Understanding the Practices and Challenges of Teaching with Data in Undergraduate Social Science Courses at Carnegie Mellon University

Melanie A. Gainey, Hannah Gunderman, Emma Slayton, Ryan Splenda

(authors listed alphabetically)

University Libraries, Carnegie Mellon University, Pittsburgh, PA, USA

Table of Contents

1. Introduction

Over the past several years, the concept of data has permeated every aspect of life. While in the past some might have erroneously assumed that data analysis or data manipulation was solely for people engaged in the "hard" sciences, it is now widely recognized that people from the social sciences and humanities are incorporating data analysis into their work. Each human has essentially become a walking dataset in our increasingly digital world, which makes understanding how to work with and critically interpret data increasingly more important. Social scientists have increasingly picked up the charge of fostering smart data practices in research, both in terms of using qualitative and quantitative data. However, while these practices are becoming more commonplace, there are still many aspects of data use or analysis that remain confusing to some students in social science courses. It is important for us as librarians who serve this community to better understand how our instructors in the social sciences at Carnegie Mellon University (CMU) explain how to use data and its associated concepts, so that we in turn can provide better support to these communities.

CMU is one of 20 universities across the United States exploring this topic under the guidance of Ithaka S+R, which is a subset of ITHAKA, a nonprofit organization supporting the academic community in using digital technologies for advancing research and teaching in a sustainable manner, and preserving the scholarly record. This particular project, titled "Teaching with Data in the Social Sciences," was launched in spring 2020 with the goal of examining instructors' practices in teaching with data in undergraduate social science courses to ultimately help universities better understand how to support the needs and challenges arising from these practices. Ithaka S+R have provided an extended blog post offering more information on the goals underpinning this particular initiative.

The general audience for this report includes anyone who is interested in supporting data literacy in undergraduate social science courses whether as the instructor or in a supporting role (such as librarians and/or consultants). This report is organized in the following manner: first, we provide an overview of the methods enlisted during the course of this project, including the theoretical methodology, recruitment of participants, and the data analysis process. Next, we detail the major themes and sub-themes which arose from our analysis of the interview content. Then, based on our findings in the interviews, we provide several recommendations for educators (including course instructors and/or librarians) on how to best support data literacy initiatives in undergraduate social science settings. Finally, we close the report with our conclusions and possible future directions for this research.

2. Methods

This section of our report includes the theoretical underpinnings of our chosen methods for this project (Methodology), our process for recruiting eligible candidates to participate in our interviews (Recruitment of Participants), and our interviewing methods and data analysis techniques (Interviews and Coding). Our methods closely align with the methods advised by Ithaka S+R and thus also mirror the methods used by the 20 other institutions taking part in this broader initiative. It is important to note that the entirety of the project took place under COVID-19 restrictions, which necessitated our use of completely virtual methods to accomplish the goals of the project. This included the use of videoconferencing software and digital collaboration tools.

2.1 Methodology

To better understand faculty's practices of and challenges in teaching with data in undergraduate social science courses at CMU, we enlisted a qualitative research methodology using semi-structured interviews and a grounded-theory coding approach to capture their perspectives which are thematically presented in this report. Semi-structured interviews contain a pre-designated set of prompts to guide the discussion, but allow for a deviation from these questions based on the participant's experiences with the topic central to the interview. The nature of the questions are "carefully tied to [a] research topic...the objective is to guide a participant in conveying an account of an experience as it relates to the topic of study" (Galletta 2013: 47). The semi-structured interview protocol was designed by Ithaka S+R and approved by CMU's Institutional Review Board (see Appendix for IRB-approved interview materials).

We used a grounded-theory coding approach for our data analysis process which involved two major phases of coding the interviews for major themes and sub-themes. The foundation of grounded-theory analysis involves coding data with the goal of discovering a "latent pattern of social behavior that explains a main issue or concern within an area of research interest" (Holton & Walsh 2017: 76). In this setting, a code can refer to "a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data" (Saldaña 2009: 3). Combining a semi-structured interview protocol with a grounded-theory coding approach allowed us to extract a wealth of in-depth information on our participants' experiences in teaching with data in social science settings at CMU.

2.2 Recruitment of Participants

The recruitment period for this project began in the summer of 2020 and lasted until early spring 2021. Our interview subject population included anyone at CMU aged at least 21 years old who teaches undergraduate social science courses that use data, including tenured and tenure-track faculty, graduate students, adjunct instructors, and staff. We sent recruitment emails to eligible candidates in two major waves, using a convenience sampling method in which we identified possible candidates based on our own professional networks at CMU and knowledge of courses and instruction across campus. We ultimately secured interviews with 11 out of 22 possible candidates, which included one postdoctoral fellow and 10 faculty members.

There is a broad interpretation of what constitutes “social sciences" and definitions of this term have varied across time and discipline (Calhoun 2002, xiii). CMU is a heavily transdisciplinary institution and the social sciences can appear in a wide diversity of colleges across the campus. This diversity is reflected in the variety of research areas and expertise in methodologies amongst our participants, which includes astronomy, Bayesian analysis, experimental design, language and culture, various economics fields, network analysis, public health, public policy, and statistics. Although the backgrounds of some of our participants were rooted in fields outside of the social sciences (such as astronomy), in these cases we focused the content of the interviews on their experiences using data to teach courses in the social sciences. The 11 participants who were recruited for this study represent two schools on campus: the business school and the college of humanities and social sciences. Within those two schools, the participants are primarily affiliated with the economics and statistics departments. Additional representation came from the areas of modern languages and politics and strategy. Our attempts to recruit academically diverse participants yielded numerous undergraduate class offerings with different foci that ranged from cloud computing, data mining, data science, digital humanities, various types of economics, global affairs, various types of statistics, sports analytics, mathematics, to network analysis. A handful of participants had experience teaching undergraduates at both the CMU Pittsburgh and CMU Qatar campuses.

