**Seeking Equilibrium in Data Reuse:**
**A Study of Knowledge Satisficing**

by

Jeremy York

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2022

Doctoral Committee:

Professor Elizabeth Yakel, Chair
Associate Professor Libby Hemphill
Associate Professor Christi Merrill
Associate Professor Ricardo Punzalan

Jeremy York

jjyork@umich.edu

ORCID iD:  0000-0001-8225-9291

## Dedication

To my wife and son, and our dear cats, Tristan (who is with us now in spirit) and Zolda

**Acknowledgements**

I would first like to thank my fellow students, who have supported me and enriched my life and research through every stage of this PhD journey. From those in our 2016 cohort—with whom, through introductory classes and discussions, I forged my foundational understandings of research methods and information science—to those in this and other cohorts who were instrumental in reading, listening, guiding, commenting on, and thus shaping my research in all of its phases. Also to those in my 2017 Museum Studies Program cohort. Our discussions and experiences were deep and had a significant impact on my life and outlook. And to the members of the MGrad running group and other running friends. The hours we spent talking as we traversed the trails in Ann Arbor were essential to my physical and mental health and introduced me to many of my favorite places. Thank you all. I would literally not have reached the point I am at now without all of you.

I'd also like to thank all of the faculty who have talked with me, taught and mentored me, or given words of encouragement in passing during my time in the PhD program. Your lectures and criticism, discussions, consultations, examples, and kind words have opened my mind, enriched my scholarship, and helped me persevere when the going got tough. I'm so grateful for the knowledge, experiences, and time you so generously shared with me.

Within and beyond these categories are colleagues with whom I have collaborated on projects or research groups of various kinds over the years. From the Archives and Digital Curation research group to Translation Networks and Collect/Connect, Library as Research Lab, MICA, Data Discovery and Recommendation, my practicum at the Detroit Institute of Arts, and

others. These opportunities gave me the chance to explore new ideas, apply what I have learned in practice, and learn and grow in ways that only come through working closely with others over periods of time.

I'd like to thank staff in counseling, finances, research support, technology, marketing, administration, and facilities at UMSI. I have relied on you time and again over the years and been well-served by your patience, expertise, and understanding. Thank you for all you have done to make my time at UMSI a success.

I want in particular to express my gratitude to my advisor, Beth Yakel, and to other members of my dissertation committee. I wish I could say I have been an easy student to guide through the choppy waters of the PhD. Fortunately, I have had those close to me who were equal to the task and have managed to support, encourage, guide, and challenge me all at the same time. I'd like also to thank Faye Polasek for the excellent work she did to assist me in the data gathering phase of my dissertation.

None of my research would have been possible without the many researchers who took the time in the midst of a pandemic to respond to my survey and to be interviewed. Your honesty and generosity have made my dissertation what it is and I am immensely grateful.

Finally, I would like to thank my friends (including those who may neither have been students, faculty, or collaborators) and family. You reminded me, at times when I was lost in my own thoughts and issues, that the world is much bigger than my PhD, and you were there for me at all times in all ways. This includes, of course, my wife and son, to whom I have dedicated this work. Words simply cannot express my gratitude for all you have given of yourselves to see me through this process. I am blessed beyond measure to wake up every day to the two of you in my life.

## Table of Contents

## List of Tables

## List of Figures

## List of Formulas

## List of Appendices

## Abstract

Government funding agencies and commissions have proposed that sharing, preserving, and providing access to more scientific research data will lead to increased reuse of data in academic research and result in greater knowledge and new discoveries. However, researchers encounter significant logistical, theoretical, methodological and ethical challenges to reusing data that hinder the achievement of these goals. One of the challenges researchers face is obtaining sufficient knowledge about data and the context of data creation to make a decision to reuse the data in their research.

In this dissertation, I report on a mixed methods study to investigate how researchers set limits on the types and amounts of knowledge they obtain about data, and what influences them to do so. A more nuanced understanding of how and why researchers determine such thresholds can inform strategic measures to enhance support for data reuse.

My study included a survey and semi-structured interviews and was conducted on a sample of researchers who reused data from the ICPSR data archive. I used Donna Haraway's theory of situated knowledges and Herbert Simon's theory of satisficing to develop conceptualizations of data and means of evaluating thresholds of knowledge that researchers obtained about data. I defined a concept called "reuse equilibrium"—when researchers determine data are sufficient to reuse to meet their research goals—and examined whether satisficing was a means by which researchers obtained knowledge to reach reuse equilibrium.

I found that researchers lacked knowledge they desired about data and that this lack of knowledge frequently had a negative impact on their research. The type of knowledge

researchers most often desired but were unable to obtain was "supplemental" knowledge that was not archived with the data and may never have been collected. While researchers lacked knowledge about the data they desired, I found that satisficing did not accurately represent their behavior in knowledge attainment. Instead, researchers sought to maximize their knowledge of data to meet personal aims (i.e., to reach "personal reuse equilibrium") in environments characterized by pressures and incentives that favored the achievement of social norms and requirements (i.e., "social reuse equilibrium").

I concluded that an important way to improve the environment for reuse was to assist researchers in obtaining supplemental knowledge about data they desired, thus supporting their achievement of personal equilibrium. This could be done by facilitating more structured and intentional "conversations" between data creators and data reusers with the purpose to influence the data that are created in the first place.

My findings about the knowledge researchers lack about data and the ways they seek to obtain it will be of interest to data reusers to gain a broader perspective on their colleagues' experiences. They will also be of interest to data creators, as well as data stewards, publishers, and other data intermediaries, to understand the knowledge researchers desire about data and the role they can play in helping researchers obtain it. Such findings, in addition to those about pressures and considerations in the reuse environment, will be of interest to funders and policy makers to gain insight into the ways current policies, practices, and incentives could be enhanced or changed to maximize the return on investment in primary research.

# Chapter 1 Introduction

There is a heightened hope that collecting, preserving, and providing access to data resulting from scientific research will lead to increased reuse of data within and across disciplinary boundaries and result in new discoveries and the production of new knowledge (e.g., National Research Council, 2003; National Science Foundation, 2007; Holdren 2013, National Academies of Sciences, 2018). Towards these and other ends—including bolstering public trust and improving the return on investment in science—much effort has been expended in recent years to increase the availability of scientific research data. Efforts include research (see York et al., 2016 and the literature review in chapter 2), the development of policies (e.g., National Institutes of Health, 2016, National Science Foundation, n.d.; Johns Hopkins Libraries, 2017), training and education (e.g., DCC, n.d.; Barbrow et al., 2017), and the building of technological infrastructure (e.g., Michener, 2015; Lee and Kang, 2015; Marcial and Hemminger, 2010; Kindling et al., 2015). In addition, researchers have sought to understand the circumstances under which data (whether publicly available or not) are reused in research (Enke et al., 2012; Arzberger et al. 2004; Dallmeier-Tiessen et al. 2014) and the extent of reuse that occurs (e.g., Rung and Brazma, 2012; Piwowar, Carlson, and Vision, 2010; Pienta et al., 2010; Peters et al., 2016). While some of this research has cut across multiple disciplines (e.g., Arzberger et al., 2004; Kuipers and van der Hoeven; 2009; Peters et al., 2016), much has been limited to particular domains of research (e.g., Wallis et al., 2013; Donaldson et al., 2017; He and Nahar, 2016; and Faniel and Yakel, 2011).

Through their work, researchers have identified a variety of factors that play a role in whether or not a given dataset is reused in research. These include determinations about the quality and relevance of the data for the intended use, the trustworthiness of the data source, the knowledge and experience of researchers reusing the data, the network of researchers who are aware of the data, and many others (e.g., Faniel et al., 2016; Curty, 2016; Zimmerman, 2008; Yoon, 2017b). Researchers have also found, however, that these factors, and the ways they interact, can vary across researchers, disciplines, and specific reuse instances (e.g., Medjedović, 2011; Yoon, 2017a; van den Berg, 2005; Stvilia et al., 2015). As a result, there is still difficulty in knowing how much of what kinds of information about data are needed to facilitate the greatest reuse for different research purposes (van den Berg, 2005; Faniel and Jacobson, 2010; Mauthner, 2014).

My dissertation research was designed to investigate these issues and address gaps in the research literature in particular relating to:

- The amounts and kinds of knowledge about data researchers obtain to determine that data are sufficient to reuse
- Factors that affect researchers' determinations about whether the amount of knowledge they have obtained about data is sufficient
- The relative priority that researchers assign to different types of knowledge about data in particular reuse instances, and why
- The ability to compare research findings across multiple instances of reuse

In my study, I used theories of feminist objectivity and situated knowledges (primarily from Haraway, 1991), and theories from information science—especially satisficing (Simon, 1994)—to develop a theoretical framework, methodology, and research project that contributes to filling these gaps. The study used a mixed method design, consisting of a quantitative phase of research (an online survey) followed by a qualitative phase involving background research and in-depth interviews. The specific research questions I asked were the following:

1. How do researchers determine the boundaries of the knowledge they obtain about data in order to reuse them in their research?
    a. What knowledge about data is most important to researchers to reach reuse equilibrium and why?
    b. How do researchers determine how much of what kinds of knowledge is enough?
    c. What do researchers report influences these determinations?
    d. How do researchers obtain the knowledge they desire?
2. Do researchers satisfice in the knowledge they obtain about the data they reuse?
    a. If so, how can researchers' satisficing be characterized?
    b. What factors are associated with knowledge satisficing?
    c. What reasons do researchers give for why they satisfice?
    d. What is the perceived impact of knowledge satisficing:
        i) On the outputs of research?
        ii) On researchers' achievement of their goals for reusing data?
    e. What do researchers believe could mitigate knowledge satisficing?

My research into the knowledge researchers obtain about data when deciding whether to reuse them contributes to the ability of stakeholders in data reuse (researchers, data managers, policy-makers, funding agencies) to make strategic decisions about the investments of time and resources needed to gather or link to information that will enable the broadest and most efficient reuse of data. Strategic decision-making is critical if aspirations to collect and steward reusable data and maximize scientific discoveries are to be achieved. This is especially the case given the significant time and effort that curators spend to prepare data for reuse (Niu, 2009; Borgman et al., 2007; Enke et al., 2012; Medjedović, 2011; Smioski, 2010) and that researchers spend to reuse them (Warwick et al., 2009; Curty et al., 2017; Faniel and Jacobson, 2010; Faniel et al., 2012; Zimmerman, 2008).

## 1.1 Organization of the Dissertation and Use of Pronouns

### *1.1.1 Organization of the Dissertation*

This dissertation is organized into five chapters. In the first, I introduce the research project and my motivations for pursuing it. I describe the central problem the research is

designed to investigate, the theoretical framework, methodology, and specific methods I employ, and the significance of my study.

In the second chapter I review literature related to the knowledge researchers obtain about data for reuse and identify gaps that my study seeks to address. I also describe the analytical model I employed and the hypotheses I tested in the quantitative portion of my study.

In the third chapter, I present the methodology of the mixed methods study in detail, including the quantitative and qualitative phases and how I integrated them during the development of the research methods, data analysis, and interpretation of results.

In the fourth chapter I present findings from the quantitative and qualitative portions of the study. In the fifth chapter I present my discussion and conclusions.

### *1.1.2 Use of Pronouns*

In my dissertation, I refer to research participants generically by their role as much as possible (e.g., "a researcher"). I use third-person pronouns in cases where it aids in the clarity of the sentence or the flow of language (i.e., reducing repetition). When I do so, I have chosen to use the pronoun "they," in part to be in alignment with guidelines from the American Psychological Association (American Pyschological Association, n.d.; Lee, 2019) and in part to maintain the confidentiality of the researchers who participated in my research. I do this, however, with an awareness of the way that singular categories can mask or subvert identities. Particularly in the context of my dissertation, where, in keeping situated knowledges, I seek to promote tensions and resonances that exist between categories, I am conscious of the potential of using "they" to minimize differences or suggest equality among genders when it does not exist.

Therefore, as I write, I keep and promote an awareness of the many different people of different identities and perspectives who conduct scholarly research. This is important so that

individuals who are considering or on the path to becoming scholars see themselves in those who are already conducting research. It is also important so we all remain mindful of the diversity of perspectives that exist in society and the benefit we can all obtain by including these perspectives in our institutions of science and scholarship, including data archives that support reuse of data in research. Therefore, while I use the pronoun "they," I would like for readers to keep in mind and visualize researchers of all of the following genders and identities, as well as others that may not be not listed here, who are striving to improve our understanding of ourselves and our world, because this is, in truth, the reality: women, men, transgender women, transgender men, trans men, trans women, cisgender women, cisgender men, people who identify as gender-fluid, gender-nonconforming, genderqueer, gender-nonbinary, gender-creative, agender, two-spirit, or transgender (American Psychological Association, n.d.).

### 1.2 Motivation

Over the last several decades, scholarship in the social studies of science has critically examined assumptions and biases that underlie the projected authority and objectivity of Western science (e.g., Haraway, 1991; Bloor, 1991; Wajcman, 2002; Freundlich, 2016; Harding, 2001). In recent years, the relevance of such critiques has extended to archives of scientific data and reuse of these data (Mauthner et al., 1998; Mauthner, 2014; Broom et al., 2009; Ribes and Jackson, 2013; Moore, 2007; Coltart et al., 2013). As digital technologies have become ubiquitous in academic research, greater attention has been placed on what outputs of scientific research are collected (due to their scientific, historical, or cultural significance), what information accompanies or is included in what is collected, and the social and technical processes that influence both of these. Information that accompanies what is collected (often referred to as "context") can be essential to interpreting and understanding what is collected

(often referred to as "data," although the "data" may include "context") and to verifying or otherwise supporting the scientific or scholarly integrity of the research.

Critical studies of data and context in data archives have focused primarily on questions of epistemology and ontology: is it theoretically feasible for a researcher to use data collected by someone else for one purpose with a specific methodology and underlying theoretical framework for another purpose (e.g., Hammersley, 1997; Mauthner et al., 1998; Moore, 2007; Van den Berg, 2005)? If so, to what ends and with what limitations? How do assumptions about what the data "are" affect practices for data sharing and archiving and what does this mean for what data "should" be archived to enable the broadest reuse of data (Mauthner, 2014)?

My interest in research to support data reuse grows out of such critical thinking about archives. However, in my dissertation I took a very practical and empirical approach to researching questions of reuse (even while placing my research in dialog with theory more broadly). That is, rather than seeking to understand whether it is feasible to reuse data or what constitute the "data" to be reused, I investigated the knowledge that researchers who engaged in data reuse sought to obtain about the data they reused. My results have implications for epistemologies of data reuse and what "should" be archived, but by investigating knowledge (as opposed, for instance, to assumptions about what the "data" are) I sought to cut directly to questions that revealed, from the researcher's perspective, limitations in data reuse and ways these might be reduced. This is one reason in my research that I took a position that viewed "data" and "context" together rather than separately (see sections 1.3 and 1.6).

My practical orientation is heavily influenced by my training in historical methods as an undergraduate and my subsequent career as an information professional in digital preservation and digital repository management. In my historical training, I learned the extent to which

sources of all kinds can be used as evidence if analyzed critically, but also the extent to which analysis, and what we can know about history, is limited by the sources that are available. I later became a project manager and then assistant director for a partnership of institutions that manages a large archive of digital books: HathiTrust. In my roles at HathiTrust, I gained experience with the social and technical processes that shape what is collected (mentioned above), and also the important roles archives play in the practical task of making resources available. I come to my research on one hand, then, from the perspective of a repository manager motivated by the desire to enhance the ability of historians, social scientists, and researchers of all kinds to use and develop new knowledge from archives of digital data. On the other, I come with interests and desires to enhance archives to conduct research of my own. Each of these is important to the goal, reflected in my dissertation, of conducting empirical research that will have a practical impact on research capabilities for data reusers.

There are many contexts in which data are reused: in teaching and education (Margaryan and Littlejohn, 2008; Bishop and Kuula-Luumi, 2017; Borgman et al., 2007a; Frank et al., 2017), organizational management (Pine et al., 2016), consumer analysis (Dargentas and Le Roux, 2005), the development of business intelligence (Wolski et al., 2017), professional development (Specht, 2015; Frank et al., 2017) and governmental administration (De Vocht et al., 2014; Podesta et al., 2014). I have chosen to study reuse of data in the context of academic research for several reasons. First, I am interested broadly in questions of how people (individuals, groups, and societies) make decisions about what to collect and preserve and the impact these decisions have on what we can know about ourselves and our past. The academic community is ideal for investigating these questions because it has a long-standing culture of inscription (e.g., Latour, 1986; Latour, 1990). Scientists and scholars communicate the results of their research in writing,

leaving evidence of what they have learned from their research. Studying the results of research that involved the use of previously collected data thus yields evidence of what can be known using data that has been collected and preserved.

Second, I have a specific interest in the academic enterprise and the role that science and scholarship have played, in connection with industrial, governmental and administrative concerns, in shaping the modern world. What we know, and the ways that we know through science and scholarship have come under scrutiny. On one hand, this is a result of critical postmodern and feminist perspectives that have exposed the biased and discriminatory contexts in which academic research is conducted (e.g., Haraway, 1991; Bloor, 1991; Wajcman, 2002; Freundlich, 2016; Harding, 2001). On the other, it is a result of high-profile cases of falsified or biased research that have reduced trust in the academy in some sectors of society and precipitated calls for strategies to ensure that the results of research can be adequately replicated or reproduced (Vogel, 2011; Open Science Collaboration, 2015; Winerman, 2017; Baker, 2012; Wicherts et al., 2011). Scrutiny of science and scholarship (from outside and inside the scientific community) is not new from a historical perspective (see, e.g., Shapin and Schaffer, 1985; Collins and Pinch, 1993; Collins, 1998). However, understanding how scientists and scholars are responding to debates and potentially changing the ways they use data in their research can provide insight into the nature of academic research at the current moment. It can also provide a benchmark for eventual comparative analysis of how research has been conducted in the past and how it will be conducted the future.

Finally—and primarily for this study—I am interested in ensuring that data that are shared from academic research are maximally useful to other researchers and the public at large. Currently, a variety of strategies, including research funder policies requiring data management

plans (see, e.g., University of New Mexico Libraries, 2018; DCC, 2018), journal policies requiring data to be deposited with publications (Piwowar and Chapman, 2008; Stodden et al., 2013; Sturges et al., 2014), and a constellation of data repositories (see, e.g., Kindling et al., 2017), tools, standards, and infrastructure are being mobilized to these ends. These strategies reflect a belief that is widely held among stakeholders in science and scholarship that the more data that are available—and in particular, that are findable, accessible, interoperable, and reusable (FAIR) (Wilkinson et al., 2016)—the greater the quantity, quality, and efficiency of academic research, and the more innovation, new discoveries, new knowledge, and trust in science that will result (Organization for Economic Co-operation and Development, 2015; Allison and Gurney, 2015; Holdren, 2013).

I have come to take a critical view of these beliefs, however—reflected in my current research—and to question whether more data sharing and data availability (even in reusable form) will result in improved science and scholarship. I raise this question in light of numerous studies that detail logistical, theoretical, ethical, and other challenges to reusing data that I discuss further in the literature review (e.g., Oleksik et al., 2012; Niu, 2009a; Parezo, 1996; Andersson and Sørvik, 2013; Carlson and Anderson, 2007; Coltart et al., 2013). Promoting critical accounts of accepted beliefs is a key intervention of feminist and postmodern theories (e.g., Haraway, 1991) that are at the heart of my interest in digital archives.

My concern is that taking the value of data sharing as a given makes assumptions about researchers, data, and data reusers that may limit the utility of the data available for reuse (e.g., by not critically examining the scope of the information desired by researchers when reusing data). I have a specific goal in my research to critically examine such assumptions—assumptions that could lead to infrastructure and policy initiatives for data archives that undermine ambitions

to broaden reuse of data and have unforeseen and potentially negative consequences for academic research.

For instance, using a NASA-funded mission as an example, Vertesi and Dourish (2011) described how the imposition of sharing norms in a scientific community without an understanding of local culture and values related to the production of data caused complications and divisions. Borgman et al. (2007a) identified concerns related to ownership of data, ethics, misuse of data, and "scooping" others' research that should be addressed in efforts to expand data sharing in a "little science" initiative. Mauther (2014) described her concerns about a new policy on data sharing by the Economic and Social Science Research Council (ESRC), the "largest funder of social science research in the UK" (Mauthner, 2014, p.177). She explained how the new policy and the new claim on data it gave to funding agencies, universities, and the public "seemed to be reconfiguring relationships between researchers and researched along new axes of power…reshaping the nature of these relationships" (Mauthner, 2014, p.178). Her concern was that under the new sharing policy it would "become harder to build the kinds of relationships based on trust, confidentiality, privacy and protection that researchers like myself had been used to doing" (Mauthner, 2012b, p.178). Finally, in their comparison of technical and social contexts that enable reuse in four e-Science communities, Carlson and Anderson expressed the concerns of social scientists that the "demand for data archiving and re-use might force qualitative materials into quantitative forms and logics (or favor the quantitative side of qualitative studies) and thereby jeopardize the specificity of qualitative approaches" (Carlson and Anderson, 2007, p.640).

To reduce the risk of making uncritical assumptions, I wanted to take a bottom-up approach to understanding challenges in data reuse, considering important questions from the

perspective of data reusers rather than policy-makers or repository administrators. I aimed to investigate in particular whether there was knowledge researchers desired about data that they lacked (and if so what it was), how researchers determined the thresholds of knowledge about data that were sufficient for them to decide to reuse the data in their research, and what influenced those decisions.

To deepen my analysis, I selected a theoretical framework that challenged me to conceive of the "data" that a researcher reused from an archive as the product of complex social and technical processes, and the means by which a researcher determined adequate thresholds of knowledge about data as a process of adaptation and boundary-making. The underlying motivations of my dissertation, then, have been to design a study capable of producing nuanced insights into the experiences researchers have when making decisions about reusing data and consider strategies for how to better support research conducted with secondary data.

### 1.3 Overview of Problem: Data and Context

Many investigators of data reuse have conceived of data that is reused in two parts: the "data" (which may have been accessed from a data archive, requested from a researcher, or obtained in some other way) and the "context"—e.g., the metadata, documentation, or provenance information the researcher believes is necessary to understand and reuse the data (see the constructions used and discussed in, e.g., Dicks et al., 2006; Holstein and Gubrium, 2004; the description of data reuse literature in Faniel et al., 2013 (p.297); Anderson and Sørvik, 2013; Carmichael, 2017; Bishop 2007; Borgman et al., 2007a; Borgman et al. 2007b; Carlson and Anderson, 2007; Warwick et al., 2009; Corti, 2005; Curty et al., 2017; Faniel and Yakel, 2011; Faniel and Yakel, 2017; Faniel et al. 2016; Faniel and Jacobson, 2010; Huggett, 2018; Irwin, 2013; Irwin and Winterton, 2011a; van den Berg, 2005; Kim and Yoon, 2017). A

conceptualization of reused data in terms of "data" and "context" has produced valuable insights into the information researchers need to effectively reuse data. However, there is a need to enhance this conception in light of findings that have identified more complex and entangled relationships between data and context. This is the case for several reasons.

First, a conception of the duality of "data" and "context" does not offer a means to account for an important dimension of data reuse—that there is not a predetermined amount or kind of context that is sufficient to understand data or determine whether they are, e.g., of sufficient quality, trustworthiness, or accuracy when using them for a given purpose (Tyler et al., 2020; Lee, 2011; Moore, 2006; Dicks et al., 2006, Holstein and Gubrium, 2004). Researchers themselves determine the amount of context, degree of quality, or degree of trustworthiness of data that is "good enough" to achieve their reuse goals (Curty, 2016; Faniel et al., 2013; Huggett, 2018; Niu, 2009a; Temple et al., 2006; Thompson, 2000; Wan and Pavlidis, 2007; Zimmerman, 2008; Carlson and Anderson, 2007).

Second, the conception lacks sensitivity to the reality that a set of data may serve as the "data" for one research scenario and the "context" for another (Stvilia et al., 2015; Wallis et al. 2007; Huggett, 2018), confounding efforts of researchers and other stakeholders to determine what "contextual" information should be recorded to enable broad reuse. Finally, as calls for "full documentation" of provenance or contextual information make clear (Dicks et al., 2006; Hills et al, 2015), "context" can in fact consist of a nearly endless regression of information (i.e., context needed to interpret the context, etc.) (van den Berg, 2005; Carlson and Anderson, 2007).

A more fluid conception of data is needed to support empirical investigation of data reuse that understands context not as something that is objectively sufficient for a given purpose or not, but rather something that is bounded along with all of the other evidence that a researcher gathers

to investigate and make claims about a behavior or phenomenon (Mauthner et al., 1998; Mauthner, 2014). Furthermore, a conception is needed that acknowledges, with regard to data reuse (and, some would argue, primary data collection as well), that all data and context are partial and incomplete (Moore, 2007; Carmichael, 2017; Gillies and Edwards, 2005). Aside from information about the original research being unavailable to data reusers as a result of reusers' not "being there" when the data were collected (see the discussions in Corti, 2000; Medjedović, 2011; and Mauthner and Parry, 2009), data may need to be anonymized (Andersson and Sørvik, 2013), there are time and resource constraints to recording relevant details from the original research (Borgman et al., 2007a; Pampel and Dallmeier-Tiessen, 2014), and some procedural or other tacit information can be difficult or impossible to record (Kelder, 2005; Roland and Lee, 2013, Niu, 2009a).

A notion of the "data" in data reuse that was sensitive to and accounted for these complexities could enable new kinds of analyses across instances of data reuse—both within and perhaps even across disciplines—about how and why researchers draw boundaries around the totality of the knowledge they obtain about data in order to reuse them. Knowledge about data is but one component in a larger ecosystem of knowledge and resources researchers mobilize to reuse data. Other components in this ecosystem include data archives, standards, analysis software and techniques, data curation practices, and cultures of intra- and inter-disciplinary communities. Considering all of these elements as a whole can guide us in assessing whether and how we might most effectively intervene in traditions and conventions of scholarship and their current trajectories to improve the efficiency and outputs of scientific research. Examining researcher knowledge about data is a vital step in this direction.

## 1.4 Theoretical Framework

The ideas necessary to construct a framework that is sensitive to and accounts for the complexities mentioned above exist in feminist theories that have been used in debates about data reuse, and theories from information science. In my dissertation, I draw specifically on the theories of Donna Haraway and Herbert Simon. I first provide an overview of what each person's theories contribute to my study and how, in my usage, the theories intersect (sections 1.4.1-1.4.3). I then go into the theories in greater detail (sections 1.4.4-1.4.7) and describe (in section 1.4.8) how I bring Simon's and Haraway's theories into conversation to develop the theoretical framework for my study. In the final section (1.4.9) I bring the theoretical framework together with considerations about epistemology and ontology in data reuse.

### *1.4.1 Donna Haraway*

Donna Haraway's theory of situated knowledges provides a way of understanding the boundaries of data as contingent on social factors and situated knowledge. She argues that in order to approach more 'realistic' accounts of the world we have to acknowledge that our perspectives are partial. In the case of data reuse, this translates into those who determine what data are deposited into the archive acknowledging that there is not one perspective (i.e. their own) from which the archived data are "complete." In other words, there is not one body of archived data that encompasses all that happened in the research or all that researchers will desire to know about the data in order to reuse them. Different researchers see data in different ways and have different needs. If we acknowledge that the perspectives held by stakeholders in data archives are partial, this raises questions of how and in what ways their perspectives are limited. These questions, in turn, lead to consideration of what knowledge should be included in

or made available from the archive—i.e., by facilitating access to desired knowledge about data—in order to better support reuse.

### *1.4.2 Herbert Simon*

I use Herbert Simon's articulation of the sciences of design of adaptive systems (the Sciences of the Artificial) to inform an understanding of data archived in a repository as a design "artifact." This framing connects the specific research I perform on knowledge satisficing and knowledge bounding to larger conversations about what data and knowledge about data are preserved and made available from archives. It also enables a consideration of data reuse itself as an adaptive process, where researchers take steps to determine whether, by enhancing their own ability to reuse data or changing the goals of their research, they can reach a state of "reuse equilibrium" where the data satisfy the needs of their research. Simon's notion of satisficing, in particular, offers a way of understanding and characterizing the thresholds of knowledge about data that researchers determine are sufficient to decide to reuse data in their research.

### *1.4.3 Intersection of Theories*

Applying Haraway's theory of situated knowledges to the design of data "artifacts" can enhance both the design of these artifacts and the strategies we use to study their design (i.e., the perspectives and factors we consider to be important in evaluating the process of design). At the same time, considering situated knowledges in the context of design can provide insight into how "conversations" among situated perspectives—which Haraway asserts are necessary to sustain rational discourse—might proceed to improve the design of data archives and better support data reuse. In the sections below, I elaborate on my use of Haraway's and Simon's theories, and how I bring them together in my study.

### *1.4.4 Objectivity and Boundaries*

One of the primary debates surrounding data reuse—qualitative data reuse in particular—has to do with the degree to which data from previous research that are archived or shared can be taken by themselves as an objective representation of a behavior or phenomenon, or whether (and to what degree) additional information about the context of creation and other details might be necessary to understand and reuse the data (see, e.g., Mauthner et al., 1998; Irwin, 2013). Within these debates are threads that relate directly to notions of feminist objectivity (e.g., Mauthner, 2014; Broom et al., 2009; Ribes and Jackson, 2013), some of which cite Donna Haraway, a feminist scholar who writes about this concept, directly (e.g., Moore, 2007; Coltart et al., 2013).

Below, I describe Haraway's theories of feminist objectivity and situated knowledges in more detail as one component in a new theoretical framework for thinking about and investigating data reuse.

In *Simians, Cyborgs, and Women,* Haraway criticizes prior efforts in the sociology of science and feminist studies to show the subjective nature of science—to portray science as a game of rhetoric, where rhetorically generated knowledge is used as a means of obtaining power (Haraway, 1991, chapter 9, "The Reinvention of Nature."). Such efforts have become conceptually untenable, she argues, and a new metaphor is needed to acknowledge and identify *some* level of reality. As Haraway describes,

> We unmasked the doctrines of objectivity because they threatened our budding sense of collective historical subjectivity and agency and our 'embodied' accounts of the truth, and we ended up with one more excuse for not learning any post-Newtonian physics and one more reason to drop the old feminist self-help practices of repairing our own cars…So, I think my problem and 'our' problem is how to have *simultaneously* an account of radical historical contingency for all knowledge claims and knowing subjects…*and* a no-nonsense commitment to faithful accounts of a 'real' world... [emphases in original] (Haraway, 1991, p.187)

The new metaphor Haraway proposes for achieving accounts of a 'real' world is vision. Haraway's vision is not the illusory, disembodied vision that purports to see objective, universal truths (which she argues has pervaded the conduct of science and refers to as the "god-trick"). Haraway proposes instead an embodied vision, one that "allows us to construct a usable, but not innocent, doctrine of objectivity" (Haraway, 1991, p.189).

Haraway's notion of vision has two key facets. The first is that it is "situated" and "partial." Haraway says, "[w]e need to learn in our bodies, endowed with primate colour and stereoscopic vision, how to attach the objective to our theoretical and political scanners in order to name where we are and are not, in dimensions of mental and physical space we hardly know how to name" (Haraway, 1991, p.190). Objectivity comes not from a detachment from our surroundings, our beliefs, or ourselves, but from a recognition of the "partial" perspectives we hold based on where and how our perspectives are located in relation to others'. Haraway summarizes:

> Feminist objectivity is about limited location and situated knowledge, not about transcendence and splitting of subject and object. In this way we might become answerable for what we learn how to see. (Haraway, 1991, p.190)

Haraway emphasizes that in her view, limited location is not about dichotomies, but about tension and resonance. For instance, one might think that in the face of a universal truth or universal world system, limited location would advocate for its opposite: local knowledges. However, referring to a chart of such opposites, Haraway says that these dichotomies

> [misrepresent] in a critical way the positions of embodied objectivity which I am trying to sketch. The primary distortion is the illusion of symmetry in the chart's dichotomy, making any position appear, first, simply alternative and, second, mutually exclusive. A map of tensions and resonances between the fixed ends of a charged dichotomy better represents the potent politics and epistemologies of embodied, therefore accountable, objectivity. For example, local knowledges have also to be in tension with the productive structurings that force unequal translations and exchanges – material and semiotic –

> within the webs of knowledge and power…Feminist accountability requires a knowledge tuned to resonance, not to dichotomy. (Haraway, 1991, p.194)

Scholars who have applied this or other postmodern perspectives to data view data as situated knowledges representing partial perspectives that are deeply entangled with the circumstances of their creation, as opposed to disembodied knowledges that can be understood separately from their constitutive context. Moore (2007) makes this point when discussing differences other scholars have described between qualitative and quantitative data. Reflexivity as it is used in the quote below refers to data being co-constructed (reflexively) by the researcher and research participant, as opposed to being gathered by the researcher from a passive participant:

> For Mauthner et al. reflexivity distinguishes qualitative data from quantitative data (1998). However this account of quantitative and qualitative research relies on accepting discourses of the objectivity of quantitative research at face value. While it is of course qualitative research that has foregrounded questions of the situatedness of knowledge, contrary to the 'god trick of seeing everything from nowhere' (Haraway 1988: 571), the comparison, perhaps unwittingly, upholds the notion of the objectivity of quantitative research. While researchers may commonly behave as if quantitative data are objective, ultimately, all data, qualitative and quantitative, are 'situated knowledges' which can only be understood as and through 'partial perspectives'. The distinction is not so much that qualitative data are situated and context bound and quantitative are not, but that reflexivity reveals all research qualitative and quantitative as situated; however the qualitative researcher is more likely to open up the process of knowledge production to scrutiny, and to attend to the conditions of production of the research. (Moore, 2007, p.7)

Coltart et al. (2013) take a similar position while disputing a claim that a researcher's distance from the original research in secondary analysis actually affords greater analytical abilities because it allows the researcher to examine "more datasets, perspectives, and evidence" (Coltart et al., 2013). Coltart et al. argue that "this reflects a quantitative epistemological position (the myth of the omniscient researcher) which has been soundly critiqued (HARAWAY, 1991; HENWOOD & PIDGEON, 1995)" (Coltart et al., 2013, capitalization in original). Coltart et al., Moore, and others call for a greater acknowledgement of and accounting for the entanglement

that exists between researchers and contexts of creation in data that are reused for subsequent research (Mauthner et al., 1998; Mauthner, 2014; Irwin, 2013; Ribes and Jackson, 2013; Broom et al., 2009).

Haraway continues her argument by highlighting the second key facet of situated knowledges: things formerly seen as passive objects from which knowledge was obtained or extracted can regain their agentive status:

> Feminists, and others who have been most active as critics of the sciences and their claims or associated ideologies, have shied away from doctrines of scientific objectivity in part because of the suspicion that an 'object' of knowledge is a passive and inert thing…Situated knowledges require that the object of knowledge be pictured as an actor and agent, not a screen or a ground or a resource, never finally as slave to the master that closes off the dialectic in his unique agency and authorship of 'objective' knowledge…Indeed, coming to terms with the agency of the 'objects' studied is the only way to avoid gross error and false knowledge of many kinds in these sciences. (Haraway, 1991, p.197-198)

Haraway argues that recognizing the agency of 'objects' of knowledge—not only people, but also other 'objects' of study in the social sciences and sciences more broadly—transforms the way we understand the knowledge gained from research. Instead of knowledge being 'objective,' faithful accounts of the world depend on "a power-charged social relation of 'conversation'" (Haraway, 1991, p.198). If we understand all knowledge to be partial and situated, there is no one perspective from which objects of knowledge can be totally known. Such a recognition requires—if more realistic accounts of objects are to be made—a consideration of other perspectives and the factors that lead to those perspectives (including attributes of the 'object' itself). When other perspectives are partial, the 'object' thus regains agentive status as a player in the associations and interactions—the conversation—involved in its own production and interpretation.

Haraway articulates a means of understanding the production and reproduction of "bodies and other objects of value in scientific knowledge projects" in this more interactive conception of research. She draws on Katie King's metaphor of the "apparatus of literary production" (King, 1987, cited in Haraway, 1991) to propose an "apparatus of bodily production" (Haraway, 1991, p.200). As in King's model, the apparatus is a matrix of associations out of which the body is born. The key feature of this generation process is that the body or object is never finalized, but rather in continual production. Its boundaries materialize in conversation or social interaction with the associations around them:

> 'objects' do not pre-exist as such. Objects are boundary projects. But boundaries shift from within; boundaries are very tricky. What boundaries provisionally contain remains generative, productive of meanings and bodies. (Haraway, 1991, p.201)

If we apply Haraway's understanding of bodily production to data, a more organic understanding of the 'object' of data selected for data reuse emerges: one where the data that are reused are determined from ongoing interactions between attributes of the researcher and research process (e.g., researcher knowledge and experience; research goals, methodology, and resources), attributes of the 'object' of knowledge (e.g., size, complexity, format, or documentation of the data), and associations or relationships that the researcher or research 'object' bear with one another and other entities and forces that affect the data reuse process. We could imagine, for example, that the totality of data selected for a research project would be affected by the research purpose, the ease with which the researcher is able to use the data, and other factors such as funding priorities that guide the trajectory of the research, interests and skills of research collaborators, and factors affecting whether and how much of the data are available to use. All of these together determine the boundaries of the body of evidence (the data) that researchers ultimately investigate and use to support their research claims.

Haraway's theory offers a way of understanding the boundaries of data as contingent, emerging differently in different reuse scenarios from ongoing interactions between attributes of the researcher and the research process and the relationships that the researcher or research data bear with one another and other entities and factors that affect the research process. The notion that all knowledge is partial opens the door to questioning more specifically at what points knowledge about an 'object' is limited, and why. In the context of data reuse, where researchers make decisions about the level of knowledge that is sufficient to reuse data, Haraway's theory provides a framework for investigating how and why researchers bound the breadth and depth of knowledge they believe is necessary to reuse data in their research. These are the aims of my proposed study.

Much is already known about factors affecting the selection of data for reuse (see Chapter 2). Understanding data that are gathered or produced to study an 'object' of knowledge (including reused data) as a boundary project provides a means of expanding what is known in several respects. While we know, for example, that quality and accuracy are important to decisions about reusing data (Birnholtz and Bietz, 2003, Curty and Qin, 2014; Stvilia et al., 2015; Sherif, 2018), less is known about how researchers determine the boundaries of what constitutes sufficient quality or accuracy and why. How much do researchers need to know about data to determine that the data quality is "good enough?"

