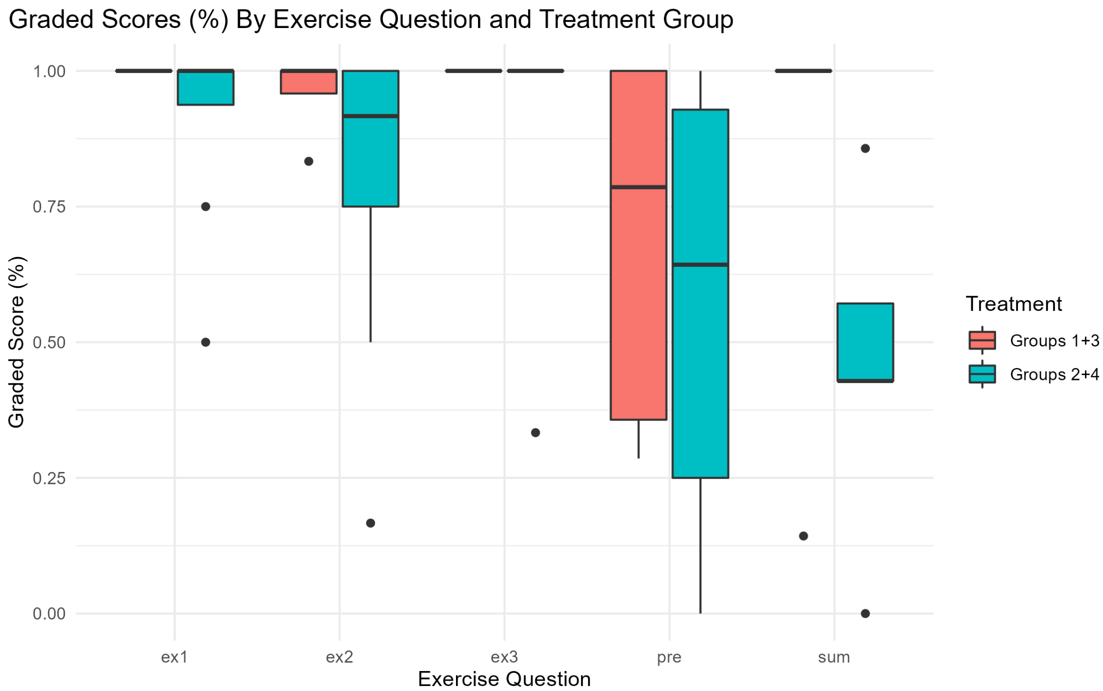


Figure 4.3: Distribution of exercise scores with the treatment groups combined by blank exercises (Groups 1 and 3) and faded examples (Groups 2 and 4). These groups were collapsed together since there was no way to track whether or not participants used the auto code grader. Preliminary results show that the faded examples do not differ from non-faded examples during the formative assessment questions, but the groups that used faded examples performed worse than those who were not given a faded example during the summative assessment question when all groups were provided an empty text field for the solution.

4.3.3 Time to Complete

Next, we wanted to see if there were any differences between types of exercises and amount of time to complete an exercise. Faded examples provide a skeleton of the code for learners to fill-in instead of writing all of the code from scratch. There did not seem to be any difference between time to exercise completion between the non-faded groups with the faded example groups. However these results suggest more about the amount of time learners needed to

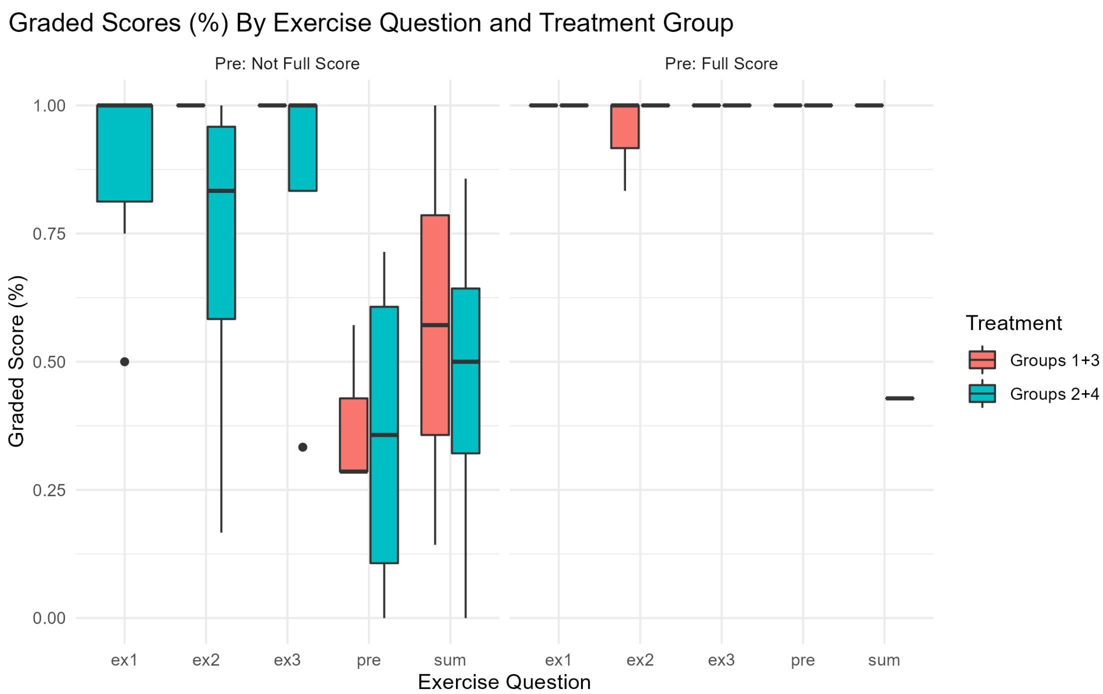


Figure 4.4: Distribution of scores for combined treatment groups across each exercise of respondents who received and did not receive a full score in the pre-workshop exercise. Participants who scored well on the pre-workshop exercise also did well in all the other exercises. The 1 participant in Group 4 scored less than 50% in the summative assessment question.

work on exercises.