2.3 Interviews and Coding

The 11 faculty participated in an up-to-60 minute interview conducted over Zoom by the researchers on this team. Each interview was recorded to provide transcripts of each conversation. Interviewees had the option to answer the questions based on their pre-pandemic research process, or their research process as impacted by COVID-19 precautions. The 11 resulting transcripts were then de-identified, and sent to Rev (a transcription service). Once returned, the team identified three interviews for each team member to individually code using the open-source qualitative coding software Taguette. From that initial coding process, we developed a codebook of open codes that represented our best understanding of the content in each of the three interviews. The team then discussed these codes and condensed them into final themes and sub-themes using LucidSpark, a collaborative virtual whiteboard space. Each team member then individually coded all 11 interviews through the lens of their respective assigned themes and sub-themes in Taguette, and summarized the codes into a narrative comprising the remainder of this report.

3. Findings

This section of the report reflects our major findings from the interviews, organized into the following five themes: finding and accessing data, data methods and techniques in course curriculum, developing skills for programmatically working with data, teaching methods and instruction support, and student attitudes and motivations. In some content areas, there may be minor overlap across themes as certain topics arose multiple times in our interviews.

3.1 Theme: Finding and Accessing Data

Since there is a heavy emphasis on using data in these undergraduate classroom scenarios, the major theme of finding and accessing data for use in class was of interest for our team to explore. It came as no surprise that several classes taught by our interviewees rely heavily on publicly available data and datasets in order to make it as easy as possible for their students to find and access these data. Additional findings, trends, and challenges are described below.

3.1.1 Finding Data

Every interviewee was asked if students generate their own datasets, search for and select pre-existing datasets, or work with faculty-provided datasets. Almost all interviewees indicated that they provided data and/or datasets as a part of their class. Most interviewees provide datasets as a starting point for students to approach learning data concepts. Depending on the disciplinary area and the goals of the class, the faculty-provided datasets are either intended for students to use for class assignments or are possible data sets to use for specific research question assignments. For example, those who operate within the field of economics usually provide a listing of popular data sites and sources where students can find data sets that could possibly answer a research question. Whereas those in the field of statistics tend to provide actual, finalized data sets for specific assignments in class.

In addition to instructor-provided datasets, the vast majority of interviewees indicated that students also find their own pre-existing datasets. This also varied by class goals and sometimes by discipline. Many students will have to find additional data sets on their own that can be used for specific-research questions that they are studying within the class structure. When this is the case, it is still preferable to find publicly available datasets for ease of use reasons in order to finish assignments and projects. However, there are times when the research question dictates the need for primary research through survey creation and deployment.

Lastly, most interviewees indicated that students generate their own datasets. The majority of students in this case do this either to answer specific-research questions where other data sources do not suffice, or as a part of a class assignment that is relevant to the overall program. Students will go through the entire process of survey design, deployment, and pre-testing in some cases when generating their own data. As one interviewee explains, “They [students] get the experience of seeing how data collection goes and what you think is an obvious [and] easy question may be or may not be.” In these cases where students generate their own surveys, the most popular software used is Qualtrics since CMU has a site-wide license to this software.

There are occasions where students will gather data from historical, primary sources that are only available in paper form or PDFs. This gathering and encoding of the data has been described loosely as generating datasets by two of the interviewees. For example, one interviewee notes, “the data collected was from coding up historical, as in, early 20th century documents and extracting data from those documents that are available either in paper form or PDFs.” It should also be noted that the majority of students who are generating their own datasets are at the more advanced stages of their undergraduate academic careers.

3.1.1.2 Trends

Use of Public Data Sources

A major observed trend is the almost universal use of publicly available data and data sets that mainly come from governmental organizations, repositories, and websites (this is further described in section 3.1.2.1 Data Sources). This is true both of the faculty-provided data and the additional pre-existing data that is searched for by students. The reliance upon public data is mainly twofold: to reduce access and data use restriction barriers, and to ensure the use of high-quality, reliable data from authoritative sources. Since there is a wealth of data available through various government agencies, this can present challenges to students in finding appropriate data for their projects and assignments. Hence the reason why many faculty members provide data sets and sites for students to explore as a part of the class(es). Public data are free of the access and downloading restrictions that many subscription-based data providers enforce. Avoiding these barriers is ideal for a classroom environment that is primarily focused on finishing class assignments and projects, both from the faculty and student perspectives.

Cleaning/Pre-Processing

Another trend that emerged is data cleaning by instructors before providing it to students. Almost half of the interviewees indicated that they do this for the students in their classes. The primary intention in these cases is that the classes are not focused on working with the data per se, but rather statistical learning. Students using pre-cleaned data does seem to go against the trend of working with data in the statistics and data science communities. As one interviewee notes, “I assume that there will be problems working with data. And so I do data pre-processing...so that they [students] do not have to. This stands in contrast to some trends you will see in statistics and data science [departments] at other large universities where they will basically have students do a lot of work with the data themselves.” There was evidence of some interviewees including the data cleaning process as a part of the class objectives as well. Another interviewee states, “I had to walk them through, show them the repository, how they can download the data, and how they can clean the data….They did a mini demo [of this process] during the lecture, and then I gave them a code snippet to actually clean the data.” Some interviewees do see the data cleaning process as integral and important to student’s abilities of working with data.

3.1.1.3 Challenges

Finding appropriate data

A primary challenge that interviewees expressed is finding appropriate data and data sources. Although there are some challenges involved with faculty finding data, the main challenges lie with students finding additional data. This can be somewhat explained by the vast number of publicly-available, government data resources. The tremendous wealth of available data can make it difficult for students to pin down appropriate data sets, or know which ones are indeed reliable and authoritative. An additional and related challenge is that many governmental data sites use old database or website architecture that can make it difficult to find and extract data on their sites. One interviewee claims, “...finding data that undergraduate students can work with to answer very good questions...is not straightforward. We really have to show them how they can find these things [data].” Many instructors rely on data sources that are popular within their field, but can find it difficult once research questions need data that rely on fields outside of their expertise. Another interviewee explains, “I pretty much go to [the] sources I know because of my field, but finding data that goes to the right [question] formulation is hard.”