Similarly, while researchers have characterized many types of context (e.g., interactional, situational, institutional) (Bishop, 2006) and know that what counts as context changes depending on the object of study (Andersson and Sørvik, 2013), we know less about how researchers determine which contexts count and why (Kelton, 2008; Curty and Qin, 2014; Curty et al., 2016), and how much of these contexts are adequate to use the data as evidence of the

behavior or phenomenon they are investigating (Borgman et al., 2007a; Mauthner, 2014; Irwin and Winterton, 2011b; Sherif, 2018). Understanding the production of data as a boundary project provides a framework for going beyond identifying types and sources of context and types of factors affecting decisions about data reuse (though these do not lose their importance) to understanding at what points, why, and to what ends researchers admit particular contexts and amounts of context as valid evidence of their 'objects' of study. In my research, I aim to use a measure of how much researchers seek to know about data in order to use them as a way of evaluating these dimensions.

Third, understanding the data used in research as a boundary project heightens our sensitivity to the co-constructed nature of data and the multiple forces involved in defining or 'materializing' the data's boundaries. Researchers do not determine the boundaries of data by their wills alone. Boundary construction is rather a process affected by the agency of the data (e.g., how easy they are to use) and a variety of other associations and relationships that are present or emerge in the circumstances of reuse. An analysis of what contexts researchers consider and gather information about when reusing data, and the factors that affect the amounts and kinds of knowledge about these contexts, can lead to deeper reflection on and conversations about what knowledge about data is collected and made available from data archives to support data reuse. These conversations are integral and critical to the dynamic Haraway envisions and argues is necessary to achieve "rational knowledge." She says,

> Above all, rational knowledge does not pretend to disengagement: to be from everywhere and so nowhere, to be free from interpretation, from being represented, to be fully self-contained or fully formalizable. Rational knowledge is a process of ongoing critical interpretation among 'fields' of interpreters and decoders. Rational knowledge is power-sensitive conversation (King, 1987a) (Haraway, 1991, p.196).

Finally, a key limitation of studies of data reuse has been the difficulty, due to the complexities and particularities of specific instances of data reuse, of relating the findings of individual studies to other studies of data reuse or generalizing about behavior (for instance, how quality is judged) outside of specific circumstances (see Chapter 2, section 2.4.6). Examining how boundaries of data are constructed does nothing to change the particularity of instances of reuse, but it does offer a frame of analysis that may be more comparable across reuse instances. We may not need to know how quality is determined, for example, to characterize the knowledge needed by the researcher in order to determine that the level of quality is "good enough" to be used in the research. Analysis in this framework is capable of revealing interesting similarities and differences in how and why reused data (including both "data" and "context") are bounded in different reuse instances, both within and across disciplines.

A question remains of how, specifically, in a boundary framework, conceptions of how much knowledge is "good enough," and why, can be empirically measured. I address this question and elaborate the second part of my theoretical framework in the next section using ideas from information science.

### *1.4.5 Sciences of the Artificial*

In The Sciences of the Artificial, Herbert Simon (1994) develops a framework for understanding natural and "artificial" adaptive systems—i.e., systems that are either biological or synthesized by people. The framework proposes that all living or mechanical things that are adapted to a purpose or environment (either through evolution or design) can be understood as "artifacts" that exist as the interface between distinct "inner" and "outer" environments (Simon, 1994, p.9). The inner environment comprises the physical structure and computational abilities of the artifact. The outer environment is the "surroundings" in which the artifact lives or operates

(Simon, 1994, p.9). Thus, the inner environment of a person would be their physiological structure and cognitive abilities, and the inner environment of a traditional clock would be the arrangement of its gears, springs, and weights. The outer environment of either would be the outside world—characterized by temperature, light, sound, geography, time, other people and artifacts, etc.—in which they exist and function.

An additional component of Simon's framework is a purpose or goal that drives the inner environment's adaptation to the outer environment. Thus, an animal such as the artic fox, for the purpose of survival, has adapted to its environment by developing a white coat (Simon, 1994, p.10). A person with a goal to fly might adapt to physical conditions by designing and synthesizing an artifact such as an airplane. Simon applies this framework to animals (including people) and to larger adaptive systems such as business firms or other organizations, and markets that people design to enhance their abilities to make decisions in environments of uncertainty and complexity.

Simon uses this model to articulate strategies by which complex systems can be studied without a knowledge of all of the details of their functioning. For instance, if an airplane were not able to fly at high altitude, shortcomings of its inner environment (i.e., attributes of its engine and other mechanical parts) in the face of the outer environment would be exposed or "show through" (Simon, 1994, p.16).

There are two main applications of Simon's model that I make in my research. The first is to understand data that exist in an archive as an artifact that people design to achieve certain purposes, including supporting new discoveries and improving trust in science. This understanding links my research and research results to a broader conversation about policies and practices that shape the data that are archived for future reuse. The second application, and

the one that relates most immediately to my empirical study, is to understand data reuse as an adaptive process, where researchers take steps to determine whether data are adequate to help them achieve their specific research goals in environments of uncertainty and complexity. I discuss the second application immediately below, and return to the first later in discussing the results and implications of my research.

### *1.4.6 Reuse Equilibrium*

Extrapolating from Simon's arguments, we can reasonably understand research, conceived of broadly—including data reuse—as a strategy that humans use, given the limitations of our attention, memory, and other features of our inner environment, to adapt to an outer environment characterized by uncertainty and complexity. This characterization of research and adaptation allows us to identify when research is imperfectly adapted for particular research goals—when, for instance, the methodology, instrumentation, calibration, or operationalization of research concepts imperfectly measure the phenomenon or behavior of interest. In these cases, errors may "show through," e.g., in the form of inaccurate results or unexplainable anomalies.

If we expand on this analogy, the degree to which the tools and strategies of research are adequately adapted to the outer environment can be evaluated through the data that result from research. Researchers use triangulation, verification, and other strategies to validate the results of their analyses. Data reuse relies on the data collected from previous attempts to adapt the inner environment to the outer (i.e., to understand particular behaviors or phenomena). Thus, using Simon's construction, instances of data reuse provide an opportunity to study how researchers make decisions about whether the data they select for research are sufficient enough representations of the outside world to be used to study their subjects of interest.

In his discussion of adaptation and inner and outer environments, Simon uses the term "homeostasis" from biology to describe a state where the inner environment is insulated against effects of the outer environment. In this state, "an invariant relation is maintained between inner system and goal, independent of variations over a wide range in most parameters that characterize the outer environment" (Simon, 1994, p.8).

Drawing on this idea, for the purposes of my research design I introduce the concept of "reuse equilibrium"—a concept similar to homeostasis. Reuse equilibrium can be said to exist when a researcher determines that a particular body of data is sufficient to reuse to study their subject of interest and decides to reuse the data to accomplish a specific purpose (in other words, the researcher believes they have, by selecting data to reuse that provide an accurate representation of the behavior, phenomenon, or other subject of research, sufficiently insulated themselves from errors that might "show through" in the research process).

In this formulation, and in the context of data reuse, the process of selecting which data to reuse is one component in a larger "research equilibrium" that includes the selection of methodology, instruments, calibration settings, technological infrastructure, and other resources (as well as data) to conduct the research. Data reuse can thus be seen as its own adaptive process where a researcher explores whether they can adapt their inner environment to the outer environment of the data to be reused.

How do researchers adapt to data in order to reach reuse equilibrium? I argue that there are two dimensions to adaptation—which I refer to as configuration and reformulation—and one modulating factor: satisficing.

**Configuration.** I draw my ideas about configuration from Grint and Woolgar's (1997) description of ways early computers were intended to "configure" their users—i.e., to convey,

through their documentation and design, the limited ways in which the machines could be used. In configuration as it relates to reuse equilibrium, researchers mobilize, activate, or respond to factors internal to themselves or their research team (i.e., the inner environment) or external to themselves or their research team (i.e., the outer environment) in order to reach reuse equilibrium.

Examples of factors or attributes of the inner environment are skills, knowledge, attitudes, beliefs, and assumptions. A researcher might mobilize or enhance certain skills, for example (perhaps with a statistical program or programming in a computer language), in order to analyze a dataset. The researcher might also respond to a personal belief or assumption about data reuse (for instance, that data collected by someone else are less trustworthy) by requiring a higher level of verification or knowledge about the way the original data were collected before admitting the data as evidence.

Attributes of the outer environment include research culture or traditions, disciplinary standards for metadata or documentation creation, the status accorded to researchers, publications, or repositories that might be consulted in the process of data reuse, or pressures that affect the time, staffing, or funding available for research. A researcher may implicitly trust data if it comes from what is considered to be a reputable source, for example, or be influenced in what they determine is necessary to know about data by conventions for data reuse in their field. Attributes of the outer environment also include documentation and support for reuse that may be offered by libraries or repositories, or through workshops on data reuse. All of these are attributes that, operating outside the will of the researcher, could be seen to be acting to "configure" the researcher in their reuse of data.

In my research, I examined how researchers configured themselves to reach reuse equilibrium specifically through their attainment of knowledge about data. I also investigated attributes of researchers' inner and outer environments that they believed contributed to or constituted the reasons why they bounded knowledge about data the way they did. For my purposes, it was less important to determine whether factors were acting to configure data reusers than it was to assess what data reusers believed influenced their decisions to bound their knowledge about data (i.e., to configure themselves through knowledge attainment) in order to achieve reuse equilibrium. It was also less important which specific attributes to consider as part of the inner or outer environment than, as above, to assess what influenced decisions to bound data. I discuss this further in section 1.4.7 below.

**Reformulation.** The framework I introduce also suggests that researchers and data may undergo processes of reformulation that bring researchers closer to reuse equilibrium. This might occur when a researcher adapts to circumstances by reformulating aspects of their inner environment. An example would be a case where a researcher finds that the data they were hoping to reuse are insufficient for the desired purpose. The researcher might decide not to reuse the data, or determine that the data can be used for a different purpose and change the topic or goals of their research (Bishop, 2007; Doolan and Froelicher, 2009).

Data reformulation might happen if a data steward decided to migrate data to a new format, or enhance data or services to provide new searching, browsing, or aggregation capabilities. It could also happen if errors in data were corrected or new techniques were developed to reduce or prepare data for analysis. The main criterion would be that the data changed independently of the researcher in ways that brought the researcher closer to reuse equilibrium.

The main way that reformulation enters my research is I evaluated whether researchers had previously established research questions that they sought data to investigate, or whether their research questions developed as they explored data to reuse.

**Satisficing.** The modulating factor involved in adaptation to reach reuse equilibrium is an additional concept from the information sciences: that of satisficing. Satisficing is a term coined by Simon to describe the ways that our actions are oriented towards achieving acceptable solutions to problems as opposed, necessarily, to optimal ones (Simon, 1994). The concept of satisficing could explain why a researcher would select a data set to reuse that was not ideal for the desired purpose (for instance, missing data points) but nevertheless, the best available (Langkamp et al., 2010). The central issue would be whether the data met an acceptable level of quality (or accuracy, reliability, etc.) for the researcher's intended purpose. According to Simon, satisficing can be measured through a comparison of aspiration levels and achievements along different dimensions. In the case of my research, these would be the different dimensions of knowledge about data that were important to researchers:

> Aspiration levels provide a computational mechanism for satisficing. An alternative satisfices if it meets aspirations along all dimensions. If no such alternative is found, search is undertaken for new alternatives. Meanwhile, aspirations along one or more dimensions drift down gradually until a satisfactory new alternative is found or some existing alternative satisfices. (Simon, 1994, p.30)

### *1.4.7 Departures from Simon's Theories*

There are several ways in which my use of concepts to define "reuse equilibrium" depart from or extend Simon's original meanings. The first relates to the composition of the inner environment. As I noted above, in Simon's framework, the inner environment strictly comprises physiological and cognitive properties—for instance, limits on how humans process information and the amounts of information we can store in short-term memory (Simon, 1994). In fact, one of Simon's (1994) central arguments is that humans as behaving systems (referring to features of

our inner environment that underlie our behavior) are relatively simple; it is complexity in our environment that largely explains the complexity of our behavior.

Simon considers long-term memory to constitute a second outer environment, and learning and skill-development to be adaptations of this environment (i.e., we learn or develop skills by storing additional data or information about processes in our long-term memory). On a conceptual level, regardless of specific neurobiology, whether skills or knowledge are a part of an "inner" our "outer" environment, they still have a bearing on decisions researchers make about reusing data (Niu, 2009a; Corti and Bishop, 2005; Curty, 2016; Faniel et al., 2013; Zimmerman, 2008). Moreover, the skills or knowledge a researcher develops can help them to process information more quickly (i.e., by learning more efficient ways of chunking or representing information in order for it to be accessed or computed quickly). The same could be said for attitudes, preferences, or biases that a researcher has. Whether these are conceived of as being part of the inner or outer environment, a variety of characteristics that are personal to a researcher may have a bearing on decisions about reuse that they make.

Therefore, in my research, I was more permissive and less concerned, strictly, with what comprises the inner or outer environment. What was of primary concern was what aspects—whether of the inner or outer environment—researchers perceived to impact their decisions about data reuse. This view, in which lines between inner and outer environment are blurred, is heavily influenced by Haraway's theory of situated knowledges; in particular her assertion that limited location is not about dichotomies, but about tension and resonance. I considered: how might our outlook change if instead of splitting inner and outer, we considered resonances between them and avenues of mutual construction? This view falls within the range of perspectives that have been taken on the composition of the inner and outer environments Simon proposes, including

that there is no clear way to draw a line distinguishing between the two (see Wheeler, 2020). It also coincides with arguments Kahneman (2003) has made about the difficulty of excluding emotions or the beliefs and preferences that underlie them when evaluating factors in decision-making.

My use of the term “satisficing” differs also from what Simon describes. Satisficing is a decision optimization strategy that occurs in the context of bounded rationality, which is the idea that people are limited in their decision-making by aspects of both the inner and outer environments. Simon (2000) describes bounded rationality as follows:

> Bounded rationality is simply the idea that the choices people make are determined not only by some consistent overall goal and the properties of the external world, but also by the knowledge that decision makers do and don't have of the world, their ability or inability to evoke that knowledge when it is relevant, to work out the consequences of their actions, to conjure up possible courses of action, to cope with uncertainty (including uncertainty deriving from the possible responses of other actors), and to adjudicate among their many competing wants. Rationality is bounded because these abilities are severely limited. Consequently, rational behavior in the real world is as much determined by the "inner environment" of people's minds, both their memory contents and their processes, as by the "outer environment" of the world on which they act, and which acts on them. (Simon, 2000, p.25)

Simon describes concerns affecting rationality that are due to the outer environment as “substantive rationality.” These are considerations that affect what decision is made. He describes considerations due to the inner environment as “procedural rationality,” which relates to how decisions are made (Simon, 1994, 2000). A main development Simon notes in modern economic theory is the incorporation of consequences of procedural rationality into traditional understandings of substantive rationality (Simon, 2000).

To further illustrate these concepts, if a decision is not so complex or demanding that it overloads a person’s cognitive abilities to compute an optimal decision, the decision outcome is a result of the person’s goals in making the decision, and attributes of the outer environment (i.e.,

what the choice options and attendant considerations are) (Simon, 1994, 2000). If a decision taxes a person's ability to evaluate all possibilities in a problem space and compute the most optimal outcome, limitations of the inner environment begin to affect the decision. Simon argues that in such circumstances, people use heuristics to reduce possible options and make decisions that, by their nature, are geared toward "acceptable" rather than optimal solutions (Simon, 1994). Satisficing is one of these heuristics. Satisficing in its traditional sense is thus a phenomenon that takes place in the context of procedural, rather than substantive rationality: it is our inability to represent or compute optimal solutions (limitations of our inner environment) in the face of limitations of the outer environment (e.g., time or knowledge) that causes us to seek heuristic or satisficing solutions.

The question arises in my research of how to characterize a circumstance in which a person who is not necessarily limited by procedural rationality (i.e., in how to solve a problem) makes a choice that is less than optimal. It is a question, essentially, of whether or how satisficing can apply to substantive rationality. A researcher may not be aware of all possible data that are available to reuse to perform their research and consequently choose a satisficing strategy to identify relevant data (e.g., select data to reuse from a well-known repository). A researcher might similarly—if lacking desired knowledge about the quality of a particular dataset and faced with limitations in computational or other resources to compare the dataset with all similar datasets—choose a satisficing strategy. For instance, the researcher might seek information about the reputation or research practices of the person who created the dataset. In this case, according to the theory of satisficing, the researcher's aspirations have been adjusted down in the face of insurmountable procedural limitations and the researcher has found an "acceptable" solution.

However, a researcher might also, in the end, choose to reuse a dataset that is less than optimal for their research purpose even when there are no such procedural limitations. Table 1.1 represents some of the knowledge about data that respondents to the pilot survey indicated were lacking or limited in some way (11 respondents overall, out of 45, indicated there was knowledge about data they desired but was lacking or limited).

*Table 1.1 Pilot Survey Responses to Question About What Desired Knowledge Was Lacking*

| Case | Limited knowledge |
|---|---|
| 1 | specific methods for selecting prisoners; specific methods for oversampling; specific survey methods in prison settings |
| 2 | whether mental disorder was diagnosed in prison or before |
| 3 | information about women convicted of homicide related charges |
| 4 | the item content from commercially published surveys included in the data set (omitted due to copyright, etc.) |
| 5 | Why decision to alter some item/scale content were made |
| 6 | zip code |
| 7 | recoded participant groups |
| 8 | state identifiers |
| 9 | transgender status; specific sexual orientation status (e.g., bisexuals specifically) |

*Note*. The full question text was "Please describe the three most important things you would have liked to know about the data but were not able to learn to the desired degree."

In most or perhaps all of these cases, the optimal state is computable or known (e.g., the researcher knows that they desire knowledge about survey methods or sexual orientation) but not optimally obtainable. Furthermore, it is not clear that there are any limitations in cognitive abilities, time, or knowledge (e.g., where to seek information) that prevent the researcher from obtaining the desired knowledge. In fact, while it is possible that such limitations could apply, evidence from the literature and my pilot study suggest that even when researchers' knowledge about data is limited, they often make decisions about reusing data with specific understanding of how that knowledge is limited (e.g., Curty, 2016; Faniel et al., 2013; Huggett, 2018; Niu, 2009a) and how any limitations will affect their research outcomes. For instance, if a researcher chose to reuse data that were suboptimal, they must have determined that the effect on their research outcomes would not be so severe as to make the research irrelevant. To put it another way, the researcher must have adapted in order to reach reuse equilibrium.

In situations such as these it becomes possible—and desirable if the purpose, like mine, is to study the thresholds of knowledge researchers determine are necessary to reach reuse

equilibrium—to measure an optimal state in relation to an achieved state. In other words, it becomes possible to measure aspiration levels and, significantly, the difference in aspiration levels at an initial point and at reuse equilibrium. This is what I seek to do in my study and is one component of what I refer to as the phenomenon of "knowledge satisficing." The other component captures satisficing in its traditional sense such that in my usage, knowledge satisficing has two components. It describes situations where researchers opt for "acceptable" outcomes due to limitations either in procedural rationality or limitations in substantive rationality.

In his article of 2000, Simon notes on one had that a complete theory of bounded rationality will be as much concerned with procedural as substantive rationality (p.25) and on the other that there did not exist (as of his writing) "a ready-made bounded rationality theory that could frame formally the discussion of such central topics as markets, the firm, or uncertainty" (p.34). While my primary aim is to develop a framework for understanding knowledge bounding and satisficing in data reuse, my conception of satisficing may contribute to a theory of bounded rationality that spans both procedural and substantive considerations.

### *1.4.8 Simon In Conversation With Haraway*

Haraway's and Simon's ideas intersect in my research because Haraway's theory of situated knowledges can deepen our understanding of the process of adaptation researchers undertake when deciding whether to reuse data. We can use this understanding, in turn, to enhance the design of data "artifacts" that are preserved and made available from data archives. Conversely, Simon's articulation of the sciences of the artificial can help us understand factors that are important to enabling and sustaining the "conversations" Haraway envisions in the context of design. I discuss each of these intersections in turn below.

In his description of sciences of the artificial, Simon (1994) acknowledges that system ("artifact") design is a two-way street. Designers make systems, but users are also designers, and determine how best to use (or design) the systems available to them to meet their own goals. However, the study of how users adapt to designs is not a core element of the curriculum Simon specifies for studying system design. Ackerman (2000) points this out most clearly in his critique that Simon's conception of the sciences of the artificial does not account for the "social-technical gap," or the gap between "what we know we must support socially and what we can support technically" (Ackerman, 2000, p.179).

Ackerman (2000) demonstrates in his paper that technical solutions will likely always fall short of social needs for Computer Supported Cooperative Work (CSCW). He uses an example of the ways CSCW falls short of supporting user needs to address problems in online privacy as an example, writing:

> To summarize, there are no current HCI mechanisms to straightforwardly mechanize the naturally occurring, everyday social activity of handling personal information in its entirety. We must necessarily restrict the problem from what we know is appropriate to the social circumstances. This is the social–technical gap. (Ackerman, 2000, pp.186-187)

Ackerman argues that the way to address this reality is to make the social-technical gap a specific area of study in his field. He proposes CSCW as an example of a "science of the artificial" described by Simon, but an example that is updated to address this gap:

> CSCW is inherently a science of the artificial, as Simon (1969/1981) meant the term: CSCW is at once an engineering discipline attempting to construct suitable systems for groups, organizations, and other collectivities, and at the same time, CSCW is a social science attempting to understand the basis for that construction in the social world (or everyday experience).
>
> CSCW's science, however, must centralize the necessary gap between what we would prefer to construct and what we can construct. (Ackerman, 2000, pp.193-194)

I propose that the challenges of assembling and archiving data to satisfy the needs of a broad range of researchers in the way stakeholders imagine is a challenge similar to the one

Ackerman describes surrounding online privacy. The difficulty of fully representing in a package of deposited data all that occurred in, or all that potential data reusers might like to know about the original research is so great that we will likely never accomplish it. This does not mean, however, that we cannot learn what information is most important to data reusers, uncover areas where they were not able to obtain all the knowledge about data they desired, or discover reasons researchers believe they were not able to obtain all the knowledge they desired. Rather, as Ackerman argues in the field of CSCW, we must make this gap, and how researchers negotiate it—how, in the course of data reuse, data reusers adapt themselves to data—an area of study.

Haraway's theory of situated knowledges informs this study because it provides a framework for understanding boundaries of data as contingent on social factors and situated knowledges. This framework allows us to consider that different researchers set different thresholds on the knowledge they obtain about data when adapting to reach reuse equilibrium, and do so for different reasons. This consideration, in turn, leads to the premise of my research, which is that it is possible to study the ways researchers configure themselves to reuse data by examining where they set these thresholds and why.

Haraway's theory additionally expands our conception of the range of factors that may influence researchers in this configuration, especially when the configuration is viewed in the context of knowledge satisficing. The premise of situated knowledges is that we are not separate from our bodies or our environments. These factors shape the perspectives we have—in the context of science, they shape our perspectives on objects of knowledge—and we must account for them if we are to avoid the "god-trick" of unlocated claims on reality and objectivity. The theory of situated knowledges suggests that if I am to use satisficing as a mechanism for understanding the ways that researchers bound knowledge about reused data, the concept of

satisficing must be expanded. It must allow for the possibility that factors other than those inherent to Simon's definition of an inner environment (specifically, cognitive constraints) could affect an individual's decision to accept a satisfactory rather than an optimal solution (put another way, it must be flexible to admit limitations attributed to both procedural and substantive rationality). It must also allow for the evaluation of differences between initial and terminal aspiration levels when initial aspirations are known.

At the same time that situated knowledges informs my understanding of adaptation and satisficing in the context of Simon's sciences of the artificial, his conception of the sciences of the artificial can inform the application of situated knowledges to the design of data archives. Haraway's vision of situated knowledges includes more than individually situated perspectives. It is through the interactions and conversation between these perspectives that rational knowledge is produced. She says, as quoted above:

> Rational knowledge is a process of ongoing critical interpretation among 'fields' of interpreters and decoders. Rational knowledge is power-sensitive conversation (King, 1987a) (Haraway, 1991, p.196)

This view is seen as well in the contestation and "webbed connections" she envisions:

> So, with many other feminists, I want to argue for a doctrine and practice of objectivity that privileges contestation, deconstruction, passionate construction, webbed connections, and hope for transformation of systems of knowledge and ways of seeing. (Haraway, 1991, p.194)

Haraway does not provide parameters for or guidance on how these conversations might proceed, however—guidance that I argue is particularly important in the context of design, and in the context of a science of the artificial). In the environment where adaptive designs are synthesized to meet specific goals, it is not always possible to maintain tensions between multiple perspectives or allow boundaries to materialize organically. In designing a data archive, for example, and the data it contains, there are very real limitations on time, resources, and

technology that tend to close rather than open discussions. Researchers only have so much time to prepare data. Repositories can only devote so much staff time to checking and validating datasets. Decisions about how to curate, preserve, and enhance datasets or repository capabilities for accessing and using datasets are limited by considerations of cost, regulatory restrictions, domain guidelines, and visions and plans of repository administrators, which include desires to meet perceived needs of the user community.

How, in such an environment, where discursive inertia is embedded in policies, practices, infrastructure, and funding models, can the conversations Haraway envisions be enabled to take place? One of the things Simon does in *The Sciences of the Artifical* (1994) is develop a curriculum for the science of design—elements that should be included in an agenda for studying design. There is a component of his curriculum for social design that speaks directly to the challenge of supporting open conversation in the context of design. This is the notion of "designing without final goals" (Simon, 1994, p.190). Simon states:

> The idea of final goals is inconsistent with our limited ability to foretell or determine the future. The real result of our actions is to establish initial conditions for the next succeeding stage of action. What we call "final" goals are in fact criteria for choosing the initial conditions that we will leave to our successors. (Simon, 1994, p.187)

Simon presents two desiderata when choosing these initial conditions, and one mechanism for achieving them. One desideratum is to create "a world offering as many alternatives as possible to future decision makers, avoiding irreversible commitments that they cannot undo" (Simon, 1994, p.187). The second is "to leave the next generation of decision makers with a better body of knowledge and a greater capacity for experience." He continues: "The aim here is to enable them not just to evaluate alternatives better but especially to experience the world in more and richer ways" (Simon, 1994, p.187). The mechanism is to be

guided in design by general but desirable heuristics. He cites his own discussion of search heuristics to note that

> search guided by only the most general heuristics of "interestingness" or novelty is a fully realizable activity. This kind of search, which provides the mechanism for scientific discovery, may also provide the most suitable model of the social design process. (Simon, 1994, p.186)

Could a heuristic such as "designing to enable conversation" be a principle around which to organize the design of data archives and data artifacts? If so, a qualification to Simon's desiderata could be to think of decisions not in terms of the next generation who will make them, but the next decisions regardless of who makes them. In today's world of rapid technological change, enabling conversation seems to necessitate more rapid and more inclusive decision-making practices.

A question remains of what designing to enable conversation might mean in the context of digital archives and data reuse. I discuss this more after presenting the findings from my study, which include findings about challenges that limit the knowledge researchers obtain about data and how they might be overcome.

### *1.4.9 Epistemology and Ontology*

Questions of epistemology and ontology are central to assumptions about data reuse and design that I seek to address in my research, as well as to my theoretical framework and methodology. An assumption that underlies data archiving, for example, is that data can accurately represent research that took place in the past. Hammersley (1997), Mauthner et al. (1998) and others have argued, however, that the utility of data reuse is limited due to the inadequacy of archived data for conveying all that is crucial for data reusers to know about the research context. These arguments are based on an epistemological understanding of the way data are created: namely, that they are inextricably embedded in their contexts of creation and

that in some kinds of research (e.g., research conducted in an interpretivist framework) this context is more embedded than others.

In a similar way, Simon's conception of design rests on the assumptions that it is possible to isolate an inner from an outer environment, and design solutions where inner environments are well-adapted to outer environments. However, an epistemological position (from Haraway) that sees researchers' perspectives as being situated in their bodies and experiences, and boundaries of data emerging out of the associations of which the data are a part, problematizes this view. It blurs distinctions between inner and outer environments and "data" and "context." This blurring opens inquiry into data reuse as an adaptive process in which researchers seek to obtain knowledge about data for reuse, and into the factors that affect that process. Ackerman's (2002) conception of the social-technical gap exposes another of Simon's epistemological assumptions: that it is possible, at least eventually, to design technical solutions that fully address social challenges. Haraway's theory is a key basis for my investigation of this social-technical gap in the context of data reuse.

As the previous paragraph shows, issues of epistemology are closely associated with issues of ontology. For instance, the idea of data reuse depending on appropriate "context" in archived data makes immediate distinctions between what are "data" and what are "context;" system and "artifact" design (in Simon's paradigm) involve immediate distinctions between inner and outer environments.

I did not design my study to understand whether it is feasible to reuse data or what constitute the "data" to be reused. However, my results speak to larger issues about epistemology and ontology in data reuse. This is because they are grounded in assumptions at the same level: an epistemology that understands data reuse as a process of adaptation and an ontology based on

resonance (between inner and outer, data and context) as opposed to dichotomy. Situated knowledges is ultimately about rational knowledge claims and the way we arrive at them (i.e., by being accountable for the positions we hold and through "power-sensitive conversation" (Haraway, 1991, p.196). It is my hope that my results will spur conversations about the ways we conceive of data reuse and data, and action to better support the construction of rational knowledge claims in data reuse.

## 1.5 Summary of Theoretical Framework and Methods

### *1.5.1 Summary of Theoretical Framework*

The foregoing ideas can now be put together to describe the particular focus of my research and research methods. To summarize, there are two main components to the theoretical framework I used to investigate researcher knowledge in data reuse. The first is an understanding of the process of a researcher deciding that data are sufficient to reuse in their research (i.e., of reaching reuse equilibrium) as a process of adaptation involving boundary formation around the data expected to serve as evidence of the behavior or phenomenon under investigation. The process of adaptation is important because it offers a way of studying the social-technical gap in data reuse (what I also refer to as the "distance" between data reusers and the data). The data serving as evidence (i.e., the "artifact") include any main bodies of "data" being reused, and all of the data and information researchers use to reach reuse equilibrium.

The second is the idea that researchers satisfice in light of integrated attributes of their inner and outer environments in their efforts to reach reuse equilibrium. This framework provides a means of empirically investigating the different levels of contextual information researchers use when reusing data. In my study in particular, it offers a way of understanding at what points researchers bound the knowledge they need to reuse data and why. My research was

premised on the notion that we may be able to better support researchers in their efforts to reuse data if we examine the design of the data "artifact" and/or the informational networks of which they are a part in light of a social-technical gap in data reuse, with the ultimate goal of facilitating greater access to desired knowledge.

### *1.5.2 Summary of Methods*

An important aspect of my research into knowledge bounding was to investigate whether, and if so, why, researchers satisficed in the knowledge they obtained to make decisions about reusing data. The quantitative phase of my study was focused on the first part of the question—whether researchers satisficed. In cases where they did, the survey investigated the conditions under which satisficing took place and the effect of satisficing on the outcomes of the research and researchers' attainment of their research goals. The qualitative phase was designed to understand the process of knowledge bounding regardless of whether satisficing was in evidence.

These two phases (the quantitative and qualitative) and the stages of integration between them constitute the mixed methods study. Both phases investigated samples of researchers who, between January 2014 and September 2019, produced journal publications or other works that cited data held in the Inter-university Consortium for Political and Social Science Research (ICPSR) data repository.

In the quantitative phase of my study, I surveyed 2,909 researchers in three strata to understand the extent, nature, and impact of knowledge satisficing in data reuse. The strata consisted of 1) a random sample of researchers who reused quantitative data; 2) a complete sample of researchers who reused qualitative data; and 3) random samples of researchers who reused one of several purposively sampled data studies. The results of the survey facilitated the selection of particular data studies and reuse instances about which I interviewed researchers in

the qualitative phase. In the third stratified sample, I selected several data studies with different levels of reuse ranging from high to low. For each of the selected data studies, I randomly selected researchers who had reused the study, so as to obtain responses from multiple researchers who had each reused the same data study. I constructed the samples using information I obtained about researchers and the data studies they reused from the ICPSR Bibliography of Data-related Publications (ICPSR, 2019b).

The idea of surveying multiple researchers who reused the same data study derived from Haraway's theory of situated knowledges. In particular, it was based on the reasoning that if all perspectives were partial, understanding the boundaries of an object—or in my case the knowledge researchers determined was necessary about the object—required investigating the object, or knowledge about it, from multiple perspectives. Understanding *why* boundaries were placed the way they were in this "boundary project" similarly required an investigation from multiple perspectives, which led to the design of the qualitative study.

In the qualitative phase, I used information about respondents and their research taken from the stratified samples in the quantitative phase to select researchers to interview in the qualitative phase. I selected researchers with the goals of maximizing the number of researchers who reused the same data and maximizing variation among the researchers in factors I identified from the literature (and asked researchers about the in survey) that affect the way researchers bound knowledge they obtain about data. I explain more about these factors in chapters 2 and 3.

The goals of the qualitative phase of research were to map, in diverse individual (per researcher) and collective (per group of researchers) circumstances, how and why researchers bounded the knowledge they obtained about data to reach reuse equilibrium. I explain the design of my study in more detail in Chapter 3.

In preparation for my research, in November 2019 I conducted a pilot of the survey I used in the quantitative phase. I reference results from this pilot several times in chapters 2 and 3. A description of the pilot survey is included in Appendix C.

### 1.6 Data, Information, and Knowledge

The terms data, information, and knowledge are used frequently throughout this dissertation and bear some explanation as their usage can vary. Ackoff (1989) proposed a hierarchy of data, information, knowledge and wisdom (DIKW). In this hierarchy, he defined data as "symbols that represent the properties of objects, events and their environments" (Ackoff, 1989, p.3), information as being "contained in descriptions" and "inferred from data" through processing of the data (Ackoff, 1989, p.3; emphasis in original), and knowledge as "know-how, for example, how a system works" and "what makes possible the transformation of information into instructions" (Ackoff, 1989, p. 4; emphasis in original). Ackoff notes that knowledge (as well as information and understanding) are focused on efficiency—they are based on logic that can be programmed and automated—in contrast to wisdom, which operates on the basis of judgement and adds value. Ackoff's model has been criticized for being narrow in its conceptions of data, information, and knowledge, and for not acknowledging the fluidity between the concepts (e.g., that data and information or information and knowledge are not always separate and distinct) (Frické, 2009, Weinberger, 2010).

Like others, I have a more fluid understanding of the terms' relationships to one other. In particular, for the purposes of my research, I understood the "data" a researcher used to comprise both data that provided evidence relevant to their subject of interest (what Lee (2011) refers to as the "target entity"), and data that allowed them to make determinations about those data (e.g., that the data are of quality, accurate, etc.). I considered the assessments and determinations

researchers made using the data, information, and knowledge they obtained about data, to ultimately constitute what researchers knew, both about their subject of interest and about the evidence and information they used to study their subject of interest. In this framing, where data, information about data, and knowledge were subjected to evaluation by a researcher in the context of a decision to reuse data, the key question in my research became how that knowledge was bounded, rather than which parts of the data, information, or knowledge counted as "data" or "context."

As will be seen in the literature review, many studies of data reuse refer to information about the data as "context" or "contextual information." Some studies investigate the "information" researchers obtain in order to reuse data (e.g., information relevant to determinations about quality) while others investigate what researchers desired or needed to "know" about data. Some studies refer to additional factors that are important to researcher decisions about data, which may include the attainment of information or knowledge. Because my research focused on any knowledge about data that researchers determined was necessary to reach reuse equilibrium, and because I considered data and context together to comprise the "data" about which researchers sought to obtain knowledge, I considered references in the literature to any of these elements—data, information, and knowledge—as relevant to my investigation of how researchers bound the knowledge they obtain when reaching reuse equilibrium.

I chose to study satisficing in relation to knowledge, specifically—as opposed, for instance, to information—for several reasons. First, knowledge is something that can be evaluated by asking respondents in a study (Fowler, 1995). While some types of knowledge (e.g., tacit knowledge) may be difficult for researchers to articulate (Kelder, 2005), the selection

of data to reuse is a conscious decision for which, as part of a sound research design, researchers must be prepared to account. It was this expressed knowledge, and the ways it is bounded, that I was interested in investigating in this study.

This knowledge stands in contrast to the specific information researchers might have used to obtain their knowledge (i.e., information about how to use a search engine or how to get in touch with a researcher), which may be more multi-faceted and difficult to recall. I believed measuring the knowledge that researchers sought and obtained would provide more reliable and comparable results than measuring the information they sought and obtained. Therefore, throughout the paper, while I refer to both information and knowledge about data when referencing relevant literature, the intent of the proposed study is to measure the knowledge about data that researchers sought, as opposed to the information.

Second, knowledge seemed to lend itself well to Simon's model for measuring satisficing (i.e., measurement in relation to aspirations or goals) because it was possible to ask researchers what they were able to know about data they selected for reuse in relation to what they would ideally have liked to know. For instance, a researcher might have ideally liked to know the temperature at which certain readings were taken, or the race or native language and class of a person who conducted an interview. It stood to reason that comparing the difference between the optimal and the acceptable would provide evidence of satisficing.

Finally, knowledge is based on information sources, and resources (e.g., time and effort) are involved in accessing and processing information sources or mobilizing or activating knowledge (e.g., activating knowledge held in memory or experience). In my study, I asked about the sources researchers used to obtain knowledge about data. I believed that identifying sources of knowledge and associating them with degrees of satisficing could lead to the

identification of areas where researchers were most in need of support when reusing data. This, in turn, could lead to strategies to better assist researchers in their reuse of data in the future. While I ultimately found that satisficing was not a useful frame through which to view researchers' behavior, my findings related to sources of knowledge and amounts of knowledge researchers lacked about data are nonetheless useful.