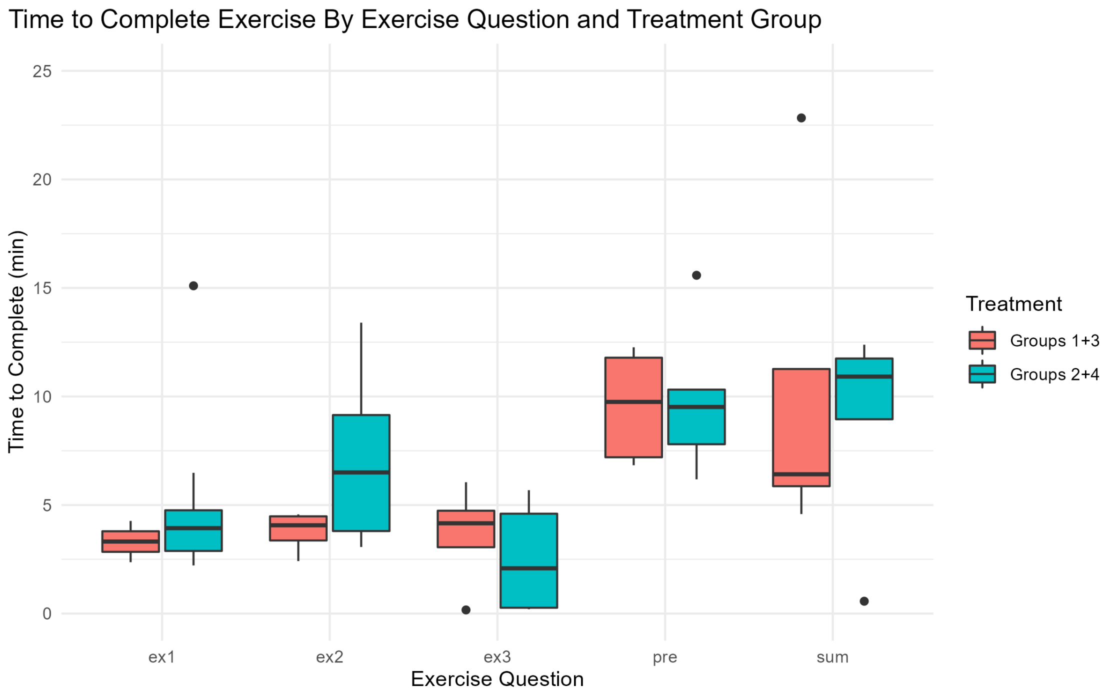


Figure 4.5: Distribution of time to complete exercises between treatment groups combined by blank exercises (Groups 1 and 3) and faded examples (Groups 2 and 4). These groups were collapsed together since there was no way to track whether or not participants used the auto code grader.

4.4 Discussion

This was a pilot study looking at how different kinds of assessment questions and how they can be used to refine workshop content in a backward design approach. Some of the code submissions suggested that not all students utilized all parts of the coding platform, so the use of the auto-grader could not assumed it was used in treatment arms 3 and 4. The analysis was run with all 4 treatment groups, and with just 2 groups comparing the blank question with the faded question. Even with the low sample size from the study, there are still findings that are applicable to data science instructors.

4.4.1 Formative Assessments Engage Students In Remote Workshops

The workshops were given in an online setting via Zoom. The observation from the instructor of the workshop was there was almost zero interaction of any kind during the workshop. The vast majority of chat messages came from the instructor posting the relevant links for each part of the workshop. Very few questions or discussions occur in the zoom chatting platform. There were more participants who took the exercises during the workshop than questions and comments in the Zoom meeting room chat. However, a surprisingly high number of students accomplished the exercises. The amount of attrition was less than expected, especially when comparing it to attrition from workshop registration to workshop attendance.

This finding suggests that even without grades as an incentive, participants who volunteering opted in to participate in the workshop were engaged with the materials, even in an online zoom setting.

4.4.2 Give Learners Time to Practice and Learn Asynchronously

Results from Figure 4.5 suggest more about how much more time it takes a novice learner to complete exercises compared to experts. Learners almost took the full 5-minutes for the formative assessment questions and almost the full 15-minutes for the summative assessment question. Using a conservative estimate for the instructor to go over the formative assessment solutions, learners take about 4-times as long to complete formative assessment questions, and almost 10-time as long to complete the summative assessment question.

During 1-hour of instruction, this means about 15-minutes would be needed for formative assessments, leaving about 45-minutes to complete the main teaching materials. In a workshop setting over multiple hours or sessions, lessons can be balanced across other parts of the workshop. However, in individual workshop settings, additional time for setup would need to be considered for every lesson.

This suggests that curating additional worked-out examples as formative assessment questions should to be provided to learners for asynchronous supplemental learning outside the main instructional period.

4.4.3 Conclusion

Although this was a pilot study, the results were able to show how pre-requisite knowledge plays a role in learning new skills. This study also showed how much more time learners need on answering formative and summative assessment questions compared to what an instructor imagines. Both of these findings translate to live teaching sessions where prior knowledge can affect the learner's ability to pick up new information, and planning how much content can be covered during an instructional period. This study did not have a treatment arm that did not use any formative assessment, but the participation rates from this study did hint that having formative assessments did force learners to be actively engaged with the materials. Using faded examples can reduce the amount of time spent on formative assessment questions when compared with a blank question box, but it is possible that faded examples reduce too much of the cognitive load for solving problem from scratch. This finding suggests that multiple types of formative assessment question types should be used in practice to balance teaching time, student engagement, and student cognitive load.

4.5 Supplemental

Supplemental materials for the "Refining Feedback and Guidance in Data Science Workshops: Making Time for Formative and Summative Assessments Engages Students and Refines Lesson Content".

4.5.1 Survey Questions

Participants were asked to create a unique identifier to track them throughout all studies without having to capture any identifiable information.