3.1.2 Accessing Data

In the ever-increasing digital world, electronic access to data and datasets becomes a higher priority. This has been exacerbated by the COVID-19 pandemic where most classes transferred to an online-only environment during the peak time of the pandemic. Although the final structure of how classes will operate in a post-pandemic world is still unclear, there is reason to believe that many will operate in a hybrid model which may become the new norm. Therefore, it was important for our group to examine how students access the data sets that are provided to them.

All 11 interviewees were asked an open-ended question about accessing data sets. Almost all interviewees indicated that they do indeed provide some type of access to the provided data sets.

The most common access method is posting datasets or links to datasets in the campus' learning management system, Canvas. About half of instructors indicated that this was their preferred method of access. A smaller portion of instructors choose to provide access simply using email. Lastly, some instructors make use of the popular repository hosting service, GitHub, to distribute data sets for class use. While there appears to be no dominant mechanism or method for accessing datasets from our sample of interviewees, it is clear that the goal is to make access as clear and easy as possible for the students in their class(es). All three of the major access methods that were discussed (email, Canvas, and GitHub) are very simple and effective ways to accomplish the question of access.

3.1.2.1 Data Sources

Since all of the interviewees rely on publicly available/governmental data and datasets, it was a point of interest for our team to learn which sources are being used. It came as no surprise that a wide variety are being used, some of which are very popular within particular disciplines. **Table 1** lists all of the different sources that are used and the frequency of use per instructor.

Table 1: Data Sources and Frequency per Interviewed Instructor

Source	**Frequency**
Corporate Data	2
COVID-19 Data/New York Times	2
Federal Reserve Economic Data (FRED)	2
International Monetary Fund (IMF)	2
Kaggle	2
World Bank	2
Bureau of Labor Statistics (BLS)	1
Integrated Public Use Microdata Series (IPUMS)	1
International Movie Database (IMDb)	1
National Climate Information Centre (NCIC)	1
National Health and Nutrition Examination Survey (NHANES)	1
Popular Media Print Sources	1
Sloan Digital Sky Survey	1

3.1.2.2 Trends

Use of GitHub

The increasingly popular repository hosting service, GitHub, was identified by a few instructors who use it for distribution and access to various datasets in their classes. The use of GitHub for various academic research projects has increased over the years as the trend of open science/research has also increased. The popularity of using GitHub to host public and open source projects started in computer science but has now spread to other disciplines, including the social sciences. Since its use has increased in the private sphere as well, instructors have been exploring it within the classroom environment to expose students to the software. One interviewee noted, "The other [mechanism] which I've started to lean towards more is GitHub. I'm not a big GitHub user, but I do use GitHub to upload CSV files. And then students...can just immediately link to it [CSV files] without having to do the downloading step."

Use of Dominant Disciplinary Sources

The interviewees that teach primarily within the economics discipline highlighted several heavily used, well respected, go-to data sources that dominate the field. These include BLS, FRED, IMF, IPUMS, and the World Bank. Within the economics discipline, it is crucial to tap into historical and longitudinal data in order to study patterns and changes that affect the economic conditions of individuals and groups at the micro and macro levels. The resources mentioned above provide this capability and are also free to use and easy to download/export for research purposes.

Nevertheless, it should be noted that there are plenty of additional data sources that are oftentimes used in these classroom environments in addition to the ones mentioned. The primary reason for this is that students will study research questions that require more advanced or niche data that is not primarily included in the previously mentioned sources. As one interviewee notes, "I've been in a couple situations where students are given a data set and they say, 'Oh, this information is not enough, I'd really like additional information to go out and find publicly available data or other data from the company or the researcher that's also available.'"

Low Use of COVID-19 Data

Given that our interviews took place during the COVID-19 pandemic, we were surprised to find out that there was not more use of the various COVID-19 datasets as sources for a number of classes. Only a couple of instructors indicated that they used or provided access to some sort of COVID-19 dataset for students to consult. Both of these instructors relied on the New York Times COVID-19 datasets, and one of those instructors indicated that they used this source because of its simplicity. The instructor notes, "It's [NYT COVID-19 dataset] literally 5 columns: state, county, the FIPS [Federal Information Processing Standards] code, the fatalities, and the date. There's really no ambiguity in the dataset." It's possible that many instructors were not able to pivot quickly enough to redesign entire projects and assignments that could use

COVID-19 datasets, but for those that also guide and allow students to find additional datasets on their own, these resources were not mentioned in the interviews. It was also somewhat surprising that one of the most respected and authoritative sources of COVID-19 data was not mentioned at all; the Johns Hopkins Coronavirus Resource Center data dashboard and raw data (also available in GitHub).

3.1.2.3 Challenges

Use of GitHub in China

A challenge that was identified by one of the instructors involved the use of GitHub in China. Due to the COVID-19 pandemic, many students studied from home in a remote environment. Additionally, a large portion of CMU's students are international with a large percentage of them coming from China. The Chinese government heavily regulates internet traffic and blocks many international internet companies. This happens to be the case with GitHub, and one of the few workarounds is the use of VPNs. As one instructor notes, "...one issue with the pandemic...was that students who are in China and are learning remotely; they cannot access GitHub. This actually created...issues for us. In the end, they [students] had to use VPN to connect. It was a mess because I couldn't get a server, a private server to upload them [datasets]."

Inconsistencies with Public Data Sources

Even though the wealth of reliable and authoritative public/governmental data sources is of great benefit to instructors, students, and researchers, they also come with many challenges. Many of these sources maintain old database architectures that are in need of an update. Some of them have older graphic user interfaces (GUI) that can complicate and frustrate users. There are also plenty of circumstances where data is no longer collected as a whole, or granular level data points are no longer supported or collected which can lead to interruptions in longitudinal research and analysis. Even something as simple as not being able to find the download option on pages adds to the frustration of many. These complications were expressed by some of our interviewees. One instructor summed up their frustrations saying, "Sometimes the links are not stable. Sometimes [data] variables are not stable. So from one year to another a variable that was used has disappeared. Maybe the length of time that was provided for has disappeared." Another instructor laments, "Some challenges [include] where to download the data because [on] each website they have their own UI. And some of these are not transparent...they don't follow the same user experience practices to show you where you can download the data."