## 1.7 Significance and Impact of Research

### *1.7.1 Significance of Research*

The research I propose is significant because it addresses gaps in our knowledge about data reuse that are critical to fill if we are to maximize the reuse potential of data from past research. A growing body of literature describes the challenges researchers encounter in reusing previously collected data and the kinds of information about data that is necessary for reuse. While questions of how much of what kinds of information about data is "enough" for researchers to reuse them have been raised in the past (Mauthner, 2014; Irwin and Winterton, 2011b), my research into how and importantly why researchers bound the knowledge they obtain in order to reuse data is one of the first to focus on these questions specifically. It establishes an empirical basis of investigation by combining well-known theories of feminist objectivity and information science in a novel way. Empirical evidence of knowledge bounding and the reasons behind it are vital for stakeholders in data reuse to critically evaluate existing policies and practices and take meaningful steps to advance the sharing and archiving of research data.

My study is also significant because it was specifically designed to enable the transfer of research results into practice. The approach I took involved a survey of recent data reusers from a single large and well-known data repository (at ICPSR) and subsequent qualitative research on individuals in the same population. The quantitative survey characterizes satisficing behavior and

gathers information about associated attributes of researchers and the research they conducted at a broad level. The qualitative interviews and background research yield details about reuse of specific data studies from multiple different perspectives (e.g., from the perspective of researchers with a variety of different goals, levels of experience, and levels of knowledge about the original data). The integration of results from the two phases of the mixed methods study provide actionable information about users of the ICPSR repository and their experiences reusing data that has implications for how reuse of data in the repository can be supported. The information spans from the very minute (e.g., how a particular type of knowledge desired by a researcher was limited) to the very general (e.g., how common it is for ICPSR data reusers to experience such limitations and under what circumstances).

This combination of specific and generalized information about a single population can be difficult to glean from existing data reuse literature because of differences in the scales, populations, and specific topics that studies investigate (this is discussed further in section 2.4.6). It is information, however, that is essential to strategic decision-making about investments of time and resources to enable the broadest and most efficient sharing, stewardship, and reuse of data for scholarly research. Being strategic and having a strong empirical basis for decision-making are especially important given the increasing amounts of scientific data that are being produced and the substantial time and effort required both to prepare data for reuse by others, and to reuse them (e.g., Niu, 2009; Warwick et al., 2009; Borgman et al., 2007; Faniel and Jacobson, 2010).

My research also makes contributions to theory through an integration of theories of situated knowledges, sciences of the artificial, bounded rationality, and satisficing. The question of how researchers determine how much knowledge about data is "enough" is underexplored in

the data reuse literature. However, investigating the question can help us better understand the social-technical gap in data reuse, whether knowledge satisficing is an appropriate way to characterize this gap (and if not, what is), and how to think about situated knowledges and the production of rational knowledge in the context of design (i.e., how we maintain and integrate contributions from multiple perspectives in the process of designing data "artifacts" and data archives).

### *1.7.2 Impact of Research*

I expect my research to have several impacts following from what I have discussed above. First and foremost, my research may spur stakeholders including universities, research funders, data stewards, government agencies, data creators, and data reusers to work collectively to support the attainment of knowledge about data that researchers desire by (a) facilitating access to knowledge or overcoming barriers to knowledge access and (b) facilitating conversations between data creators and data reusers about what data can be collected to maximize the impact and value of reuse. This kind of support for data reuse could have several downstream impacts, including:

- increasing the prestige ascribed to conducting secondary research and to sharing data that facilitate valuable secondary research
- increasing incentives to create and share data that are well-curated for secondary reuse
- helping researchers more easily find data that meets their research needs, since greater access to knowledge about data can aid researchers in determining more efficiently whether research data will support their reuse needs;
- helping data creators better define and explain terms of consent, as findings about how much of what kinds of knowledge about data are important to subsequent researchers can help current researchers characterize to their study participants which data are likely to be of interest;
- encouraging researchers and other stakeholders to think critically about research conducted with reused data, by highlighting what kinds of knowledge about data are lacking and the ways limits to specific kinds of knowledge could systematically limits the results of research; and

- helping policymakers create more targeted policies for data sharing, based on empirical evidence from data reusers about the knowledge they desire. Such evidence could help reduce or better support the labor expended to document data;
- encouraging research on the cultures of data creation and reuse in academia, how cultures might differ in different communities, how these cultures impact data reuse, and other strategies to maximize the value and impact of research with secondary data;
- through all of these, helping to bring the vision of making new discoveries and creating new knowledge with secondary data closer to a reality.

In this introduction I have outlined the purpose of my study and research motivations, the core problems and gaps in the literature I seek to address, a theoretical framework for my investigation, the methods I intend to employ, and the significance of my study. I turn now to a review of relevant literature about the knowledge researchers obtain to determine the sufficiency of data from prior research, and the different dimensions of data reuse that must be accounted and controlled for in my proposed research.

## Chapter 2 Literature Review

In the first chapter I described assertions that greater sharing of data from academic research will result in new discoveries and greater trust in science. I also described concerns that attend the application of generalized policies about data sharing, and ways that a conceptual separation of "data" and "context" limits the ability to conduct more specific and nuanced investigation into information that could made available to enable the broadest possible reuse of data for research. I described my theoretical framework in which data and context were considered together and outlined my project to empirically investigate how and why researchers bound the knowledge they obtain about data when determining whether to reuse data in their research. Whether, and if so to what degree, researchers satisficed in the knowledge they obtained about data, what factors influenced satisficing, and how satisficing affected the outcomes of research were important questions in this investigation.

In this chapter, I review literature that has examined what researchers seek to know about the data they select for reuse and what influences their decisions. I discuss gaps in this research and identify key influences and factors from the literature that I proposed were significant to characterizing knowledge bounding and satisficing in data reuse. These include factors such as a researchers' goals, their "distance" from the data they reuse, their sources of knowledge about data, and their experience with data reuse.

The literature review is organized into nine sections. In the first section (2.1 Introduction) I give an overview of the literature on data reuse and the portion of the literature I focus on in particular. In the second (2.2 Defining Data Reuse) I review definitions of data reuse and

differences researchers have identified between primary research and research involving data reuse, including the importance of a researcher's "distance" from the data in these discussions. In the third (2.3 Purposes of Data Reuse) I discuss six main purposes for reusing data that are prevalent in the data reuse literature. In the fourth (2.4 Factors In Data Reuse Decisions) I describe factors researchers have found to be important to data reusers' decisions to reuse data. In the fifth (2.5 The Importance of "Context") I discuss literature more specifically related to the contextual information that is important to researchers when deciding to reuse data.

In sections 2.2 to 2.5 I first identify and explicate the main concepts, debates, and issues relevant to research on knowledge satisficing in data reuse. I then include a "Discussion" section in which I identify limitations in the literature where further research can substantially improve our understanding of data reuse and knowledge about how to best support researchers who desire to reuse data from prior research. My goals through this explication of terms and limitations are to illustrate the scope and significance of gaps relating to knowledge satisficing in data reuse, and to identify parameters and considerations discussed in the literature that were relevant to the design of my study.

In section 2.6, I review research in information science that has been conducted (explicitly or implicitly) within the frameworks of bounded rationality and/or satisficing. In section 2.7, I give a summary of the gaps in the literature I identified in the previous sections and in section 2.8, I describe the connections between the gaps I have identified in the literature, my theoretical framework, and my research questions. In the final section (2.9) I explain, based on the literature and prior to a fuller discussion of methodology, the assumptions and hypotheses that underlie the quantitative and qualitative phases of my study.

## 2.1 Introduction

### *2.1.1 Overview*

Literature related to the knowledge researchers obtain to determine the sufficiency of data for reuse in academic research spans a variety of research domains and methodologies. Domains of research investigated for this dissertation include astronomy, ecology, anthropology, zoology, HIV/AIDS research, space physics, neuroscience, archaeology, proteomics, engineering, epidemiology and social science (e.g., Borgman et al., 2007a; Borgman et al., 2007b; Wallis et al., 2013; Wynholds et al., 2013; Carlson and Anderson, 2007; Faniel and Yakel, 2017; Birnholtz and Bietz, 2003; Wan and Pavlidis, 2007; Faniel et al., 2013; Huggett, 2018; Fear and Donaldson, 2012; Faniel and Jacobsen, 2010; Trevelyan, 2016; Roland and Lee, 2013; Heaton, 2004; Faniel et al., 2012; Medjedović, 2011; Jackson et al., 2007; Yoon, 2016; Niu, 2009a; Niu, 2009b; Bishop, 2006; Bishop, 2007; Corti, 2005; Coltart et al., 2013; Thompson, 2000).

Studies employ a variety of methods including interviews (e.g., Borgman et al., 2007a; Curty and Qin, 2014; Faniel et al., 2012; Faniel and Jacobson, 2010; Faniel et al., 2013; Faniel and Yakel, 2017; Fear and Donaldson, 2012; Frank et al., 2017), focus groups (Broom et al., 2009; Donaldson et al., 2017; Redman-Maclaren et al., 2014; Yardley et al., 2014), surveys (Enke et al., 2012; Faniel et al., 2016; Faniel, 2009), data re-analysis (Curty et al., 2017; Fielding and Fielding, 2000), bibliometric analysis (Bishop and Kuula-Luumi, 2017), syntheses of literature (Corti, 2005; Dicks et al., 2006; Doolan and Froelicher, 2009; Kelton, 2008; Sherif, 2018; Straughn-Navaro, 2016; Huggett, 2018; Coltart et al., 2013, Gregory et al., 2019), and combinations of these. The literature includes self-studies and analyses of specific instances of reuse (Bishop, 2007; O'Connor and Goodwin, 2010; Thompson, 2000) as well as broader studies of data reusers (Enke et al., 2012; Curty et al., 2017; Kim and Yoon, 2017, Gregory et al., 2020) and theoretical discussions that draw on concrete reuse examples (e.g., Dicks et al., 2006;

Wästerfors et al. 2014). There are clear imbalances in the literature (noted as well by Gregory, 2021) with substantial work done in some fields such as the social sciences, astronomy, ecology, and archaeology, and much less in other fields such as proteomics, neuroscience, and epidemiology.

There are three overarching categories that much of the literature falls into. The first is studies geared toward facilitating data access and reuse (e.g., Borgman et al., 2006; Borgman et al., 2007a; Broom et al., 2009; Curty, 2016; Curty et al., 2017; Enke et al., 2012; Faniel and Yakel, 2017; Frank et al., 2017; Faniel et al., 2013; Faniel et al., 2016; Kim and Yoon, 2017; McKay, 2017; Medjedović, 2011; Medjediovic and Witzel, 2005; Rolland and Lee, 2013; Wallis et al., 2013; Atici et al., 2013; Dicks et al., 2006; Niu, 2009a, Gregory et al., 2020). These studies focus on many different aspects of reuse (e.g., infrastructure needs; user support; issues of context, documentation, and what information should be archived; factors and challenges involved in data reuse; and assessments of data credibility, quality, trust and other dimensions of data). However, findings and conclusions are oriented towards improving data archiving or the support available to researchers for data reuse. A large number of these works are based on empirical research, though some are the result of reviews and syntheses of literature.

The second category comprises works that contain debates about, reflections on, or guidelines for how data reuse is or should be conducted (e.g., Mauthner et al., 1998; Moore, 2006; Moore, 2007; Savage, 2005; Parezo, 1996; Doolan & Froelicher., 2009; Wästerfors et al. 2014; Parry and Mauthner, 2005; Vogt, 2008; Walters, 2009; Coltart et al., 2013; Heaton, 2004; Fielding and Fielding, 2000; Irwin, 2013; Sherif, 2018; Church, 2001). These works investigate and describe limitations on data reuse, differences between primary and secondary research, and particularities of reusing quantitative or qualitative data. Some also self-consciously demonstrate

the value and utility of data reuse (Thompson, 2000, Savage, 2005; Medjedović and Wetzel, 2005). Some works in this group are based on empirical research (often self-studies of reuse) and others are more theoretical.

A third category includes assessments of reuse, such as how much reuse is taking place, by whom and for what purposes, and how reuse is facilitated (Bishop and Kuula-Luumi, 2017; Corti, 2005; Wan and Pavlidis, 2007; Rung and Brazma, 2012; Smioski, 2010; Medjedović, 2011; Piwowar, Carlson, and Vision, 2010; Piwowar, and Vision, 2013; He and Nahar, 2016; Pienta et al., 2010; Chao, 2011; Peters et al., 2016; Tenopir et al., 2015). Works in this category tend to look broadly across resources, measurement statistics, and behaviors to understand the wider landscape of reuse, both within and across different research domains.

This dissertation reviews literature primarily in the first two categories, though some works in the third category contain information that is relevant to the study of the knowledge researchers obtain about data in order to reuse them. This literature is particularly strong in identifying challenges to and enablers of data reuse, as well as factors that affect researchers' decisions and their abilities to reuse data (e.g., Curty, 2016; Anderson and Sørvik, 2013; Gilles and Edwards, 2005; Jackson et al., 2007; Irwin and Winterton, 2011b; Medjedović, 2011; O'Connor and Goodwin, 2010; Sherif, 2018; Thompson and Holland, 2003; Yoon, 2016; Kim and Yoon, 2017; Rung and Brazma, 2012; Donaldson et al., 2017). It is also strong in the kinds of evaluations of data that data reusers make (e.g., Faniel and Jacobsen, 2010; Fear and Donaldson, 2012; Frank et al., 2017; Donaldson and Conway, 2012; Stvilia et al., 2015; McKay, 2014; Stanley, 2013; Yoon, 2017a; Sherif, 2018; Zimmerman, 2008) and the types of context that are important to researchers (e.g., Carlson and Anderson, 2007; Temple et al., 2006; Enke et al., 2012; Wallis et al., 2007; Wallis et al., 2013; Wynholds et al., 2012; Faniel and Jacobsen,

2010; Faniel et al., 2013; Broom et al., 2009; Warwick et al., 2009; Corti, 2005). The literature also has significant weakness in several areas, which are discussed further below.

## 2.2 Defining Data Reuse

There are a variety of ways researchers define and understand data reuse, also referred to in different contexts as secondary data analysis, secondary use, replication, reanalysis, third-party use, repurposing, integration, and a variety of other ways. As definitions are not consistent across the terms used to describe it, I will refer to the phenomenon primarily as "data reuse," using other terms (such as "secondary analysis" or "secondary research") as appropriate.

### *2.2.1 Definitions*

Most researchers agree as a baseline in their definitions that data reuse involves the use of data created in prior research (Kwek and Kogut, 2015, Doolan and Froelicher, 2009; Heaton, 2008; Rolland and Lee, 2013). Many researchers qualify this by specifying that in order to count as data reuse, the data involved must be used for a purpose different from the purpose for which they were originally created (Bishop and Kuula-Luumi, 2017; Carmichael, 2017; Curty and Qin, 2014; Faniel et al., 2016; Faniel and Jacobson, 2010; Heaton, 2004, cited in Sherif, 2018; UK Data Service, n.d.; Zimmerman, 2008; Yoon, 2017a; Fielding and Fielding, 2000; Notz, 2005). Some restrict reuse further to use of data for a purpose not even conceived of in the original study (Yardley et al., 2014). Others allow that data reuse can involve analysis for either the same or different purposes (Wynholds et al., 2012). Still others note the importance of researchers at least employing new or better techniques of analysis than those used in the original research (Curty and Qin, 2014; Glass, 1976, cited in Sherif, 2018).

Another qualification some make to the baseline criterion is that reuse must be conducted by researchers different from those who carried out the original research (Niu, 2009a; Church,

2002, Curty et al., 2017; Huggett, 2018; Medjedović, 2011). At the same time, some admit research by original researchers as reuse if a concept was present in the original investigation but not analyzed, or if the reuse involved a change in analytical focus (Trevelyan, 2016). Other definitions of data reuse require use both by different authors and for different purposes (Pine et al., 2016), considerations of the source of the data (e.g., whether the data have been deposited in an archive prior to reuse) (Huggett, 2018; Wästerfors et al., 2014), and whether the data are gleaned from published articles or are the "raw" data resulting from the original research. Church (2002) includes use of both data from published articles and original data in data reuse, while Heaton (2008) excludes meta-analyses and systematic reviews as categories of reuse because they do not involve original data.

Mauthner et al. (1998) illustrate one basis for such differing views of data reuse in their description of how, after sufficient time has passed, original data can seem almost "foreign" to the original researchers. This happens as memories fade and involvement with the original research fade, especially if care was not taken in documenting or processing the data (see also Dicks et al., 2006 and Atkinson, 1992, cited in Wästerfors, 2014). In such cases, they argue, issues and challenges similar to those involved in primary analysis can arise.

These kinds of blurred lines between use of data in primary research and reuse in secondary analysis are echoed by other researchers as well. For instance, Heaton (2004) differentiates secondary analysis from documentary and conversational analysis on the basis of whether the analysis involved the use of non-naturalistic data (secondary analysis) or naturalistic data (documentary and conversational analysis) (Heaton, 2004; Coltart et al., 2013). For Heaton, non-naturalistic data are data that are intentionally "produced" as part of previous research, such as interview recordings, transcripts, field notes, and observational records. Naturalistic data are

data that are "found" such as diaries, essays, photographs, or film. At the same time, Heaton acknowledges that the distinctions are not always clear and that some non-naturalistic data (such as a life story or diary recorded for research) may later be "found" and used in primary analysis.

Other studies describe the similarities between secondary analysis and the "primary" analysis that is done in cross-language research (i.e., when interpreters or translators are involved in the research) (Temple et al., 2009), collaborative research (Andersson and Sørvik, 2013; Ribes and Jackson, 2013; Mauthner and Doucet, 2008; Carlson and Anderson, 2007; Coltart et al., 2013; Trevelyan, 2016; Bishop, 2007; Carmichael, 2017), and qualitative longitudinal research (Thomson and Holland, 2003; O'Connor and Goodwin, 2010).

For the purposes of my research, I defined data reuse as the use of data created in prior research in subsequent research. I defined it this way regardless of who created the data or for what purpose. That is, I considered use of data created previously as reuse, regardless of whether the data were being reused by the researcher or researchers who created them, or for the same or a different purpose.

In the next section I address the specific issues involved in secondary analysis and the differences researchers have described between primary and secondary analysis.

### *2.2.2 Difference Between Primary and Secondary Research*

#### *2.2.2.1 "Distance"*

The difference between primary and secondary research that is most commonly described across the literature has to do with the greater "distance" those reusing data have from the original research, or the greater "proximity" of the original researchers to the underlying data. Kwek and Kogut (2015) point out, for example, with regard to qualitative research, the "special relationship" that can exist "between researcher and the research participant, or between

researcher and research data, given the personal nature of data production for many researchers" (Kwek and Kogut, 2015, p.17). This relationship is asserted by numerous other researchers as well (Bishop 2007, 2009; Coltart et al., 2013; Katsanidou et al., 2016; Irwin and Winterton, 2011b; Hammersley 1997; Jackson et al., 2007; Mason, 2007, cited in Jackson et al., 2007; Irwin, 2013, and Wästerfors et al. 2014; Pasquetto et al., 2019; Koesten et al., 2021).

#### *2.2.2.2 "Distance" in Quantitative and Qualitative Data Reuse*

Some researchers believe this special relationship has more significant implications for the reuse of data gathered through qualitative than quantitative research. For instance, in focus groups with qualitative researchers oriented toward the establishment of a qualitative data archive, researchers expressed their belief that archiving quantitative data was less problematic than qualitative data "because of the depersonalized, abstracted, and ultimately transportable character of the data" (Broom et al., 2009, p.1168). In addition, in a seminal 1998 article, Mauthner et al. described the problems, when reusing qualitative data, of not adequately accounting for the deep entanglement between researchers and research subjects in the construction or "making" of data. This includes, for instance, the ways that researchers' attitudes and beliefs can influence the questions they ask in interviews and the ways responses are interpreted. They noted, however, that an important aspect of quantitative research in contrast to qualitative research is that in quantitative research, "the methodological paradigm which informs collation and analysis of statistics is one which seeks to minimise [sic] the influence of the researcher in [the data collection] process" (Mauthner et al., 1998, p.743).

Niu's (2009a) findings on the reasons researchers seek information outside of what is documented or known to them when reusing a dataset support this view. She found that a higher percentage of qualitative researchers sought hard-to-document tacit knowledge than quantitative

researchers. Other research counters findings that the role of the researcher or research environment (social, historical, physical, etc.), and thus the "distance" present in data reuse, is greater or more important in qualitative research. This research provides evidence of the importance of information about the context of data creation to quantitative researchers and the barriers to reuse that a lack of such information presents (Carlson and Anderson, 2007; Ribes and Jackson, 2013; Rung and Brazma, 2012; Wan and Pavlidis, 2009; Yardley et al., 2014).

Savage's (2005) perspective provides a means of reconciling these opposing views. He argues that all research, whether based on qualitative or quantitative analysis represents an abstraction of the data it is based upon. A key difference between archived qualitative and quantitative sources in this view is that in quantitative sources, the level of abstraction can be so complete as to entirely cover over the details involved in collecting the data. In qualitative research, on the other hand, such details are difficult to obscure. The knowledge that such details exist but are not present in archived data is the basis of some objections to the reuse of archived qualitative data (Mauthner et al., 1998; Parry and Mauthner, 2005).

Thompson (2000), argues that qualitative studies should be conducted alongside large longitudinal quantitative studies because of what a combination of both kinds of data could do to facilitate reuse:

> While the danger in quantitative research has always been to impose meanings of social behaviour without the evidence which comes from listening sufficiently to informants, the failing of much qualitative work in the postmodern or narrative modes has been to make the interactive research process the centre of study in itself, and forget what can be learnt from the stories which are told. If we could convince both traditions of the value of re-use, and then move forward towards creating the kinds of linked data which would be of the greatest mutual value, I believe that we would release a powerful reinvigorating new force in social research. And that is my hope for the future. (p. 14)

This perspective aligns closely with the view I take in my research and motivates my interest in identifying knowledge about data that could be expressed to the greatest effect in the "linked data" Thompson mentions.

#### *2.2.2.3 The Importance of "Distance"*

Just as there is disagreement about the importance of "distance" from the original data in different reuse contexts, there is disagreement about the importance of "distance" in distinguishing primary from secondary research, and the implications of "distance" for the methodological approach researchers should take in reusing data. On one hand, some scholars see researchers' "not being there" when data were collected as engendering challenges that are particular to data reuse, such as having to contend with or account for deficiencies in documentation and a lack of contextual information (Medjedović, 2011; Bornat, 2005, cited in Parry and Mauthner, 2009; Irwin and Winterton 2011b; Coltart et al., 2013; Dicks et al., 2006; Jackson et al., 2007; Wästerfors et al. 2014; Pasquetto et al., 2019). Other scholars note that documentation and contextual information are just as essential to primary research (for instance, primary research that is conducted in teams or over time) as to research involving data reuse (e.g., Niu, 2009b; Medjedović, 2011). Still others contend that distinctions between primary and secondary research are false or overstated because all research mobilizes evidence of some kind to achieve its ends and all data used in research are constructed and contextualized in the context of the research process (e.g., through the coding of interviews and other forms of analysis) (Moore, 2007; Bishop, 2007; Irwin and Winterton, 2011b; Irwin, 2013; Gillies and Edwards, 2005; Heaton, 2000, cited in Gillies and Edwards, 2005; Fielding, 2004, cited in Broom et al., 2009). Seen in this way, issues of "distance" take second place to the importance of rigorous

methods and defensible research practices that justify the validity of the research in light of the research questions and goals.

#### *2.2.2.4 Other Similarities and Differences Between Primary and Secondary Research*

Researchers have described additional similarities between primary and secondary research with regard to the research process and ethics. In her reflexive self-study of data reuse, Bishop (2007) notes similarities in the processes of "defining questions, locating data, and sampling" (Bishop, 2007, section 11.1). Meanwhile, Jackson et al. (2007) and Moore (2010) encountered ethical issues similar to primary research having to do with the handling of sensitive and potentially damaging information about research subjects (even when they are deceased) uncovered during the research process.

Researchers have also pointed out ways aside from "distance" that secondary analysis differs from primary analysis. For instance, researchers conducting secondary analysis must consider aspects of the reused data's context of creation (e.g., the research question(s), purpose(s) of research, and methodology employed), and evaluate and justify the data's "fit" to the current research questions (Bishop, 2007; Gillies and Edwards, 2005; Kelder, 2005; Kwek and Kogut, 2015; Curty, 2016; Doolan and Froelicher, 2009; Moore, 2006; McKay, 2014). The scope of desired research may also be constrained when reusing "finite" secondary data (i.e., data from the past that it may not be possible to supplement) in ways that it is not when using data collected for purpose (Bishop, 2007; Doolan and Froelicher, 2009; McKay, 2014).

A third difference is the particular ethical issues that arise in the conduct of secondary research. For instance, how does one ensure informed consent and confidentiality in cases where researchers may not be able to contact research participants (Bishop, 2007; Kwek and Kogut, 2015; Curty, 2016; Gillies and Edwards, 2005)? How should consent that was obtained pre-

Internet, when expectations of public sharing were more limited, be interpreted in today's networked environment? Should researchers involved in the original collection of reused data be granted special authority in interpreting research results (Jackson et al., 2007)? Researchers have also noted logistical challenges resulting from requirements that reused data be kept anonymous (Kelder, 2005; Gillies and Edwards, 2005; Medjedović, 2011) such as contending with anonymization strategies that involved the falsification of information (Medjedović, 2011). In the midst of these challenges, there is also a positive difference between primary and secondary research, which is that reusing data can enable research of human behavior that may be difficult due to ethical concerns in primary research (McKay, 2014).

A fourth difference is that though there are exceptions (Henderson et al., 2006; O'Connor and Goodwin, 2010), in most cases the subjects of secondary analysis do not participate in analysis and interpretation of results (Bishop, 2009).

The differences between primary and secondary research are indicative of possible types of knowledge that might be needed and sought out by secondary researchers. As such, they represent possible types of knowledge that may need to be facilitated by stakeholders interested in enabling reuse of data.

### *2.2.3 Discussion*

Many of the debates about "distance" and differences between primary and secondary research have taken place in relation to the reuse of qualitative social science data. They are most prevalent in arguments based on reviews of literature or an author's experience supporting or conducting reanalysis of research data (Dicks et al., 2006; Moore, 2007; Irwin and Winterton, 2011a; Parry and Mauthner, 2005; Kwek and Kogut, 2015; Coltart et al., 2013; Fielding, 2004). They are also prevalent in case studies, many of which are self-studies of reuse (Bishop, 2007;

Irwin, 2013; Irwin and Winterton, 2011b; McKay, 2014; Gillies and Edwards, 2005; Jackson et al., 2007; Thompson, 2000).

Large survey studies of qualitative (Medjedović, 2011; Heaton, 2004) and quantitative (Niu, 2009b) data reusers in the social sciences and in other disciplines (Curty et al., 2017) confirm findings of smaller studies that "distance" is an issue and that researchers struggle with inadequate documentation in secondary research. However, these larger studies do not investigate the epistemological positions of the data reusers in the studies. Debates about whether reanalysis involves bridging contextual "distance" or constructing new conceptualizations or abstractions of data in the context of the new research are thus conducted with a relatively narrow base of empirical evidence. While some researchers have taken positions on how data reuse should be conceived of or conducted, a more relevant question is how researchers are in fact conducting research (Carmichael, 2017) and more particularly, how they are conceiving of and negotiating the distance between themselves and the data they reuse. On this topic there has been little investigation.

Evidence suggests that data reuse is well-established in quantitative social sciences (Kwek and Kogut, 2015; Faniel and Yakel, 2017) and more and more researchers are reusing qualitative data (Bishop and Kuula-Luumi, 2017). Thus, however it is conceived, reuse of both qualitative and quantitative data in academic research is happening (there is ample evidence of the reuse of quantitative data outside the social sciences as well—e.g., Enke et al., 2012; Rung and Brazma, 2012). Researchers across numerous disciplines are using data from prior studies as evidence of phenomena or behaviors and attempting to make credible arguments using findings based on those data. This is happening in ways that are in many respects similar to the ways primary researchers assemble evidence and make arguments.

Given the reuse that is taking place and the lack of relevant research there is an important need for empirical investigation into how researchers negotiate the gap or "distance" between themselves and the original data: in other words, how they reach reuse equilibrium. My analysis indicates that one aspect of reuse equilibrium that lends itself to investigation and may be particularly valuable to explore is how much of which kinds of knowledge about reused data researchers determine is necessary to make credible arguments (i.e., to reach reuse equilibrium). This research, including investigation into the factors that affect the amounts and kinds of knowledge researchers determine is necessary, can be used to facilitate reuse of data in research (e.g., by facilitating access and reducing barriers to knowledge).

### 2.3 Purposes of Data Reuse

Researchers have identified numerous ways that researchers reuse data; most apply to both qualitative and quantitative data. One common way data are reused is for background or contextual purposes: to provide information that supports research question definition, helps to identify important topics and themes, aids in calibration of instruments, or provides background knowledge that is used to understand or analyze data (Notz, 2005; Heaton, 2008; Wallis et al., 2013; Wynholds et al., 2012; Bishop and Kuula-Luumi, 2017; Thompson, 2000, Gregory et al., 2020). A second way is to answer new research questions or questions that are modified from the original (Heaton, 2008; Fielding and Fielding, 2000; Wynholds et al., 2012; Sherif, 2018; Wästerfors et al., 2014; Savage, 2005, Gregory et al., 2020). These include questions that focus on the investigation of phenomena and behaviors, and on the methodologies of investigation themselves (Rung and Brazma, 2012; Heaton, 2008; Sherif, 2018; Wästerfors et al., 2014; Irwin and Winterton 2011b).

Third, researchers also perform secondary analysis for verification and validation purposes. In these instances, data are reused to verify, corroborate, or refute findings or interpretations from original studies (Rung and Brazma, 2012; Heaton, 2008; Wallis et al., 2013; Walters, 2009; Corti and Bishop, 2005; van den Berg, 2005; Faniel, 2009; Savage, 2005). Data are also reused to validate or support theoretical models (Sherif, 2018; Faniel et al., 2012; Donaldson et al., 2017; Faniel and Jacobson, 2010). One of the main differences between reuse purposes for quantitative and qualitative data is that reuse of qualitative data for verification purposes is more controversial (Heaton, 2008). Some researchers view verification of qualitative research as fraught due to the complexity of this research and the difficulty of fully understanding the original research context (Heaton, 2008; Corti and Bishop, 2005; Fielding and Fielding, 2004; Hammersley, 1997).

Fourth, researchers reuse data in a variety of combinatory and comparative ways (Pasquetto et al., 2019, Gregory et al., 2020). For instance, researchers gather data from multiple studies to enlarge their sample size or analyze the results of multiple studies all together (e.g. meta-analysis) (Trevelyan, 2016; Heaton, 2008; Rung and Brazma, 2012; Wan and Pavlidis, 2007; Bishop and Kuula-Luumi, 2017; Mauthner et al., 1998). Researchers additionally compare results from multiple studies, as in the case of re-studies or follow-up studies (Wynholds et al., 2012; Corti and Bishop, 2005; Wästerfors et al., 2014). Comparisons are especially important for studies that investigate social change (Walters, 2009; Gillies and Edwards, 2005; Faniel et al., 2017; O'Connor and Goodwin, 2010; Corti and Bishop, 2005; Henderson et al., 2006).

Fifth, data are reused in efforts to develop and test algorithms and tools. Wan and Pavlidis (2007) describe, for example, how researchers might compare different algorithms for gene expression analysis using data from past research. Rung and Brazma (2012) similarly

describe the reuse of data to develop new or more efficient methods of microarray data analysis. In her survey of earthquake engineers, Faniel (2009) found that more than half of respondents indicated they were most likely to reuse someone else's data to "develop and validate computational models or tools" (p.35).

Gregory et al., 2020, also identified additional uses of data for purposes that may or may not be associated with reuse for publication purposes. These include use to experiment with new methodologies and data analysis techniques, or identifying trends and or creating summaries of data.

### *2.3.1 Discussion*

The knowledge we have about the purposes of reuse comes from a variety of disciplines. Those reported on here include the social sciences (Notz, 2005; Heaton, 2008; Thompson, 2000; Fielding and Fielding, 2000, Sherif, 2018, Wästerfors et al., 2014; Savage, 2005; Corti and Bishop, 2005, van den Berg, 2005; Walters, 2009; Bishop and Kuula-Luumi, 2017; Mauthner et al., 1998; Faniel and Yakel, 2017; Gillies and Edwards, 2005; O'Connor and Goodwin, 2010; Henderson et al., 2006), ecology (Wallis et al., 2013; Wynholds et al., 2012), neutron science (Donaldson et al., 2017), earthquake engineering (Faniel, 2009, Faniel and Jacobson, 2010), neuroscience (Wan and Pavlidis, 2007), archaeology (Faniel and Yakel, 2017), zoology (Faniel and Yakel, 2017), and studies in the biological sciences involving gene expression data (Rung and Brazma, 2012).

The literature indicates that there are variety of general purposes for which researchers reuse data, and that each of the purposes is evidenced across a diversity of disciplines. The literature rarely explores, however, which general purposes are more prevalent within or across different disciplines. Gregory et al., 2020 surveyed purposes of reuse across multiple disciplinary

domains and identified particular uses that were most significant in each domain. However, they did not analyze purposes of reuse within particular domains. Gregory et al., 2018 identified reuses purposes associated with particular disciplinary domains, but their study was a review of research on reuse rather than an empirical study and they did not examine associated frequencies within or across disciplines. Other studies have examined purposes of reuse on smaller scales, but these sometimes produce conflicting results. For example, Wallis et al. (2013) found in their study of ecology researchers at the Center for Embedded Network Sensing that researchers only reused data produced by others for "background" purposes. In a different study focusing on the knowledge ecology researchers use to locate and reuse data, all of the examined reuse cases were ones where data were being used to investigate new research questions (Zimmerman, 2007). These and other studies provide windows into data reuse in ecology and other fields, but a broader view of the purposes of reuse in the reuse landscape is lacking.

Moreover, existing research has not sufficiently explored similarities and differences in the amounts and types of knowledge about data researchers need to determine the sufficiency of data that is reused for different purposes—for instance, when data are used as background for their research, for validation purposes, when combining multiple sets of data to increase sample size, or reusing data to answer new research questions. While many studies discuss the contextual knowledge researchers need to reuse data (e.g., Faniel, Frank, and Yakel, 2019; Niu, 2009a; Curty et al., 2017; Carlson and Anderson, 2007; Borgman et al., 2007b, Gregory et al., 2020), only two studies that I reviewed did so explicitly across multiple instances of reuse in light of a specific research purpose. The first was a study by Faniel and Jacobsen (2010) of the contextual information earthquake engineers use when validating theoretical models. The second was an investigation of 41 studies in health and social care that reused qualitative data to

investigate new questions or verify previous studies (Heaton, 2004). While Heaton developed a typology of reuse instances, however, she did not explore the knowledge researchers needed to reuse data or how they negotiated the "distance" from the original research.

Our understanding of data reuse could benefit not only from research that explored the amounts and kinds of knowledge researchers need to achieve reuse equilibrium, but also the knowledge needed to reach reuse equilibrium for different purposes of research. Such research could begin to uncover similarities or differences in the types and amounts of knowledge that repository managers and other stakeholders need to obtain or link to in order to support different kinds of reuse (e.g., validation or the ability to combine data with other data). This information could be used to guide policies about the contextual information researchers need to deposit along with data.

## 2.4 Factors in Data Reuse Decisions

There are a wide variety of factors that researchers have found are at play in researcher decisions to reuse data in academic research. These factors can be separated into groups based on whether they relate to perceptions and beliefs researchers have, assessments researchers make about data, resources available to researchers, and social and other aspects of the general environment in which researchers conduct their research.

### *2.4.1 Perceptions and Beliefs*

Numerous researchers discuss benefits of or concerns about data reuse in a general way or as lessons learned from research (Church, 2002; Gillies and Edwards, 2005; Rung and Brazma, 2012; Savage, 2005; Travers, 2009; Wästerfors et al. 2014; Wallis et al., 2007). However, fewer consider the impact of beliefs about benefits and risks of data reuse on reuse behavior. Information about perceptions and beliefs in this review is drawn primarily from three

empirical studies: one based on interviews with registered users of social science data archives in the United States (Curty and Qin, 2014; Curty, 2016), and two large surveys of researchers in multiple disciplines (Curty et al., 2017; Kim and Yoon, 2017). Findings from these studies indicate that some of the perceptions and beliefs that impact researcher decisions to reuse data are:

1. perceived benefits of reusing data (e.g., beliefs in the potential to make new discoveries, the scope and richness of data, the reliability or credibility of data, and that reusing data will require less time or effort than primary research) (Curty and Qin, 2014)
2. perceived risks in reusing data (e.g., fears of receiving less credit for work involving reused data, lack of clarity about ethical and intellectual property issues surrounding reuse, concerns about misusing or misinterpreting data, and concerns about hidden errors in the data that researchers would not be able to detect) (Curty and Qin, 2014)
3. perceived effort involved in reusing data (e.g., beliefs in the ability to make innovative discoveries compared with primary research, in the abilities to easily search for and access data, and in the relative ease of matching data to current research questions, preparing data for analysis, and understanding data) (Curty and Qin, 2014)

Curty et al. (2017) found in their survey research that beliefs in the efficiency and efficacy of data reuse as well as perceptions about the importance of data reuse to advance science or the researcher's career predict data reuse by researchers. They also found that perceptions of social pressures against data reuse predicted less reuse and that agreement with concerns about the trustworthiness of data did not predict less reuse of data.

Kim and Yoon (2017) had similar results in their survey, including finding positive significant relationships between perceived usefulness of data and intentions to reuse data, and between the availability of human or technical support for researchers in an institution and intentions to reuse data. They found a negative relationship between perceived risks (e.g., about copyright or the ability to publish results) and intentions to reuse data, and no significant relationship between perceived effort and intentions to reuse data.