The question they were asked was adapted from the same question used in The Carpentries workshop feedback questions [62, 182–185, 251, 252]. Below is a reproduction of the question presented to participants:

> Please create a unique identifier. This unique identifier will be used to link your survey responses but keep your personal information anonymous.
>
> To create an identifier type in:
>
> 1. Number of siblings (as numeric) +
> 2. First two (2) letters of the city you were born in (lower-case) +
> 3. First three (3) letters of your current street (lower-case).
>
> E.g., Sherlock Homes has **1** brother, was born in **Po**rsmoth, and lives on **Bac**ker street would have the ID: **1pobac**

Chapter 5

Conclusion

The work in this dissertation started out exploring how to create better data science educational materials by creating and using learning personas. The learning materials provided the ability to self-study, but the research component focused on the in-class (virtual or in-person) learning. We discovered that the workshop was beneficial when looking at self-reported confidence of competing learning objective tasks. Having more instruction and guidance can help with getting over the activation energy to get started and learn these skills and the learning personas and relevant teaching examples can help with internal motivation to get over the initial learning curve.

However, we also discovered that there was a drop in confidence of competing learning objective tasks in the long-term study (at least 4 months later). This has lead us to conclude that more efforts should be spend on long-term learning, rather than creating more training materials in an already crowded market. For our audience of interest, working professionals in the medical field, this points to a problem costs to attending workshops and classes are wasted. These costs can be monetary (i.e., paying for the instruction), but there is also a time cost. The decrease in long-term confidence suggests that the value of these data sciences courses are short-term.

One hypothesis is from the lack of using and practicing the skills from the workshop. This plays into Malcolm Gladwell's "1,000 hour rule" and Ericsson's notion of "deliberate practice" [253–255]. Instead of simply providing datasets for examples, the personas we identified can be used to create and curate examples. This can help with deliberate and focused practice to actually maintain and build on knowledge and skills.

Providing datasets to work on data skills is one mechanism where learners can practice skills. However, these datasets need to be curated as learners may not always know where to find them let alone loading them for practice. The OpenX community can help with the curation of datasets, and communities like TidyTuesday publish weekly datasets where participants can explore datasets on their own. The personas we identified in this dissertation work can be used to refine these public datasets by providing different levels of questions to explore in a dataset. Combining focused case-study examples and the community gives a way for more experienced learners to mentor newer learners, and spreading out case-study examples can serve as a model for spaced repetition to practice foundational skills, while motivating learning a new skill.

While the work in this dissertation did create its own introductory workshop materials, it is possible that the effort in curating the learning materials could have been better served creating exercise questions to be used during and after the actual workshop. It may be possible that the materials used to teach need not be as domain focused to motivate learners, if more relevant examples come in the form of exercise questions and long-term case-study examples. New materials can link to existing materials and create a learning path, rather than re-creating the same materials. This leverages the plethora of existing data science materials, and puts a long-term focus for specific domains by working on domain-specific exercises. Focusing the domain-specific learning materials on exercises, instead of the introductory text, can help with the "reusability paradox" where the potential for reuse clashes is inversely related to pedagogical value because the specific context needed for pedagogical value, but reduces its ability to transfer to other contexts [256].

What may be more important for new learners is focusing on long-term learning and practice. This can be done with exercise questions and case-studies. Having a centralized location where datasets and exercise questions tagged with what skills are being practiced, extend the current work with the TidyTuesday project, and can be more beneficial to educators and learners without having to recreate yet another introductory text. These suggestions are mainly focused on the population of working adults. K-12, university, and higher education programs typically expose their students over the degree program to new topics and knowledge to build skills.

Communities of practice are one mechanism to help balance the amount of resources needed to teach. Auto-graders can help alleviate resources to check example solutions. Learning materials can be published in an open source platform, and communities can work on the long-term maintenance of these materials. These communities can also be centralized hubs where exercises can be posted. If these exercises were tagged with skill and domain information, educators can better filter teaching exercise for learners, and learners can explore examples on their own.

One of the main goals as educators is to inspire the next generation. Having compassion not only impacts and benefits those we interact with, but is also beneficial to ourselves [257]. Greg Wilson lists "The Rules" for teaching [70], and it begins with: "Be kind: all else are details".

Bibliography

[1] Patricia Zagallo, Jill McCourt, Robert Idsardi, Michelle K Smith, Mark Urban-Lurain, Tessa C Andrews, Kevin Haudek, Jennifer K Knight, John Merrill, Ross Nehm, et al. Through the eyes of faculty: Using personas as a tool for learner-centered professional development. *CBE—Life Sciences Education*, 18(4):ar62, 2019.

[2] American Medical Association. American Medical Association. https://www.ama-assn.org/.

[3] American Nurses Association. ANA Enterprise | American Nurses Association. https://www.nursingworld.org/.

[4] U.S. Bureau of Labor Statistics. U.S. Bureau of Labor Statistics. https://www.bls.gov/.

[5] The Comprehensive R Archive Network. The Comprehensive R Archive Network. https://cran.r-project.org/.

[6] Computer Science Teachers Association. CSTA. http://www.csteachers.org.

[7] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Pers-

son, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016.

[8] GAISE College Report ASA Revision Committee. Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016. Technical report, 2016.

[9] National Library of Medicine. About Us | NNLM. https://nnlm.gov/about.

[10] Observational Health Data Sciences and Informatics. OHDSI – Observational Health Data Sciences and Informatics.

[11] Observational Health Data Sciences and Informatics. OMOP Common Data Model – OHDSI.

[12] Hilary Mason and Chris Wiggins. A Taxonomy of Data Science, September 2010.

[13] Lee S. Shulman. Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2):4–14, February 1986.

[14] Index | TIOBE - The Software Quality Company. https://www.tiobe.com/tiobe-index/.

[15] John Pruitt and Tamara Adlin. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufmann, Amsterdam ; Boston, 1st edition edition, April 2006.

[16] Bruce Tognazzini. *Tog on Software Design by Bruce Tognazzini*. Addison-Wesley Professional, January 1748.