3.2 Theme: Data Methods and Techniques in Course Curriculum

Our participants described several of the specific data methods taught in their courses as well as the overarching structure of coursework aimed at increasing students' skills in manipulating and interpreting data. Many interviews reflected disciplinary shifts that are prioritizing teaching students quantitative skills even in primarily social science environments, and the curriculum described in the interviews supports those priorities. Interviewees reflected on the broader goals for student learning they hoped to achieve by teaching these data methods and techniques, including helping students develop better skills in research design critique, critical analysis of data (where does it come from? Who created it? Who interpreted it? Is it accurate?), data analysis workflows, and communicating the results of a data analysis and visualization to a wide audience. Particularly, several interviewees reiterated their desire for their students to progress past simply being able to work with data and conduct data analysis, and actually be able to communicate the significance and importance of their work to others.

The specific types of data methods (both foundational and advanced) are summarized in **Table 2** below.

Table 2: Types of Data Methods Taught in Sampled Courses

Foundational Techniques	**Advanced Techniques**
Understanding data structures such as vectors and data frames Basics of empirical analysis as applied to problem-solving Pre-processing data/data cleaning Merging data from different sources Basic manipulation of datasets including creation of extra variables, and extraction of data points of interest Introductory basics of visualizing data Introductory basics of data modeling Collecting and analyzing survey data	Advanced statistical and data science techniques (including linear regression, survival analysis, machine learning techniques, clustering, network analysis) Augmented replication Data mining Understanding advanced survey sampling techniques and statistical programming Cloud computing Making compelling data visualizations and interpretation of visualizations Working with unstructured text data and text encoding

3.3 Theme: Developing Skills for Programmatically Working with Data

While many skills for working with data were discussed during the interviews, using code to clean, view, and analyze data was a major theme that emerged. The open source languages, R and Python, are commonly used. In this section, we explore how students develop skills for working with these languages both in and out of the classroom and the challenges they face with learning how to work with data programmatically. We also discuss the expectations that instructors have for their students in this area.

3.3.1 Programming Languages and Software

Reflecting CMU's strength in technical and computational disciplines, all of the instructors talked about the use of programming to manipulate and analyze data in their classes. R, specifically, is used by the majority of instructors in their classes, with Python being used by a couple of instructors. One instructor has students use Shell scripts to clean data. Excel was used in some classes for viewing data, and a few instructors noted that students have an option to use Excel, or other types of point and click statistical software, in their classes if they are not comfortable with coding. The following software is used by a single instructor: GitHub, Stata, Oxygene, Visual Studio Code, Google Cloud, SQL, and Tableau. A couple of instructors have students use Qualtrics or Mechanical Turk to collect survey data (Table 3).

Table 3: Programming Languages and Software for Data Collection, Cleaning, or Analysis used in Sampled Courses

Programming Languages and Software	**Frequency (per Instructor)**
R	9
Excel	3
Python	2
Qualtrics	2
Mechanical Turk	1
Oxygene	1
Shell	1
GitHub	1

Stata	1
Visual Studio Code	1
Google Cloud	1
SQL	1
Tableau	1

While instructors had a variety of reasons for choosing specific software for working with data in their classes, most were related to a perception that R is the superior language for statistical analysis and data visualization and is therefore ubiquitous in their disciplines. Instructors in two different departments mentioned that their departments have explicit expectations that instructors will use R in their classes. Similarly, a couple of instructors mentioned a move away from using Stata or SPSS to R in recent years in their classes. Some instructors allow students to choose the software or language that they use, noting that the major of the student often influences whether they will use R or Python. Finally, some instructors choose to teach R because they have expertise in it.

Some instructors think that learning R itself is pedagogically important since it will prepare them for their future careers, while others think that it is important for students to be able to programmatically analyze data, but do not think that the language itself is pedagogically important. One interviewee said that some of their students have familiarity with Python from taking classes in the School of Computer Science at CMU and that has influenced their decision to use R in their classes:

> “And on one hand you might say, well, that's the reason you should use Python, from my point of view is the reason you shouldn't use Python, is to give them the diversity and see how some things are implemented in similar and very different ways. And it's less about being tied to a language, but it's more about thinking things through. So for that reason, I wanted to do something different than what they might have seen before.”

In spite of the fact that almost all of the instructors are using open source languages in their classes, only one instructor mentioned cost as a motivating factor when choosing software. They suggested that one reason that students commonly learn R in the social sciences at CMU is because it is free. The instructor that is using a proprietary language is using Oxygene because they have experience using it in their own research.

3.3.2 Skill Development in the Classroom

Instructors mentioned many learning objectives related to working with data that map onto the research process such as finding appropriate data sets for a given research question and interpreting and critically thinking about data. A skill that was consistently mentioned in all of the interviews, however, was the ability to programmatically clean, manipulate, analyze, or visualize data.

Although most of the instructors do not expect that their students know how to program coming into their courses, many of them said that a majority of students do have prior experience with programming, albeit to differing degrees. Much of the variation in student abilities is related to their prior coursework and whether the course is related to their own major. For example, one instructor mentioned that their students come from several different departments including statistics, computer science, business, mathematics, economics, industrial engineering, and chemical engineering. This large breadth in subject area background is not surprising given the emphasis on interdisciplinary scholarship at CMU. Instructors that have higher expectations for coding abilities are generally teaching upper-level elective courses. Importantly, one instructor also noted that even incoming first-year students have varying abilities depending on the high school that they attended.

There were varying perceptions of how difficult it is for students without coding experience to get up to speed, with some instructors finding that the students pick it up quickly and others acknowledging a steep learning curve. A couple of instructors discussed using surveys at the beginning of the course to try to better understand the variation in skill sets. Many instructors spend some time in the beginning of the course going over basic skills needed for the class. Since these classes are not coding classes per se, the instructors are often teaching specific packages needed for data analysis or visualization, such as ggplot, rather than focusing on the fundamentals of programming.