### *2.4.2 Assessments of Data*

Research shows that a wide variety of assessments about data are important to researchers' decisions to reuse data. Some of these include assessments about:

- data relevance (Faniel and Jacobson, 2010; Faniel et al., 2013; Faniel et al., 2016; Kelton, 2008; Sherif, 2018)
- data completeness (Faniel et al., 2013; Sherif, 2018; Medjedović, 2011; Yoon, 2016, Corti and Thompson, 2004)
- data quality (Curty and Qin, 2014; Curty, 2016; Enke et al., 2012; Faniel et al., 2013; Faniel et al., 2016; Frank et al., 2017; Yoon, 2016, Gregory et al., 2020)
- documentation quality (Faniel and Jacobson, 2010; Faniel et al., 2013; Niu, 2009a; Curty and Qin, 2014; Faniel and Yakel, 2017; Yoon, 2016; Stvilia et al., 2015)
- data usability (Faniel and Jacobson, 2010; Faniel et al., 2013; Yoon, 2016)
- rigor of the original study (Curty, 2016; Yoon, 2016)
- data interoperability (Yoon, 2016)
- data comparability (Enke et al., 2012)
- data "fit" to reusers' research questions (Kelder, 2005; Curty and Qin, 2014; Andersson and Sørvik 2013; Medjedović, 2011; Medjedović and Witzel 2005; Sherif, 2018)
- implications and consequences of ethical issues for data reuse (Andersson and Sørvik 2013; Kelder, 2005; McKay, 2014; Medjedović, 2011, Sherif, 2018)
- data accuracy (Fear and Donaldson, 2012; Donaldson and Conway, 2015; Kelton, 2008; Sherif, 2018)
- data credibility (Fear and Donaldson, 2012; Curty and Qin, 2014; Curty, 2016)
- data authenticity (Fear and Donaldson, 2012; Donaldson and Conway, 2015)
- data creator expertise (Fear and Donaldson, 2012)
- data understandability (Curty and Qin, 2014; Faniel and Jacobson, 2010; McKay, 2014; Medjedović, 2011)
- data trustworthiness (and multiple criteria that comprise constructs for trust such as integrity, currency, validity, objectivity, credibility, and others) (Yoon, 2017a; Curty, 2016; Donaldson and Conway, 2015; Donaldson et al., 2017; Faniel and Jacobson, 2010; Faniel and Yakel, 2017; Frank et al., 2017; Fear and Donaldson, 2012; Kelton, 2008; Sherif, 2018, Zimmerman, 2008, Gregory et al., 2020)
- depth of research (Frank et al., 2017)
- data reliability (Stanley, 2013; Faniel et al., 2016; Faniel and Jacobson, 2010; Faniel et al., 2013; Yoon, 2016; Yoon, 2017a)
- data consistency (Curty, 2016; Enke et al., 2012)
- timeliness of the data to be reused (Sherif, 2018)
- richness of the data (Sherif, 2018)

Researchers have found, significantly, that assessments of these kinds are driven by and made in relation to the specific goal or purpose the researcher has to use the data (e.g., Medjedović, 2011; Rolland and Lee, 2013). They are also made in relation to specific research contexts and may vary according to a researcher's role in a project or level of experience reusing data (Stvilia et al., 2015). The variety of assessments that researchers make about data are indicative of the many kinds of things researchers need to know about data (i.e., to make these assessments) in order to reuse them. I analyze the implications of literature on assessments of data in my discussion below (see section 2.4.6).

### *2.4.3 Resources*

The resources available to researchers also influence their decisions, and, additionally, their abilities to reuse data. These include such resources as data repositories that facilitate reuse of data (Curty and Qin, 2014; Faniel et al., 2013; Faniel and Yakel, 2017; Frank et al., 2017; Kim and Yoon, 2017), available funding (Enke et al., 2012; Smioski, 2010) and researchers' own time and effort (e.g., to understand and make sense of data) (Corti and Bishop, 2005; Enke et al., 2012; McKay, 2014; O'Connor and Goodwin, 2010; Sherif, 2018; Smioski, 2010; Zimmerman, 2008; Andersson and Sørvik 2013; Curty, 2016; Huggett, 2018; Kim and Yoon, 2017). It is noteworthy that Kim and Yoon (2017) did not find a significant relationship between perceived effort and intentions to reuse data. However, they noted that further study was "necessary to confirm how the effort involved in data reuse affects actual data reuse process" (Kim and Yoon, 2017, p. 2716). Researchers are also affected in their reuse of data by the extent of assistance available to them from data creators, data archives, and other researchers (Carlson and Anderson, 2007; Corti, 2005; Curty and Qin, 2014; Curty, 2016; Donaldson et al., 2017; Frank et al., 2017;

Kim and Yoon, 2017; McKay, 2014; Medjedović, 2011; Wolski et al., 2017; Yoon, 2016; Yoon, 2017b).

Researchers draw on these and other resources, including prior publications (Curty et al., 2016; Faniel and Yakel, 2013; Wallis et al., 2013; Donaldson et al., 2017; Faniel and Jacobson, 2010) and their own internalized knowledge and skills (Niu, 2009a; Corti and Bishop, 2005; Curty, 2016; Faniel et al., 2013; Zimmerman, 2008), as well as data documentation and other sources of provenance and contextual information (Carlson and Anderson, 2007; Faniel and Yakel, 2017; Corti and Bishop, 2005; Fear and Donaldson, 2012; Medjedović, 2011) to acquire sufficient knowledge to make assessments about and reuse data.

### *2.4.4 Social Factors*

In addition to perceptions and beliefs, assessments of data, and resources used to make assessments and aid reuse, a variety of social factors also play a role in researchers' decisions to reuse data. One of these is the disciplinary environment for reuse. Factors in this arena include whether or not there is a tradition of reuse or history of practices such as investments in infrastructure that favor reuse (Kim and Yoon, 2017; Rung and Brazma, 2012; Carlson and Anderson, 2007; Faniel and Jacobson, 2010; Smioski, 2010); the receptiveness to data reuse in the discipline (Curty, 2016), the existence of standards for documenting, curating, and archiving data (Enke et al., 2012; Faniel and Yakel, 2017; Carlson and Anderson, 2007); and the existence and applicability of policies related to data sharing and reuse (Katsanidou et al., 2016). The availability of data is another important factor in data reuse—e.g., the formats of data that can be accessed and whether raw data are available (Corti and Bishop, 2005; Enke et al., 2012; Yoon, 2016). The ability to access data can depend on aspects of the social environment such as policies governing data access and access to the software and tools needed to reuse data (Corti

and Bishop, 2005; Yoon, 2016; Kim and Yoon, 2017). In addition, various kinds of support for data reuse, to the extent that they involve or depend on social relationships (e.g., between data reusers and repository staff or among networks of researchers) (Frank et al., 2017; Yoon, 2017b) can affect researcher's knowledge about what data sources are available or their confidence in reusing data (see, e.g., Zimmerman, 2008).

To this last point, researchers have found that the reputation of the data creator or custodian impacts researchers' decisions to reuse data (Curty, 2016; Faniel et al., 2013; Faniel and Yakel, 2017; Faniel et al., 2016, Frank et al., 2017; Wallis et al., 2013; Fear and Donaldson, 2012; Faniel and Jacobson, 2010; Enke et al., 2012), as does the authority of the recommendations to use data that a researcher might receive (Kelton, 2008). To the extent that they are influenced by disciplinary norms, the resources researchers draw on and the relative importance of the resources can be seen as an influenced by social factors as well.

### *2.4.5 The Importance of Knowledge*

A commonality among the different factors that affect researcher's decisions to reuse data is that ultimately, because decisions to reuse data are conscious decisions, the factors are expressed in what researchers know about data. This is the case even for perceptions and beliefs, as these are based on the knowledge researchers have (e.g., about data creator reputations, acceptance of data reuse as a methodology in their field, etc.). There are thus certain things researchers need to know about data in order to feel comfortable admitting them as evidence for their research. The goals of my proposed research are to identify the contours and extents of the knowledge researchers obtain to reuse data, the reasons they obtain the knowledge, and the factors researchers believe affect their determinations about how much of what kinds of knowledge are sufficient to reuse data. Research on the amount researchers need to know, or the

amount of contextual information that should be made available to them has been called for by a number of researchers (Borgman et al., 2007a; Mauthner, 2014; Irwin and Winterton, 2011b; Sherif, 2018).

### *2.4.6 Discussion*

The literature on factors that affect decisions about reuse is extensive. Findings about the importance of perceptions and beliefs, assessments about data, resources used to obtain contextual knowledge, and social factors involved in making decisions about reusing data appear to a large degree to be consistent across studies and across multiple disciplines. There are some important gaps in the literature, however, which are illustrated by conflicting findings in different studies. The conflicts have to do with differences in perceptions or general assessments of data and actual reuse behavior.

For instance, Curty and Qin (2014) found in thirteen interviews with social science researchers that the perceived effort in reusing data was a factor in reuse decisions. However, Kim and Yoon (2017) found in a multi-disciplinary survey (n=1,237) that there was no significant association between perceived effort and reuse intentions. Moreover, Niu (2009a) found in the interview portion of a mixed methods study consisting of interviews and a large survey that data reusers persisted in their reuse of datasets even when the datasets were poorly documented. She did not investigate this issue further in the survey.

Second, multiple studies have found the importance of the trustworthiness of data to data reusers' decisions to reuse data (Donaldson and Conway, 2015; Faniel and Jacobson, 2010; Wallis et al., 2007). At the same time, Curty et al. (2017) found in a reanalysis of a worldwide survey (n=595) that agreement of respondents with concerns about the trustworthiness of data did not predict less reuse of data.

These conflicts demonstrate two main limitations in the literature. The first is the difficulty of comparing results between different reuse studies. This may be partly due to the difficulty of comparing results from qualitative and quantitative research, where the levels of description, extents of context, and purposes of research (e.g., to understand a phenomenon deeply as opposed to obtaining generalizable results) can differ greatly. When there are discrepancies, such as in the findings of Curty and Qin (2014), Kim and Yoon (2017), and Niu (2009a), it can be difficult to determine on what terms and in what ways results might be compared or understood.

This difficulty is compounded, however, by a lack of specificity and detailed characterization of study populations. For instance, numerous studies investigate reuse in broad settings such as the "social sciences" (Curty and Qin, 2014; Kim and Yoon, 2017; Faniel et al., 2012; Niu, 2009a), "biological sciences" (Kim and Yoon, 2017), or "biodiversity science" (Enke et al., 2012). At the same time, researchers who describe these broad categories recognize the particularities of sub-disciplines and more focused areas of research (e.g., Zimmerman, 2007; Curty, 2016, Gregory et al., 2020). Some even select broad sampling criteria like "social science" with an expectation of diversity in subareas of the discipline (e.g., Curty, 2016). It can thus be difficult, if not impossible, to determine which findings or aspects of findings in larger quantitative studies relate to findings in smaller qualitative studies in fields like ecology or genetics, or sub-domains of social science like health and social care (Heaton, 2004).

A second gap in the literature is a lack of studies that examine nuances in the knowledge and contextual information needed to reuse data. Findings indicate that trustworthiness is important to decisions of reuse but that researchers may reuse data despite concerns about trustworthiness (Curty et al., 2017). Findings also indicate that researchers reuse data despite

gaps in knowledge about the data (e.g., from inadequate documentation or ability to understand the documentation) (Niu, 2009b; Heaton, 2004). This passage describing findings from the interview portion of Niu's (2009b) study illustrates this phenomenon:

> Some uncertainty [about reusing data] could be partially solved after seeking outside information. When uncertainty cannot be solved, users would need to make a decision whether to tolerate the uncertainty or give up using data. Interview #5 described how she dealt with uncertainties in using secondary data: if she is really not sure about something and can't get help about it, she won't proceed on it. Sometimes she had to rely on her best interpretation of what to do and leave it to peer review later. (Niu, 2009b, p. 78)

Such findings suggest that there are levels of trustworthiness, levels of quality, and likely levels of other criteria that secondary researchers evaluate subjectively when making decisions about reusing data. It would be reasonable to assume from these findings that some or all of the factors described above (perceptions and beliefs, assessments, resources, social factors) influence the levels of data quality, trustworthiness, credibility, and other characteristics that secondary researchers are willing to accept. Unfortunately, however, the literature does not examine these levels or the factors that affect how these levels are set.

We therefore know a lot about the characteristics that are important to researchers generally in selecting data to reuse (quality, accuracy, authenticity, etc.) and factors that might influence researchers' decisions (e.g., perceptions, disciplinary traditions, time, effort, and support available for reuse). But we know these primarily in specific cases that are hard to generalize (i.e., from case- or self-studies of reuse or researcher interviews using small sample sizes), or in general cases (e.g., in or across research domains) where findings pertinent to specific disciplinary or sub-disciplinary areas are difficult to isolate. Gregory et al., 2020 is nearly an exception to this, but while they examined knowledge researchers needed to reuse data and conducted other analyses by disciplinary domain, they did not analyze the knowledge researchers desired by domain. There is furthermore, even in studies involving cross-disciplinary

comparisons (e.g., Faniel and Yakel, 2017; Faniel et al., 2016; Carlson and Anderson, 2007, Gregory et al., 2020), a lack of research into the amounts of knowledge researchers use to determine the sufficiency of data and what influences how those amounts are defined.

In light of these gaps, multiple trajectories of research are needed. Research is needed first into the levels of knowledge researchers desire and are able to obtain about reused data, i.e., at what points along which dimensions researchers decide that secondary data are suitable for their purposes. Second, research is needed into the influence of perceptions and beliefs, the availability of resources, and social factors on the knowledge needed to reach reuse equilibrium at the level of individual instances of reuse. Understanding factors important to researchers in one discipline or another may not be specific enough to describe the factors that influence a particular researcher or research team.

At the same time, research is needed that is able to produce comparable findings about what researchers need to know to determine data are sufficient to reuse across multiple different instances of reuse. While some research has examined multiple different reuses of the same dataset (van den Berg et al., 2004; Atici et al., 2012), most case studies of reuse investigate individual instances of reuse by one researcher or research group. This mode of study limits the ability to compare instances of reuse across different research purposes.

The kinds of research conducted to date about data reuse provide a wealth of information about data reuse in different contexts. There are significant gaps in the research, however, which, if explored, could provide critical insight into how to overcome challenges to conducting research with secondary data. These include, particularly for the proposed research, challenges of facilitating access to the kinds and levels of knowledge researchers need in light of their skills,

experience, disciplinary traditions, and numerous other factors, to make the most of data from prior research. My study is a step toward engaging in these needed areas of research.

### 2.5 The Importance of "Context"

Researchers have noted that data used in secondary analysis should be adequate to the purpose of the secondary analysis (Irwin, 2013; Notz, 2005; Sherif, 2018; Doolan and Froelicher, 2009). But what is the breadth of information that researchers need to know to determine that data are adequate? Many researchers refer to the information about data that is needed to reuse them as "context." Researchers use context to assess and produce knowledge about data. We can therefore better understand the knowledge researchers need by investigating attributes of context.

Research has revealed a variety of aspects of contextual information. It has found, for example, that information about context is important to data reuse (Warwick et al., 2009; Dicks et al., 2006; Faniel and Jacobson, 2010; Faniel et al., 2013; Faniel et al., 2019; Sveinsdottir et al., 2013; Bishop, 2007, citing Fielding, 2004, Heaton, 2004, Fielding and Fielding, 2000, and Hammersley, 1997; Shankar, 2007, cited in Oleksik et al., 2012; Andersson and Sørvik, 2013; Wan and Pavlidis, 2007; Zimmerman, 2008; Pasquetto et al., 2019; Koesten et al., 2021), that data are embedded within their contexts of creation (Broom et al., 2009; Dicks et al., 2006; Gillies and Edwards, 2005; Mauthner et al., 1998; Irwin, 2013), that the "full" context of reused data is never available (Andersson and Sørvik, 2013; Carlson and Anderson, 2007; Dicks et al., 2006; Bishop, 2006; Parry and Mauthner, 2005; Yoon, 2016), and that the type and level of contextual information needed to reuse data depend both on the purpose of the research (Andersson and Sørvik, 2013; Huggett, 2018; Bishop, 2006; Bishop and Kuula-Luumi, 2017; Faniel et al., 2013; Jackson et al., 2007; Medjedović, 2011; Fear and Donaldson, 2012; Rolland

and Lee, 2013; Gregory et al., 2020) and on the stage of research (Faniel and Jacobson, 2010; Jackson et al., 2007; Gregory et al., 2020).

Types and levels of context also affect the kinds of reuse that can be conducted (Travers, 2009), as do the methodology, theoretical frameworks, and other aspects of the research design used in the original study (Carlson and Anderson, 2007; Wan and Pavlidis, 2009; Irwin and Winterton, 2011a; Brody, 2011, cited in Faniel et al., 2013). Researchers have found, moreover, that the data that are "data" to some can be "context" to others (Borgman et al., 2007a; Dicks et al., 2006), that the contextual information that is important to consider is determined by the researcher (van den Berg, 2005), that experience and skill affect researcher's facility in reusing data (Niu, 2009; Zimmerman, 2008), and that metadata may be inadequate to document contextual information that is needed (Borgman et al., 2007; Dicks et al., 2006; Sveinsdottir et al., 2013; Fear and Donaldson, 2012, Duff and Johnson, 2002; Keene, 2005, cited in Faniel et al., 2013; Faniel and Jacobson, 2010, citing Birnholtz and Bietz 2003; Koesten et al., 2017).

Research has additionally shown that a lack of desired contextual information does not necessarily prevent researchers from reusing data (Faniel et al., 2013; Niu, 2009a, Bishop and Kuula-Luumi, 2017). As Irwin and Winterton (2011a) point out, researchers find ways to "make do" with what is available. It is when the information available is not sufficient to determine that data are fit for the intended research purpose (when researchers "don't know what they don't know") that risks of systematic misinterpretation increase. Researchers must determine, using comparisons e.g., with other evidence and existing theories, whether there are sufficient sources of verification and grounds on which to accept the results of their secondary analyses (Irwin and Winterton, 2011a).

### *2.5.1 Discussion*

Like knowledge about the purposes of reuse, what we know about the contextual information needed to reuse data comes from a variety of domains of research including earthquake engineering (Faniel and Jacobsen, 2010), neuroscience (Wan and Pavlidis, 2009), proteomics (Fear and Donaldson, 2012) epidemiology (Rolland and Lee, 2013), archaeology (Faniel et al., 2013) and social science (e.g., Medjedović, 2011; Jackson et al., 2007; Yoon, 2016; Niu, 2009a). Multi-disciplinary studies have also investigated contextual information in particle physics, astronomy and astrophysics, health and clinical research, bioengineering, environmental research, archaeology, anthropology, and zoology (Sveinsdottir et al., 2013; Carlson and Anderson, 2007; Faniel and Yakel, 2017).

Within and across these domains, many of the gaps in the literature that have been mentioned above are also in evidence. First, research has uncovered a substantial amount of information about the contextual information researchers use when reusing data, but there is a lack of empirical investigation into how researchers negotiate the "distance" between themselves and the original data when the desired information is not available, including a lack of research into the levels of knowledge researchers determine is necessary to reuse data, and factors that affect these determinations.

For instance, Zimmerman (2007) found that one measure ecologists take to ensure they understand data created by others is to select data to reuse that are similar to data they have collected themselves. She does not discuss, however, *how* similar the data need to be. Furthermore, Zimmerman found that ecology researchers are influenced in data reuse decisions by their desire to conform to scientific norms. They strive, for example, to draw data to reuse from representative samples, reduce the potential for error, and defend data collection decisions publicly. However, Zimmerman does not examine how researchers determine what is

representative enough or how much reduction in the risk of error is sufficient. There is also little discussion of factors that influence these decisions. This is the case, although an excerpt from one of the thirteen in-depth interviews conducted for the study illustrates the kind of satisficing and decision-making that researchers engage in (in this instance, to obtain a more representative sample). Zimmerman reports a researcher's discomfort with the way one of her students justified the inclusion of local data in a research project:

> She's got a sentence in here that I feel very uncomfortable with. OK. "To obtain sources, we searched Biosis between 1995 and November 1999 using the keywords... In addition, we added references from our files." I said, "Lora, you're going to get creamed on that one." So, she added the sentence: "Although not the most systematic approach, this increased the time period from which references were drawn, and it increased the number of ecological relative to agricultural studies. "So, those are both important things to do, and I agree in this case that we should use them. Ah, but I just... that sounds so... "We added references from our files..." is sort of like, "Well, we happened to have it around." (Zimmerman, 2007, p. 10)

Similar instances can be found across the literature related to contextual information, where researchers identify contextual information or contextual knowledge important to data reuse but do not explore the extent of the information or knowledge needed and what influences that extent.

Second, while much research has examined the contextual information important to researchers, less has considered the types of contextual information that are most important to or prioritized by researchers and why. Only one of the studies reviewed for this proposal included results with a researcher's indication of the priority of different kinds of contextual information to his research. This was a study by Faniel et al. (2013) in which the authors investigated the challenges archaeological data reusers faced when different kinds of contextual information were not available. Understanding the priority of different kinds of contextual information was not a focus of this research, however, and the topic received little attention.

In general, little work has been done (within or across disciplines) to understand why and under what circumstances researchers might be willing to satisfice in certain types of knowledge they obtain about secondary data over others. The result is that researchers have developed typologies and characterizations of context they believe are necessary to reuse data, such as situational, interactional and institutional context (Bishop, 2006); macro context (Walters, 2009) and micro context (Irwin and Winterton, 2011a; McLoughlin and Miller, 2006; Walters, 2009); substantive context (Dicks et al., 2006); scientific and social context (Chin and Lansing, 2004); and proximal and distal context (Holstein and Gubrium, 2004). They have also made projections about the kinds and characteristics of data that are most likely to be reused (Bishop and Kuula-Luumi, 2017; Thompson, 2000, Wan and Pavlidis, 2009; Kwek and Kogut, 2015), and developed specific guidelines for the types of context that must be considered in order to reuse data (van den Berg, 2005). But there has been little empirical research testing these notions or comparing the kinds of contextual information or kinds of knowledge researchers prioritize obtaining from one research purpose or discipline to the next.

Characterizations of the types of context that are or could be important to researchers (e.g., Lee, 2011, Bishop, 2006) provide rich background for analyzing findings in studies of contextual information. However, research is needed to better understand how researchers determine which contexts matter in specific reuse instances—which types of knowledge about reused data are more important than others—and what influences those decisions.

It is to facilitate such research that I understand "data" in my dissertation to encompass what might be considered as both data and context. Given the different ways these terms are defined and understood in different circumstances, a broader, more fluid conception of data is

appropriate to understand how researchers bound the knowledge they obtain about data in order to decide to reuse them.

## 2.6 Satisficing in Research

### *2.6.1 Review of Literature*

Satisficing is one of many phenomena that have been identified in studies of rationality. Simon (2000) traces studies of rationality back to at least the time of classical Greek philosophers when rationality was seen as part of the process of reasoning that occurred, for example, in the construction of logical proofs. Simon argues that this conception of rationality, which encompassed both products of reasoning and processes of reasoning, extended nearly through the 19th century when it was replaced with the theory of utility maximization (Simon, 2000). This theory took as its focus what Simon refers to as "substantive rationality"—"the quality of the adaptation to the external environment in the light of the decision maker's utility function" (Simon, 2000, p. 27)—to the exclusion of "procedural rationality," the process by which decisions are made.

It was not until after the Second World War that economic theories began to explore rational behavior in the context of uncertainty, both about circumstances in the world and about decisions that others will take (Simon, 2000). New treatments of uncertainty led to developments in probability theory, game theory, the theory of rational expectations, and experimental economics (Simon, 2000). These developments increasingly showed the inadequacy of utility theory for explaining actual human behavior. It is in this context, Simon argues, that a return to ideas of bounded rationality occurred, incorporating both procedural and substantive limits to rationality in the study of human behavior.

While Simon's conception of bounded rationality incorporates both of these elements (substantive and procedural), for a considerable time, much theoretical work and empirical research focused on procedural elements; specifically constraints on time, cognition, and at times knowledge (Gigerenzer and Goldstein, 1996). This was as much the case in the field of information science (IS) as in other fields, where many studies of bounded rationality and satisficing examined ways that users satisficed in the context of information seeking, and what factors or "stopping rules" influenced when a user terminated their search (e.g., Agosto, 2002; Bilal, 1998; Hirsch, 1999; Kafai and Bates, 1997; Zach, 2005; Duggan and Payne, 2011). I describe some of these studies further below when reviewing Agosto's (2002) work in more detail.

A more recent strain of scholarship has investigated stopping behavior in greater depth and distinguished judgements that users make prior to coming to a decision from the decision itself. For instance, drawing on Hogarth (1980), Berryman (n.d.) describes two kinds of judgements users make:

> Judgement may be either predictive or evaluative and, whereas predictive judgements are used in sequential decision making [Berryman locates Simon's theory of satisficing here], evaluative judgements reflect personal preferences—and possibly, personal interaction with contextual factors (Hogarth, 1980, p. 3). (Berryman, n.d., p. 33)

This framing greatly expands the scope of factors that might influence decision-making (i.e., the judgements that happen beforehand) and opens inquiry into how much information is enough to make a decision (i.e., to terminate a search) and what affects these decisions (Zach, 2005; Berryman, 2006). For example, Berryman (2006) studied how policy-makers in Australia decided they had obtained enough information in the work they performed, particularly in the preparation of briefing reports. She found that a necessary step for these workers in deciding how much information was enough was to establish a framework for the task: e.g., the purpose for the

information gathering or the organization's overall goals. She found that factors that subsequently affected decisions about how much information was enough included the structure of the problem (i.e., the scope of the work to be done); uncertain and dynamic environments in which work took place; shifting, ill-defined or competing goals; action, feedback loops (i.e., iterative feedback from superiors); time stress; the high-stakes nature of the work; whether multiple players were involved; and organizational goals and norms (Berryman, 2006).

In a different study, Zach (2005) found that administrators in arts organizations were influenced largely by their own level of comfort with obtained information and the amount of time they had when making determinations about how much information was enough to perform managerial tasks. Administrators engaged in iterative evaluation of acquired information as their decisions proceeded, and were influenced primarily in the amount of effort they put into gathering information for a decision by their own experience, the importance of the task, and the perceived importance or potential impact of a decision (Zach, 2005).

Berryman (2006, 2008) has suggested that because the ways people make decisions in practice has been found to differ from rational choice theory (i.e., utility theory) and behavioral decision theory (Simon's theory of bounded rationality), decisions in the real world can be more thoroughly analyzed using a more naturalistic conceptualization of decision making (e.g., as articulated in Kerstholt and Ayton, 2001 and Zsambok and Klein, 1997). This conceptualization emphasizes decision making as a process of ongoing assessments and judgements, and views contextual factors as critical in shaping decision making behavior.

Berryman prefers this conceptualization in particular because of its focus on decision making as a process rather than a single choice between two options (as in behavioral choice theory), and the attention it pays to the role of judgements in decision making as opposed to

decisions themselves alone (Berryman, 2006, 2008). She supports an argument for shifting to a naturalistic model by demonstrating the correspondence between the factors identified in her study (those listed above including the structure of the problem, uncertain and dynamic environments, etc.) and those identified by researchers in naturalistic decision making (i.e., Zsambok and Klein, 1997) (Berryman, 2006). One of Berryman's key arguments is that a naturalistic model allows for a more holistic conceptualization of factors that affect decision making (Berryman, n.d.; 2008). This overcomes limitations she observes in much research on information seeking, which has been limited to the study of a small number of pre-determined factors (Berryman, n.d.).

Berryman's holistic conception of factors affecting decision making and emphasis on evaluative judgements resonate with the approach I have taken to knowledge bounding and knowledge satisficing in data reuse. However, I based my research on Simon's model due to several important differences. First, Berryman is concerned with any reasons why information seeking processes are terminated (i.e., stopping rules), whereas I was concerned with satisficing as a stopping rule in particular. In other words, to the extent that any evaluative judgements took place—in addition to predictive judgements—I was interested in understanding them in light of satisficing rather than other possible stopping rules.

Second, in the information seeking scenarios Berryman used, there is not necessarily a known end state, though this may be understood after the fact to be the sum of multiple iterative evaluations of obtained information that occur throughout the search process. I am interested in knowledge researchers desired about data but could not obtain or were only able to obtain in a limited way as they adapted to a design "artifact" (i.e., the data). This is different from investigating a scenario where a subject engages in a search with unknown termination, and

endeavoring to understand the reasons why they stopped their search. Measuring the initial goal state in relation to the achieved goal state was central to my research, at least insofar as it concerned knowledge satisficing.

At the time I designed my study, it was not clear to me whether understanding decisions according to a naturalistic framework, rational choice theory, or behavioral decision making framework had a bearing on measurements of initial versus achieved goals for knowledge attainment. It seemed an open question whether a holistic analysis of decision-influencing factors within this framework diminished if, as I proposed, environmental constraints that took place in substantive rationality were extended to procedural rationality. For all of these reasons, I chose to conduct my study within Simon's framework of behavioral rationality.

Agosto's (2002) study, mentioned above, indicates some of what has been learned about rationality by using this paradigm. I review her study here in some depth because it provides examples of studies of satisficing that have been conducted in IS and because of the similarities it bears to my own study.

Agosto examines the role of bounded rationality, satisficing, and personal preferences in young people's Web-based decision-making (Agosto, 2002, p. 16). She leads her literature review with a quote from Marchionini (1995) demonstrating the applicability of the notion of satisficing to research in library and information science (LIS):

> Satisficing is essential to most information seeking because all pertinent information for open-ended problems can seldom be assembled and assimilated optimally. (Marchionini, 1995, p. 63-64, quoted in Agosto, 2002)

Agosto references three other studies of satisficing or bounded rationality in LIS research. The first is Schwartz (1989) who examined decisions about and proposed a model for book selection in libraries based on the concept of bounded rationality. The main extensions he

made in his model were to include tacit knowledge as a factor in decision making (e.g.., the importance of the "experienced glance" of the expert practitioner (Schwartz, 1989, p. 338)) and to allow that rational decisions could be neither substantive (about the end decision) nor procedural (concerned with the efficiency of the process). Instead, as in book selection, the point of the decision may be the symbolic importance of the process itself (e.g., to show that the library is well-managed). This finding, along with Agosto's described below, illustrates ways that the concept of bounded rationality can continue to be applied and extended based on the evaluation of specific experiences and scenarios.

The second is Chu (1994), who considered library reference services in light of bounded rationality and satisficing. Chu's main contribution was the observation that in a situation where librarians have multiple ways of representing reference responses and what constitutes an appropriate response is a function of the person asking (i.e., what is satisfactory for an undergraduate student may not be so for a graduate student), a response that is "good enough" depends on a negotiation occurring and common understanding being reached between the two parties. While this observation does not play into Agosto's research, it lends credence to my investigation of satisficing in the context of adaptation (i.e., configuration and reformulation).

The third is Higgins (1999). In experiments where undergraduate business students were put in the position of graduate admission officers, Higgins found that the students used the credibility of the sources (in this case, institutions) that referred the applicants as a heuristic to select students for admission. This was the case whether or not students' decisions were forced to occur under time constraints, though students were less definite about their preference for the students they chose to admit when under time constraints. Higgins' study is an example of

studies that focus on testing or identifying types of satisficing behavior within limits of time and cognition.

Agosto reviews a number of studies of young people's Web-based information-seeking behaviors, which fall into a similar category (Bilal, 1998; Hirsch, 1999; Kafai and Bates, 1997; Schacter, Chung, and Dorr, 1997; Fidel et al., 1997). Although the contexts of each study were different, the results were similar in showing that young people exhibited satisficing behavior, employing strategies that reduced their cognitive load while browsing the Web. These included skimming resources rather than reading in detail (Hirsch, 1999); reading only a few of many search results (Bilal, 1998); assembling directories of sites based on evaluation of visual rather than textual content (Kafai and Bates, 1997); using browsing rather than pre-meditated search strategies (Schacter, Chung, and Dorr, 1997) or preferring keyword searching over browsing (Bilal, 2000) depending on which strategy reduced cognitive load the most; and quickly abandoning unsuccessful searches (Fidel et al., 1999) (these studies are summarized in Agosto, 2002).

Agosto herself conducted group interviews of young women who participated in 50-minute Web-surfing sessions. In the sessions they were asked to visit three sites for as long as they desired and to spend the rest of the time browsing the Web as they pleased (Agosto, 2002, p. 19). The women were asked to consider three questions about the three websites: what they liked, what they did not like, and what they would change if they could. In the group interviews, Agosto used several strategies to encourage participants to reflect broadly on their experiences and to avoid influencing the types of responses they gave. First, she did not give participants any parameters for how to evaluate the websites (e.g., to think about design or accuracy). Second, in the group interviews, she asked participants to give their opinions about the Web in general, not

limiting discussion to the Web-surfing session in particular. She also intervened only rarely in the group discussions in order to allow participants to pursue directions in the conversation that were of interest to them.

From the group interviews, Agosto found that the women experienced both self-imposed and externally-imposed time constraints in their use of the Web; cognitive constraints brought on by information overload, textual overload, and outcome overload (too many options to choose from); and physical constraints due discomfort from poor posture or eye fatigue, or exertion—e.g., the reduced labor involved in finding resources on the Web instead of obtaining information resources from a physical location (Agosto, 2002). She also found that participants evaluation of sites were influenced by personal preferences including design preferences (color and aesthetic design preferences), preferences for certain topics and tones used to present those topics, and participants' own personal convictions.

In terms of satisficing, Agosto found that participants exhibited two types of satisficing behaviors. The first were methods of "reduction" to reduce the number of web sites to be evaluated. These included, "returning to known sites, relying on site synopses, and using indexing categories to dismiss sites from deeper consideration" (Agosto, 2002). The second were methods of termination, or "stopping rules." A key feature of satisficing in Simon's view is that someone exhibiting satisficing behavior will take the first decision that satisfies their aspirations along all applicable dimensions (see quote from Simon (1994) on satisficing in section 1.4.6). Agosto found that while participants took decisions based on satisficing, they also took decisions before a satisficing option had been reached. They did this for reasons of physical discomfort, boredom, time constraints, or information "snowballing" (i.e., when site after site presented the same information) (Agosto, 2002).

### *2.6.2 Satisficing in Data Reuse*

References to satisficing are relatively rare in the reuse literature itself. In their study of factors that affect researcher satisfaction with data reuse Faniel et al., 2016 selected data relevancy as one of the factors and framed the concept they used to measure relevance in terms of satisficing:

> Data reusers cannot always identify data that are a perfect match. For instance, the original study may have limitations inherent in the research design (Chin & Lansing, 2004; Faniel & Jacobsen, 2010). This leads to a choice: collect data, halt the research until the perfect data are found, or modify research objectives and reuse the available data. Given the time, money, and effort to collect data along with the likelihood that no data will be perfect, researchers often satisfice with the available data and shape reuse projects around what is possible given their access. The more reusers have to modify their research objectives to reuse the data, the less satisfied they are going to be with data reuse. Reusers who settle for data less relevant to research objectives are likely to have lower levels of satisfaction. (p. 1406)

Interestingly, although the authors found a number of factors to be significantly associated with reuse satisfaction—including data accessibility, data completeness, documentation quality, data credibility, and data ease of operation (usability)—data relevancy was not. Faniel et al. suggested this might have been the case because researchers avoid data that are not highly relevant (meaning that data that were not highly relevant did not enter their sample, which consisted of data known to have been reused).

Borrás (2020) investigated how and why data reuse happens, given the challenges researchers face. I include some extended quotes and explanation from Borrás' work below because of the similarities to my own. Borrás developed two theoretical models that underpinned her analysis of ten case studies of reuse (with each case being an individual case of reuse by a researcher in a health discipline). The bounded individual horizon model explains researchers' ways of working and the data-reuse mechanism model provides a causal explanation for why and how data reuse takes place. Borrás theorizes that while researchers desire to satisfy their

curiosity and contribute to the advancement of science, they also want and need rewards for their work. They therefore pursue secondary research in the service of professional goals and are affected in their reuse by a wide variety of influences. She says:

> researchers self-allocate a goal –a scientific contribution [as evidenced by a publication] or a career milestone–, which they expect to achieve within a limited period and with a limited amount of available material and cognitive resources by keeping in mind both their personal and professional situations, their values, beliefs, and feelings, their discipline's epistemic norms, and the reward system they belongs [sic] to." (pp. 40-41)

Borrás uses satisficing as the underlying model to explain how researchers operate in these conditions:

> I suggest that researchers, like all human beings, have limited time to achieve their goals, hence they do not maximize their goals but accept options that are good enough or satisficing, which is a better option than optimizing (Simon, 2000). (p. 38)

In particular, she notes that because a researcher has limited time and resources to conduct their research, they have a limited capacity to understand the efforts and resources needed to make a scientific contribution in the given amount of time. They therefore may not make the best choices in their decisions about data reuse, but rather satisficing choices. Borrás sees these decisions as a cascade of choices over time—each involving a satisficing solution—that affect the scientific contribution, the time or resources required to make it, or any or all of these. (p. 50)

Borrás hypothesized five initial conditions for her explanatory model (p. 50):

1. the researcher knows that secondary data exist
2. data have to be accessed or obtained by the researcher
3. secondary data are a satisficing option for the researcher
4. the idea of collecting particular primary data is not a satisficing option, and
5. an expected scientific contribution exists and the researcher finds its potential rewards satisficing. However, changes in these conditions, namely in condition C4, may still lead to the use of secondary data as evidence of scientific claims as I expound later in this section.

Borrás used ten case studies of data reuse to test whether, in light of initial conditions or combinations of conditions, the explanatory model (the data-reuse mechanism) led to specified outcomes (with sub-outcomes if primary data collection is a satisficing option):

> 1) "Use of secondary data does not happen at all after having tried or considered the option"; 2) "Use of secondary data happens but reuse is not shared with the research community and the data do not end up being evidence of scientific claims. Thus, secondary data end up serving as widening the researcher's background knowledge and triggering new research hypotheses." 3) "Use of secondary data happens and only secondary data are used as evidence of scientific claims." (p. 55)

Ultimately, Borrás found that the data-reuse mechanism did not explain the link she hypothesized between (a) a motivation to reuse data to make scientific contributions or career advancement and (b) reuse of secondary data as evidence of scientific claims. In particular, she found that the third condition (that data are a satisficing option for the researcher) is not a binary yes or no. Researchers make adjustments over time so that even if data are perceived to be less satisficing or not satisficing at all at a particular point in time, researchers "deploy several strategies with the data and/or the research question(s) in order to perceive data satisficing again" (p. 185). They do this as long as there is a research gap and prospects for making a contribution (evidence by a publication) exist.