[17] John W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

[18] Gil Press. A Very Short History Of Data Science. https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/, May 2013.

[19] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Pub. Co, 1977.

[20] John W Tukey. The technical tools of statistics. *The American Statistician*, 19(2):23–28, 1965.

[21] Thomas E Bradstreet. Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician*, 50(1):69–78, 1996.

[22] Association for Computing Machinery. ACM Software System Award. https://awards.acm.org/award-winners/CHAMBERS_6640862, 1998.

[23] William S. Cleveland. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review / Revue Internationale de Statistique*, 69(1):21–26, 2001.

[24] Data Science Journal. http://datascience.codata.org/about/.

[25] Journal of Data Science. https://jds-online.org/journal/JDS/information/about-journal.

[26] Hal Varian on how the Web challenges managers, January 2009.

[27] Drew Conway. The Data Science Venn Diagram. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, September 2010.

[28] Thomas H. Davenport and D. J. Patil. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, October 2012.

[29] Megan Smith. The White House Names Dr. DJ Patil as the First U.S. Chief Data Scientist. https://obamawhitehouse.archives.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist, February 2015.

[30] Saurabh Tyagi. How Fortune 500 is embracing open source technology, July 2016.

[31] James Guszcza. Data science and open source analytics | Deloitte US. https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/open-source-analytics.html, May 2015.

[32] Sebastian Kirsch. How Open Source is Driving the Future of Data Science, April 2021.

[33] Arfon M. Smith, Kyle E. Niemeyer, Daniel S. Katz, Lorena A. Barba, George Githinji, Melissa Gymrek, Kathryn D. Huff, Christopher R. Madan, Abigail Cabunoc Mayes, Kevin M. Moerman, Pjotr Prins, Karthik Ram, Ariel Rokem, Tracy K. Teal, Roman Valls Guimera, and Jacob T. Vanderplas. Journal of Open Source Software (JOSS): Design and first-year review. *PeerJ Computer Science*, 4:e147, February 2018.

[34] National Institutes of Health. NIH Strategic Plan for Data Science | Data Science at NIH. https://datascience.nih.gov/nih-strategic-plan-data-science, September 2020.

[35] Sean Kross, Roger D. Peng, Brian S. Caffo, Ira Gooding, and Jeffrey T. Leek. The Democratization of Data Science Education. *The American Statistician*, 74(1):1–7, January 2020.

[36] Coursera. Top Free Courses - Learn Free Online. https://www.coursera.org/search?query=free.

[37] Udacity. Data Science Online Courses & Programs. https://www.udacity.com/school-of-data-science.

[38] EdX. Data Analysis Courses. https://www.edx.org/learn/data-analysis.

[39] Bookdown. All books on bookdown.org | Bookdown. https://bookdown.org/home/archive/.

[40] Executable Book Project. Gallery of Jupyter Books. https://executablebooks.org/en/latest/gallery.html#.

[41] R: The R Project for Statistical Computing. https://www.r-project.org/.

[42] Richard A. Becker. A Brief History of S. In Werner A. Müller, Peter Schuster, Peter Dirschedl, and Rüdiger Ostermann, editors, *Computational Statistics*, pages 81–110. Physica-Verlag HD, Heidelberg, 1994.

[43] Kurt Hornik. Announce: CRAN. https://stat.ethz.ch/pipermail/r-announce/1997/000001.html, Wed Apr 23 08:40:28 CEST 1997.

[44] David Smith. Over 16 years of R Project history. https://blog.revolutionanalytics.com/2016/03/16-years-of-r-history.html, March 2016.

[45] Ross Ihaka. The R Project: A Brief History and Thoughts About the Future. page 34.

[46] Hadley Wickham. *Practical Tools for Exploring Data and Models.* PhD thesis, Iowa State University, 2008.

[47] JJ Allaire. RStudio, PBC. https://rstudio.comhttps://www.rstudio.com/blog/rstudio-pbc/, January 2020.

[48] RStudio. About RStudio. https://www.rstudio.com/about/.

[49] David Smith. Hadley Wickham on why he created all those R packages. https://blog.revolutionanalytics.com/2015/07/hadley-profile.html, July 2015.

[50] Hadley Wickham. Tidy Data. *Journal of Statistical Software*, 59(1):1–23, September 2014.

[51] Institute of Mathematical Statistics | COPSS Presidents' Award: Hadley Wickham, September 2019.

[52] Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, December 2016.

[53] Max Kuhn and Julia Silge. Tidy Modeling with R. https://www.tmwr.org/, July 2021.

[54] Charles Severance. Guido van Rossum: The Early Years of Python. *Computer*, 48(02):7–9, February 2005.

[55] IPython Development Team. History. https://ipython.readthedocs.io/en/stable/about/history.html.

[56] Dask. Keynote - Peter Wang and Travis Oliphant | Dask Summit 2021, June 2021.

[57] The Apache Software Foundation. Apache Arrow. https://arrow.apache.org/, 2021.

[58] Ursa Labs. https://ursalabs.org/.

[59] Voltron Data. https://voltrondata.com/.

[60] WebAssembly. https://webassembly.org/.

[61] The Carpentries: How We Operate. https://carpentries.github.io/instructor-training/21-carpentries/..//21-carpentries/index.html.

[62] Kari L. Jordan. Carpentries 2020 Annual Report. https://carpentries.org/annual-report-2020/.

[63] Daniel Chen. Point of contact for each lesson · Issue #11 · carpentries/maintainer-RFCs. https://github.com/carpentries/maintainer-RFCs/issues/11, 2021.

[64] The Carpentries. Carpentry Trainer Training Program. https://carpentries.github.io/trainer-training/.

[65] Russell Shackelford, Andrew McGettrick, Robert Sloan, Heikki Topi, Gordon Davies, Reza Kamali, James Cross, John Impagliazzo, Richard LeBlanc, and Barry Lunt. Computing Curricula 2005: The Overview Report. In *Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '06, pages 456–457, New York, NY, USA, March 2006. Association for Computing Machinery.