Related to the idea that the students are often learning how to work with data programmatically in an applied way rather than through a dedicated coding class, instructors often noted gaps in the programming abilities of their students. One instructor mentioned that students coming into their classes often know how to use code to answer specific questions but are not able to transfer that knowledge to other problems due to an inadequate understanding of programming. In response to this issue, they have specifically tried to bolster foundational coding skills in their students. Another instructor discussed how their students can often perform fairly advanced analyses but are missing basic skills. For example, they might not know how to load a dataset if previous instructors have provided them.

Some instructors leverage team projects to address the differing abilities in programming among students. Teams of a handful of students allow for peer learning, and more experienced programmers on the team can compensate for lack of advanced coding skills in other students.

One instructor talked about how team projects also help prepare the students for their future careers which will likely involve group work.

3.3.3 Student Challenges

A trend that emerged from the interviews is that students often face challenges in troubleshooting their code. A common strategy for fixing errors when programming is to look for solutions on Google or online community forums such as Stack Overflow, but the students often rely on instructors to fix their errors. One instructor mentioned that being able to use Google effectively to fix errors is a skill in itself, saying:

> “It's inherent with programming, in general, that a lot of things are googleable and you go onto these forums where someone's like, "I ran into this error. What do you do?" And do I think that is a skill. How to google things appropriately is a skill for programming and I really do think when I think of myself as a good programmer, I don't just mean that I can write code, I more so mean, when I get stuck I can google things appropriately, such that I can find the solution. And I definitely make that explicit to the students, where I do this in all my classes, is I say, "Look, you can always ask me and the TAs for questions, but one of our goals is that we do not send you a 'Let me google that for you,' link, during the semester.”

Another student challenge that a couple of instructors discussed was anxiety or fear of coding. These instructors are intentional in trying to ease these negative emotions: “For the [x] class, I explicitly say, "I'm going to make the assumption that none of you have programmed before, that none of you have used R before, but guess what? We're using R for the whole semester and we are programming, and I'm sure that makes you very anxious, but it's going to be okay."

3.3.4 Extracurricular Learning and Career Preparation

Some students have strengthened their data skills outside of their classes, with internships being the most common type of extracurricular activity mentioned, followed by online tutorials. It was also noted that the students at CMU are well poised to get internships because of their experiences working with data in their classes. The internships might not necessarily be teaching them new skills but instead providing an opportunity to try their skills in different contexts and with new pieces of software. While none of the instructors actively encourage internships, they largely found them valuable. In contrast, there were mixed opinions on the usefulness of online tutorials. A few instructors suggested that the tutorials are often not aligned enough with the content of the course to be useful and at worst can confuse students.

Most of the instructors mentioned that teaching students how to work with data was important for preparing them for future careers, with many particularly highlighting the value of data visualization. One instructor, who teaches in a subject area in which working with data is less common, indicated that their students will have a competitive advantage on the job market with their data skills. One interviewee emphasized the vocational nature of working with data and expressed a desire to see departmental support for students in creating a portfolio of their work that they could provide to potential employers.

3.4 Theme: Teaching Methods and Instruction Support

The sections above detail the programmatic or content questions that the social science instructors in our interviews consider as they design their courses and help students work with data. This is only half the equation. The other important element to course design next to content is application, or what methods instructors employ to educate students on how to use or interact with data. This section will detail specifics about these instruction choices and offer suggestions to other educators on what pedagogy elements to include in their courses.

3.4.1 Instruction Strategies

Just as there are many different types of data that can be used within an education environment, there are also many different pedagogical methods that can be used to teach them. Our interviews showed that the most common methods used for in-class data instruction included group projects, individual exploration of data, in-class discussion, or applied tool application/data analysis training (or some combination of all four).

Many of our interviewees rooted their teaching of the data or code techniques described in Section 3.2 within these major course design frameworks. In the project-based curriculum where students collect and/or access data and design a project with problem-solving implications, students were largely able to focus on their own interests while developing core skills of teamwork and data manipulation. Other examples of this model include courses focused on these extended projects where students collaborated with companies from outside the university, to have the opportunity to apply their developing data skills to research questions from a “client” in real time. For the project-based curriculum, one interviewee described a capstone course in which students could be paired with researchers in the city of Pittsburgh to collect/access and analyze data to tackle a real-world issue, choosing appropriate analysis and visualization techniques suited to the problem at hand. In both cases, the idea is for students to have a hands-on approach to learning data skills from a holistic perspective. Here, the goal for instructors was not to introduce students to only one data concept of analysis practice, but rather describe the series of steps required to make larger data projects successful.

Courses based on problem sets where students spend the semester learning how to answer sets of data analysis and visualization questions focused more on the use of pre-packaged data or assignments with pre-built code, and heavily guided by the instructor. In multiple examples of this technique shared by interviewees, this pedagogical style is meant to train students in a particular practice that would serve as a building block for their future work with data-driven research questions.

Other teaching strategies included working with outside instructors, teaching assistants (TAs), or community members with professional skills as guest lecturers or helpers in their courses. Typically, TAs were able to offer students an additional perspective on the data projects they were working on by providing another avenue for questions to be answered within the class. Often these TAs were hand-selected by the faculty instructor to make use of their experience working with data in similar ways to those addressed in group projects. As such, TAs were able to provide both technical support and advice on how to successfully complete assignments. TAs were also used as a method for course grading, allowing faculty to have more time to work directly with students to overcome any issues they were experiencing. Faculty and guest instructors were often used to provide examples of data expertise in active use, and how these members explored using data in their research, work, or daily lives. These outside perspectives provided additional inspiration to students, who were then often challenged by the interviewees to apply this inspiration to their class projects or weekly discussions.