Borrás found a similar effect (non-binary possibilities) for data being accessed by researchers as well (i.e., research may proceed with less than the desired amount of data). She also found that changes can occur during data reuse, for instance in researcher skills and knowledge or in data themselves (e.g., if data become available in a more processed form) that affect the order in which the five conditions she proposed are met in time. Finally, Borrás found a relationship between the amount of time and resources expended in reusing data and making the expected scientific contribution. Researchers who had expended a lot of effort were very determined make a contribution and effort being low was associated with deciding not to pursue

a contribution (though this did not rule out the possibility of contributions being made with low expended effort).

### *2.6.3 Discussion*

Agosto's findings are important to my study for several reasons. Her findings about the role of physical constraints and personal preferences in decision-making empirically demonstrate the embodied nature of decision making in the context of bounded rationality and satisficing. They also demonstrate the difficulty of clearly separating features of the inner environment from those of the outer environment (i.e., rational considerations were bounded by body and eye fatigue and evaluation decisions were influenced by preferences and emotions). Moreover, her findings about preferences are similar to what I expected to find in my study, in that personal preferences held sway once satisficing strategies in their traditional sense (e.g., reducing options in a problem space) had been accomplished. At that point, considerations of substantive rationality were more important to the participants than considerations of procedural rationality in coming to a decision or evaluation.

Agosto does not go so far as to call suboptimal decision-making in the context of substantive rationality "satisficing" as I proposed; and she clearly distinguishes between satisficing as a termination strategy that leads to an acceptable outcome and other termination strategies that lead to suboptimal outcomes. However, it is noteworthy that the methodology she uses—particularly her intentional strategy to leave questions and group discussions as open as possible, was adequate to detect all of these (i.e., both satisficing and non-satisficing behaviors and what affected them).

In my research, I investigated scenarios where I knew researchers came to acceptable decisions because they went on to publish their research—I knew they reached reuse equilibrium

and decided to reuse the data. I pre-selected, then, for "satisficing" outcomes, but my intent was similar to Agosto's in that I wished to gather information about what affected researchers' decisions to bound knowledge about data they obtained both in the context of traditional satisficing and in a context of substantive rationality where constraints of the inner environment (strictly interpreted in Simon's sense) may not control.

Agosto's study is thus encouraging as a model for how complexities of decision-making in the contexts of bounded rationality, satisficing, and other considerations (e.g., for Agosto, preferences) can be studied. The ability of the methods she used to distinguish between optimal outcomes (through satisficing) and suboptimal outcomes (which occurred in her study before satisficing took place) is also encouraging. I did not conduct group interviews so my participation in discussions with participants was necessarily more involved. As I explain in section 3.4.4, however, I used an approach that directed participants' responses as little as possible and encouraged them to reflect broadly on their experiences, while still gathering information on my topics of interest.

Borrás' research is similar to my own in many regards including that, with Agosto, Borrás allowed for a wider variety of factors that could influence researchers in decisions about data than are typically considered in rational choice theories. Her research is noteworthy, likewise, for not just considering but centering the professional pressures researchers face in reusing data and the relationship between those pressures and satisficing behavior.

My research differed from Borrás in questioning the assumption of satisficing as a foundation for researchers' decisions and in investigating knowledge bounding in reuse decisions in particular. I did not seek to develop a causal mechanism for conducting secondary analysis but

rather to understand, in situations where reuse has happened, how and why researchers determined the knowledge they obtained about data was enough.

## 2.7 Summary of Gaps in Literature

Gaps identified in the literature fall into three categories:

- knowledge bounding and knowledge satisficing
- importance of different kinds of knowledge
- comparability and generalization

These are summarized below, with a description of how the current study contributes to addressing them.

### *2.7.1 Knowledge Bounding and Knowledge Satisficing*

The literature identifies numerous criteria that researchers assess and kinds of knowledge they obtain when deciding to reuse data. It also identifies that researchers "make do" with the contextual information available. It does not address, however, how much of which kinds of knowledge researchers need when making these determinations or whether and how they satisfice.

The literature also identifies numerous factors that play a role in researchers' decisions to reuse data but does not discuss how those factors (e.g., knowledge, experience, resources, skills, data policies) affect determinations about how much knowledge is enough to determine that data are sufficient to reuse.

The literature additionally does not investigate knowledge satisficing, the factors that affect knowledge satisficing across different general purposes of reuse, or the impact satisficing might have on researchers' achievement of the desired outcomes of their research or their goals for reusing data.

### *2.7.2 Importance of Different Kinds of Knowledge*

The literature identifies the kinds of knowledge about data that are important to researchers, but there is little investigation into how researchers determine, of the many types and extents of context that could be investigated, which types of knowledge are more or less important in particular reuse instances and why. Further, the literature does not address how satisficing behavior might be evidenced across the kinds of knowledge about data that are more or less important to researchers.

### *2.7.3 Comparability and Generalization*

The literature investigates contextual information and knowledge needed for reuse across a variety of reuse research topics, methodologies, and scales (e.g., studies of individual reuse cases, particular groups of researchers, and of disciplinary domains). However, findings from these studies can be difficult to compare or generalize from except at high levels (e.g., that quality or trust is important in decisions about reuse). Even at that level, there can be conflicting findings between studies that are hard to reconcile.

### *2.7.4 Addressing Identified Gaps*

In my dissertation, I address these gaps by investigating:

- how researchers bound the knowledge they obtain in order to decide to reuse data (including whether or not researchers satisfice in obtaining desired knowledge),
- factors that affect knowledge bounding and knowledge satisficing,
- the relative priority that researchers assign to different types of knowledge about data in particular reuse instances, and why, and
- the impact of knowledge satisficing on the outcomes of research and researchers' attainment of their researcher goals.

I examined these across different reuse purposes and domains of research, and in light of particular attributes of researchers and research circumstances (e.g., researcher's experience with reuse and how research questions were developed). I also sought to address issues of

comparability and generalizable findings through a survey of satisficing behavior across a wide range of data reuse cases, and through the qualitative analysis of multiple instances of reuse.

## 2.8 Connection Between Theoretical Framework, Gaps in Literature, and Research Questions

### *2.8.1 Components of Theoretical Framework*

My theoretical framework has two main components with several underlying ideas. The first is an understanding of the process by which researchers determine data are sufficient to reuse (i.e., of reaching reuse equilibrium) as a process of adaptation. The adaptation involves researchers setting boundaries or thresholds on the kinds and amounts of knowledge they obtain about data in order to reuse them. This understanding emerges directly from my application of Haraway's and Simon's theories to data reuse. The underlying ideas here include:

1. An understanding of archived data as a design "artifact" (from Simon)—i.e., a type of adaptive system humans design to be a representation of past research in order, among other purposes, to help subsequent researchers understand phenomena in a world characterized by uncertainty and complexity. In this understanding, a wide range of policies, administrative processes, and other social and technical phenomena interact and shape the data that are ultimately preserved in and made available from a data archive. Researchers engage in activities to configure themselves—specifically in my study through the acquisition of knowledge—to use the data for their research purposes.
2. An understanding that, because of a social-technical gap, there will always be a gap between the data that are available and the knowledge researchers desire about the original research. Consequently, there is an opportunity to enhance support for data reuse by making this gap, and how to design data "artifacts" and archives themselves in light of this gap, a focus of study.
3. An understanding of boundaries of data being contingent on social factors and situated knowledge (from Haraway). This understanding allows us to consider that different researchers set different thresholds on the knowledge they obtain about data when adapting to reach reuse equilibrium, and do so for different reasons. This notion leads to the premise of my research, which is that it is possible to study the ways researchers configure themselves to reuse data by examining where they set these thresholds and why.

The second component is the idea that researchers satisfice in the knowledge they obtain about data in light of integrated attributes of their inner and outer environments. The ideas underlying this component are:

4. A recognition of the usefulness of the conception of "inner" and "outer" environments in understanding adaptive systems, but also that the lines between these environments may not be as clear as Simon's theory makes out. In fact, there may be resonances and interactions between attributes of the environments that blur the lines of what might be considered to belong to either. This stance expands the range of factors that might be considered to influence researchers in the process of configuration to reuse data. Such expansion is necessary when evaluating factors that affect knowledge satisficing, since knowledge satisficing may take place in the context of substantive rationality (i.e., limited by aspects of the outer environment) or procedural rationality (i.e., limited by aspects of the inner environment).
5. An understanding of satisficing (in the context of both procedural and substantive rationality) both as a process of adjusting aspirations to an equilibrium state (which occurs in the context of procedural rationality) and as a process that can leave a gap between a known optimal state (desired knowledge) and sub-optimal state (actual achievement of knowledge).
6. An understanding of the investigation of this gap between desired and achieved states in the context of satisficing and in the adaptive process more generally as a component of the social-technical gap in data reuse.

My identification of gaps in the reuse literature, the ways I sought to address them (see section 2.7.4), and my research questions all derived from this framework and its underlying ideas. I provide my research questions below for reference, and then explain the relationship between each of the gaps identified in section 2.7.4 and the research questions.

1. How do researchers determine the boundaries of the knowledge they obtain about data in order to reuse them in their research?
    a. What knowledge about data is most important to researchers to reach reuse equilibrium and why?
    b. How do researchers determine how much of what kinds of knowledge is enough?
    c. What do researchers report influences these determinations?
    d. How do researchers obtain the knowledge they desire?
2. Do researchers satisfice in the knowledge they obtain about the data they reuse?
    a. If so, how can researchers' satisficing be characterized?
    b. What factors are associated with knowledge satisficing?
    c. What reasons do researchers give for why they satisfice?
    d. What is the perceived impact of knowledge satisficing:
        i) On the outputs of research?

ii) On researchers' achievement of their goals for reusing data?

e. What do researchers believe could mitigate knowledge satisficing?

### *2.8.2 Gaps in the Literature*

**How researchers bound the knowledge they obtain in order to decide to reuse data (including whether or not researchers satisfice in obtaining desired knowledge).** This is the primary phenomenon I am investigating and proceeds directly from ideas 1-3, 5 and 6 in the framework above. If data reuse is in fact an adaptive process and the attainment of knowledge is one way that researchers adapt, I propose that a deeper understanding of the way researchers determine the thresholds of knowledge about data that are sufficient to reuse data (how they bound data to reach reuse equilibrium) can give insight into that process. I have chosen to investigate knowledge satisficing as a possible means of characterizing knowledge bounding.

**The relative priority that researchers assign to different types of knowledge about data in particular reuse instances, and why.** Understanding the kinds of knowledge that are most important to researchers provides a context for investigating knowledge bounding and knowledge satisficing. In a larger conversation about design of data "artifacts" and support for data reuse it is important to understand whether knowledge bounding and satisficing occur in areas of knowledge that are important to researchers. If they do not, there is likely little to be gained by designing solutions in those areas. Questions of how researchers bound knowledge and whether and in what ways they satisfice are reflected in research questions 1, 1.a, 1.b, 2, and 2.a.

**Factors that affect knowledge bounding and knowledge satisficing.** I propose that if we are to improve the design of archived data "artifacts," we must better understand the reasons why researchers set thresholds of knowledge about data the way that they do. The fourth idea in the framework in section 2.8.1 provides a conceptual basis for leaving the scope of influencing factors as open as possible in order to consider improvements to the design of data "artifacts"

along as many dimensions as possible. Questions about the factors that affect knowledge bounding and knowledge satisficing are reflected in research questions 1.c, 2.b, and 2.c.

**The impact of knowledge satisficing on the outcomes of research and researchers' attainment of their researcher goals.** Similar to above, if researchers satisfice in the knowledge they obtain, but satisficing does not ultimately affect the outcomes of their research, there are likely other areas where bolstering support for data reuse would be more beneficial. The impact of satisficing on research outcomes is thus important to assess. Questions about the impact of knowledge satisficing are reflected in research questions 2.d.i and 2.d.ii

**Comparability and Generalizability.** I use several strategies to enhance capabilities for comparing knowledge bounding and knowledge satisficing across diverse instances of data reuse, and for generalizing the behavior I observe in my study to data reuse more broadly. These strategies relate to my theoretical framework to the extent that they might support its applicability to a wider range of circumstances than those I am investigating. If they do not support such applicability, they could point to ways that the framework needs to modified, either in light of my specific findings, or in order to be apply to different reuse circumstances and scenarios. I do not have specific research questions about comparability or generalizability. Rather, I seek to address these gaps through strategies in my methodology, which I discuss in Chapter 3.

### *2.8.3 Connections With Research Questions*

In order to investigate my research questions, I sought a methodology that would be able to detect knowledge bounding and knowledge satisficing, and also a wide range of factors that might affect these, whether the factors belonged to the inner environment, outer environment or resulted from interactions between them. To address the complexities involved, I developed two

phases of research: one quantitative and one qualitative. In the quantitative phase (a survey), I operationalized the concept of satisficing to examine whether and to what extent it might occur in my study population. I also gathered information about specific aspects of researchers (e.g., involvement in the original collection of the data, experience with reuse) and their research circumstances (e.g., their research motivation and research process) that were found in the literature to impact decisions about data reuse. I did this to both characterize and analyze satisficing in the survey population and to facilitate the selection of interview subjects in the second phase of my research.

In the second phase, I used qualitative methods, including interviews and background research, to achieve my goal of examining the widest possible range of factors that could affect knowledge bounding and knowledge satisficing. To achieve my goal, it was not as important to me to understand which factors were associated with which environment (inner or outer) as it was to have a protocol that could detect them. In this phase, I asked study participants (selected from survey participants in the quantitative phase) to reflect more deeply on their reuse experiences in a less structured setting.

A crucial aim I had initially in this phase was to gain a variety of perspectives and experiences from groups of researchers who all reused the same data. I saw this as being important both to the comparability and generalizability of my research and to collecting data that allowed me to engage with aspects of my theoretical framework that spoke to the design of data "artifacts" and repositories (Simon) and ongoing conversations about how we should proceed in this design (Haraway). As I explain in the methodology, the significance of researchers reusing the same data was less important to my findings than I first imagined.

However, I do not believe this took away from the comparability or generalizability of the study, or my ability to engage with my theoretical framework.

In the section below, I discuss the specific hypotheses I investigated in the quantitative phase of my research and the bases for those hypotheses. Then, in chapter 3, I go into specific details about the quantitative and qualitative phases of my research and how I integrated results between them.

## 2.9 Research Assumptions, Hypotheses, and Analytical Model

### *2.9.1 Research Assumptions and Hypotheses*

My study was based on two main assumptions drawn from the reuse literature. The first was that, to the extent that data and their context are essential to data reuse and deeply entangled (e.g., Dicks et al., 2006; Bishop, 2006; Anderson and Sørvik, 2013), they can be considered together to constitute the totality of "data" about which researchers seek to obtain knowledge. The second assumption was that researchers make decisions about how to bound the data they select to reuse in their research. They determine which data to reuse, and how much about the data (both "data" and "context") is "enough" to know in order to reuse them (e.g., Curty, 2016; Faniel et al., 2013; Huggett, 2018; Niu, 2009a).

These assumptions were the basis of inquiry for both the quantitative and qualitative phases of my study. Importantly, they undergirded the hypotheses I investigated in the quantitative phase, which I explain below. I formulated the hypotheses in relation to nine concepts comprising three dependent variables and six independent variables. I also included three control variables in the testing of the hypotheses, which I discuss after explaining the hypotheses. The nine concepts were:

- Dependent variables
    - knowledge satisficing

    - impact of satisficing on research outcomes
    - attainment of data reuse goals
- Independent variables
    - whether the data were produced through quantitative or qualitative means
    - researcher "distance" from the data
    - researcher experience with data reuse
    - researcher data reuse ability
    - researcher motivation for reusing data
    - research process

The control variables were

- the number of prior citations to the data
- researchers' beliefs about whether the data were created with the purpose of being reused
- whether the reused data were obtained from ICPSR or a different source

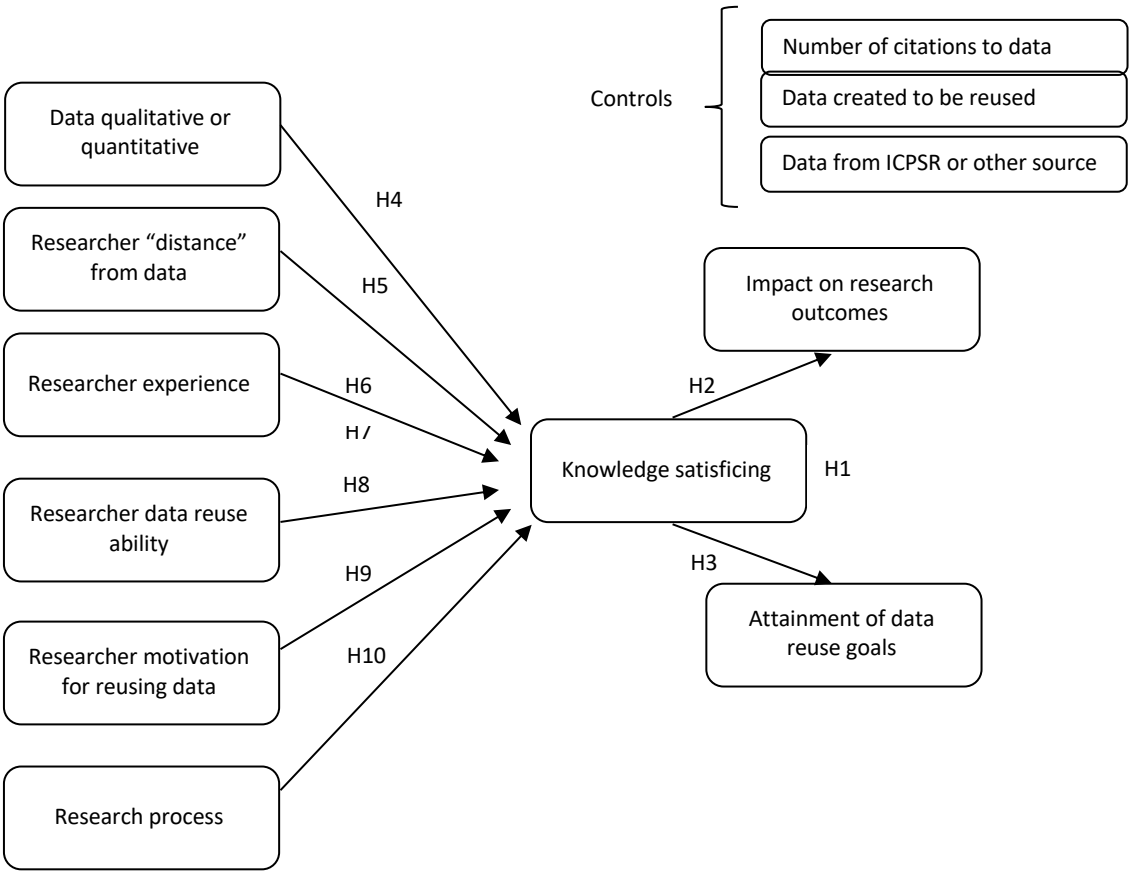
The independent variables were dimensions of data reuse that I anticipated from my review of literature would significantly affect the extent and nature of knowledge satisficing. They, and the control variables, were also factors that had the potential to vary significantly from one instance of data reuse to the next, and were therefore important to account and control for in order to achieve the most accurate interpretation of my results.

I selected these factors to measure in the survey phase, as opposed to other factors discussed in the literature (i.e., having to do with perceptions and beliefs, assessments of data, or social factors) because the selected factors lent themselves more readily to examination in the context of a survey. I discuss considerations in formulating survey questions more fully in section 3.3.1.3. In short, however, based on the literature and my testing of the survey instrument, I believed that the survey questions I formulated (e.g., related to "distance," motivation, experience, and reuse ability) were ones that could be answered readily and accurately by survey respondents and that respondents' answers could be easily interpreted. Questions I believed or had found to be more complicated to explain to researchers, to answer, or

to interpret answers for (e.g., related to perceptions and beliefs, assessments, and social factors), I addressed in the qualitative interviews.

I describe each of the nine concepts and associated hypotheses in the sections below and illustrate them in Figure 1.1.

*Figure 1.1 Relationship Between Research Concepts and Study Hypotheses*




*Note*. Research Model showing the relationship between the nine concepts and associated hypotheses to be investigated in the quantitative phase of the proposed research.

#### *2.9.1.1 Satisficing*

According to Simon (1994), humans do not necessarily seek optimal solutions to problems, but rather acceptable solutions. He calls this tendency toward acceptable solutions "satisficing." I proposed that in the context of knowledge satisficing in data reuse, it could be measured through a comparison of aspirations in relation to achievements. In light of research findings detailing the challenges researchers must overcome when reusing data (particularly those related to the researcher's "distance" from the data, discussed in section 2.9.1.5 below) and findings that it can be difficult for researchers to obtain data that are ideal for fulfilling their research goals (e.g., Pine, 2016; Faniel et al., 2016; Stvilia et al., 2015; Niu, 2009b), I hypothesized that researchers did satisfice in the knowledge they obtain about data.

*H1*. Researchers satisfice in the knowledge they obtain when determining whether to reuse data.

I operationalized the concept of satisficing in my survey through five questions. The first question was whether, when the researcher was considering reusing data, there was additional knowledge about the data they desired. The second was whether, when the researcher made the decision to reuse the data, there was knowledge about the data they desired but were not able to obtain, or obtain only in part. I asked the first question because if the researcher selected data precisely because they knew everything about the data that they wanted, they would not report satisficing in the second question. The first question therefore established a baseline for interpreting the second question. The third question asked about the three most important things the researcher would have liked to know about the data but was not able to learn or learn to the desired degree. The fourth asked how much of each type of desired knowledge was obtained on a scale of 0 to 100, and the fifth asked about the importance of the knowledge that was not obtained or that was limited to deciding to reuse the data.

The question areas related to the concept of satisficing—and all other concepts—are represented in Tables 3.7 and 3.8. Full questions are included in Appendix A.

#### *2.9.1.2 Impact of Knowledge Satisficing on Research Outcomes*

Since evidence of satisficing, as operationalized, would indicate that researchers did not know everything they would have liked about the data they reused, I hypothesized that knowledge satisficing had a negative impact on research outcomes.

*H2*. Knowledge satisficing has a negative impact on research outcomes.

I operationalized the impact of knowledge satisficing on research outcomes through a question that asked researchers to indicate on a five-point scale from "not at all" to "extremely" whether a lack of knowledge negatively impacted the outcomes of their research. I did not give specific guidance to respondents on what constituted a research outcome. I was most interested in whether there was an effect and the magnitude of the effect. In order to help interpret this response, I included an additional question that asked researchers why a lack of knowledge did or did not affect the outcomes of their research.

#### *2.9.1.3 Impact of Satisficing on Attainment of Data Reuse Goals*

Similar to the impact on research outcomes, and for a similar reason, I hypothesized that knowledge satisficing had a negative impact on researchers' ability to attain their goals in reusing the data.

*H3*. The probability of researchers' attaining their goals for reusing data is lower in the presence of knowledge satisficing.

I operationalized the impact of knowledge satisficing on research outcomes through a question that asked researchers to indicate on a five-point scale from "much worse than expected" to "much better expected" for each of their expressed reuse purposes, how well reuse of data met their reuse goals.

These three hypotheses encompassed my expectations for the dependent variables in my study. I discuss the remaining six concepts—all independent variables that I intended to use to characterize satisficing behavior, if it existed—and my corresponding hypotheses in the sections below.

#### *2.9.1.4 Data Produced Through Quantitative or Qualitative Methods*

I investigated satisficing in cases where the data were gathered in either quantitative or qualitative research because of differing assumptions and empirical findings in the literature about the knowledge needed to interpret and reuse data produced using these different methods (e.g., Mauthner et al., 1998; Savage, 2005; Broom et al., 2009; Niu, 2009a). I obtained information about the methods used in each data study from the ICPSR Bibliography of Data-related Publications (ICPSR, 2019b).

Based on my understanding from the literature that challenges in overcoming "distance" are in evidence in both qualitative and quantitative research, I hypothesized that the methods used to collect or produce data was not positively or negatively associated with knowledge satisficing.

*H4*. There is no association between knowledge satisficing and reuse of qualitative or quantitative data.

#### *2.9.1.5 Researcher "Distance" from the Data*

Researchers have identified the data reuser's "distance" from reused data as a significant factor in conducting research with secondary data (e.g., Mauthner et al., 1998; Medjedović, 2011; Niu, 2009b; Heaton, 2004). This "distance" is most often characterized as a "problem of not having 'been there'" (Heaton, 2004, p. 58), referring to the privileged position researchers have in interpreting data when they were intimately involved in the data's collection. While some researchers see "distance" as an inhibitor to reuse, others view "distance" as a positive factor that can enable more objective analysis (e.g., Weaver and Atkinson, 1994, cited in Heaton, 2004; Mason, 2007). By either interpretation, "distance" is a factor that researchers encounter and must negotiate in some way when reusing data.

Based on the importance of contextual information to reuse of data from quantitative research (Carlson and Anderson, 2007; Ribes and Jackson, 2013; Rung and Brazma, 2012), I believed that researchers could be affected by issues of "distance" as much in quantitative data reuse as in qualitative and thus included it as a concept in my research. I further hypothesized that where there was more "distance" between researchers and the data reuse, there was a higher likelihood of knowledge satisficing.

*H5*. The greater a researcher's "distance" from data, the higher the probability of knowledge satisficing.

I operationalized the concept of "distance" primarily through two questions: one that asked about the researcher's involvement in the original research and another that asked about the knowledge of the respondent's research team (including the respondent) about aspects of the original data collection that were relevant to reusing the data. Through this second question, I included the possibility that researchers might not have been involved in the original research but

might nevertheless have had important knowledge about the original data collection (e.g., from being a close associate of someone who was) that was relevant to the data's reuse.

#### *2.9.1.6 Researcher Experience with Data Reuse*

Research has shown that data reusers leverage knowledge gained from their own experience conducting primary research (e.g., Borgman, 2007; Zimmerman, 2007; 2008) and their previous experience reusing data (Niu, 2009b) when evaluating data for reuse. On this basis, I included questions related to the concept of experience. There is overlap between the concepts of "distance" and experience that arises, for instance, because knowledge that a researcher obtained from their own experiences collecting data could help interpret data collected by others (Zimmerman, 2007) and thus bridge the "distance" between the research and reused data that was lacking from not "being there."

I kept the two concepts separate at least initially, however, because knowledge gained from experiences that allowed a researcher to bridge "distance" could be distinct from knowledge about data gained from having actually "been there" or obtaining a high level of knowledge in some other way.

I operationalized experience through two sets of questions. First, I asked about the researcher's departmental affiliation, primary domain of research, and years of experience in their primary domain of research. I also asked whether the research being reported on was conducted in the researcher's primary domain of research, and in what domain of research the reused data were produced. These questions were intended to provide indicators of how much a researcher might have been going outside their own experience (e.g., domain of expertise)—thus potentially needing to fill gaps in knowledge related to terminology, methodology, or other domain-specific conventions—in order to reuse the data.

The second set of questions were about the researcher's experience with data reuse. Prior research has investigated data reusers' experience and expertise using measures of professional status (Niu, 2009b; Kriesberg et al., 2013) and the number of projects or articles a researcher has worked on or published (Faniel et al., 2012). Because I was unsure which measures of experience might be most relevant to knowledge satisficing, I included both of these measures in my survey and added measures of frequency of reuse and years of reuse experience as well. In sum, then, the second set included questions about the researcher's professional position, years of experience reusing data produced in the same domain as the reused data, years of experience reusing data produced in any domain, frequency of reuse, and frequency of reuse for the same purpose as the data being reused. I also asked about the number of reuse projects the researcher had worked on, the number of these reuse projects where reuse considerations were substantially similar to ones involved in reusing these data, the number of authored or coauthored published papers describing research involving data reuse, and the number of authored or coauthored published papers describing research involving data reuse where conditions were substantially similar to those involved in reusing these data.

My hypothesis in both cases was that the more experience a researcher had, the less knowledge satisficing would be in evidence.

*H6*. The more experience a researcher has in their primary domain of research, the lower the probability of knowledge satisficing.

*H7*. The more experience a researcher has with data reuse, the lower the probability of knowledge satisficing.

#### *2.9.1.7 Researcher reuse ability*

Much research has linked the importance of data reusers' training and skills to their ability to reuse data (Niu, 2009a, 2009b; Corti and Bishop, 2005; Curty, 2016; Faniel et al., 2013; Zimmerman, 2008). I viewed training and skills as highly related to experience, but included an additional concept for analysis—reuse ability—to account for the possibility that there was knowledge researchers gained from past experience that was not related either to the length of time or frequency with which they reused data. To operationalize reuse ability, I asked researchers to characterize their level of ability in reusing data produced both inside and outside their primary domain of research. I asked them to make this characterization on a scale from beginning (just learning about considerations in conducting research with secondary data) to intermediate (comfortable conceiving of and executing research using secondary data) to advanced (advanced knowledge and experience conceiving of and executing research using secondary data). My hypothesis, similar to experience, was that reuse ability was inversely associated with knowledge satisficing.

*H8*. The greater a researcher's perceived ability to reuse data, the lower the probability of knowledge satisficing.

### *2.9.2 Researcher Motivation for Reusing the Data*

Under the concept of researcher motivation I included both the purpose(s) for which a researcher reused data and how important it was—at the time the researcher decided to reuse the data—to reuse the data for that purpose (or purposes, if they were multiple). I included researcher motivation as a concept based on the substantial amount of research that has found a dependency between the kinds and amounts of contextual information sought by researchers (i.e., what researchers seek to know) and the purpose of reuse (Andersson and Sørvik, 2013;

Huggett, 2018; Bishop, 2006; Bishop and Kuula-Luumi, 2017; Faniel et al., 2013; Jackson et al., 2007; Medjedović, 2011; Fear and Donaldson, 2012; Rolland and Lee, 2013). The specific purposes I investigated in the survey were derived from the literature described in section 2.3 and through testing I conducted on my survey instrument. They included:

- to provide background or context for the research (e.g., to develop a questionnaire or obtain calibration information)
- to validate or corroborate research results
- to answer a different question from the original research
- to replicate or reproduce the original research
- to combine with other data
- to compare with other data
- to test a theory
- to develop or test an algorithm or tool
- other (please specify)

Although the purpose of the research has been shown to influence the type and amounts of context researchers seek, my hypothesis was that none of the research purposes had an impact on knowledge satisficing. I made this hypothesis based on my reasoning that researchers were not limited in obtaining knowledge about data simply by the reason they sought to use it. The way researchers bounded the knowledge they desired about data might be affected by research purpose, but there did not seem to be a reason why the ability to obtain the desired knowledge should.

*H9*. There is no association between researcher's motivation for reusing data and knowledge satisficing.

### *2.9.3 Research process*

My theoretical framework accounts for variations described in the literature in the way researchers arrive at their research questions (Doolan and Froelicher, 2009; Bishop, 2007)—specifically that researchers might change their research questions after their research had begun

or develop questions as they explored available data. These findings were supported in a pilot survey I conducted of 114 researchers in November, 2019. Of 43 respondents, 34 indicated that they either changed their initial research questions based on knowledge they obtained about data or determined their research questions as they explored the data. My hypothesis was that the probability of knowledge satisficing would be lower when researchers changed their research questions or developed research questions as they explored data. This was because the chance to proactively change the knowledge about data that was desired gave researchers the possibility of intentionally minimizing the amount of desired knowledge that was limited or unobtainable.

*H10*. When research questions change or develop over time there is a lower probability of knowledge satisficing.

I included as part of the research process concept the source that was most important to researchers in obtaining desired knowledge about data. I did not have expectations about which sources might be more or less associated with knowledge satisficing, but asked about this parameter on the possibility that associations might exist. The sources were drawn from literature (see section 2.4.3) and include:

- original data creators
- colleague(s)
- advisor(s)
- data repository staff
- data documentation
- the data themselves
- personal knowledge
- team knowledge
- literature
- other (please specify)

### *2.9.4 Control Variables*

I used three control variables in the statistical analyses. These were the natural log of the number of citations (in my sample) to the data study that was reused, researchers' beliefs about whether the data were created with the purpose of being reused, and whether the reused data were obtained from ICPSR or a different source. I used these variables as controls because regardless of what other factors were involved, they could affect whether the researcher was able to obtain knowledge they desired (Heinze et al., 2018; Wojtkiewicz, 2016). For instance, if the data were well cited (and therefore well used) there could be more information about them in circulation, or an accepted threshold of the amount of knowledge about the data that was needed to reuse them. More information or knowledge could similarly be available if the data were created to be reused (e.g., a government-run national survey) in contrast with data that were not. Finally, differences in curation practices at ICPSR and other sources from which researchers obtained data could result in different amounts of desired knowledge being obtained. These rationales are similar to those used by Curty et al. (2017) in the selection of control variables for their study on attitudes and norms that affect data reuse.

I obtained information about the frequency of data citations from the ICPSR Bibliography of Data-related Publications (ICPSR, 2019b). I gathered information about whether the data were created to be reused and the source of the data from the survey.

I turn now to describing the research design I used to conduct my study of knowledge bounding in data reuse.

## Chapter 3 Research Design

In this chapter, I describe the mixed methods study I conducted to measure the ways researchers bound the knowledge they obtain about data in order to reuse them as evidence in their research. A fundamental aim of my study was to explore whether and how information about the ways researchers bound knowledge about data can inform current efforts to archive and prepare data for reuse.

The chapter is separated into eight sections. In section 3.1 (Overview of Research Design) I give an overview of the mixed methods study, including a description of the goals of the quantitative and qualitative phases. In section 3.2 (Mixed Methods) I define mixed methods research, describe its benefits, and explain why I have selected it for my research project. I also include a table (Table 3.2) of the quantitative, qualitative, and integration phases of the study. The phases depicted in Table 3.2 serve as an outline for the subsequent parts of the chapter.

In section 3.3 (Phase 1. Quantitative Study) I describe data collection and analysis procedures for the quantitative phase. I include information on the survey population, sampling strategies I employed, concepts I operationalized in the survey, survey hypotheses, considerations in the development of the survey instrument, and survey analysis strategies.

Section 3.4 (Phase 2. Qualitative Study) mirrors section 3.3 in many ways, covering data collection and analysis procedures for the qualitative interviews. I describe the development of the interview protocol and how I used results from the survey to finalize the protocol and select interview participants. I also give details about the conduct of the interviews and procedures and strategies I used to analyze the results.

In section 3.5 (Integration of Survey and Interview Results) I explain how I integrated the results of the survey, interviews, and background research and triangulated among them. Section 3.6 is a detailed discussion of how I strove to ensure the inference quality (validity) of my research. In section 3.7, I describe the connections between the particular data I collected and my research questions and theoretical framework, and section 3.8 is the chapter conclusion.

### 3.1 Overview of Research Design

I completed a sequential mixed methods study (Creswell et al., 2003) conducted in two phases: one of quantitative data collection and analysis and one of qualitative data collection and analysis. The purpose of the quantitative phase (a survey of researchers who have reused data) was to assess 1) whether researchers satisficed in the knowledge they obtained about data they reused, 2) the impact of satisficing on research outcomes, and 3) how well researchers' experiences reusing data met their expected goals. I also sought to characterize knowledge satisficing, if it was in evidence. These aims correspond to research questions 2, 2a, 2b, and 2d. I made these assessments in light of numerous attributes of researchers and their research including the researchers' purpose(s) in reusing the data, how the research questions were developed, researchers' "distance" from the data, their experience with data reuse, and their ability in reusing data. An additional purpose of the quantitative phase was to gather information about the attributes mentioned above to use in the selection of researchers to interview in the qualitative phase. More information about this and the qualitative phase is given below.

The aims of the second, qualitative phase (consisting of interviews and background research) were dependent to some degree on the results of the quantitative phase. I knew that whether or not there was statistically significant evidence of satisficing in the results of the survey, the interviews would focus on understanding and characterizing:

- the kinds of knowledge about data that were important to researchers;
- how researchers determined how much of what kinds of knowledge were enough to decide to reuse the data;
- what researchers reported influenced these decisions; and
- how researchers obtained the desired knowledge.

I intended, if there was statistically significant evidence of satisficing in the survey results, for the interviews to focus more deeply on understanding and characterizing the nature and extent of the knowledge satisficing that took place, why researchers reported that they satisficed, and whether and in what ways satisficing might have affected the results of the research or researchers' achievement of their reuse goals. If there was not statistically significant evidence of satisficing, I intended to still investigate these questions to some degree—in part to confirm whether satisficing was not in evidence or whether the survey was not effective in measuring it—but not at the same depth. As I discuss below, although about a quarter of researchers lacked knowledge about data that they desired, I found that satisficing did not well characterize researchers' behavior. For this reason, the interviews focused primarily on understanding how researchers bounded the knowledge they obtained about data, whether or not they lacked knowledge they desired about data. In keeping with my theoretical framework (see section 1.5), the basic design of the interview portion of my study was to interview clusters of researchers who all reused the same data study.

### 3.2 Mixed Methods

The term "mixed methods" has been used to describe a diversity of research designs involving multiple research phases and/or methods (Creswell, 2015). After conducting a study in 2007 where they identified 19 different definitions of mixed methods research from interviews with 21 researchers, Johnson, Onwuegbuzie, and Turner (2007) defined mixed methods research as:

> the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the purposes of breadth and depth of understanding and corroboration. (p. 123).

In practice, mixed methods are employed when researchers wish to gain a more comprehensive understanding of a phenomenon (National Institutes of Health Office of Behavioral and Social Sciences (NIH OBSSR), 2018), explore the results of quantitative studies more fully (e.g., Ivankova, Creswell, and Stick, 2006), test or attempt to falsify a hypothesis (Leckenby and Hesse-Biber, 2007), employ techniques of triangulation (i.e., use data gathered from multiple methods to answer a single research question) (Morse, 1991), or build or improve research instruments or interventions (NIH OBSSR, 2018; Creswell and Creswell, 2018). Tashakkori and Teddlie (2003) note that three advantages of mixed methods research are (a) it "can answer research questions that other methodologies cannot," i.e., those that seek to both generate and verify theory at the same time; (b) it "provides better (stronger) inferences" and (c) it provides the opportunity for presenting greater diversity of divergent views (Tashakkori and Teddlie, 2003, pp. 14-17). All of these advantages are applicable to my study, although with regard to the first, I am testing hypotheses and seeking to apply (and potentially extend) theory, rather than generating and/or verifying theory.