[66] CC2020 Task Force. *Computing Curricula 2020: Paradigms for Global Computing Education*. ACM, New York, NY, USA, November 2020.

[67] ACM Data Science Task Force. Computing Competencies for Undergraduate Data Science Curricula. Technical report, 2021.

[68] Computer Science Teachers Association. CSTA K-12 Computer Science Standards, 2017.

[69] Susan A Ambrose, Michael W Bridges, Michele DiPietro, Marsha C Lovett, and Marie K Norman. *How Learning Works: Seven Research-Based Principles for Smart Teaching*. John Wiley & Sons, 2010.

[70] Greg Wilson. *Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them.* Taylor & Francis, 2019.

[71] Anna Bargagliotti, Christine Franklin, Pip Arnold, Rob Gould, Sheri Johnson, Leticia Perez, and Denise A. Spangler. *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education.* American Statistical Association, Alexandria, VA, December 2020.

[72] Bureau of Labor Statistics, U.S. Department of Labor. Occupational Outlook Handbook. https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm, 2016.

[73] Foster Provost and Tom Fawcett. *Data Science for Business.* O'Reilly Media, Inc., August 2013.

[74] Jack Z. Sissors. What is a Market? *Journal of Marketing*, 30(3):17–21, July 1966.

[75] William Winston and Art Weinstein. *Defining Your Market: Winning Strategies for High-Tech, Industrial, and Service Firms.* Routledge, New York, 1st edition edition, July 1998.

[76] Geoffrey A. Moore. *Crossing the Chasm: Marketing and Selling Technology Products to Mainstream Customers.* HarperBusiness, New York, N.Y., 1991.

[77] Lynn B. Upshaw. *Building Brand Identity: A Strategy for Success in a Hostile Marketplace.* Wiley, New York, 1st edition edition, June 1995.

[78] Sheila Mello. *Customer-Centric Product Definition: The Key to Great Product Development.* PDC Professional Publishing, Boston, October 2003.

[79] John M. Carroll, editor. *Scenario-Based Design: Envisioning Work and Technology in System Development.* Wiley, Weinheim, 1st edition edition, May 1995.

[80] Norunn Mikkelson and Wai On Lee. Incorporating user archetypes into scenario-based design. In *Proceedings of the Ninth Annual Usability Professional' Association (UPA) Conference*, Asheville, North Carolina, 2000.

[81] Janice C. Redish and JoAnn T. Hackos. *User and Task Analysis for Interface Design*. John Wiley & Sons, Inc., New York, 1st edition edition, February 1998.

[82] L.L. Constantine and Lockwood, Ltd. Personas. https://web.archive.org/web/20131111174207/www.foruse.com/newsletter/foruse15.htm, 2001.

[83] Karen Holtzblatt and Hugh Beyer. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann, San Francisco, Calif, 1st edition edition, September 1997.

[84] Karen Holtzblatt. Personas and Contextual Design. https://web.archive.org/web/20050314093552/http://www.incent.com/community/design_corner/0 2002.

[85] MF Tahir. Who's on the other side of your software creating user profiles through contextual inquiry. In *Proceedings of the Usability Professional Conference '97*, Bloomingdale, IL, 1997. The Usability Professionals' Association.

[86] Alan Cooper. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams - Pearson Education, 1999.

[87] Barry Schwartz. *The Paradox of Choice: Why More Is Less, Revised Edition*. Ecco, New York, revised ed. edition edition, May 2016.

[88] RStudio. Learner Personas, 2019.

[89] Software Carpentry. Learner Profiles. http://software-carpentry.org//audience/.

[90] Monica Chin. Students who grew up with search engines might change STEM education forever. https://www.theverge.com/22684730/students-file-folder-directory-structure-education-gen-z, September 2021.

[91] Jody Macgregor. Students don't know what files and folders are, professors say. *PC Gamer*, September 2021.

[92] Caitlin Kelleher and Randy Pausch. Lowering the barriers to programming: A taxonomy of programming environments and languages for novice programmers. *ACM Computing Surveys*, 37(2):83–137, June 2005.

[93] Christina Koch and Greg Wilson. Software carpentry: Instructor Training, June 2016.

[94] Felienne Hermans. *The Programmer's Brain*. Manning, 2021.

[95] Stuart E Dreyfus and Hubert L Dreyfus. A five-stage model of the mental activities involved in directed skill acquisition. Technical report, California Univ Berkeley Operations Research Center, 1980.

[96] Patricia Benner. Using the Dreyfus Model of Skill Acquisition to Describe and Interpret Skill Acquisition and Clinical Judgment in Nursing Practice and Education. *Bulletin of Science, Technology & Society*, 24(3):188–199, June 2004.

[97] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[98] David Didau and Nick Rose. *What Every Teacher Needs to Know About Psychology*. John Catt Educational, Melton, Woodbridge, August 2016.

[99] Karl W. Broman and Kara H. Woo. Data Organization in Spreadsheets. *The American Statistician*, 72(1):2–10, January 2018.

[100] Lorin W Anderson, Benjamin Samuel Bloom, et al. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman,, 2001.

[101] Benjamin S. Bloom. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. Addison-Wesley Longman Ltd, London, 2nd edition edition edition, June 1956.

[102] John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, 14(1):4–58, January 2013.

[103] Susana Masapanta-Carrión and J. Ángel Velázquez-Iturbide. A Systematic Review of the Use of Bloom's Taxonomy in Computer Science Education. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, pages 441–446, New York, NY, USA, February 2018. Association for Computing Machinery.

[104] Donald Clark Plan B: Bogus pyramids: Learning methods, Maslow and Bloom, July 2020.

[105] L. Dee Fink. *Creating Significant Learning Experiences: An Integrated Approach to Designing College Courses*. John Wiley & Sons, July 2013.

[106] Mary C. M. Anderson and Keith W. Thiede. Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, 128(1):110–118, May 2008.