3.4.2 Supporting Data Education by Sharing Open Educational Resources

Many of the instructors noted that they create content for their own courses, but about half of the interviewees also use materials from other instructors. Instructors found these materials in a variety of ways, with a couple finding them online, a few using materials developed by colleagues at other institutions, and some inheriting materials from other instructors in their department.

Only a couple of instructors are sharing materials publicly with discoverability and reuse in mind. Most instructors said that they would share their materials if asked for them by another instructor, although in practice some of them have never been asked for their materials. One instructor mentioned that they share their materials in Dropbox with other instructors in their discipline at CMU. Most instructors did not provide a reason for not sharing their materials more widely, apart from one saying that they do not want their students to find the materials outside of the course setting.

3.4.3 Teaching Ethics of Data

Several interviewees spoke about how ethical issues played a role in their classroom instruction. This refers to the ways in which data is analyzed, as well as issues of diversity and equity that may be posed by misuse of the data. Instructors teaching with data about individuals from a public policy or modern languages perspective made reference to the right to privacy or an interest in protecting the ethical and social implications posed by the datasets. Other instructors were more concerned with the constraints applied to data collection, or ensuring that the information used to support student project work were properly collected or sampled in order to provide a sound basis for any analysis results.

3.4.4 Outside Classroom Resources

In addition to resources that instructors provide in-class, several interviewees mentioned they promote online resources for students to engage with outside of class hours. This included online resource videos which explore how to use specific software (such as Excel or Tableau) or coding languages (such as R or Python). These materials are shared with students to help them prepare for the rigors of the course, become familiar with concepts or tools that would be used, and to offload some of the course preparedness from instructors to students. Other resources shared with students include online forums or communities that enable students to crowdsource questions to a larger group. Primarily, this process referred to instructors working with students to learn how to effectively search for their questions online so they could then independently find the help they needed to overcome obstacles in coding or data manipulation. As referenced earlier in the report, and by this quoted instructor, the mark of a strong digital data user is more than just someone who knows how to code but "when [they] get stuck [they] can Google things appropriately, such that [they] can find the solution". This is particularly important in the context of this report, as students may often enter social science data courses with different levels of preparedness or data skills. By sharing these resources with students, it enables them to have agency in their learning process and take leadership over their own progress with data analysis and associated techniques.

3.4.5 Instructor Preparedness and Outside Support

Instructors often mentioned that they feel generally comfortable using data within the confines of their course goals and when shepherding class projects. As mentioned above, many of these instructors pull example data sets from their own research. Yet, these same instructors often allow students to find their own data as a base for their projects, or have companies offer data sets for a student capstone project. Thus, having instructors understand how to develop research questions for multiple disciplinary foci could help them provide better support to their students who seek out unique data that fits student interests for their projects.

While many instructors we spoke to recognize the increasing importance of data-driven courses within their departments, there was a consensus that efforts to support their work in data education by their departments and the broader university are still lacking. Many instructors referenced how they wished they could have had access to some form of instructional course offered by the university or department that would address recommended practices for teaching about or with data. Instructors often stated there were some portions of their courses that were difficult to run due to student questions regarding aspects of working with data that the instructors were not familiar with. One instructor even offered the familiar academic adage that they are staying “one step ahead of their students' learning”. In these cases, having a training opportunity that prepares instructors for answering student questions around data or digital skills while also incorporating issues of ethics or data access might prove beneficial. Furthermore, some interviewees mentioned that they work with their departments to assess the effectiveness of their course syllabi or learning outcomes to ensure they meet departmental guidelines or goals. Having departments play a role in the training of their faculty, therefore, is a needed component of making sure instructors are prepared to train their students in recommended data practices.

Other instructors stated that they did not require any extra support from their institutions to teach data in the social sciences. These instructors mentioned receiving some form of prior instruction on how to develop pedagogy and work with students, and/or develop the skills needed to work with data as a part of their graduate work or personal research. However, it is possible that these instructors would still benefit from training on instruction techniques.

Instructors mentioned several challenges they faced when instructing undergraduate students in data use or analysis. The primary challenge faced was dealing with students who had varying levels of familiarity and ability with data analysis, coding, and recommended data practices. While none of the interviewees mentioned that their courses did not function because of this mix of student experience with data, it did lead to some extra work or consideration in pedagogical development by the instructors.

3.5 Theme: Student Attitudes and Motivations

When asking interviewees to address how they viewed their own teaching practices and familiarity with data, we also asked them to speak to any challenges that their students faced in their courses, as well as their general motivations or attitudes when dealing with data. Overall, interviewees were very concerned about their students' ability to accomplish the tasks set to them in these data-focused courses. Indeed, many issues they addressed were in some way related to how students either interacted with data, their level of skill, or whether students were prepared to undertake the course. This section of the report will dive into specifics of what instructors noticed regarding student comfort with course subjects, students’ motivation when learning about data, as well as any challenges students faced during their courses.

A common thread in all responses was the level of student awareness of and motivation to work with data. Every interviewee, regardless of disciplinary affiliation, noted that the students they taught had differing skill levels. In many cases, the knowledge (or lack thereof) of these students was revealed through the process of explaining group projects or in early assignments required in the course. Only one interviewee made reference to any type of pre-preparation or evaluation of students’ skills prior to the start of the class (often in relation to capstone style projects). As such, instructors had to develop solutions on the fly to ensure student understanding and engagement within these courses remained high.

Student confidence also played a role in their success, and their instructors overall need to support them in that particular area. In many cases, instructors referenced having to spend additional time with students who were concerned over their abilities to work with data. Coming primarily from the more qualitative-focused fields in the social sciences, these students not only had to overcome the lack of pre-knowledge but the insecurities of often not being at the same level as their peers in data comfort. Often these issues could be resolved within the timeframe of the course with support from TAs or other students who had more experience with these skills. Instructors also mentioned that they were able to work with students who found their coursework challenging by encouraging these students to practice their skills outside of the classroom.