According to the NIH Office of Behavioral and Social Sciences Research (2018), mixed methods studies can differ in several dimensions, including:

- whether the quantitative and qualitative data are integrated for analysis or whether one set of data builds on the results of the other;
- whether the datasets are gathered at the same time or sequentially;
- whether the quantitative and qualitative data are given equal importance in the study;
- at what point the data are combined or integrated (during data collection or data analysis, or after data analysis, e.g., when results of quantitative analysis might be compared with themes arising from qualitative analysis)

In my research, the qualitative study came sequentially after the quantitative study and built on the results of the quantitative study. That is, the focus of the interviews was informed by the results of the survey and the interviews sought to elicit more information from respondents to the survey. However, the quantitative portion of my study investigated its own set of hypotheses and I integrated the results of the two studies at multiple stages. For instance, data about the nature and prevalence of satisficing informed the development of the interview protocol and selection of interview candidates. The results of the quantitative study also provided important context for understanding and interpreting the results of the interviews. Moreover, I used attributes of researchers and research projects obtained in the survey in the selection of interview candidates in the qualitative analysis.

In relation to Creswell, Plano Clark, Gutmann, and Hanson's (2003) description of six major mixed methods designs, the current study most closely fits the definition of a "sequential explanatory design" (p. 223). This type of design is characterized "by the collection and analysis of quantitative data followed by the collection and analysis of qualitative data" (Creswell et al., 2003, p. 223). A sequential explanatory design typically prioritizes collection and analysis of quantitative data. However, Creswell et al. note that,

> In an important variation of this design, the qualitative data collection and analysis is given the priority. In this case, the initial quantitative phase of the study may be used to characterize individuals along certain traits of interest related to the research question. These quantitative results can then be used to guide the purposeful sampling of participants for a primarily qualitative study. (p. 227)

The quantitative and qualitative methods used in the current study have been selected because of the special affordances each method has for investigating the primary research questions. Because of this, and because I integrated the results of the methods at multiple points

throughout the process, the quantitative and qualitative phases had more of an equal priority or weight compared with the typical designs described by Creswell et al. (2003).

Table 3.1 shows which research questions I investigated in each phase of the mixed methods study and Table 3.2 gives an overview of this dissertation, including phases of quantitative and qualitative data collection and analysis and integration of the methods. These phases are explicated in greater detail in section 3.3 and following.

*Table 3.1 Research Questions Investigated in Each Phase of the Mixed Methods Study*

| Research Questions | Investigated in Survey | Investigated in Interviews |
|---|---|---|
| 1. How do researchers determine the boundaries of the knowledge they obtain about data in order to reuse them in their research? | X | X |
| 1a. What knowledge about data is most important to researchers to reach reuse equilibrium and why? | | X |
| 1b. How do researchers determine how much of what kinds of knowledge is enough? | | X |
| 1c. What do researchers report influences these determinations? | | X |
| 1d. How do researchers obtain the knowledge they desire? | | X |
| 2. Do researchers satisfice in the knowledge they obtain about the data they reuse? | X | X |
| 2a. If so, how can researchers' satisficing be characterized? | X | X |
| 2b. What factors influence knowledge satisficing? | X | X |
| 2c. What reasons do researchers give for why they satisfice? | | X |
| 2d. What is the perceived impact of knowledge satisficing: | X | X |
| 2di. On the outputs of research? | X | X |
| 2dii. On researchers' achievement of their goals for reusing data? | X | X |
| 2e. What do researchers believe could mitigate knowledge satisficing? | | X |

*Table 3.2 Overview of Mixed Methods Used in This Dissertation*

| Phase | Procedure | Product |
|---|---|---|
| Quantitative data collection | Cross-sectional web-based survey<br>Stratified random sample<br>Quantitative and qualitative data<br>Data with high, medium, and low numbers of citations | Numeric and textual data |
| Quantitative data analysis | Frequencies<br>Logistic regression<br>Model development<br>Qualitative coding | Descriptive statistics<br>Inferential statistics<br>Model for lacking knowledge<br>Results of hypothesis testing |
| Integration (implications of quantitative results for qualitative analysis) | Examination of results and quantitative analysis | Interview protocol<br>Interview selection |
| Qualitative data collection | Background research<br>Maximum variation sampling<br>Interviews with clusters of researchers who reused the same data and researchers who reused different data | Textual data (interview transcripts and documents) |
| Qualitative data analysis | Qualitative coding<br>1st and 2nd cycle coding<br>Thematic analysis | Codes and themes<br>Cross-thematic analysis |
| Integration of quantitative and qualitative results | Interpretation and explanation of quantitative and qualitative results | Discussion<br>Implications<br>Future research |

*Note*. The overview is based on the diagram in Ivankova, Creswell, and Stick, 2006, p. 16.

Ultimately, the study employed mixed methods

- to test hypotheses related to knowledge satisficing, the impact of knowledge satisficing on research outcomes, the association between knowledge satisficing and researcher's achievement of their goals for reuse, and a variety of factors that can be used to characterize knowledge satisficing (see sections 2.9 and 3.3.2)
- to explore quantitative results more fully (i.e., to better understand the nature and extent of satisficing, or of knowledge bounding if satisficing is not prevalent, through qualitative interviews);
- to aid in the selection of interview candidates;
- to build and improve a research instrument (the results of the survey informed the design of and questions asked in the interviews); and
- to triangulate results (e.g., to triangulate and validate responses researchers gave in the survey and interviews).

The sections below describe the proposed procedures for collecting, analyzing, and interpreting data in each phase of the research outlined in Table 3.2.

## 3.3 Phase 1. Quantitative Study (Survey)

### *3.3.1 Data Collection*

#### *3.3.1.1 Survey Population*

The survey population (from which interview subjects were also selected) was composed of researchers who reused quantitative or qualitative data stored in the Inter-university Consortium for Political and Social Research (ICPSR) data repository between January 2014 and September 2019. ICPSR is one of the oldest and largest archives of data resulting from social science research in the world (ICPSR, 2019a). It was founded in 1962 and is supported by a membership of nearly 800 institutions worldwide (ICSPR, 2019a). I chose a population of researchers who reused data in ICPSR for several reasons.

First, ICPSR maintains a Bibliography of Data-related Literature (ICPSR, 2019b) listing citations of scholarly sources that have cited data archived in ICPSR. While determining whether data that are cited in a published work were actually reused can be difficult (Parsons et al. 2010, Mayernik 2012, Mooney and Newton 2012), ICPSR undertakes this work with a high degree of reliability (E. Moss, personal communication, 9/6/2019; also confirmed in results of my pilot survey). The Bibliography is very large (many studies have reused ICPSR data), likely owing to the large amount of data stored in ICSPR and its reputation as a trusted source of social science data. As of September 14, 2019, when a sample of data from ICSPR was obtained for this study, the Bibliography listed 12,222 citations to 3,081 distinct data studies between January 2014 and September 2019. A data study in this case is either a one-time study or a distinct collection of data within a larger series (for instance, the National Survey on Drug Use and Health from 2014).

The Bibliography is also well-curated, with detailed information about the data studies that are cited, the sources that cite them, and the sources and strategies used to assemble the Bibliography itself (ICSPR, 2019b). For the purposes of my research, this detailed information facilitated targeted sampling of the survey population and granular analysis of survey results. For

example, the Bibliography includes information about the types of sources where citations are found (see Table 3.3) and about the data and methods used to collect them (e.g., survey data, administrative records data, clinical data, etc.). The information about citation sources makes it possible to exclude citation sources that are less likely to be original reports of data reuse or part of the published scholarly record (i.e., Documents, Audiovisual Material, Newspaper Articles, and Magazine Articles; I refer to formats I have included in my study as "eligible formats" below).

*Table 3.3 Count of Citations for Each Type of Source Citing Data in ICPSR*

| Type of source | Count |
|---|---|
| Journal Article | 8356 |
| Thesis | 1248 |
| Report | 947 |
| Conference Proceedings | 940 |
| Book Section | 313 |
| Document | 191 |
| Book | 90 |
| Electronic Source | 60 |
| Audiovisual Material | 33 |
| Newspaper Article | 32 |
| Magazine Article | 12 |

*Note*: The data are for sources citing data in ICPSR between January, 2014 and September, 2019.

Third, based on evidence about the bibliography development processes, I ascertained that selecting cases of reuse reported in the Bibliography could reduce the chance of variation in the survey results that might come from researchers obtaining their data from different data archives (also known as environmental variation; see Eisenhardt, 2002 and Benbasat et al., 1987). I was concerned, for instance, that if researchers reusing the same dataset obtained their data from different sources, it could be difficult to tell what impact, if any, different curation practices at different archives had on knowledge satisficing as distinct from variables measured in the survey. As it turned out, my findings were that 44% of researchers reused data from a

source other than ICPSR so the selection of ICPSR did not meet my intended goals. I talk about this more when presenting the research results.

Fourth, while ICPSR curates data produced in the conduct of social science research, the data represent work conducted in a variety of disciplines (e.g., economics, education, history, health care, and law) covering a breadth of topics including "elections, education, mental health, criminal justice, aging, gender, race, religion, medical care, war, family, and geography" (ICSPR, 2019b). This meant that I could leverage the substantial advantages that I believed would come from studying reuse of data from a single source (e.g., having a large set of identified and well documented cases of reuse), while maintaining some degree of cross-disciplinary analysis, which was important to my study. While it turned out that a substantial number of researchers reused data from other sources, I still obtained the interdisciplinarity I had hoped for.

Fifth, I thought that limiting the scope of disciplines included in my study would increase the accuracy of my analysis. The greater the diversity in the disciplines included in my study, the more difficult it would be to know whether I was interpreting data gathered from respondents correctly. I had strategies in place to address this (e.g., quantifiable measures of knowledge satisficing in the survey and background research in the qualitative phase), but I considered that placing some bounds around the disciplines I examined would help me gain greater familiarity with the populations I was studying and aid in identifying commonalities, trends, and differences in both the survey and interview portions of my study.

Sixth, as evidenced in the preceding literature review, there is a significant body of research on data reuse in the social sciences. This was valuable for several reasons. First, the existence of this literature allowed me to gain additional familiarity with research in the social

sciences that helped in the interpretation of results. Second, in some cases I was able to compare the results of my work with the results of other studies for triangulation and validity, especially in regard to the way researchers determined how much knowledge was "enough." Finally, I was able to draw on the large amount of prior work when making inferences and drawing conclusions about the results of my study.

Lastly, to the extent that my research included researchers who reused data from ICPSR (as opposed to a different source), I considered that it might produce findings and conclusions that ICPSR could use to improve its support for researchers seeking to reuse data.

The researchers I selected from the population for the survey were in most cases the first author of a work in an eligible format. There were some cases where a first author indicated a different author was a more appropriate contact (in which case I contacted that person). The sample included authors of eligible works in the Bibliography created between 2014 to 2019 and only included works if they were listed in the Bibliography as citing at least one individual data study. I excluded works that cited a data series only (and not an individual data study that was part of the series). I only included works that cited specific data studies because a key aim of my research was to study the experience of different researchers who reused the same data. Since it was not clear which particular study researchers used when only a series was cited, I did not include these.

Of the 11,954 works authored between January 2014 and September 2019 that were in eligible formats and cited data in ICPSR, 837 did not cite individual data studies. Overall, then, the survey population included authors of 11,117 works that cited 3,047 distinct data studies. Considering only unique authors, and selecting only one data study per authored work, there were 7,190 unique authors (and, therefore, unique works) and 1,557 distinct data studies. In their

studies, authors frequently reused more than one data study. When this was the case, I selected one data study at random from the several that were reused to associate with the author and work. In cases where researchers were first authors of multiple works, I selected only the most recent work. 7,063 works cited at least one of the 1,430 distinct quantitative data studies and 127 works cited at least one of the 40 distinct qualitative data studies.

To test these assumptions, I conducted a pilot study in November 2019. The pilot helped me develop and test my procedures for selecting participants. It also helped me test the effectiveness of my questions and the applicability of my theoretical framework, as well as confirm the reliability of the ICPSR Bibliography in identifying instances of data reuse) and develop hypotheses (e.g., surrounding how researchers arrive at their research questions).

While fulfilling these functions, the pilot also reduced the number of researchers I could contact in the main survey. In the pilot, I surveyed 114 researchers (103 who cited quantitative data and 11 who cited qualitative data). After subtracting these from the pool, the survey population for the main study comprised unique authors of 7,076 works (7,190 minus 114) who cited 1,469 unique data studies. 6,960 of these works cited at least one of 1,430 unique quantitative data studies cited in the population of works and 116 cited at least one of 39 qualitative data studies cited in the population (one of the qualitative data studies in the pilot was uniquely cited by one of the surveyed authors).

#### *3.3.1.2 Survey Sample*

For the dissertation, I selected participants for the survey using stratified random sampling (Kemper, Stringfield, and Teddlie, 2003) from the survey population. I stratified by the type of data that researchers cited (i.e., quantitative and qualitative) and by the number of times data studies were cited. For the purposes of my study, I considered a "reuse instance" to be an

instance where a unique participant (author/researcher) reused data held in ICPSR and produced a work in an eligible format based on their reuse that was referenced in the ICPSR Bibliography (ICPSR, 2019b).

#### *3.3.1.2.1 Stratification by Data Type: Quantitative*

##### *3.3.1.2.1.1 Determination of Appropriate Sample Size*

I used Yamane's formula (see Formula 1) to determine the number of responses needed from authors who cited quantitative data (Yamane, 1967, cited in Israel, 1992, p. 3) to measure the rate of satisficing behavior (lacking desired knowledge at the time the decision to reuse the data was made) in the population with high confidence and low error. This formula, which is designed to calculate sample sizes where the dependent variable of interest is dichotomous, considers three variables: the estimated degree of variability of the attribute being measured in the population, the desired level of confidence and the desired level of precision or error. The degree of variability refers to the distribution of the measured attribute in the population. The formula assumes maximum variability (50% or 0.5), a 95% confidence level, and I chose to perform calculations with a 5% error. To assume maximum variability yields the most conservative (largest) sample size needed to achieve 95% confidence and 5% error. This was appropriate for my study because I was unsure of the prevalence of knowledge satisficing in the population of ICPSR data reusers. In the pilot survey, 24% of respondents reported satisficing but I believed satisficing may have been underreported due to the way the relevant questions were asked. I took steps to address this issue in the main project survey.

*Formula 3.1 Yamane's Formula for Sample Size Determination*

$$n = \frac{N}{1 + N(e)^2}$$

*Note*. Formula reproduced from Israel (1992), p. 4. $n$ is the sample size, $N$ is the population size, and $e$ is the level of precision. The formula assumes a 95% level of confidence and variability of the measured attribute of 50% (0.5).

Based on this formula and tables provided in Israel (1992), I determined that I would need to receive approximately 378 survey responses (given a population of 6,960 researchers) to determine the mean occurrence of satisficing in the population of ICPSR data reusers with 95% confidence and 5% error. Assuming a 20% response rate, this meant I would need to survey at least 1,890 researchers. I planned to survey at least 2,000 in order to hedge against a low response rate and potentially raise the confidence level and lower the margin of error. A 20% response rate would thus yield approximately 400 responses.

*3.3.1.2.1.2 Estimation of Power for Sample Size*

I used the powerlog STATA program (UCLA: Statistical Consulting Group, n.d.a) to estimate the sample sizes I would need in order to detect estimated effect sizes in the logistic regressions I would use to test my research hypotheses. The program takes four inputs:

P1: the probability of an outcome at the mean value of an independent variable

P2: the probability of an outcome at one standard deviation from the mean

Alpha: the significance level, or the probability of incorrectly rejecting a correct null hypothesis

R squared: "the squared multiple correlation between the predictor variable and all other variables in the model" (UCLA: Statistical Consulting Group, n.d.a)

I used different possible values of P1 and P2 and different values of R squared to determine the sample sizes that I would need to obtain a 0.80 level of power for different analyses. These are shown for selected hypotheses in Table 3.4. I based the probabilities loosely

on the results I obtained in the pilot survey. As the table shows, the sample size needed is larger when the overall probabilities are lower (e.g., compare the rows for H3 with H6), when the difference between the probabilities is smaller (compare H5(1) and H10), or the value for R-squared is higher (compare the semi-colon separated values in the final column). I used values of zero and .2 for R-squared because I did not know the relative effect that my independent variable in each hypothesis test would have on the dependent variable in comparison with the effect of the control variables in the regression.

I did not have an alternate hypothesis for two of my hypothesis tests (H4 and H9). I have included H9 in Table 3.4 to show that I would need a sample size of 573 to detect a decrease in the rate of satisficing from .25 to .20. This would be the case under any circumstances but in relation to H9 (where I do not expect to see a significant association) it means that if the effect size in my regressions were small enough, I would miss detecting a significant association between satisficing and reuse purpose.

According to my power estimations, I would need a sample size greater than 400 to have high power for detecting small changes in association or associations with phenomena that have low incidences in my sample. As I describe in the following section, I was able to sample more researchers and received more responses than I initially planned for, which increased the power of my analyses.

*Table 3.4 Samples Sizes Needed to Achieve .85 and .9 Levels of Power in Different Scenarios*

| Hypothesis | P1 | P2 | Alpha | Sample size for .80 power at R-squared of 0 and .20 |
|---|---|---|---|---|
| H3. Probability of attaining research goals is less in the presence of satisficing (high rate of attaining goals assumed for all reuse purposes: Background, Validate, New question, Replicate, Combine, Compare, Theory, Tool, Other) | .80 | .60 | .05 | 54; 67 |
| H5(1). Probability of satisficing is higher the lower researchers' involvement in the original research | .40 | .30 | .05 | 156; 195 |
| H5(2). Probability of satisficing is higher the less knowledge on the researcher's team about the original research | .90 | .8 | .05 | 53; 66 |
| H6. Probability of satisficing is lower the more experience a researcher has in their primary domain of research | .30 | .25 | .05 | 533; 666 |
| H9. No association between satisficing and reuse for different purposes (Background, Validate, New question, Replicate, Combine, Compare, Theory, Tool, Other) | .25 | .20 | .05 | 458; 573 |
| H10. Probability of satisficing is lower if research question changes or develops as the research progresses | .40 | .32 | .05 | 242; 302 |

*3.3.1.2.1.3 Quantitative Sample*

As of December, 2019, nine of the 6,960 data studies and their citing papers had been removed from the ICPSR Bibliography. I therefore began the full survey with a sample of 6,951 papers (and an equal number of researchers). I put the researchers in random order and a research assistant looked up an email address for each researcher until they compiled a list of emails for the 2,500 researchers (the email search process was faster than expected, allowing a greater number of researchers to be included in the sample). 76 of the emails bounced, 5 failed (there was a problem in the email address), and 14 were undeliverable.

Overall, I sent emails to 2,405 (2500-76-5-14) researchers in this stratum. 868 responded to the survey (36.1%). I excluded 73 responses where the researcher indicated they had not reused the data or left the question blank and 56 who reused the data but did not provide information about satisficing (Q14 or Q15). In conducting background research prior to the

interview phase of my research, I discovered three instances where researchers had not reused the data cited in the Bibliography and responded to the survey in relation to a different set of data. Since the data were not held in ICPSR, I omitted these instances from the results. I also omitted the response from a researcher who noted they did not cite the indicated study but responded to the survey in relation to more recent research with data from ICPSR. There were thus 739 (872-73-56-4) usable responses from this stratum. This is well over the 400 responses I calculated I would need to measure mean satisficing in the population with parameters of 95% confidence and 5% error. In fact, the number of responses exceeded the sample size required to achieve a 99% confidence level with 5% error (according to Cochran, this would be approximately 633) (Cochran, 1963, cited in Israel, 1992).

#### *3.3.1.2.2 Stratification by Data Type: Qualitative*

Because there were relatively few, I surveyed all of the researchers in the population who cited qualitative data. I sent emails in the end to 107 of the 116 researchers in this stratum (email addresses for nine researchers were not found). 36 researchers responded (34.2%). Of these, seven answered they had not reused the data and one who reused data did not respond to Q14. There were thus 28 usable responses. As noted above, in order to measure mean satisficing in this population with 95% confidence and 5% error, I would have needed approximately 90 responses. Since I did not obtain this many, I was not able to conduct inferential statistics on this sample alone. However, I combined this qualitative sample with the quantitative sample for the purposes of hypothesis testing (see sections 2.9 and 3.3.2).

I drew from these samples and the sample based on number of citations to the reused data (discussed below) to select a sample of researchers to contact for interviews (See section 3.4).

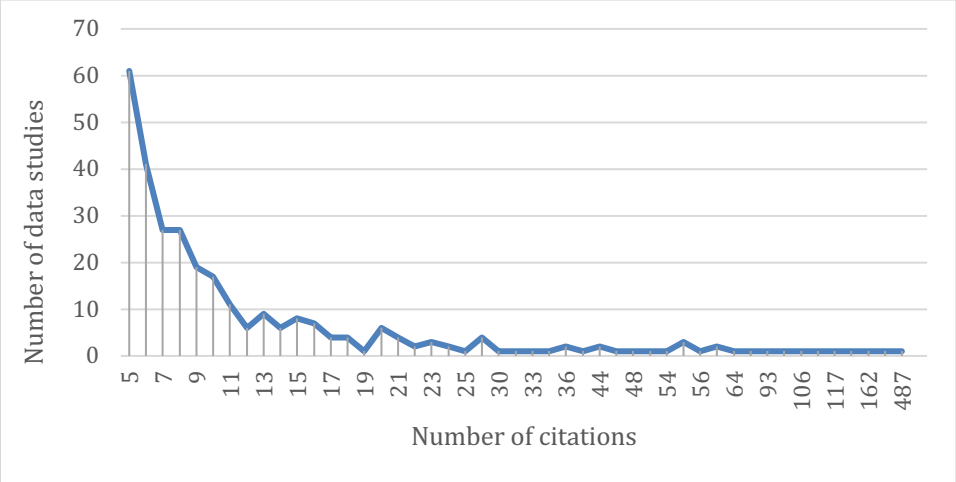
#### *3.3.1.2.3 Stratification by Citation Count*

In the interview phase of my study, I was interested in gaining in-depth perspectives about knowledge satisficing and the way researchers bounded the knowledge they obtained about data they reused. In order to accomplish this, I wished to interview researchers and research projects with a variety of attributes (such as the researcher's "distance" from the data and reuse experience, and the purpose for which the data were reused). I also wanted to investigate my phenomena of interest in relation to data with different levels of reuse. Perhaps, for instance, the way researchers approached data studies that were heavily reused was different from the way they approach data studies that were less well reused. To achieve this diversity of attributes at different levels of reuse, I divided researchers from the population of researchers who reused quantitative data into four groups based on the number of times the data they reused had been cited (I was already surveying all of those who reused qualitative data so I excluded them from this sample).

The first group included data studies with 5 to 24 citations, the second of studies with 25-50 citations, the third of studies with 51-100 citations, and the fourth of studies with more than 100 citations. I excluded data studies that had less than 5 citations because, assuming a response rate of less than half, my goal was to interview multiple researchers who reused the same data and I wanted to be sure I had enough responses.

Figure 3.1 shows a selected distribution of citation counts for data studies with five or more citations (in the survey population of those who reused quantitative data). Not shown (they would be plotted at the left of the figure) are 674 data studies that had one citation, 252 that had two citations, 129 that had three citations, and 75 that had 4 citations. These data and the figure illustrate that there were many data studies that had a lower number of citations and only a few data studies that had a higher number of citations.

*Figure 3.1 Distribution of Citations in the Population of Researchers Who Reused Quantitative Data*




I randomly selected 3 datasets from each citation tier. This yielded 12 data studies with a total of 783 citing papers. A research assistant found email addresses for 687 paper authors. Due to an error in the way the sample was drawn, 271 authors were duplicated across the sample stratified by quantitative datasets. I accidentally contacted these 271 researchers twice (once for each sample they were in) but only accepted one response per researcher (there was only one researcher who answered twice). Of the 687 total researchers contacted in this sample, 4 emails bounced and 9 emails were undelivered resulting in 674 researchers being contacted. 403 of these contacted (674-271) were unique to this stratum (they were not also in the sample stratified by quantitative data).

Table 3.5 details the data studies selected from each range of citations, the number of works citing them, the number of citing works (and authors) included in my sample, and the number of these that were duplicated across this sample and the sample of researchers who cited quantitative data.

*Table 3.5 Additional Details of Sample Comprising Data Studies Grouped by Number of Citations*

| Group | Citations | Data study faux ID | Number of citing works | Number of citing works (and authors) included in sample | Number duplicated across samples drawn by number of citations and by quantitative data |
|---|---|---|---|---|---|
| 1 | 5-24 | 00001<br>00002<br>00003 | 15<br>10<br>5 | 11<br>10<br>5 | 3<br>3<br>4 |
| 2 | 25-50 | 00004<br>00005<br>00006 | 36<br>36<br>34 | 29<br>25<br>31 | 13<br>7<br>10 |
| 3 | 51-100 | 00007<br>00008<br>00009 | 93<br>71<br>63 | 84<br>56<br>51 | 43<br>21<br>22 |
| 4 | More than 100 | 00010<br>00011<br>00012 | 162<br>141<br>117 | 148<br>132<br>105 | 56<br>47<br>42 |
| Totals | | | 783 | 687 | 271 |

Of the 403 contacted, 132 responded to the survey (32.7%). 8 indicated their use of the data was not reuse and 8 reused the data but did not provide information about satisficing (Q14 or Q15) resulting in 116 usable responses beyond those obtained in the sample stratified by quantitative data.

This number of responses was less than required to conduct meaningful inferential statistics in this sample. However, I did not include researchers in this sample in the combined samples of researchers who cited quantitative and qualitative data studies. The primary purpose of drawing this sample was to obtain as many responses to the survey as possible from researchers who reused the same data, with the goal of having as large of pool of these researchers as possible to contact for interviews. This sample fulfilled that purpose, if not a purpose of inferential analysis, and in the end I drew from all three samples (stratified by use of quantitative data, use of qualitative data and use of data stratified by number of citations) for the interview sample.

#### *3.3.1.3 Survey Instrument*

I designed the survey instrument to assess three aspects of knowledge satisficing: 1) whether it occurs, 2) what effect it has on researchers' achievement of their reuse goals, and 3) what effect it has on research outcomes. These aspects correspond to research questions 2, 2a, 2b, and 2d. I only investigated the third among the group of researchers who indicated they lacked knowledge about data that they desired. The concepts represented in these variables and descriptions of how the concepts have been operationalized in the survey are shown in Table 3.6 (the actual questions are available in Appendix A). I developed these concepts based on Herbert Simon's notion of satisficing and my extension that knowledge satisficing can be measured through a comparison of what was aspired to in relation to what was attained. I included questions about the effect of satisficing on research outcomes and researcher's goals for reusing data to understand the implications of satisficing (if any) for strategies to improve support for data reuse in the future.

*Table 3.6 Description of Information Assessed to Measure Dependent Variables*

| **Dependent Variables** | |
|---|---|
| **Concept** | **Concept operationalization** |
| Satisficing (Satisficing is also an independent variable) | Whether, when considering reusing the data, there was additional knowledge about data that was desired |
| | Whether, when the decision to reuse the data was made, there was knowledge about the data that was desired but not obtained or obtained only in part |
| | The three most important things the researcher would have liked to know about the data but was not able to obtain, or obtain to the desired degree. |
| | How much of each type of desired knowledge was obtained |
| | The importance of the knowledge that was not obtained or obtained only in part to deciding to reuse the data |
| Outcomes of research | Whether (and how much) a lack of knowledge affected the outcomes of the research |
| | Why the lack of knowledge did or did not affect the outcomes of research |
| Goal attainment | Researcher perception about goal attainment for data reuse |

My research examined how these dependent variables were affected by a variety of concepts which comprised the independent variables in my study. These included information

about the researcher and the perspective they brought to the research, whether the data were produced through quantitative or qualitative methods, the number of times the data were cited in the ICPSR Bibliography, the researcher's "distance" from the data, the researcher's experience with data reuse, the researcher's "reuse ability," the researcher's motivation for reusing the data, and elements of the research process. With the exception of the researcher's perspective, the researcher's perception of whether the data were created with an intention to be reused (part of "distance"), and the number of times data have been cited.

The information I obtained the about the researcher's perspective allowed me to assess whether data were in fact reused and whether differences in researchers' responses were associated with differences in their perspectives (i.e., their role in the research, how involved they were in determining the goals of the research, how many others were involved in the decision to reuse the data, and whether the data were reused from ICPSR). I included the researcher's assessment of whether the data were created with an intention to be reused (represented in the table as "Researcher's perception of original data creation purpose") based on Zimmerman's (2008) proposition of standardization as way to overcome "distance" between a researcher and reused data. I reasoned that if data were created with the intention to be reused (like a government-run national survey), the data would be more likely to have undergone a higher degree of curation, if not standardization, than data that were created without an intention to be reused. This hypothesis is yet to be tested, but I believed the intention of creation could be a potential confounder in my results and was important to include. I included the number of citations to data as an analysis parameter based on my reasoning that data studies that were cited more frequently might have properties that make them more amenable to reuse, including properties that potentially affect knowledge satisficing. I obtained information about whether the

data are quantitative or qualitative and the number of times data have been cited from the ICPSR Bibliography.

Descriptions of how I operationalized concepts for independent variables in the survey are given in Table 3.7 (see also the descriptions in section 2.9). Since the effect, if any, that satisficing had on the outcomes of research and goal attainment would also be assessed, satisficing (from Table 3.6) could also act as an independent variable in the analysis.

*Table 3.7 Description of Information Assessed to Measure Independent Variables*

| **Independent Variables** | |
|---|---|
| **Concept** | **Concept operationalization** |
| Information about researcher's perspective | Whether data were reused |
| | Researcher's role |
| | Researcher's involvement in the decision to reuse the data |
| | Number of people involved in the decision to reuse the data |
| | Researcher's involvement in determining the goals of the research |
| | Data obtained from ICPSR or another source |
| Data quantitative or qualitative | Information obtained from ICPSR Bibliography |
| Number of citations | Information obtained from ICPSR Bibliography |
| Researcher "distance" from the data | Researcher's perception of original data creation purpose |
| | Researcher's involvement in the original research |
| | Knowledge of research team about the way the original data were collected or produced |
| Experience with data reuse | Researcher's unit or departmental affiliation |
| | Researcher's professional position |
| | Researcher's primary domain of research |
| | Whether the research conducted was in the researcher's primary domain of research |
| | Researcher's years of experience in primary domain of research |
| | Domain of research in which the reused data were produced |
| | Researcher's years of experience reusing data produced in this domain |
| | Researcher's years of experience reusing data produced in any domain |
| | How often researcher's work involved reuse |
| | How often researcher's work involved reuse for the same purpose as these data |
| | Number of reuse projects researcher worked on |
| | Number of reuse projects where reuse considerations were substantially similar to ones involved in reusing these data |
| | Number of authored or coauthored published papers describing research involving data reuse |
| | Number of authored or coauthored published papers describing research involving data reuse where conditions were substantially similar to those involved in reusing these data |
| Reuse ability | Researcher's perceived ability reusing data produced in primary domain of research |
| | Researcher's perceived ability reusing data produced outside primary domain of research |
| Reuse motivation | Reuse purpose(s) |
| | Relative importance of each reuse purpose |
| Research process | How research questions were developed |
| | Information source(s) most important in obtaining desired knowledge |
| Satisficing (Satisficing is also an independent variable) | Whether, when considering reusing the data, there was additional knowledge about data that was desired |
| | Whether, when the decision to reuse the data was made, there was knowledge about the data that was desired but not obtained or obtained only in part |
| | The three most important things the researcher would have liked to know about the data but was not able to obtain, or obtain to the desired degree. |
| | How much of each type of desired knowledge was obtained |
| | The importance of the knowledge that was not obtained or obtained only in part to deciding to reuse the data |

An important consideration in the way I operationalized concepts in the survey was how strictly or loosely to define "knowledge about data." As an example, a strict definition might

refer only to the context of data creation (for instance, sampling procedures or other survey methods). A loose definition might include attributes of the study population itself—for instance, about gender or income or other variables like zipcode or sexual orientation (see responses from the pilot survey in Table 1.1). In addition to knowledge about methodologies used to create data, these attributes also constituted what researchers wanted to know about the data. In the end, in keeping with my theoretical framework, which problematized a separation of "data" and "context," as well as to understand from the researcher's perspective what they desired to know— whatever it was—I used a loose definition, or rather, did not define knowledge specifically at all. This led to some surprises and difficulties that I discuss further in section 3.4.8.5.

On the methodological side, there were several considerations that influenced the design of my survey. The most fundamental of these was that participants' ability to reliably report about phenomena could diminish if the event was not salient enough, the period of report was too far in the past, or the event or episode of interest was too generic (i.e., issues of "recall", see Tourangeau et al., 2000). Participants might also "satisfice" in the answers they provide depending on their ability and motivation to answer questions (Krosnick and Presser, 2010). Apart from these more basic issues, there were numerous considerations about the way questions and response options were presented, and the way the survey as a whole was structured that could impact the accuracy and reliability of the data collected. I address some of the most important considerations below.

There are two main ways that I sought to address concerns about participant recall. The first was by asking questions about a process (evaluation of data for reuse) that researchers should, in keeping with conventions of science and scholarship, be able to explicitly account for

as part of their research methodology. The second was by selecting a particular reuse instance within a specific recent time frame with a substantive outcome (e.g., a research article) that participants, as first or corresponding authors, invested significant time in and could refer to, if needed, to recall relevant information (Flanagan, 1954)

I sought to address issues of motivation by clearly communicating the ways that participation in the survey could benefit participants (i.e., that my study is designed to gather information that could facilitate reuse of data in the future) and formulating questions that were designed to obtain accurate and reliable responses, but that could be answered as quickly and with as little effort by participants as possible. Several considerations were at play in the formulation of questions and responses.

**Closed versus open-ended questions.** Most of the questions in the survey were closed-ended. These questions generally take less effort for participants to answer, and answers are easier to quantify and analyze (Bradburn et al., 2004, Schaeffer and Presser, 2003, p. 76). There were several questions where I allowed open-ended responses. These were questions where I was interested in learning about the specific numbers of people involved in decision-making about data reuse, the role(s) that the participant played in the research, the researcher's primary domain of research, and the researcher's organizational affiliation. There were also questions where I asked about specific knowledge that researchers desired but were not able to obtain or only able to obtain in a limited way, and the reason knowledge satisficing did or did not negatively impact outcomes of their research.

I used an open-ended response in the first instance because, as my survey population was quite diverse, I was not sure what response categories might be most appropriate to capture the number of people involved in the research. Fowler (1995) argues that open-ended responses are

appropriate in such cases. In addition, Bradburn et al. (2004) note that the difficulty of analyzing open-ended responses does not increase if the responses are numeric. I used open-ended responses in the other instances because I was interested in understanding how researchers described their roles, research domains, affiliations, knowledge about data, and impact of knowledge satisficing (if any) in their own words. In some of these instances—for instance if researchers did not see themselves in standard roles (such as principle investigator or data manager) or easily defined research domains—having open-ended responses may actually have reduced the cognitive load of responding and yielded more accurate responses (Bradburn et al., 2004).

**Requiring a "yes" or "no" response as opposed to "check all that apply."** There were several questions where I asked respondents to supply information about each of several response options (for instance, each of several purposes of data reuse, as well as types and sources of knowledge that were important to reusing the data). In these cases, I asked for information about each individual response, rather than allowing participants to "check all that apply" or, for instance, simply indicate all types of knowledge they would rate as having a particular level of importance. I did this on the guidance of Bradburn et al. (2004) and Fowler (1995), who argue that responses to questions are more complete when each response is answered individually. Fowler notes, for instance, that "The instruction (CHECK ALL THAT APPLY) is not consistently effective in a self-administered form" because some respondents will check one option and move to the next question" (p. 92). In this type of question, while answering all options may be less efficient for the respondent, the inconvenience is justified because of the increased likelihood of a more complete response.

**Questions involving agreement.** Schaeffer and Presser (2003) cite evidence that using forced-choice response options can result in more reliable responses than questions that merely ask for agreement or non-agreement or questions with simple yes-no or true-false response options. Schaeffer and Presser (2003) further point out the potential of respondents, when faced with questions about agreement, to "acquiesce," or agree regardless of what the question is asking (p. 80). In light of these arguments and in order to obtain more reliable responses I used questions that asked about agreement sparingly and made heavy use of questions that present response options on a scale. In addition, when asking about respondents' level of agreement with particular statements I used the adverb "Completely" instead of "Strongly" in order to minimize the mixing of respondents' emotional and cognitive evaluations (Fowler, 1995) (e.g., I used a scale ranging from "Completely disagree" to "Completely agree").

I have chosen not to follow Fowler's (1995) advice about how to assess measures of truthfulness as opposed to measures of agreement or emotion, however. Fowler (1995) argues that measures of truthfulness should be used when the goal is to measure how well perceptions align with a particular statement and to measure agreement or other emotion when the goal is to measure how well preferences align with a particular statement. I have chosen to measure agreement because I believe the distinction between "somewhat disagree" and "somewhat agree" is clearer and more meaningful than the distinction between "somewhat untrue" and "somewhat true." Moreover, a scale of agreement is a commonly used construction that will likely be familiar to survey respondents. These factors led me to believe that more reliable responses would be achieved using a measure of agreement despite the fact that my use occurred in an instance when I was measuring how well respondents' perceptions aligned with a particular statement.

**Use of rating scales.** The survey made significant use of rating scales in order to gather information about respondents' perceptions (e.g., the degree of involvement of participants in decision-making about data reuse) and attitudes (e.g., the importance of specific purposes of reusing data). Using a rating scale to identify the magnitude of responses on a continuum is a strategy that is commonly used to gather information for these types of questions (Fowler, 1995, p. 48). Questions that use a rating scale can yield more reliable information than questions that ask respondents to rank items in a particular order (Fowler, 1995, p. 98), and can be particularly appropriate if what is being measured will not necessarily apply to all respondents (e.g., if some respondents might believe that a measure is "very important" and others that it is "not at all important") (Fowler, 1995, p. 162).

Some of the main considerations when using rating scales have to do with whether the categories are verbal or numeric, the number of response categories, the way the categories are labeled, the way responses are distributed along the scale, and whether or not a "middle" category is offered. Schaeffer and Presser (2003) cite evidence that use of numeric scales leads respondents to assume that response categories are equidistant (i.e., that the same distance separates "slightly," and "somewhat," or "pretty," and "very" on a scale) (Klockars and Yamagishi, cited in Schaeffer and Presser, p. 78). They also point to evidence, however, that use of both verbal and numeric scales results in greater reliability (p. 78). Based on this evidence, and research that has found that differences in the perceived location of response categories along a continuum result in different response distributions (Schaeffer and Charng, 1991), I chose to use both numeric and verbal labels.