[107] Rahul Banerjee, Paul George, Cedric Priebe, and Eric Alper. Medical student awareness of and interest in clinical informatics. *Journal of the American Medical Informatics Association*, 22(e1):e42–e47, April 2015.

[108] American Medical Association. Student interest in informatics outpaces opportunities: Study.

[109] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[110] Robert Hoyt and Victoria Wangia-Anderson. An overview of two open interactive computing environments useful for data science education. *JAMIA Open*, 1(2):159–165, October 2018.

[111] Philip R O Payne, Elmer V Bernstam, and Justin B Starren. Biomedical informatics meets data science: Current state and future directions for interaction. *JAMIA Open*, 1(2):136–141, October 2018.

[112] American Medical Association. Accelerating Change in Medical Education. https://www.ama-assn.org/education/accelerating-change-medical-education, 2021.

[113] Randall Owen. The Ethical Intersection of Healthcare and Technology, September 2017.

[114] R.Scott Nolen. Artificial intelligence & veterinary medicine. https://www.avma.org/javma-news/2020-07-15/artificial-intelligence-veterinary-medicine, June 2020.

[115] Edward H Shortliffe. The adolescence of AI in medicine: Will the field come of age in the'90s? *Artificial intelligence in medicine*, 5(2):93–106, 1993.

[116] Elizabeth Green. *Building a Better Teacher: How Teaching Works (and How to Teach It to Everyone)*. W. W. Norton & Company, first edition, August 2014.

[117] Daniel Chen. Data science figure, December 2020.

[118] Joel Ostblom and Tiffany Timbers. Opinionated practices for teaching reproducibility: Motivation, guided instruction and practice. *arXiv:2109.13656 [cs, stat]*, September 2021.

[119] Hanan Aboumatar and Robert A. Wise. Notice of Retraction. Aboumatar et al. Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease: A Randomized Clinical Trial. JAMA. 2018;320(22):2335-2343. *JAMA*, 322(14):1417–1418, October 2019.

[120] Excel: Why using Microsoft's tool caused Covid-19 results to be lost. *BBC News*, October 2020.

[121] Lena Wallensteen, Marius Zimmermann, Malin Sandberg, Anton Gezelius, Anna Nordenström, Tatja Hirvikoski, and Svetlana Lajic. Retraction notice to"Evaluation of behavioral problems after prenatal dexamethasone treatment in Swedish adolescents at risk of CAH" [Hormones and Behavior 85C (2016) 5-11]. *Hormones and Behavior*, 103:140, July 2018.

[122] Harvey Whitehouse, Pieter François, Patrick E. Savage, Thomas E. Currie, Kevin C. Feeney, Enrico Cioni, Rosalind Purcell, Robert M. Ross, Jennifer Larson, John Baines, Barend ter Haar, Alan Covey, and Peter Turchin. Retraction Note: Complex societies precede moralizing gods throughout world history. *Nature*, 595(7866):320–320, July 2021.

[123] Barry R. Zeeberg, Joseph Riss, David W. Kane, Kimberly J. Bussey, Edward Uchio, W. Marston Linehan, J. Carl Barrett, and John N. Weinstein. Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, 5(1):80, June 2004.

[124] Mark Ziemann, Yotam Eren, and Assam El-Osta. Gene name errors are widespread in the scientific literature. *Genome Biology*, 17(1):177, August 2016.

[125] Sara Gerke, Timo Minssen, and Glenn Cohen. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, pages 295–336, 2020.

[126] Niels Peek, Carlo Combi, Roque Marin, and Riccardo Bellazzi. Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial Intelligence in Medicine*, 65(1):61–73, September 2015.

[127] <img src='/content/dam/sm-news/images/2018/01/hannon-patty-90 jpg img 620 high png/1515426935980 png' alt='Patricia Hannon'> Patricia Hannon Patricia Hannon is the associate editor of Stanford Medicine magazine in the Office of Communications Email her at phannon@stanford.edu. Researchers say use of artificial intelligence in medicine raises ethical questions. http://med.stanford.edu/news/all-news/2018/03/researchers-say-use-of-ai-in-medicine-raises-ethical-questions.html.

[128] Michael J. Rigby. Ethical Dimensions of Using Artificial Intelligence in Health Care. *AMA Journal of Ethics*, 21(2):121–124, February 2019.

[129] Nicole Wetsman. WHO outlines principles for ethics in health AI. https://www.theverge.com/2021/6/30/22557119/who-ethics-ai-healthcare, June 2021.

[130] Oscar Baruffa. Big Book of R. https://www.bigbookofr.com/, 2021.

[131] American Medical Association. Education. https://www.ama-assn.org/education.

[132] JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron

Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *Rmarkdown: Dynamic Documents for r*, 2021.

[133] Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017.

[134] Stefan Milton Bache and Hadley Wickham. *Magrittr: A Forward-Pipe Operator for r*, 2020.

[135] Coen A. Bernaards and Robert I.Jennrich. Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, 65:676–696, 2005.

[136] Taiyun Wei and Viliam Simko. *R Package 'Corrplot': Visualization of a Correlation Matrix*, 2021.

[137] Gábor Csárdi and Rich FitzJohn. *Progress: Terminal Progress Bars*, 2019.

[138] David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *Xtable: Export Tables to LaTeX or HTML*, 2019.

[139] Andrie de Vries and Brian D. Ripley. *Ggdendro: Create Dendrograms and Tree Diagrams Using 'Ggplot2'*, 2020.

[140] Sacha Epskamp. *semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output*, 2019.

[141] Sam Firke. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2021.

[142] Marek Gagolewski. Stringi: Fast and portable character string processing in R. *Journal of Statistical Software*, 2021.

[143] Marek Gagolewski. *Stringi: Fast and Portable Character String Processing in R*, 2021.

[144] Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Antônio Pedro, Sciaini, Marco, Scherer, and Cédric. *Viridis - Colorblind-Friendly Color Maps for r*, 2021.

[145] Jasper Ginn and Julia Silge. *qualtRics: Download 'qualtrics' Survey Data*, 2021.