3.5.1 Student and Instructor Expectations

The expectations instructors had for their students guided much of the student experience within the course. Many of the instructors we interviewed expected their students to have at least some experience working with data, from their own exploration or skills they gained in other classes. However, this can vary depending on the course and the level of data and/or coding skills that are needed to successfully complete the course. In our interviews, these expectations ranged from: "the assumption that students know how to use computers and can use graphical interfaces" to "weaker assumption than assuming they know how to program". These varying expectations can dictate how outwardly accessible the course is to different students across the spectrum of the social sciences (in our sample, from the modern languages to statistics).

Instructors often expected their students to spend a significant amount of their time outside of class engaging in these data concepts. As some of our interviewees expressed, this acted in counter to those students who expected to be "spoon fed" during class hours, and who would end up falling behind due to not spending enough time actively manipulating data as dictated by their assignments. In addition, the instructors also expected their students to develop some level of critical thinking skills in regard to using data. As one instructor noted, they want their students to be able to ask follow-up questions that enhance the research narrative of the data the instructor provided them.

Conversely, student expectations seemed to correlate to the level of difficulty of the course and the focus on individual or group project work versus weekly assignments. Interviewees indicated that students in project-based courses expected to develop skills that they could apply to project-based work outside of the class, and possibly could be used as foundational learning they would further develop during their other courses (or internships). This aligns with the broader CMU initiative to increase graduates' understanding of data and how to properly use it to solve a problem. Conversely, students who were in courses that focused on week-to-week assignments seemed to primarily focus on developing skills needed to complete the assignments in front of them. Based on the results of the interviews, students also seemed to be more motivated to complete assignments or engage critically with data if they were able to design the research questions or projects they completed during the course. Expanding concepts covered in the course to allow for student creativity increased student excitement about the course topics as well, leading to better outcomes at large.

4. Recommendations

In this section, we offer four major recommendations that we pose as solutions to the myriad challenges expressed by our interviewees in their experiences of teaching with data in social science undergraduate courses. These recommendations build and expand upon existing data services at CMU Libraries, and are all actionable either in the short- or long-term within the current strategies and goals for expanding data services at our institution.

4.1 Supplement Instruction for Foundational Coding

The library should engage in assessing the data competencies of students to help identify gaps in core skills, including foundational coding, in partnership with administrators of the colleges and other stakeholders at the university. As one of our interviewees suggested, first-year students will have different baseline knowledge for data literacies depending on their high school curriculum and prior experiences. Quantitatively assessing each incoming first-year class would help to surface gaps in foundational knowledge among students. Additionally, working directly with instructors from various departments to assess students in individual classes. After the initial assessment, a set of recommendations for scaffolded data literacy instruction and measurable outcomes can be created that aligns with both subject area learning objectives and career preparation goals. Qualitative data, such as longitudinal focus groups for subsets of students, can be used to supplement the quantitative data gathered through these surveys. A systematic assessment will not only help clarify the gaps in student knowledge but will also help even the playing field for students with different educational backgrounds.

One recommendation that is further explored below is to offer a dedicated data literacy program to ensure that students have the same baseline knowledge as they continue their studies. Members from the CMU Libraries currently offer a half semester course on data literacy, titled Discovering the Data Universe, which discusses core topics related to the data lifecycle and undergraduate research.

We also recommend offering periodic Carpentries or Carpentries-style workshops at a program level to supplement students' foundational knowledge. The Carpentries is a non-profit that teaches foundational coding skills with 2-3 day hands-on workshops. These workshops can provide discipline-agnostic baseline knowledge for coding in R or Python. The students that might benefit the most from this type of instruction could be surfaced through the quantitative and qualitative assessments.

Importantly, the Carpentries workshops also address the challenge of fear around coding. Many aspects of the workshops directly address the negative emotions that students and researchers can have when learning coding, including The Carpentries Code of Conduct that emphasizes a safe and welcoming learning environment for all participants, a low student-teacher ratio, and guidance on how to continue learning and troubleshoot errors as learners work on their own projects. Feedback from learners in prior workshops at CMU suggests that these workshops are effective in both filling in gaps in foundational knowledge and easing discomfort around coding. Institutions that do not have a Carpentries membership can take advantage of the open source lesson plans (see Software Carpentry lessons and Data Carpentry lessons) to create their own Carpentries-style workshops. **CMU Libraries could do more outreach to the departments and colleges to showcase these opportunities and form partnerships that allow a department to have a dedicated workshop or courses for their students**.

4.2 Increase Awareness of Open Educational Resources

Most of the instructors expressed a willingness to share instructional materials but are not systematically doing so. This is an area of support that the library, with specialists in Open Educational Resources (OER), is well-positioned to support. It will be important to first assess the barriers that are limiting sharing. One can imagine that instructors face a wide variety of barriers to both sharing their own work and reusing the work of others, such as wanting to get proper credit for their work, lack of time, not enough incentives, and not being familiar with the infrastructure, to name a few. Once the barriers are well understood, it will be important for librarians to raise awareness of the benefits of OER. Librarians should also consider increased outreach to department heads and others in positions of leadership to help create a culture that values and incentivizes sharing and reuse of instructional materials.

Librarians can partner with the teaching and learning center, as well as other academic units, to create guidance on recommended practices for sharing instructional materials; provide training for infrastructure that supports sharing and collaboration, such as Open Science Framework; and foster a community of practice. One instructor using Dropbox to share materials noted that although they think many people are reusing their materials, they do not know for sure. This suggests that supporting a platform like Open Science Framework that reports reuse metrics, such as downloads and forks, might help bolster sharing.