Bradburn et al. (2004) argue that five categories are the upper limit of the number of verbal categories that respondents can understand without visual aids. Schaeffer and Presser

(2003) cite studies finding that the reliability of rating scales increased when more than three categories were used, up to approximately seven or nine categories. They, as well as Fowler (1995) note, in general, that the number of categories chosen is a tradeoff between achieving greater distinctions along the continuum and respondents' ability to make finer distinctions. I chose to use rating scales with five categories throughout the survey except in one question where I used a rating scale with only four. In all but that one instance (described below) I included a "middle" category for both unipolar and bipolar questions.

A middle category represents a midpoint between two positions (for instance, between positions of agreement and disagreement on a bipolar scale and between degrees of intensity on a unipolar scale). According to Bradburn et al. (2004), it has been typical in survey research to exclude the midpoint in unipolar scales in order to "push" respondents' answers towards one end of the scale or the other. However, research has shown that while, when offered, more respondents choose a middle category, this eventuality does not affect the overall proportion of responses that fall on either side of the midpoint (Bradburn et al., 2004). Furthermore, O'Muircheartaigh et al. (1999) found that including a middle category reduces random measurement error without affecting response validity. Based on this research, and with consideration for consistency throughout the survey, I chose to use five categories, including a midpoint, for nearly all ratings. I decided to remove the middle category in one question in which I asked researchers' level of agreement with several statements about how their research questions developed. I did this after discovering during analysis of data from the pilot survey that in some cases, if researchers used the middle category for more than one of the response options, I was unable to clearly interpret the results.

**Unipolar and Bipolar.** Research on the relative reliability of responses from a unipolar or bipolar scale is inconclusive (a unipolar scale has been found to have more reliability in some cases but not in others) (Schaeffer and Presser, 2003, p. 77). For this reason, I selected the scale to use that made the most sense in the context of the question.

**Questions about frequency.** Fowler (1995) argues that when asking questions of frequency, it is generally better to ask about the specific number of times something has occurred in an appropriate amount of time (p. 157). This is because the use of a response scale assumes a regularity of activity that may not be reasonable to assume for all respondents (for some, activities may be irregular) (p. 157). At the same time, Schaeffer and Presser (2003) note that relative frequencies may be appropriate "when the investigator wants to give weight to the evaluative component in the respondent's perception of the frequency, when group comparisons are not a central analytic goal, or when the frequencies are too difficult to report in an absolute metric" (p. 74).

Following these two types of guidance, I employed questions that asked about both relative and absolute frequencies of data reuse. I used sets of questions that asked about the number of projects in which data had been reused, and the number of papers the respondent authored or co-authored (absolute frequency) involving data reuse. I also asked about the relative frequency with which data were reused, both for purposes indicated by the researcher and for any purpose, at the time the research was conducted. I did this in order to compare responses and obtain a more dynamic understanding of the experience with data reuse that researchers brought to the specific case of reuse under investigation.

### *3.3.2 Data Analysis*

Analysis of the survey data took several forms. I generated descriptive statistics to determine, for example, how frequently knowledge satisficing was evidenced, what sources of knowledge about data were most important to researchers, and how many researchers in the sample were at what stage of their careers or had particular levels of experience with data reuse. Descriptive statistics helped me understand the survey results at a broad level and guide further analysis. For instance, when I learned that close to half of the researchers in my random sample reused their data from a source other than ICPSR, I became interested in learning more about why in the interviews.

I originally intended to conduct principle component analysis to determine which independent variables explained the greatest amount of variance for underlying concepts in the survey and had the greatest impact on the dependent variables. However, because I found few significant relationships between the independent variables and dependent variables, my analysis consisted of descriptive statistics, hypothesis testing using logistic regression, and model development. I conducted logistic regressions using satisficing as the dependent variable to determine the predicted probability of specific variables affecting knowledge satisficing (e.g., the effect on knowledge satisficing of reuse purpose, importance of reuse purpose, how researcher questions were developed, a researcher's "distance" from the data, etc.). I also performed logistic regression using goal attainment as a dependent variable (with satisficing as an independent variable). I conducted all regression analyses on the combined samples of researchers who reused quantitative data and researchers who reused qualitative data (excluding the sample stratified by citation count).

The hypotheses I investigated included the following and are explained in more detail in Chapter 2, section 2.9.

*H1*. Researchers satisfice in the knowledge they obtain when determining whether to reuse data.

*H2*. Knowledge satisficing has a negative impact on research outcomes.

*H3*. The probability of researchers' attaining their goals for reusing data is lower in the presence of knowledge satisficing.

*H4*. There is no association between knowledge satisficing and reuse of qualitative or quantitative data.

*H5*. The greater a researcher's "distance" from data, the higher the probability of knowledge satisficing.

*H6*. The more experience a researcher has in their primary domain of research, the lower the probability of knowledge satisficing.

*H7*. The more experience a researcher has with data reuse, the lower the probability of knowledge satisficing.

*H8*. The greater a researcher's ability to reuse data, the lower the probability of knowledge satisficing.

*H9*. There is no association between researcher's motivation for reusing data and knowledge satisficing.

*H10*. When research questions change or develop over time there is a lower probability of knowledge satisficing.

For the open-ended responses in the survey, I conducted qualitative coding. These included responses to questions about knowledge that researchers desired but were not able to obtain or only able to obtain in a limited way, the reason knowledge satisficing did or did not negatively impact the results of research, and the researcher's role, professional position, and primary domain of research. I conducted a combination of descriptive and in vivo coding for responses to each of these other questions (Saldaña, 2009).

#### *3.3.2.1 Knowledge Lacking or Limited*

I coded the open ended survey responses concerning the types of knowledge researchers indicated were lacking or limited inductively and iteratively (see Table 4.8): by inductively I

mean I used the language that researchers used to describe the knowledge they lacked in my codes; by iteratively I mean I performed initial coding of the descriptions of knowledge, placing new descriptions into existing categories when appropriate and creating new categories if there was not a good match. However, over time, I made adjustments to the names of some categories to incorporate the variety of similar types of knowledge that I found. In these cases, I recoded the names of prior categories to ensure that similar types of knowledge were all coded the same way. As an example, in the end, under the code "data analysis (cleaning)" I included knowledge related to "cleaning" explicitly, but also knowledge related to transformation, transcription, and processing of missing data, though I did not code all of these as "cleaning" in my first pass.

I coded the types of knowledge in two levels (see the results and definitions in Table 4.8). The first level included 14 categories and the second level included 41 subcategories. In some cases, there was no subcategory (e.g., "data access information") so the first and second level codes were the same. In other cases (for instance, "data" and "data analysis"), where a subcategory would have been the same as the overarching category (e.g., where there was a general lack of knowledge about data or data analysis), I created a subcategory called "general."

There are some types of knowledge that I presumed or inferred would go into specific categories based on my interpretation of the researcher's response. One example is "data supplement (coverage)" (three cases), in which I included responses where the researcher desired access to restricted data. The other is "data supplement (missing)" (1 case), where I inferred from the survey response that the researcher desired access to missing data. There were also some cases where I referenced descriptions of the original data to determine in what category to place the described knowledge.

#### *3.3.2.2 Reason for Negative Impact of Knowledge Lacking or Limited*

I followed the same procedure here as for knowledge lacking or limited, coding responses inductively, iteratively, and at two levels (see Table 4.46). I created initial codes using the language of the respondents starting with the first researcher response. I both created and revised codes as I went down the list, making them more general in some cases or using more inclusive terms to capture similar responses under the same broad categories.

#### *3.3.2.3 Researcher's Role*

Through an initial round of in vivo and descriptive coding, I put the roles researchers entered into 14 categories including analyst, co-author, principle investigator, project manager, co-investigator, co-principle investigator, co-project manager, collaborator, consultant, data manager, graduate student research assistant, paper lead, researcher, and none (i.e., the person did not have a role in the research). Identifying and placing researchers into these categories was largely a matter of normalizing the free-text responses.

In a second round of coding, I placed these roles into broader categories of support, collaborator, lead, and none. Here, support included analyst, project manager, co-project manager, data manager, and graduate student research assistant; collaborator included co-author, co-investigator, collaborator, consultant, and researcher; lead included co-investigator, co-principle investigator, and principle investigator.

#### *3.3.2.4 Researcher's Professional Position*

I performed an initial round of coding that, like coding researchers' roles, was largely in vivo and largely a matter of normalizing responses into categories of undergraduate, masters, PhD student, post-doctoral researcher, assistant professor, assistant researcher, associate

professor, associate researcher, lecturer, professor, researcher, and other (see Table 4.21). In this round of coding, I established the following guidelines:

- researcher included positions of research scientist, fellow, and statistician
- professor included positions of research professor, clinical professor, and senior professor
- other included positions of consultant, director, founder, CEO, program manager, and similar designations
- when coding, if more than one position was listed, I used the one with the least experience (e.g., assistant professor rather than associate professor)

#### *3.3.2.5 Primary Domain of Research*

My final coding of researchers' primary domains is represented in Table 4.17. Researchers' responses, and therefore my coding, reflected topical areas of research. Similar to my coding of areas of knowledge, I coded topical areas or domains at two levels: a broad level and a more refined level (it is the broad level that is represented in Table 4.17). In analyzing the domains, I made heavy use of in vivo coding so as not to put too much weight on my own interpretations. In general, I coded to maximize my ability to see trends at the first level of coding and to compare domains among researchers, especially among researchers who reused the same data. A more detailed account of my coding is given in Appendix D.

## 3.4 Phase 2: Qualitative Study (Interviews and Background Research)

### *3.4.1 Introduction*

As noted in section 3.3, I had several purposes for conducting a qualitative analysis in additional to quantitative. The goals of the quantitative portion of the study were to test a variety of hypotheses related to researchers lacking knowledge about data, to aid in the selection of interview candidates, to help build an interview instrument (i.e., according to the sequential design of my mixed methods study, I intended for the results of the survey to inform the questions I would ask in interviews) and to triangulate findings between the quantitative and qualitative data in order to test the validity of my results.

The goals of the qualitative research were to better understand the nature and extent of knowledge satisficing—if I found it to exist in the survey—and how researchers placed limits on or bounded the knowledge they obtained when deciding to reuse data. Some of the larger goals of my research were to produce generalizable results related to the knowledge researchers obtain to reuse data across different instances of reuse (especially through the survey) and comparable findings about what researchers need to know to determine the data are sufficient to reuse across multiple different instances of reuse.

In the qualitative portion of the study, I employed two methods—in-depth interviews and background research—and used information about respondents and their research taken from the survey to both select researchers to interview and finalize my research protocol. A key strategy related to my goal to produce comparable findings was to interview multiple clusters of researchers, with each cluster centered around a different set of data, and researchers in each cluster having reused the same data. By doing so, I intended to gather multiple perspectives from diverse individual (per researcher) and collective (per group of researchers) circumstances about how and why researchers bounded the knowledge they obtained about data to reach reuse equilibrium. My goal was to compare these perspectives in light of Donna Haraway's theories related to situated knowledges and the formation of objects as boundary projects. My ultimate aim was to identify any insights the comparisons might reveal for how conversations about the design of data "artifacts" and repositories might proceed.

In preparation for the interviews, I conducted background research into the academic history of each researcher. I intended to use knowledge I obtained to establish rapport with research participants in the interview, better understand participants' responses to questions, and probe or ask follow-up questions to explore nuances and particularities appropriate to each

research instance. I intended also to compare what I learned in the background research with interview responses in the analysis phase as a means of triangulation. As I describe below, some of my goals related to the interviews and background research I was able to realize, and others I was not.

In the following sections, I detail my rationales for conducting in-depth interviews and background research. I then discuss how I used results from the survey to construct the interview protocol and select candidates for the interviews. I present some findings in the process but leave a full accounting of results from the survey and interviews to chapter 4. In section 3.4 I describe the methods involved in my conduct of the interviews specifically, as opposed to the way the survey results informed the development of the interviews.

### *3.4.2 Interview Protocol*

I determined that in-depth interviews were appropriate for this study because I had particular phenomena of interest (knowledge bounding and knowledge satisficing) I wanted to learn about (Hesse-Biber, 2006), and these phenomena were things about which I believed study participants, in keeping with scholarly norms of justifying research methodology, should be able to report (Fowler, 1995; Tourangeau et al., 2000). I conducted semi-structured interviews, as opposed to interviews more formally structured, because while my phenomena of interest were well-defined, there was not a lot known about them or the factors that affected them. Striking a balance between a loose and tight structure in the interviews allowed me to focus the interviews and facilitate comparison between different reuse instances, while leaving room for study participants to report and describe their experiences in their own words (Miles et al., 2014; Hesse-Biber, 2006).

Similar to survey research, there are two important challenges inherent to the use of interviews. One is that participants' ability to reliably report about phenomena can diminish if the period of report is too far in the past or an event or episode of interest is too generic (Tourangeau et al., 2000). Participants may also "satisfice" in the answers they provide depending on participants' ability and motivation to answer questions (Krosnick and Presser, 2010).

In order to mitigate the first challenge, I selected reuse instances that occurred within roughly five years (eventually six considering when I actually began the interviews) from when the research was reported (i.e., between January, 2014 and September, 2019). This timeframe was based on timeframes of three to five years were used effectively in other studies of reuse (Zimmerman, 2007, Duff, 2002). I notified researchers at the time I contacted them about the specific research I wished to ask them about and I expected that asking about specific instances associated with works written in the recent past would mitigate concerns about recall (Flanagan, 1954). Fowler (1995) argues that asking multiple questions and placing events in time can help to stimulate recall and these were strategies I used in the interviews as well.

Regarding the second challenge, Krosnick and Presser (2010) suggest that minimizing the difficulty of answering questions and maximizing participant motivation can reduce satisficing and increase the accuracy of responses. I anticipated that the strategies to mitigate the first challenge above would also reduce the difficulty of responding. I expected, however, that issues of motivation would be more difficult to overcome. For instance, it was possible participants would not want to reveal information about knowledge they lacked about data that could be embarrassing or call the integrity of their research into question (Tourangeau and Yan, 2007; Fowler 1995).

Fowler (1995) recommends several strategies to reduce the likelihood of answer distortion. The first is to assure the confidentiality of responses and make sure participants know their responses are protected. The second is to clearly communicate to participants the importance of responding accurately. This might include having participants make a verbal or written commitment to accuracy and reinforcing thoughtful and complete answers when they are given by participants. The third is reducing the role of the interviewer. This can be accomplished by using a self-administered questionnaire or using strategies that prevent the interviewer from knowing which question is being answered. Two additional strategies Fowler describes are minimizing the formation of personal relationships with participants (e.g., by not telling personal stories or expressing personal opinions) and stressing to participants that there are no right or wrong answers. Fowler tempers his recommendations for the final two strategies by acknowledging that there is debate over the role an interviewer should have. Interviewers can help to motivate respondents, for instance, or encourage accurate responses by establishing rapport.

Hesse-Biber (2006) stresses the importance of rapport in creating an environment where there is a reciprocal relationship and shared authority between the interviewer and research participant. She argues that developing rapport can help participants to feel safe, comfortable, and valued and enable an environment where meaning between the interviewer and research participant can be co-created—a goal of interpretivist research. Hesse-Biber recommends listening actively and intently, and "being prepared to drop [one's own] agenda" in the process of the interview as strategies to encourage such an environment.

In the end, I chose to follow Hesse-Biber's guidance in conducting the interviews. Rather than striving to remove myself from the situation (e.g., through a self-administered

questionnaire)—and in keeping with the paradigm of situated knowledges—I strove to motivate accurate and honest answers by being as transparent as possible about who I am and what my goals were. I explained, for example, that the goal of my research was to better understand the substantial social and technological barriers to reusing data in academic research that are reported in the literature and contribute to solutions that improve the ability to conduct research with secondary data.

By positioning myself honestly as a colleague and ally, I hoped to create a sense that we were addressing shared challenges and motivate researchers to respond to questions as fully as possible. I also strove to listen actively to the responses participants gave and have less attachment to my own agenda in the interviews to better understand how researchers viewed issues of knowledge bounding and knowledge satisficing (whether those views conformed to my own expectations or not). Following Fowler's (1995) advice, I additionally communicated clearly about measures that I would take to protect confidentiality of researchers' responses (e.g., in my storage and use of recorded, transcribed, and analyzed material) and stressed the importance of answering questions accurately.

In terms of the particular structure of the interview, there were three main categories of information I knew I was interested in learning about: 1) how researchers bounded the knowledge they obtained about the data they reused 2) why they determined the boundaries the way they did and 3) how they obtained the knowledge about data they desired. I was additionally interested in understanding whether researchers satisficed in the knowledge they obtained, how the satisficing could be characterized, why they might have satisficed, and the impact of knowledge satisficing. I decided to operationalize the collection of this information by dividing the interview questions into six information areas:

1. background information about the research
2. the process of deciding to reuse the data
3. the process of bounding knowledge about the data
4. factors that may have influenced knowledge bounding
5. the impact of knowledge satisficing (if applicable)
6. factors that might mitigate knowledge satisficing and its effects (if applicable)

I developed these areas prior to the release of the survey but I knew I would determine the final format and wording of the questions, as well as the different areas I wanted to probe during the interview, after analyzing the survey results. I knew also that, similar to Agosto's (2002) study reviewed in section 2.6, I wanted to have enough structure to obtain information about the questions I was interested in but also have participants reflect as broadly as possible on their experiences without being overly influenced in their responses by what they felt like I might want to know. In this regard, I saw my rapport with participants and the way I positioned myself to be particularly important. In addition, I believed a strong overarching articulation of my research project and my goals at the beginning was crucial to setting expectations for the depth of reflection I wish participants to engage in, and creating a safe environment for them to be honest and open with me.

In the next section, I describe how some of these desires played out in the final development of the protocol—both aspects that depended on the survey and adjustments I made after beginning to conduct the interviews.

### *3.4.3 Interview Protocol: Integration of Survey Results*

In the survey, I found that about 25% of researchers lacked knowledge they desired about data when they decided to reuse the data, and that in 87.7% of cases, the lack of knowledge had at least some negative impact on the outcomes of their research (section 4.6.1.1). I also obtained information about the kinds and amounts of knowledge researchers lacked (sections 4.2.3.1 and 4.2.3.3) and why lacking that knowledge had a negative effect (or not) (section 4.6.1.3).

The main questions I had yet to address were those about how researchers bounded knowledge about data, regardless of whether there was desired knowledge they lacked about data (these are represented in my first research question, RQ 1 through RQ 1d). I also wanted to gather information about why researchers proceeded with using data when they lacked knowledge (RQ 2c) and what researchers believed could mitigate instances where they did not have all the knowledge they would have liked about the data they reused (RQ 2e).

The final protocol I developed to address my research questions in the interviews is given in Table 3.8 below and in Appendix B (the appendix includes the introductory text). There were seven main areas in the end that I sought to collect information about: background about the research, the decision to reuse the data, bounding knowledge about data, researchers' goal attainment (related to reuse), sources of knowledge about data, factors that might mitigate satisficing and its effects, and some final closing information.

*Table 3.8 Interview Protocol Identifying the Purpose of the Questions in Relation to the Research Questions*

| **Purpose of questions and relation to research questions** | **Interview Questions** |
|---|---|
| Background to understand researcher's perspective<br><br>2b. What factors influence knowledge satisficing? | Background about the research<br>• I'd like to start with a little bit of background. I wonder if you would talk a little bit about how you became interested in the area of research you see this project as belonging to, and then this project in particular.<br>• And could you set the scene for this project and give some additional details?<br>  o I'm wondering things like how the idea for the project came up, how it took shape, and who was involved?<br>  o [If data not from ICPSR]: I note from the survey that you obtained the data from [source]. Could you talk some more about how that came about?<br>• What were some of the main considerations or pressures influencing how the research developed and proceeded?<br>• You mentioned in the survey that members of your research team had [amount of] knowledge of the original data creation. Would you talk a little bit about how that influenced your choice of the dataset and how your research took shape? |
| 1. How do researchers determine the boundaries of the knowledge they obtain about data in order to reuse them in their research?<br><br>1b. How do researchers determine how much of what kinds of knowledge is enough?<br><br>1c. What do researchers report influences these determinations?<br><br>2b. What factors influence knowledge satisficing? | Decision to reuse the data<br>• Would you talk a little bit about the specific data you reused and how you came to the decision to reuse these data in particular?<br>  o [Follow-up] What led you to choose these data over others?<br>• Did you have any concerns about reusing the data, either in general or about reusing these data in particular? Would you talk some about those?<br>• Were there any specific guidelines for data reuse that you followed or consulted, either in terms of methodology or specific instructions for using these data?<br>• Thinking about those with whom you interact most frequently about your research and who might be involved in commenting or evaluating your research: how is data reuse perceived within this community? Is it something that occurs all the time, something researchers are more reticent to engage in? If so, why? |
| 1. How do researchers determine the boundaries of the knowledge they obtain about data in order to reuse them in their research?<br><br>1a. What knowledge about data is most important to researchers to reach reuse equilibrium and why?<br><br>1b. How do researchers determine how much of what kinds of knowledge is enough?<br><br>1c. What do researchers report influences these determinations? | Bounding knowledge<br>• I'm interested in learning how researchers know when they've obtained sufficient knowledge about the data to reuse them.<br>  o [If no knowledge reported lacking] What were some of the most important things you needed to know about the data to reuse them?<br>  o [If knowledge reported lacking] There seemed to be some places where there was more you would have liked to know about the data. For instance, you mentioned [specific areas of knowledge].<br>• Would you talk some about, in each of these cases, why the knowledge was important and how you knew these data would fulfill your needs for research? How did you know you had reached a certain threshold where the knowledge you had was enough and what influenced you in making that decision?<br>• Did you have ideas about how much knowledge would be enough ahead of time or did an idea develop as you reused the data? |
| 2dii. What is the perceived impact of knowledge satisficing on researchers' achievement of their goals for reusing data? | Goal attainment<br>• In terms of how well your reuse of the data met your research goals, in the survey you gave the following responses [repeat responses]. Would you explain a little more why you responded that way? |
| 1d. How do researchers obtain the knowledge they desire? | Sources of knowledge<br>• What were some of the primary sources you used to obtain the knowledge you desired about the dataset? Would you talk some about how you went |

| | |
|---|---|
| | about obtaining your desired knowledge and what that experience was like? |
| 2e. What do researchers believe could mitigate knowledge satisficing? | Factors that might mitigate knowledge satisficing and its effects<br>• Are there particular factors you would say facilitated or got in the way of obtaining the knowledge you desired? What would have helped overcome barriers you experienced?<br>• Thinking more broadly about your experience reusing data, what are some of the major issues or shortcomings you see for those who seek to reuse data and what changes do you think would be most helpful to support those who are conducting research with secondary data? |
| | Closing<br>• Did you have any difficulty answering any of the questions and if so, if you would describe what it was? [I removed this question after the first several interviews]<br>• Is there anything else related to obtaining knowledge about data or negotiating situations where desired knowledge is not available that you would like to add? |

One of the main issues in putting the final protocol together was determining how to reconcile my desires to learn more about instances where researchers lacked desired knowledge about data and to gather information about what knowledge was most important to researchers (since areas where researchers lacked knowledge were not necessarily areas that were most important for them). I decided in the end to have two slightly different protocols: one asking those who indicated in the survey that they lacked desired knowledge to talk more about how they decided the data were "good enough" for their purposes, and one for those who indicated they did not lack knowledge to ask them first what knowledge was most important and then to ask how they determined how much was enough.

I realized that this would not give me comparable results across all of the interview participants as far as the importance of the knowledge I asked about, and that it would give me less data related to research question 1a ("What knowledge about data is most important to researchers to reach reuse equilibrium and why?"). I expected, however, that it would still allow me to compare the experience of bounding knowledge about data, which was a key goal of the interviews, and at the same time allow me to gain greater insights into instances where researchers reused data despite lacking knowledge about them. I was willing to make this tradeoff in comparability additionally because it would allow me to interview both researchers

who lacked knowledge and those who did not. For a time, I considered interviewing only researchers who did not lack knowledge in order to improve comparability. However, in this alternative, I would lose insights into the experience of those who lacked desired knowledge and also reduce the number of researchers who reused the same data. I also considered interviewing researchers who reported lacking knowledge about data, but only asking them about the knowledge that was most important to them.

A second important issue was to consider exactly what I wanted to learn about knowledge satisficing from the interviews. For this, I returned to my theoretical framework and goals related to measuring knowledge satisficing and knowledge bounding. One of my goals in studying what I had termed "knowledge satisficing" was to develop protocols that were sensitive to measure satisficing in both "substantive" and "procedural" contexts.

However, I realized in retrospect that my survey was not necessarily able to measure satisficing from the perspective of procedural rationality. This is because it is not possible to know from researchers' responses whether their decisions to reuse data were based on substantively reality (e.g., they knew how much knowledge would be optimal, they knew they didn't have it, and they made a calculated decision about how much would be enough), or made according to a heuristic in the context of procedural rationality (e.g., they did not know how much knowledge would be optimal, and not being able to "compute" this, used a heuristic such as an expectation about how much knowledge would be enough knowledge to get the research published to make a decision about reusing the data).

Despite my efforts, then, I had created a survey protocol that was able to tell me that researchers lacked knowledge—and I knew I was capturing at least some instances where the decision to reuse data was made in the context of substantive reality—but I could not be sure I

was capturing decisions made in the context of procedural rationality. In order to have done that, I would have needed to ask researchers if they knew ahead of time how much knowledge would be enough for them to reuse the data, or whether that knowledge developed as they explored the data. This is a similar concept to my question about whether research questions were fixed ahead of time or fluid, but the former has specifically to do with knowledge (i.e., knowledge satisficing) where the latter has to do with research questions (i.e., research question satisficing) and how the way they developed is associated with a researcher lacking desired knowledge.

In line with this reasoning, when I operationalized my outstanding research questions into a protocol, I added a question and appropriate follow-up questions to address the issue, i.e.: "Did you have ideas about how much knowledge would be enough ahead of time or did an idea develop as you reused the data?"

Deciding to ask slightly different questions to those who indicated they lacked knowledge and those who did not and determining what, specifically, I wanted to ask about satisficing (specifically to capture and distinguish, if possible, instances of procedural satisficing) were two important issues in the development of my interview protocol. There were several other important developments that occurred after I began conducting the interviews. Since these fall outside of the integration of survey results, I discuss them in the next section.

### *3.4.4 Interview Protocol: Additional Influencing Factors*

In putting together the interview protocol, I struggled with how structured it should be. I knew I wanted to have more of a conversation with researchers and I wanted to understand their experiences reusing data as holistically as possible. At the same time, I had specific questions I wanted to ask that differed based on whether researchers lacked knowledge or not, and whether they knew the scope of the knowledge they desired about data ahead of time or not. There were

also cases where a researcher entered only one area of knowledge that was lacking and I had to consider whether to ask them about that area of knowledge or about the knowledge that was most important to them, or both.

As a result, my first operational protocol was a rather complicated series of if/then scenarios (if they lacked knowledge, if they knew how much knowledge they needed ahead of time) with lists of follow-up questions to ask in case researchers did not provide the kind of information I was hoping to obtain. I made individual protocol documents for each respondent where I entered their responses to relevant survey questions and made notes about specific things to ask. This became a lot to manage and felt out of keeping with the looser structured I desired.

Several developments changed my approach to the interviews over time, including the way I approached asking questions and the wording of questions themselves. The first was that through my background research, which I describe below, and during the conduct of interviews, I found that most researchers lacked at least some knowledge that they desired about the data they reused, whether they reported it in the survey or not. The lack of reporting may have been due to the way I asked my questions or the idea researchers had of what I was interested in. It may also have had to do with researchers' expectations about their ability to obtain knowledge about data (i.e., problems obtaining some kinds of knowledge were normal to them and did not merit mention). Whichever it was (I discuss this more later), when I read the researchers' papers describing their reuse of data, there was nearly always some desired knowledge they lacked that was of a kind with the knowledge other researchers reported lacking in their survey responses. I also came across cases where researchers had the same difficulties analyzing data that others had indicated in the survey, or had developed strategies to deal with similar problems (e.g., missing data) but did not report them.

Since I was interested in learning more about these cases as instances of potential satisficing, I found the need to be flexible and go where the conversations led to be paramount. I paid close attention to what the researchers said and followed up on issues they raised that I thought could shed light on how they determined the thresholds of knowledge that were "enough" for them to reuse the data and others of my research questions. Over time, I came to understand in the interviews that researchers knew exactly what I was asking about when I asked about knowledge that was lacking and how they determined how much knowledge was enough, and they were happy to talk about it. Rather than being concerned about if/then scenarios or about asking the right questions, I eventually tried to set up my questions about knowledge bounding in a way that researchers could most freely talk about their experiences. This resulted in some differences in the interviews. However, what I came to focus on most centrally, whether we talked about knowledge that was lacking or knowledge that was most important, was how researchers set limits on the amount of knowledge about data they deemed sufficient to reuse the data, and what influenced where those limits were set.

An example of my attempt to improve the set-up of my questions leads into the second development that caused me to change my approach to the interviews. In my first few interviews, I asked researchers how they knew the knowledge they obtained about data was "enough" for them to decide to reuse it. I found, however, that some researchers seemed to feel challenged by this formulation of the question, as though I were questioning the integrity of their research or the validity of their results. I found that adjusting the question slightly, to ask researchers how they knew the data were sufficient to fulfill their needs, or how they knew they had reached a threshold where their knowledge was enough to reuse the data, did a lot to relieve tension and dissolve any perception that I was taking an oppositional stance to their research.

This finding or development was very interesting to me and corresponded with another finding I had as I conducted the interviews. As I talked with researchers, I came to question whether satisficing was an accurate characterization of what transpired when researchers lacked knowledge they desired about data but decided to reuse the data anyway. I present a fuller discussion in section 4.2.2 below, but what a number of researchers described explicitly, and others less so, was that in their reuse of data, they did not believe there was an "optimal" state of knowledge. In reusing data collected by someone else, researchers did not expect that the data would be an optimal fit for their research or that they would know everything they would like about them. Researchers strove to choose the "best" data for their researcher and nearly always wanted "better" data. But as one researcher described (see section 4.2.2), it was not actually the "best" they strove for, but rather the "best-ish," since there usually was no best in data reuse. Researchers' accounts of their reuse thus seemed more exploratory and creatively productive to me than satisficing scenarios—where I would expect to see researchers adjusting their expectations for goal attainment down from an imagined optimal.

One result of my developing realization about the poor fit between satisficing and what researchers described is that, while I still asked it, my question about whether researchers had an idea of what they needed to know ahead of time took on less importance. In fact, when I asked the question, more often than not it launched a new discussion or a response from researchers I was not expecting, so I did not necessarily obtain an answer that would help me determine whether there was knowledge satisficing in the way I expected.

This made sense to me when considering my theoretical framework in retrospect. On December 20, 2020, I wrote in my research notes,

> "Something to be aware of is that asking researchers about knowledge they wanted but could not obtain is a positivist way of looking at the problem (knowledge is "out there").

> A more interpretivist view would be to see them as creating the knowledge they needed and I might ask them how they created the knowledge sufficient to accomplish their task." (York, personal research notes, 2020)

It seemed obvious in hindsight, but when I started to feel that satisficing was not fitting well with the way researchers described their experiences, I started to realize that satisficing is a concept that, in pre-supposing or assuming an objective "optimal" solution to a problem, itself takes a "gods-eye" view of the circumstances; not admitting that what is optimal might vary from perspective to perspective, and that in some cases it may simply not exist. It was no wonder, then, that in the approach I was taking to understand how researchers bounded knowledge in the context of situated knowledges, satisficing was not emerging as the best fit. And yet I was gathering evidence that researchers lacked knowledge about data that they desired. This indicated that there was a change between some kind of optimal state and an acceptable state, and that something could be done to improve the experience, value, and impact of reusing data. I discuss these findings further below.

These three findings, that 1) researchers experienced a lack of knowledge (or what other researchers reported as a lack of knowledge) but did not mention them in the survey, 2) the language I was using to ask about "enough" knowledge felt oppositional to some researchers, and 3) satisficing might not be the best characterization of what happens in data reuse caused me to relax my protocol somewhat, particularly in the area of knowledge bounding but also in relation to the order of questions. I endeavored to keep the same general structure to interviews and ask the same core questions but was flexible to let researchers discuss what was most meaningful to them and ask probing questions—e.g., about what influenced them to determine sufficient thresholds of knowledge—at whatever point in the interview relevant issues came up.

The result over time was interviews that were more conversational and in which I think researchers felt more comfortable discussing issues of how they bounded knowledge about data and negotiated situations where there was knowledge about data they desired but were not able to obtain. An additional adjustment I made to facilitate a more conversational feel was to remove a question at the end of the interview about whether the researcher had any difficulty answering the questions I asked.

There are some additional points about developing the interview protocol that are important to mention. One is that, while I made adjustments to the interview protocol after I started the interviews, I also conducted four practice interviews with researchers who had also tested the survey protocol earlier in the process. Some of these indicated in the test survey that they lacked knowledge they desired and some did not. I did not make a lot of adjustments to the protocol after these tests, but they were invaluable for me to test the flow of questions in different scenarios and gain confidence that my questions could yield the kinds of responses I expected.

A second point is that I did make a couple of adjustments in addition to those I mention above. One of these was to remove a question about the impact of satisficing on the outcomes of the researcher's research. I found in the interviews that responses to this question were very similar to ones where I asked researchers to explain why they answered the question in the survey about how well their reuse of the data met their research goals.

The final point is that the introduction to the interview was particularly important to me. I wanted to introduce the interview and cover logistical aspects. As discussed above, however, I also wanted to provide some background about myself and be transparent with researchers about my motivations for conducting the research and my theoretical perspective. Although in some

cases it may have made the researcher more suspicious about what I was trying to achieve, for the most part I think being transparent conveyed my intention to understand the researcher's experience at a deep level and my hope that they would be open in the responses they gave. The full introduction is in Appendix B.

### *3.4.5 Selection of Interviewees: Integration of Survey Results*

One of my primary goals in the interviews was to produce comparable findings about how researchers determined the thresholds of the knowledge they obtained about data in order to reuse them in their research. I sought to achieve this goal by comparing results from multiple clusters of individual researchers who reused the same data study. I started my selection process by examining respondents to the survey who indicated they would be willing to be contacted with additional questions (there were 284). I then made a list of these respondents who were included in one of the samples of researchers who reused quantitative data (either the random sample based on reuse of quantitative data, or the sample of reusers selected based on the number of citations to the data studies they reused). I made a second list comprising reusers of qualitative data. There were 63 researchers in the first list who reused quantitative data (comprising multiple reusers of each of seven different data studies). The second list included 11 researchers who, I thought, had reused qualitative data (comprising three reusers of each of two data studies, two reusers of a third study, and single reusers of three additional studies). In fact, as I learned when I began to read researchers' published papers, 10 of the 11 qualitative researchers in the second list reused only quantitative portions of the data, and only one actually reused qualitative data.

I focused first on the list of quantitative data reusers and used specific criteria to identify clusters of individuals who had reused the same data to contact for interviews. I established a

priority level for different criteria, represented in Table 3.9, and then selected researchers in accordance with the criteria, striving for maximum variation and diversity in areas where diversity was desired.

I was guided in my decision to strive for maximum variation in part by literature in case study research. Case studies are often used to generate or test hypotheses, build or extend theories, or describe or analyze phenomena (Flyvbjerg, 2006, Eisenhardt, 2002, Creswell, 2007). Although I was not seeking to generate or test hypotheses or build theory in the qualitative portion of my study, I was seeking to apply theories (i.e., of Haraway and Simon) in order to gain a new perspective on phenomena (i.e., knowledge bounding, and potentially knowledge satisficing), which might contribute to or extend theory.

In writing about the selection of cases to build or extend theory, Eisenhardt (2002) notes that "it makes sense to choose cases such as extreme situations and polar types in which the process of interest is "transparently observable"" (Eisenhardt, 2002, pp. 12-13). Accordingly, in order to understand and elucidate the application of Haraway and Simon's theories to knowledge about reused data, I purposively selected instances of reuse for the interviews where researchers and research circumstances demonstrated maximum variation along criteria measured in the survey.