[146] David Gohel. *Flextable: Functions for Tabular Reporting*, 2021.

[147] David Gohel. *Officer: Manipulation of Microsoft Word and PowerPoint Documents*, 2021.

[148] Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011.

[149] Lionel Henry and Hadley Wickham. *Purrr: Functional Programming Tools*, 2020.

[150] Lionel Henry and Hadley Wickham. *Rlang: Functions for Base Types and Core r and 'tidyverse' Features*, 2021.

[151] Lionel Henry and Hadley Wickham. *Tidyselect: Select from a Set of Strings*, 2021.

[152] Jim Hester, Hadley Wickham, and Gábor Csárdi. *Fs: Cross-platform File System Operations Based on 'Libuv'*, 2021.

[153] Jim Hester and Jennifer Bryan. *Glue: Interpreted String Literals*, 2021.

[154] Alboukadel Kassambara and Fabian Mundt. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020.

[155] Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. MVN: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162, 2014.

[156] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *Cluster: Cluster Analysis Basics and Extensions*, 2021.

[157] Kirill Müller. *Here: A Simpler Way to Find Your Files*, 2020.

[158] Kirill Müller and Hadley Wickham. *Tibble: Simple Data Frames*, 2021.

[159] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014.

[160] Jeroen Ooms. The jsonlite package: A practical and consistent mapping between JSON data and r objects. *arXiv:1403.2805 [stat.CO]*, 2014.

[161] Jeroen Ooms. *Writexl: Export Data Frames to Excel 'xlsx' Format*, 2021.

[162] Patrick O. Perry. *Utf8: Unicode Text Processing*, 2021.

[163] Gilles Raiche and David Magis. *nFactors: Parallel Analysis and Other Non Graphical Solutions to the Cattell Scree Test*, 2020.

[164] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[165] William Revelle. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2021.

[166] Yves Rosseel. Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

[167] Terry Therneau and Beth Atkinson. *Rpart: Recursive Partitioning and Regression Trees*, 2019.

[168] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *Dplyr: A Grammar of Data Manipulation*, 2021.

[169] Hadley Wickham. *Forcats: Tools for Working with Categorical Variables (Factors)*, 2021.

[170] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[171] Hadley Wickham, Jim Hester, and Jennifer Bryan. *Readr: Read Rectangular Text Data*, 2021.

[172] Hadley Wickham. *Rvest: Easily Harvest (Scrape) Web Pages*, 2021.

[173] Hadley Wickham. *Stringr: Simple, Consistent Wrappers for Common String Operations*, 2019.

[174] Hadley Wickham. *Tidyr: Tidy Messy Data*, 2021.

[175] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[176] Yihui Xie. *Dynamic Documents with R and Knitr*. Chapman and Hall/CRC, Boca Raton, Florida, second edition, 2015.

[177] Yihui Xie. Knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014.

[178] Yihui Xie. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*, 2021.

[179] Yihui Xie, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook*. Chapman and Hall/CRC, Boca Raton, Florida, 2020.

[180] Yihui Xie, J.J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida, 2018.

[181] Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2021.

[182] Kari Jordan, François Michonneau, and Belinda Weaver. Analysis of Software and Data Carpentry's Pre- and Post-Workshop Surveys. Technical report, Zenodo, July 2018.

[183] Kari L. Jordan and François Michonneau. Analysis of The Carpentries Long-Term Surveys (April 2020). Technical report, Zenodo, March 2020.

[184] Kari Jordan. Analysis of The Carpentries Long-Term Impact Survey. Technical report, Zenodo, July 2018.

[185] Kari L. Jordan, Ben Marwick, Belinda Weaver, Naupaka Zimmerman, Jason Williams, Tracy Teal, Erin Becker, Jonah Duckles, Beth Duckles, and Elizabeth Wickes. Analysis of the Carpentries' Long-Term Feedback Survey. Technical report, Zenodo, October 2017.

[186] Qualtrics. Qualtrics, 2005.

[187] Wan Nor Arifin. Exploratory factor analysis and Cronbach's alpha. https://wnarifin.github.io/workshop/qvw2017/efa.pdf, October 2017.

[188] Timothy A. Brown. *Confirmatory Factor Analysis for Applied Research, Second Edition*. The Guilford Press, New York ; London, second edition edition, January 2015.

[189] Theodoros A. Kyriazos. Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General. *Psychology*, 09(08):2207, August 2018.

[190] Erika J. Wolf, Kelly M. Harrington, Shaunna L. Clark, and Mark W. Miller. Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and psychological measurement*, 76(6):913–934, December 2013.

[191] Daniel J. Mundfrom, Dale G. Shaw, and Tian Lu Ke. Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing*, 5(2):159–168, June 2005.

[192] Michelle C. Dunn and Philip E. Bourne. Building the biomedical data science workforce. *PLoS Biology*, 15(7):1–9, July 2017.

[193] National Institutes of Health. Big Data to Knowledge. https://commonfund.nih.gov/bd2k, June 2013.

[194] Stefan Milton Bache and Hadley Wickham. *Magrittr: A Forward-Pipe Operator for r*, 2020.

[195] Hanbo Chen. *VennDiagram: Generate High-Resolution Venn and Euler Plots*, 2021.

[196] David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *Xtable: Export Tables to LaTeX or HTML*, 2019.

[197] Sam Firke. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2021.

[198] Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Antônio Pedro, Sciaini, Marco, Scherer, and Cédric. *Viridis - Colorblind-Friendly Color Maps for r*, 2021.

[199] Jasper Ginn and Julia Silge. *qualtRics: Download 'qualtrics' Survey Data*, 2021.

[200] Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011.

[201] Lionel Henry and Hadley Wickham. *Purrr: Functional Programming Tools*, 2020.

[202] Lionel Henry and Hadley Wickham. *Rlang: Functions for Base Types and Core r and 'tidyverse' Features*, 2021.

[203] Jim Hester, Hadley Wickham, and Gábor Csárdi. *Fs: Cross-platform File System Operations Based on 'Libuv'*, 2021.