4.3 Provide Support for Building Critical Data Literacy Skills

As our findings reflected, many of our interviewees prefer their students to learn data skills through project-based methods, including projects using data to provide solutions to real-world challenges and issues. At the crux of these projects is the goal for students to critically think about their processes in finding/collecting, analyzing, visualizing, and communicating data; all which can be summarized through the phrases “critical data literacy skills.” Academic libraries have long been a hub for providing information literacy education to their respective campuses, and can further support their campus instructors who are building data skills in students in social science courses through data literacy workshops, written resources such as LibGuides, and in-class instruction. Academic libraries can also become involved in their campus iterations of Responsible Conduct of Research (RCR) training programs, which often include sessions on data ethics. On the CMU campus (and this may be the case on other university campuses), RCR sessions can be custom-designed to include other relevant elements of research education, such as data literacy. Academic librarians stand in an excellent position to help design supplemental RCR curriculum revolving around critical data literacy. **Further, CMU Libraries is developing a more formal program for data literacy, which will focus on creating pedagogy accessible to beginning learners who need to become familiar with the basics of data and the data lifecycle.** We recommend that other academic libraries consider pursuing the development of a similar program to support data literacy education on their campuses.

4.4 Support Collaborative Efforts in Finding Data Sources

It was mentioned by several interviewees that students face significant challenges when finding appropriate data sources. Some interviewees also indicated that they can struggle at times to find appropriate data sources that fall outside of their disciplinary expertise. This is an area where the library should engage more with both instructors and students to help minimize frustrations and provide educational and instructional support surrounding various open/government data sources. The library houses many specialists that are well-versed in specialized data sources for many disciplinary areas. **A recommendation to consider is faculty working with library specialists to help students navigate and query the vast array of publicly available data sources that are appropriate for research projects and class assignments**. Library specialists can also assist faculty in the designing of class projects and assignments that may use specialized data sources. This collaborative effort can lead to better class outcomes and can help reduce the frustrations that both faculty and students experience.

5. Conclusions and Future Directions

While the recommendations we have offered here were developed based on the content of our interviews and thus come from a CMU context, we hope that they will be beneficial to any academic library hoping to better understand the practices and challenges of those teaching with data in the social sciences on their respective campuses. We hope that uptake of these recommendations will support a more collaborative, open, and engaged campus effort in supporting data skill development in students.

In the future, we will continue to see academic libraries be increasingly engaged with data education and data literacy efforts with their campus users, and we encourage future research directions that use the methodology we applied here to interview other niche populations on their campus. Academic librarians (and academic libraries) are striving to provide effective services to campus users despite shrinking budgets and limited bandwidth, and taking the time to conduct these interviews can allow for a more targeted and effective service development and delivery.

6. References

Calhoun C. (Ed.). (2002). *Dictionary of the social sciences*. Oxford University Press.

Galletta, A. (2013). *Mastering the semi-structured interview and beyond: From research design to analysis and publication*. New York University Press.

Holton, J. A., & Walsh, I. (2017). *Classic grounded theory: Applications with qualitative and quantitative data*. SAGE Publications, Inc.

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). SAGE Publications.

7. Appendix A: Interview Protocol

This semi-structured interview guide was developed by Ithaka S+R and utilized in this study under full approval from CMU's Institutional Review Board under STUDY2020_00000228 granted on 2020-07-28.

Semi-Structured Interview Questions

Note regarding COVID-19 disruption: I want to start by acknowledging that teaching and learning has been significantly disrupted in the past year due to the coronavirus pandemic. For any of the questions I'm about to ask, please feel free to answer with reference to your normal teaching practices, your teaching practices as adapted for the crisis situation, or both. Please do not say anything during the interview that is both identifiable and private, and please ensure you are in a quiet environment where others who have not consented to this study are not inadvertently included in the recording.

Just a reminder, this interview is being recorded. I am going to start the recording now.

Background

Briefly describe your experience teaching undergraduates.

- How does your teaching relate to your current or past research?
- In which of the courses that you teach do students work with data?

Getting Data

In your course(s), do your students collect or generate datasets, search for and select pre-existing datasets to work with, or work with datasets that you provide to them?

If students collect or generate datasets themselves, describe the process students go through to collect or generate datasets in your course(s).

- Do you face any challenges relating to students' abilities to find or create datasets?

If students search for pre-existing datasets themselves, describe the process students go through to locate and select datasets.

- Do you provide instruction to students in how to find and/or select appropriate datasets to work with?
- Do you face any challenges relating to students' abilities to find and/or select appropriate datasets?

If students work with datasets the instructor provides, describe the process students go through to access the datasets you provide. Examples: link through LMS, instructions for downloading from database

- How do you find and obtain datasets to use in teaching?

- Do you face any challenges in finding or obtaining datasets for teaching?

Working with Data

How do students manipulate, analyze, or interpret data in your course(s)?

- What tools or software do your students use? Examples: Excel, online platforms, analysis/visualization/statistics software
- What prior knowledge of tools or software do you expect students to enter your class with, and what do you teach them explicitly?

To what extent are the tools or software students use to work with data pedagogically important?

Do you face any challenges relating to students' abilities to work with data? How do the ways in which you teach with data relate to goals for student learning in your discipline?

Do you teach your students to think critically about the sources and uses of data they encounter in everyday life?

Do you teach your students specific data skills that will prepare them for future careers?

Have you observed any policies or cultural changes at your institution that influence the ways in which you teach with data?

Do instructors in your field face any ethical challenges in teaching with data?

- To what extent are these challenges pedagogically important to you?

Training and Support

In your course(s), does anyone other than you provide instruction or support for your students in obtaining or working with data? Examples: co-instructor, librarian, teaching assistant, drop-in sessions

- How does their instruction or support relate to the rest of the course?
- Do you communicate with them about the instruction or support they are providing? If so, how?

To your knowledge, are there any ways in which your students are learning to work with data outside their formal coursework? Examples: online tutorials, internships, peers

- Do you expect or encourage this kind of extracurricular learning? Why or why not?

Have you received training in teaching with data other than your graduate degree? Examples: workshops, technical support, help from peers, etc.

- What factors have influenced your decision to receive/not to receive training or assistance?

Do you use any datasets, assignment plans, syllabi, or other instructional resources that you received from others? Do you make your own resources available to others?

Considering evolving trends in your field, what types of training or assistance would be most beneficial to instructors in teaching with data?

Wrapping Up

Is there anything else from your experiences or perspectives as an instructor, or on the topic of teaching with data more broadly, that I should know?