[204] Jim Hester and Jennifer Bryan. *Glue: Interpreted String Literals*, 2021.

[205] Matthias Kohl. *MKpower: Power Analysis and Sample Size Calculation*, 2020.

[206] Kirill Müller. *Here: A Simpler Way to Find Your Files*, 2020.

[207] Kirill Müller and Hadley Wickham. *Tibble: Simple Data Frames*, 2021.

[208] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014.

[209] Jeroen Ooms. The jsonlite package: A practical and consistent mapping between JSON data and r objects. *arXiv:1403.2805 [stat.CO]*, 2014.

[210] Jeroen Ooms. *Writexl: Export Data Frames to Excel 'xlsx' Format*, 2021.

[211] David Robinson, Alex Hayes, and Simon Couch. *Broom: Convert Statistical Objects into Tidy Tibbles*, 2021.

[212] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *Dplyr: A Grammar of Data Manipulation*, 2021.

[213] Hadley Wickham. *Forcats: Tools for Working with Categorical Variables (Factors)*, 2021.

[214] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[215] Hadley Wickham, Jim Hester, and Jennifer Bryan. *Readr: Read Rectangular Text Data*, 2021.

[216] Hadley Wickham. *Rvest: Easily Harvest (Scrape) Web Pages*, 2021.

[217] Hadley Wickham. *Stringr: Simple, Consistent Wrappers for Common String Operations*, 2019.

[218] Hadley Wickham. *Tidyr: Tidy Messy Data*, 2021.

[219] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[220] Allison Horst. Allisonhorst/stats-illustrations: R & stats illustrations by @allison_horst. https://github.com/allisonhorst/stats-illustrations, 2021.

[221] Monica Chin. File not found. A generation that grew up with Google is forcing professors to rethink their lesson plans.Students who grew up with search engines might change STEM education forever. https://www.theverge.com/22684730/students-file-folder-directory-structure-education-gen-z, September 2021.

[222] Carnegie Mellon University. Formative vs Summative Assessment - Eberly Center - Carnegie Mellon University. https://www.cmu.edu/teaching/assessment/basics/formative-summative.html.

[223] Sealed Envelope Ltd. Create a blocked randomisation list | Sealed Envelope. https://www.sealedenvelope.com/simple-randomiser/v1/lists, 2021.

[224] JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *Rmarkdown: Dynamic Documents for r*, 2021.

[225] Hanbo Chen. *VennDiagram: Generate High-Resolution Venn and Euler Plots*, 2021.

[226] David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *Xtable: Export Tables to LaTeX or HTML*, 2019.

[227] Sam Firke. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2021.

[228] Jasper Ginn and Julia Silge. *qualtRics: Download 'qualtrics' Survey Data*, 2021.

[229] Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011.

[230] Lionel Henry and Hadley Wickham. *Purrr: Functional Programming Tools*, 2020.

[231] Lionel Henry and Hadley Wickham. *Tidyselect: Select from a Set of Strings*, 2021.

[232] Jim Hester, Hadley Wickham, and Gábor Csárdi. *Fs: Cross-platform File System Operations Based on 'Libuv'*, 2021.

[233] Jim Hester and Jennifer Bryan. *Glue: Interpreted String Literals*, 2021.

[234] Kirill Müller. *Here: A Simpler Way to Find Your Files*, 2020.

[235] Kirill Müller and Hadley Wickham. *Tibble: Simple Data Frames*, 2021.

[236] Jeroen Ooms. The jsonlite package: A practical and consistent mapping between JSON data and r objects. *arXiv:1403.2805 [stat.CO]*, 2014.

[237] Jeroen Ooms. *Writexl: Export Data Frames to Excel 'xlsx' Format*, 2021.

[238] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *Dplyr: A Grammar of Data Manipulation*, 2021.

[239] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[240] Hadley Wickham, Jim Hester, and Jennifer Bryan. *Readr: Read Rectangular Text Data*, 2021.

[241] Hadley Wickham and Jennifer Bryan. *Readxl: Read Excel Files*, 2019.

[242] Hadley Wickham. *Rvest: Easily Harvest (Scrape) Web Pages*, 2021.

[243] Hadley Wickham. *Stringr: Simple, Consistent Wrappers for Common String Operations*, 2019.

[244] Hadley Wickham. *Tidyr: Tidy Messy Data*, 2021.

[245] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[246] Yihui Xie. *Dynamic Documents with R and Knitr*. Chapman and Hall/CRC, Boca Raton, Florida, second edition, 2015.

[247] Yihui Xie. Knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014.

[248] Yihui Xie. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*, 2021.

[249] Yihui Xie, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook*. Chapman and Hall/CRC, Boca Raton, Florida, 2020.

[250] Yihui Xie, J.J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida, 2018.

[251] Kari L. Jordan, Ben Marwick, Jonah Duckles, Naupaka Zimmerman, and Erin Becker. Analysis of Software Carpentry's Post-Workshop Surveys. Technical report, Zenodo, July 2017.

[252] Kari Jordan. Data Carpentry Assessment Report: Analysis of Post-Workshop Survey Results. Technical report, Zenodo, October 2016.

[253] Malcolm Gladwell. *Outliers: The Story of Success*. Back Bay Books, New York, 1st edition edition, June 2011.

[254] K. Anders Ericsson and Kyle W. Harwell. Deliberate Practice and Proposed Limits on the Effects of Practice on the Acquisition of Expert Performance: Why the Original Definition Matters and Recommendations for Future Research. *Frontiers in Psychology*, 10, 2019.

[255] K. A. Ericsson and A. C. Lehmann. Expert and Exceptional Performance: Evidence of Maximal Adaptation to Task Constraints. *Annual Review of Psychology*, 47(1):273–305, 1996.

[256] David Wiley. The Reusability Paradox. https://opencontent.org/docs/paradox.html, August 2002.

[257] Stephen Trzeciak and Anthony Mazzarelli. *Compassionomics*. Studer Group, Pensacola, FL, 1st edition edition, May 2019.