Table 3.9 gives an overview of the prioritized criteria I used to select researchers to contact for interviews. My highest priorities were to include researchers:

- who reused the same data as other researchers in my sample
- who finished the survey
- who both had and had not lacked knowledge about data
- for whom the knowledge they lacked was important, or whose research was negatively impacted by the lack of knowledge
- who had a spectrum of involvement in the decision to reuse the data
- who had a spectrum of involvement in determining the goals of the research

- who both believed the data they reused were created explicitly to be reused, and who did not
- who obtained data from ICPSR and who did not
- whose research team (including themselves) had varying levels of knowledge about the original creation of the data
- who had a spectrum of professional positions (PhD student, assistant professor, full professor, etc.)
- who reused data for diverse purposes (and for whom those purposes were important to their research)
- whose research questions developed in a variety of ways (i.e., whose questions were set before they reused the data and whose questions developed as they reused the data)

*Table 3.9 Description of Prioritized Criteria for Selection of Interviewees*

| **Dependent Variables** | | |
|---|---|---|
| **Concept** | **Concept operationalization** | **Priority and Selection Criteria for Interviewees** |
| Satisficing (Satisficing is also an independent variable) | Whether, when considering reusing the data, there was additional knowledge about data that was desired | Not used in selection |
| | Whether, when the decision to reuse the data was made, there was knowledge about the data that was desired but not obtained or obtained only in part | Priority 1; mix of Yes and No |
| | The three most important things the researcher would have liked to know about the data but was not able to obtain, or obtain to the desired degree. | Priority 2; diversity of types of knowledge lacking |
| | How much of each type of desired knowledge was obtained | Priority 2; spectrum of amounts obtained |
| | The importance of the knowledge that was not obtained or obtained only in part to deciding to reuse the data | Priority 1; knowledge lacking must be important, or if not important, have a negative impact |
| Outcomes of research | Whether (and how much) a lack of knowledge affected the outcomes of the research | Priority 2; spectrum of impact (based on the maximum impact across types of knowledge lacking) |
| | Why the lack of knowledge did or did not affect the outcomes of research | Priority 2; diversity of impact reasons |
| Goal attainment | Researcher perception about goal attainment for data reuse | Priority 2; spectrum of goal attainment |
| **Independent Variables** | | |
| Administrative | Survey completion | Priority 1: all respondents must have completed the survey |
| | Type of reference (Journal article, etc.) | Priority 3: spectrum of reference types |
| | Type of data reused (survey, clinical, etc.) | Priority 3: spectrum of data types |
| | Type of data reused (quantitative or qualitative) | Priority 2: spectrum of types |
| | Number of citations to data | Priority 2; diversity of high and low citations |
| Information about researcher's perspective | Whether data were reused | Priority 1; all data must be reused |
| | Researcher's role | Priority 1; priority given to lead roles (PI, Co-PI), spectrum of roles after that |
| | Researcher's involvement in the decision to reuse the data | Priority 1; spectrum of involvement |
| | Number of people involved in the decision to reuse the data | Priority 2; spectrum of numbers of people |
| | Researcher's involvement in determining the goals of the research | Priority 1; spectrum of involvement |
| | Data obtained from ICPSR or another source | Priority 1; diversity of data reused from ICPSR and from other sources |
| Researcher "distance" from the data | Researcher's perception of original data creation purpose | Priority 1; spectrum of opinions about data creation purposes |
| | Researcher's involvement in the original research | Priority 1; spectrum of involvement |
| | Knowledge of research team about the way the original data were collected or produced | Priority 1; spectrum of knowledge |
| Experience with data reuse | Researcher's years of experience in primary domain of research | Priority 2: spectrum of experience |
| | Researcher's unit or departmental affiliation | Not used in selection |
| | Researcher's professional position | Priority 1; diversity of positions |
| | Researcher's primary domain of research | |

| | | |
|---|---|---|
| | Whether the research conducted was in the researcher's primary domain of research | Priority 3; diversity, but responses were nearly all Yes |
| | Domain of research in which the reused data were produced | Not used in selection |
| | Researcher's years of experience reusing data produced in this domain | Priority 2: spectrum of experience |
| | Researcher's years of experience reusing data produced in any domain | Priority 2: spectrum of experience |
| | How often researcher's work involved reuse | Not used in selection |
| | How often researcher's work involved reuse for the same purpose as these data | Not used in selection |
| | Number of reuse projects researcher worked on | Not used in selection |
| | Number of reuse projects where reuse considerations were substantially similar to ones involved in reusing these data | Not used in selection |
| | Number of authored or coauthored published papers describing research involving data reuse | Not used in selection |
| | Number of authored or coauthored published papers describing research involving data reuse where conditions were substantially similar to those involved in reusing these data | Not used in selection |
| Reuse ability | Researcher's facility with reuse of data produced in primary domain of research | Not used in selection |
| | Researcher's facility with reuse of data produced outside primary domain of research | Not used in selection |
| Reuse motivation | Reuse purpose(s) | Priority 1; spectrum of purposes |
| | Relative importance of each reuse purpose | Priority 1; purpose at least somewhat important to the researcher |
| Research process | How research questions were developed | Priority 1; diversity of processes of research question development |
| | Information source(s) most important in obtaining desired knowledge | Priority 3; diversity of sources |

The reason some of these criteria where of high priority is clear in some cases (e.g., that the knowledge researchers desired but lacked was important to them) and perhaps less clear in others. The variables I identified as possible confounders in my survey analysis were whether or not the data had been created to be reused; whether the data were reused from ICPSR or another source; and the number of citations to the data (I discuss these in sections 2.9.4 and 4.1.3). The first two I included as first priorities in the selection and I added the third (number of citations) as a priority after interviewing an initial wave of researchers who had all reused highly cited data (I talk about this further in my description of the qualitative methods below).

Knowledge about the original data collection and the way research questions developed were variables I found in the survey had significant relationships with lacking desired

knowledge, as were several of the purposes of research. I thus included these as first priorities. While I did not find significant relationships in the survey, I felt that looking across a diversity of researchers' involvement in the decision to reuse data and involvement in setting goals for the research was important for obtaining maximum variation and included them as top priorities as well. Similarly, although I did not find significant relationships between measures of experience and researchers lacking desired knowledge, I wanted to explore why in the interviews. I saw professional position as an efficient way to do this since it captured differences both in experience with reuse and experience in the profession.

I made decisions about the second and third priority criteria using similar logic: considering the statistical significance of the variables in the survey analysis and my interest in examining the variables further with an eye to better understanding the survey results and my phenomena of interest.

### *3.4.6 Interview Preparation*

In addition to preparing the interview protocol and selecting participants for the interviews, an essential component of the interview process was my preparation ahead of each interview. This preparation involved several components related to each researcher:

- assembling background biographical information and publication history
- reading and coding the publication in which the researcher reported their reuse of the selected data
- creating an interviewee profile composed of elements from the survey, biographical information, publication history, and research publication
- preparing a custom interview notes sheet for each interviewee that included elements from the profile that I thought were important to have in mind during the interview, keep in mind for follow-up questions, or ask researchers about specifically

I intended the background research to provide me with knowledge I could use to establish rapport with research participants in the context of the interview, better understand participants' responses to questions, and probe or ask follow-up questions to explore nuances and

particularities appropriate to each research instance. I also intended the research to provide a means of triangulating multiple sources of evidence in each case (Yin, 2014). Triangulation is discussed further in sections 3.5 and 3.6.2.

#### *3.4.6.1 Background Information*

I had some information already about participants from the survey, including their primary domain of research, academic position and department at the time the research was conducted, as well as a variety of information about their experience. I wanted to have more current information about the researchers, however, and understand any details about their current work and position that might help me understand where the critical case I would be asking them about fit in to their broader research trajectory. For instance, was the publication typical of the kind of research they performed—in terms of topic, collaborators, method, or other features—or more of a departure or one-off? Was the publication a central research interest or something more peripheral? Were there other details the researcher shared about their personal interests or experiences in their professional biography that might be relevant to understanding their motivations for the research reported in the publication?

To obtain this information I conducted searches online using the name of the researcher, the email address I used to contact them for the survey, or the email address at which they asked to be contacted for the interview. As part of my effort to learn about where the critical case fit in their research trajectory, I visited sites like Academia.edu, ResearcherGate.net, and Google Scholar and assembled lists of citations and abstracts for papers researchers had published over their careers. I counted the number of publications before and after the one I had selected for my study in which they were one of the first several others. I also noted how extensively they had worked on this topic, with these data, and with specific collaborators.

While some of the information I was able to obtain through these searches was helpful, I encountered some difficulties. It was difficult, for instance, to find the same information for all interviewees. Biographies had some common information, but some were much more descriptive than others and I was not able to find one for several researchers. It was hard to determine how important one bit of information I found about one researcher was to my research questions and to their story let alone make a comparison with a different bit of information in a different researcher's story. In a similar way, information about publications was available for most participants but I found differing information posted on sites like Academia.edu and researcher's personal web pages, for example, and differing levels of granularity in the publication history for different researchers. Since different amounts of information were available and I didn't have a basis for ascertaining which aspects of researcher's biographies might be most important for knowledge satisficing and knowledge bounding, I determined that the best means of understanding what affected researchers' behavior in these regards would be by asking them in the interviews. I therefore put more weight on the interviews than biographical research in my analysis.

#### *3.4.6.2 Research Publication*

I next obtained the full text of the publication that was the basis for my selection of each researcher and coded it for several elements. These were:

- what's the point (of the article)
- what data were used
- where the data fit in (how they were used to answer the research question(s))
- limitations of the study
- knowledge about the data the researcher desired but did not obtain

For each publication, I wanted to understand the main point of the research and what data were used, including the specific portion of the data when applicable and if provided. I captured

particular measures that were important and how the researcher used the data to accomplish their purpose. I also wanted to understand the limitations the researcher identified or encountered. I found that in the limitations section of papers researchers often listed both limitations of the data and of their research—for instance that the original data did not allow generalization, or the researchers made particular choices about what data to use and avenues to investigate—as well as limitations of their knowledge about the data such as not knowing how attributes of the sample they selected from the data compared to the general population or details about measures that were masked in the data that were shared.

From the limitations sections and other portions of the publications I constructed lists, where they were relevant, of knowledge that researchers desired about the data but were not able to obtain. Where possible, I compared knowledge from the lists with responses to the survey. I then made notes about particular knowledge I wanted to ask researchers about in the interviews, even in cases where researchers said they did not lack knowledge, or where they said did but did not provide details.

#### *3.4.6.3 Interviewee Profile*

I put all of the information I gathered into a document organized by interview wave where I kept a profile of each researcher. The profile contained the information I could assemble from the survey, background research, and the article.

#### *3.4.6.4 Interview Notes Sheet*

Before each interview, I took selected information from the profile and added it to a bespoke copy of the interview protocol I created for each researcher. For instance, I wrote notes about the knowledge I identified as lacking from reading the publication next to my questions about how the researcher determined the thresholds of knowledge about the data that were

sufficient to reuse them, and included information about the amount of knowledge about the original data collection on their research team and how well their reuse of the data met their goals at relevant points in the protocol. I then used that information to customize questions and jumping off points for follow-up questions. Aside from helping me to have more in-depth discussions with researchers about issues that were relevant to my research questions, I wanted to have this information on hand and make each interview unique so that researchers knew I had invested time in learning about their research so I could make the most out of the time I had with them. I hoped this would give them confidence in my research and increase their willingness to speak freely in response to my questions.

### *3.4.7 Interview Conduct*

I conducted and recorded five waves of interviews between January 27, 2021 and April 8, 2021. I contacted six researchers in the first wave, 10 in the second, 11 in the third, 11 in the fourth, and 15 in the fifth. I received three, four, five, seven, and seven responses for each, respectively. The first three waves comprised clusters of researchers who reused the same quantitative data. The fourth wave was made of up of researchers who reused qualitative data (including some groups who reused the same data). For the fifth, I drew on researchers who agreed to be contacted but had not reused the same data in order to select researchers who reused data that had been cited less frequently. I contacted 53 researchers in all, conducting 26 interviews for a response rate of 49%.

### *3.4.8 Interview Analysis*

I sent the recordings to a third party for transcription. After verifying them, I performed qualitative coding using two applications: NVivo and Taskpaper. I report on my analysis of the interviews below.

#### *3.4.8.1 Interview Analysis Strategy*

To develop my strategy for analyzing the interviews I drew in part on case study research. My research differed from typical case study research in that I was not observing a phenomenon in its natural setting (i.e., I did not observe data reuse as it happened) (Flyvbjerg, 2006) and I did not collect or analyze a diversity of sources (i.e., I did not use observations of reuse or documentation directly related to reuse instances other than the resulting publication) (Benbasat, 1987). However, a key similarity between my research and typical case study research is the importance of context to each reuse instance, and the value, therefore, of understanding each reuse instance on its own terms before seeking broader comparisons with other reuse instances.

In his guidelines for analyzing multiple case studies, Stake (2006) argues for maintaining a healthy tension between the analysis of a phenomenon of interest as it is manifest within individual cases and the analysis of the phenomenon as it is manifest across multiple cases as a whole (Stake, 2006, p. 46). He refers to findings that pertain to individual cases as Findings and findings that cut across cases as Themes. Stake argues that researchers should not hurry to the analysis of Themes and should only make comparisons on a limited number of Themes across cases. This preserves a "dialectic" between generalizations across cases and Findings in individual cases. According to Stake, coincidences between Findings and Themes consolidate and extend the understanding of the phenomenon.

I adopted this conception of the relationship between individual and multiple cases in analyzing the reuse instances in my study and employed it at several levels. Although adopting this conception did not work out as I planned, I describe it here because it was fundamental to the design of my process for conducting and analyzing the interviews. More specifically, as I have related, I sought through the interviews to produce comparable findings about how

researchers determined the thresholds of the knowledge they obtained about data in order to reuse them in their research. My primary strategy to achieve this was to compare results from multiple clusters of individual researchers who reused the same data study.

Accordingly, and in alignment with case study analysis, I planned, first of all, to understand each reuse instance on its own terms—including the very specific history, context, and circumstances that came to bear on the presence, extent, and type of factors that affected knowledge bounding and knowledge satisficing in each situation. Second, I planned to analyze similarities and differences in patterns across instances of reuse of the same data study. Finally, I planned to analyze similarities and differences across instances of reuse of different data studies (i.e., across clusters of researchers who reused the same data). My goal in taking this multi-step approach was to preserve the particularity of the individual contexts in which reuse of data took place while making the most of the opportunity to examine, compare, and perhaps generalize about knowledge bounding and knowledge satisficing across reuse instances.

I ultimately performed three cycles of coding on the interviews, employing a variety of coding strategies. In the first cycle, I performed attribute and structural coding, in the second I performed initial, descriptive, in vivo, values, and process coding, and in the third I performed pattern and focused coding. These are methods, with the exception of process coding (which I included to track activities that had a bearing on the way researchers bounded knowledge about data), that Saldaña (2009) suggests as a set of "generic" coding strategies (p. 48).

I provide Saldaña's description of each type of coding and how I employed it in Table 3.10. While these coding strategies were effective for my analysis, I took an overall approach to coding that Saldaña (2009) calls "pragmatic eclecticism." This is an approach where researchers "believe in the necessity and payoff of coding for selected qualitative studies, yet wish to keep

themselves open during initial data collection and review before determining which coding methods—if any—will be most appropriate and most likely to yield a substantive analysis" (Saldaña, 2009, p. 47). In keeping with this approach, I did not completely pre-determine what I would code or limit what I coded to the definitions in Table 3.10, though these strategies predominated.

*Table 3.10 Analysis Methods for In-Depth Interviews and Background Research*

| Stage of Coding | Type of coding | Purpose of coding | Use in research |
|---|---|---|---|
| First Cycle | Attribute coding | Used to record contextual information about the research setting including attributes of research participants and other information of interest | To record details about each reuse instance, such as what dataset was reused, the format and other characteristics of the data, and attributes of the researcher and research that are drawn from the quantitative survey and background research. |
| | Structural coding | Used to apply a specific concept or topic to a segment of data that relates to research questions used to frame the interview | To code responses to particular questions I asked in the interviews that were associated with specific research questions (e.g., how researchers obtain their desired knowledge) |
| Second Cycle | Initial coding | Describes an initial summarization, categorization or condensation of ideas contained in a segment of qualitative data | To collect, process and understand ideas from the interviews |
| | Descriptive coding | Used to summarize topics that are identified | To identify potential stages or attributes of knowledge bounding and areas of knowledge satisficing, as well as perceptions and beliefs, resources, and other factors that might affect both phenomena |
| | In Vivo coding | Use of coding terms that are drawn from the language of the researchers themselves | To strive as much as possible to capture the meanings researchers ascribed to concepts that were important to their reuse of data |
| | Values coding | Used to identify attitudes, values, and beliefs | To record specific dimensions of attitudes, values, and beliefs expressed by participants |
| | Process coding | Used to identify actions | To identify activities researchers mentioned in interviews that were associated with knowledge bounding or satisficing |
| Third cycle coding | Pattern coding | Used to group summaries and other codes into themes, concepts, causes, or explanations | To identify concepts, categories, and themes from the other coding processes across reuse instances and reuse clusters. |
| | Focused coding | Use to identify the most frequent or significant codes from first cycle coding to use in higher-level analysis | To identify concepts, categories, and themes from the other coding processes across reuse instances and reuse clusters. |

*Note*. All types and purposes of coding are taken from Saldaña (2009).

### *3.4.8.2 First Cycle Coding*

The first kind of coding I applied to the interviews was structural (Saldaña, 2009), using my interview questions (primarily) as a basis for developing codes to identify important high-level concepts throughout the interview (see Table 3.11). I did all of the coding in NVivo 1.5.1 (2021). I created 8 high-level codes: Background, Research Context, Knowledge-bounding, Reuse expectations, Knowledge bounding, Reflection-specific, Reflection-general, Closing, and

Conversation. The last code, Conversation, is the one code not based on the interview questions. It is something that I noticed coming up repeatedly as I conducted the interviews that related directly to the conversations that are part of my theoretical framework and I wanted to capture them as a main first-cycle code.

The next level of codes related more specifically to my interview questions. The questions provided a broad outline for the interview, but researchers would sometimes respond to one question with information that was relevant to many questions or provide information relevant to a single question throughout the interview. Both of these occurred especially in the first three parts of the interview (related to Background, Research context, and Knowledge bounding). For example, sometimes researchers described considerations and pressures or concerns when they talked about the motivation for the research or gave additional information about sources of knowledge about the data when they talked about how they negotiated gaps in their knowledge. I used the second-level codes to capture these more granular areas of discussion wherever they occurred in the interview.

*Table 3.11 Codes and Definitions Used in First Cycle Structural Coding*

| Level 1 code | Level 2 code | Coding definition |
|---|---|---|
| 1-Background | | |
| | 1-Motivation research | What influenced the person to investigate this research area. |
| | 2-Motivation paper | What influenced the person to pursue the research for this paper |
| | 3-Motivation data | What influenced the person to choose the data; could have been prior research or another person; includes description of data, discussion of other data that were considered, and discussion of how they gained access to data (for instance, these data were easier to access than other data). Description of challenges they faced in accessing data (e.g., privacy restrictions) go elsewhere, unless they talk about them as part of their general response to questions about what motivated them to choose the data. |
| | 4-Who involved | Who was involved? Includes anyone involved in the research: consultants, those who helped with analysis, and those who were sources of knowledge (but only if they were involved in the research; if a person who is a source of knowledge was not involved in the research—for instance, the researcher contacted the original data creator for information—that person should be coded under 5-Sources of knowledge about the data). |
| | 5-Sources of knowledge about the data | Sources (including people) from which the researcher gained knowledge about the data (including about validity, context of creation, etc.) |
| | 6-RQ Development | Information about how the research question(s) developed—for instance, if the person had a research question ahead of time or if it developed as they worked with the data |
| | 7-Guidelines or training | What guidelines the person used or training they obtained related to data reuse (how did they learn to use the dataset?). Includes guidelines from IRB, ICPSR, or others about proper use of data. Sometimes researchers obtained knowledge about how to use the data from literature or websites. In these cases, code the literature and websites as 5-Sources of knowledge about the data in addition to 7-Guidelines or training. |
| | 8-Reuse purpose | Information about the researcher's purpose for reusing the data, e.g., to compare with other data or to answer a new question. Full list from the survey: background, combine, compare, new question, replicate, test theory, tool creation, validate. |
| | 9-Subsequent research | Discussions about research subsequent to the paper of interest. Sometimes researchers talk about research they did after this paper, or research they plan to do in the future. |
| 2-Research context | | |
| | 1-Considerations and pressures | What kinds of considerations and pressures the person experienced related to data reuse. Some people talk about issues of access to data here (e.g., R14; for instance, if there was pressure because gaining access to the data caused a time delay or they reused the data because there would be less of a delay. Considerations and pressures that had a bearing on the person's decision to reuse the data should be coded as 3-Motivation data. |
| | 2-Concerns | Any concerns, including knowledge lacking about the data; include concerns about age of data. |
| | 3-Perception of reuse | How reuse is perceived by people in the researcher's immediate community. Only code this when it is the response to my question. |
| 3-Knowledge bounding | | |
| | 1-Knowledge desired | Information about what specific knowledge about the data the researcher desired (information about the "impact" of the lack of knowledge should be coded under "4-Reuse expectations"). I often ask about this explicitly with questions about the most important things the researcher wanted to know about the data or I recall specific answers they gave in the survey. Sometimes they mention additional knowledge they desired about the data (like more granular location information) and when they do, I want to code those also. A concern about the age of the data or |

| | | |
|---|---|---|
| | | other things they know about the data would go under concerns, not knowledge desired (because, e.g., in this case, they know how old the data are). |
| | 2-Gap negotiation | How researchers negotiated gaps in their knowledge, decided that they knew enough about the data, or decided the data met the threshold to be "enough" for their research; include negotiations related to concerns about the age of the data, and discussions about publication review. Excludes negotiations related to knowledge about analysis techniques or software. |
| 4-Reuse expectations | | Responses to my question about reuse expectations OR questions about the impact of the knowledge that was lacking. I often code these as 2-Gap negotiation as well because there is often discussion about why the researcher understood the data to meet or exceed expectations (i.e., determine whether the data were sufficient for their purposes). |
| 5-Reflection-specific | | What facilitated or hindered obtaining desired knowledge in this project? Code mainly in response to the question about what facilitated or hindered attainment of knowledge in their reuse of the data for this paper. Code if in other parts of the interview they talk about things that would have facilitated attainment of knowledge. |
| 6-Reflection-general | | What hinders or could facilitate obtaining desired knowledge in general? Code only in response to the question about what facilitated or hindered attainment of knowledge generally. |
| 7-Closing | | |
| | 1-Difficulty with questions | The person's response to my question about whether they had any difficulty answering the questions |
| | 2-Anything to add | The person's response to my question about whether they had anything to add |
| | 3-Final notes | Final chatter after the questions have finished |
| 8-Conversation | | Discussion of conversations related to enhancing analysis of data by getting variables of interest included in future studies. |

#### *3.4.8.3 Inter-Rater Reliability*

After my first cycle of coding across the interviews, I and a research assistant conducted inter-rater reliability (IRR) testing to evaluate the understandability of the codes and how consistently they could be applied. There were a couple of components to the testing. I asked the research assistant to pick three interviews to code, choosing randomly between 1 and 26 (the numbers of the interviews I conducted). I excluded the longest and shortest interviews from selection because they were a little bit out of the norm, and I did not think comparing our coding of them would be as good of a test of the codes as others that were closer to the norm.

I first asked the research assistant to code one interview, following which I calculated IRR statistics, we discussed differences in our coding, and I revised the codebook based on her and my own recommendations. In order to discuss the differences, I made a spreadsheet of all disagreements and, for the entries I had coded, included a rationale of why I had coded selections

the way I did. After we had talked about and understood the differences, I identified areas that the research assistant coded differently from me that I agreed with and re-calculated the IRR statistics counting these as agreements (I did not change instances where the research assistant agreed with my coding because I was interested in what could be coded, not what I could convince others of afterwards). I then asked the research assistant to code the remaining two interviews and I repeated the processes of calculating the IRR statistics, listing instances where we disagreed in our codes, and updating the IRR statistics after re-counting instances where I agreed with their coding.

I used two formulae to calculate inter-rater reliability: Holsti's Coefficient of Reliability (Holsti, 1969, pp. 138-141) and Scott's Pi (Scott, 1955 discussed in Holsti, 1969, pp. 138-141). The formula for Holsti's Coefficient divides twice the number of total judgements on which the coders agree (M) by the number by the total number of judgements each coder makes ($N_1$ and $N_2$). This is shown in Formula 3.2:

*Formula 3.2 Calculation of Holsti's Coefficient*

$$Holsti's\ Coefficient = \frac{2M}{N1 + N2}$$

Holsti points out that this formula has been criticized because it does not account for there being some level of coding agreement due to chance. Scott's Pi accounts for this by considering both the frequency with which categories are used ($N_1$ and $N_2$ above) and the total number of coding categories. Scott's Pi is equal to the ratio shown in Formula 3.3.

*Formula 3.3 Calculation of Scott's Pi*

$$Scott's\ Pi = \frac{(\%\ observed\ agreement) - (\%\ expected\ agreement)}{(1 - \%\ expected\ agreement)}$$

In this equation, percent observed agreement is equivalent to Holsti's Coefficient and percent expected agreement is the number of times a particular code is used divided by the number of times all codes are used. Scott's Pi is calculated by summing the squares of the percents of expected agreement.

Table 3.12 shows the initial coefficients and those I revised after re-designating instances the research assistant coded as agreements rather than disagreements. In both the initial and revised calculations, Scott's Pi makes a downward correction on the Holsti's Coefficient (< 0.03). This is expected since Scott's Pi accounts for chance agreement. While there were other differences, I attributed the lower agreement in the third coded interview primarily to one code (Motivation data) that I applied several more times than the research assistant.

*Table 3.12 Initial and revised inter-rater reliability coefficients for three test interviews*

| | Initial Holsti's | Initial Scott's Pi | Revised Holsti's | Revised Scott's Pi |
|---|---|---|---|---|
| Test 1 | .724 | .743 | .816 | .802 |
| Test 2 | .735 | .712 | .792 | .774 |
| Test 3 | .667 | .638 | .72 | .697 |

Krippendorff has said that agreement above 80% is good but that 67-79% is acceptable (Krippendorff, 1980) and Bernard (2012) recommends that a statistic of 0.7 (70%) be achieved for Scott's Pi. I took the fact that both the initial and revised statistics were in an acceptable range as an indication that the codes I had developed could be applied to the interviews with reasonable consistency. I interpreted the data this way for a couple of reasons. First, I found in the process of listing out and talking about coding disagreements with the research assistant that several of the disagreements we had could be explained by the additional context and experience I had about the interviews. For instance, researchers sometimes referenced data studies that they had not reused without distinguishing those from data they had reused. As a result, for someone coding who did not have full context, information referring to an unused dataset could be

interpreted as relevant when in fact it was not. Researchers also sometimes talked about multiple times that they reused the data, and knowledge relevant to a previous instance that was not relevant to the current instance could be coded.

In addition, as we conducted IRR, I was in the process of performing a second cycle of coding. In that cycle I was finding sections of text that I believed I should have been coded with a first cycle code. This, in addition to my agreement with codes the assistant researcher applied that I had not, made me believe that the coding could continue to be refined in the future (and would be as I continued the second cycle coding), but we had demonstrated that the codes and definitions could be applied to the interviews with a reasonable baseline level of agreement. Since greater agreement on one code in particular in the third test (Motivation data) would have made a large difference in the IRR coefficients, I saw this as a more transitory discrepancy rather than one indicative of larger problems in the coding scheme or its application.

#### *3.4.8.4 Second Cycle Coding*

Having organized the interviews into high-level categories, I began a deeper dive into the text to better understand, and code, researchers' experiences reusing data. In this process I used a combination of initial, descriptive, in vivo, values, and process coding Saldaña (2009), and I moved from using NVivo to another tool called TaskPaper to analyze the data. TaskPaper is a text editor with enhanced features that enable the user to create virtually infinite hierarchies of bullet points and use tags to categorize and organize text. TaskPaper made it easier for me to accomplish several tasks that were essential to my analysis strategy.

Because of the centrality of situated knowledges to my theoretical framework, it was important for me to try to understand as much as I could about each individual researcher's experience that had a bearing on the way they determined the boundaries of the knowledge about

data that were sufficient for them to decide to reuse the data in their research. In order to accomplish this, I wanted as much as possible to understand each researcher's experience on its own terms before comparing it to the experiences of others. Practically speaking, this translated into heavy use of in vivo coding—using the language and concepts that the researchers used to describe, for example, their concerns about the data, considerations and pressures involved in reusing the data, how they negotiated situations where they did not have all the knowledge about data they desired, or values they held and processes they engaged in related to data reuse.

For the second cycle of coding, then, I systematically went through each interview, pulling up all the text I had coded under each first cycle code sequentially, starting with Motivation research. I then summarized concepts, such as what motivated the researcher's research and their use of the data in short phrases. I also included quotes as bullets under the phrases when I felt the quotes were particularly useful to illustrate or explain a phrase.

I used TaskPaper's tagging feature to tag the descriptive phrases with the number of the interview and the first-cycle coding area the phrase belonged to. I also began to tag concepts, feelings, values, processes, considerations, and a variety of other categories of information that were relevant to my research questions, especially how researchers bounded knowledge they obtained about data, what influenced those decisions, and what researchers believed could mitigate situations where more knowledge about data was desired than could be obtained.

If different researchers described their feelings or experiences in the same way, I used the same tags; if they described their situations in different ways, in many cases I used different tags that reflected the way they spoke about their circumstances. In some cases I made up tags to reflect common ideas researchers were expressing. One example of this is the tag "sunk_cost." A number of researchers described the large investment of time required to get to know a dataset

and ways that investment influenced the data they selected and how they proceeded with their research. While many of these experiences were different, and I simultaneously used additional codes when relevant, I used "sunk_cost" as a way to identify these situations across the interviews and aggregate them together at the end.

Once I had conducted second level coding on all of the interviews, I used pattern coding and focused coding to collect the long and disparate list of codes I had created into themes and more focused categories. I kept the original codes I had applied but organized them under broader headings. For instance, I organized the different codes I had assembled under the heading "Considerations and pressures" into categories of individual, data, and social.

It is through this preservation of the particular in the aggregation to the general that I tried to preserve the experiences of individual researchers in my analysis. The tags or codes I applied take into account my understanding of the individual experience and situation of each researcher, which I brought together, but preserved, in the higher level themes and headings I created based on similarities across the interviews.

#### *3.4.8.5 Analysis Considerations: What is Knowledge?*

In discussing the development of my interview protocol, I described my experience that the wording of my question about lacking knowledge, i.e., how researchers knew they had "enough" knowledge about data to reuse them, seemed to come across as oppositional to some researchers. This led me to change the wording of the question—to ask instead how researchers knew the data were sufficient to fulfill their needs, or how they knew they had reached a threshold where their knowledge was enough to reuse the data. It also caused me to begin to question whether the notion of satisficing accurately characterized what my interview participants experienced as they determined the limits of knowledge about data they deemed

sufficient to decide to reuse the data. An additional finding I related was that it was evident from the publications I read prior to the interviews that many researchers who did not report lacking desired knowledge in the survey did in fact lack knowledge—at least, the same kinds of knowledge that other researchers I surveyed reported as lacking. I discuss these findings in more detail here to provide context for additional findings from the interviews below.

At the heart of all three issues above—how questions about knowledge were phrased, whether satisficing was an appropriate frame to understanding knowledge that is lacking in the context of reuse, and whether researchers reported lacking knowledge—is the question of how "knowledge" is defined. What does "knowledge about data" mean, and does that differ from knowledge about the context of data? If data were not collected during the original study but data reusers wanted that knowledge (e.g., what I have called data supplement (coverage) and data supplement (detail)), does it still count as "knowledge about data?"

When I created the survey, one of the issues I considered was whether to try to define for researchers exactly what I meant by "knowledge about data." In keeping with my theoretical framework—which problematizes a separation of "data" and "context"—and to understand from the researcher's perspective what they desired to know, whatever it was, I did not. This led to some surprises, however, and difficulties I had not anticipated.

For instance, if a researcher suspected that some data were missing but did not know for sure, I would consider that to be knowledge a researcher lacked "about" the data (i.e., something about the context of data creation). However, if the researcher knew that data were missing but wanted to know what the specific missing values were, would that constitute knowledge "about" data, or something else, like a limitation of using the data? I came to think of this as a distinction between knowledge "about" data and knowledge "of" data. Knowledge "about" data I

characterized as knowledge that might typically be thought of as context, such as how data were collected or analyzed, or other knowledge about the data. Knowledge "of" data I characterized as knowledge of a sample or population itself (including specific values when data were missing). I discuss knowledge "of" data further in section 4.2.3.1. Since researchers indicated they lacked knowledge about both in the survey, and the distinction between the two was sometimes times hard to draw, I included both in the scope of "knowledge" about data for the purposes of my study.

Unfortunately, not all researchers viewed the scope of "knowledge" in the same way. During an exchange, one researcher asked:

> Okay, so when you say what I know about the database [what is included in the collected data], to me, that's different than [knowledge about] the data, right? So does that make sense? To me, I distinguish between what limitations of the data and the limitations of the database or data collection. And I think what you're asking are what's the limitations of the data? (D-11)

This was the only researcher who raised a question in the interviews, but I think this discrepancy in interpretation of "knowledge" was likely a reason for variable reporting in the survey about knowledge researchers desired but were not able to obtain (which I observed when reading researchers' publications and when conducting interviews, as I discuss below).

In this case, the researcher eventually understood why I was including both knowledge "about" data and knowledge "of" data in my study and the difficultly at times of distinguishing between them.

Another example of a difficulty related to knowledge "about" or "of" data that I encountered somewhat frequently were cases where a researcher desired additional knowledge about the population or phenomenon they were studying—perhaps additional characteristics of the people involved or additional detail about the phenomenon of interest. Even well into my

interviews, I had trouble distinguishing whether certain types of knowledge researchers discussed were in the scope of my study or out.

A final example of a difficulty defining "knowledge" occurred in cases where researchers described not knowing whether the variables available in the data adequately measured the concept they intended to study. In one sense, there was not any knowledge about the data that they lacked—i.e., they knew what variables were available and how to interpret them. In another, there was something the researchers did not know about how, or to what extent the data reflected the phenomenon they were researching. This, it seemed to me, was reasonable to interpret as knowledge that was lacking about the data (whether and to what extent the data provided evidence of the phenomenon). Furthermore, there were concrete steps the researchers took to reach an equilibrium concerning this lack of knowledge (e.g., searching literature for the ways others had measured the concept or talking to an advisor or trusted colleague), and I felt this negotiation was important evidence of how researchers bounded the knowledge they obtained about data (or how they knew the data were sufficient for their purposes).

Ultimately, one part of the question of defining "knowledge"—either knowledge "about" or "of"—and lacking knowledge seemed to be about whether the knowledge that was lacking was something the original data creators could have provided but did not (i.e., they collected the data but did not make them available) or whether the researchers were wanting something that had never been collected. Either way, I reasoned, in the larger frame of my study, the researchers might be seen to be satisficing, either in their knowledge about the data or in the data themselves. Either way, also, it seemed to me, researchers were bounding the knowledge they obtained, and studying how and why they did so was relevant to my research. I described the two cases this way:

- when researchers did not know all they wished about the collected data, they drew boundaries around how much they needed to know about the data.
- when researchers did not know all they wished in order to conduct their research (i.e., details about the population or additional variables they might include in the study), they drew boundaries around the scope they would investigate.

I was less interested in the second question than the first because the goal of my project was to explore how we might enhance data that are archived from research projects in order to facilitate reuse, not how to improve the research projects themselves so the data were more advantageous to reuse, though I thought it was interesting to understand what data researchers needed to pursue their interests. On the other hand, the two types of knowledge were difficult to separate. As I wrote in my research notes:

> Take knowledge about whether data are valid or how valid or reliable the data are. Is that something that the researcher could have provided or is it knowledge about something that was not collected? Either way, it is a concern for the reuser, and they had to determine how much knowledge was enough. (York, personal research notes, 9/11/2020)

I concluded in the end that the main questions for my research were "what kind of knowledge did the research want (whether it had been collected or not)? And "how did they negotiate or come to terms with—i.e., reach equilibrium with respect to—how much knowledge was enough?

While focusing on these broader questions involves a certain lack of precision about the way "knowledge" is defined and measured—since different researchers might have understood the question in different ways—it matches the scope of the types of knowledge researchers reported and discussed. In addition, and very importantly, it is sufficient to the task of analyzing the scope of factors that may have influenced how researchers determined the boundaries of the evidence necessary to achieve their reuse goals and publish an article about their research. Thus, while my study involves inaccuracies in the overall kinds and prevalence of the knowledge researchers desired but lacked about data they reused, what I measured can be considered to be a

floor rather than a ceiling (i.e., my study underreports knowledge that is lacking rather than overreporting or reporting inaccurately). Similarly, the influences on how researchers bounded knowledge can be seen at worst to under-analyze how researchers set thresholds on the amounts and types of data they obtain (since at worst I may be missing analysis of knowledge researchers lacked about data).

There is an important effect that the way "knowledge" was defined (or rather, not defined) had on my analysis that went hand-in-hand with something else I discovered as I conducted the interviews. Similar to my experience conducting background research, I found that researchers gave differing levels of detail in their answers to my questions (for instance, about how they became interested in this area of research, what concerns they had about reusing the data, etc.). Some researchers were very detailed and open e.g., about considerations surrounding their research, while others were less so. The effect of these developments was to disrupt my initial plans to compare my findings by clusters of researchers who reused the same data prior to comparing findings across all researchers.

I did analyze the interviews in clusters of researchers who had reused the same data, beginning with the data studies that were the most cited. However, after analyzing four of these clusters (comprising 13 of the 26 researchers), I found it difficult to distinguish whether differences in my findings resulted from reuse of a particular study. In fact, in my second-cycle coding, my natural inclination was to draw similarities considering results across all the interviews I had conducted, regardless of the data study. I was not entirely successful, in other words, as Stake (2006) suggests, of holding off on drawing similarities across interviews until a later stage of analysis—I could not help seeing connections and coding similarities between interviews as I analyzed them, and even as I conducted them.

I mentioned previously that after the first three waves of interviews (which included the four clusters of data studies and first 13 interviews) I became interested in gathering data from researchers who used a wider variety of data, especially data that were less frequently cited. Unfortunately, after the first three waves I was not able to obtain responses from more than one researcher who had reused the same data, so it was not possible to look at variation in researchers' experience with the same lower-cited data. In addition, while my findings from the fourth and fifth waves (comprising lower-cited data) expanded the reuse stories and scenarios I obtained, given the differences in the depth of responses to different questions and the different ways I asked about and researchers interpreted knowledge, I could not be sure if differences I encountered were due to levels of citation separate from any number of other factors. Therefore, in the end, I felt most comfortable analyzing the interviews as a whole, rather than separating them into or comparing them by clusters, to understand how researchers bounded knowledge about data and what influenced their decisions.

Although this was the case, I tried to remain true to my intentions to understand and reflect the situatedness of each individual researcher perspective through the pains I took to understand each interview on its own terms and code researcher-specific responses to questions before analyzing those codes and grouping them into larger categories.

To give a deeper sense of my experience and process, and of the interviews themselves, I present in Appendix E summaries of the interviews of five researchers who reused the same data study. I also draw similarities between the interviews, which I present as summary findings. My intention through the presentation of cases and findings is to illustrate some of the depth, variation, and similarities of the interviews and demonstrate the process by which I began to identify themes and make comparisons across the interviews. The findings represent the

categories or larger groups into which I placed codes from my secondary coding, and I use quotes from the interviews at different points to illustrate ideas that are particularly resonant across the larger collection of interviews.

In Appendix E and throughout the findings, I use the same convention to reference interviews. An example of a reference is D1-01. I use the first part of the reference (e.g., D1) to refer to a specific data study and the second part (e.g., 01) to refer to a researcher who reused that study. For instance, D2-03 would indicate a third researcher who reused the D2 data study. The data studies go from D1 to D4 and there are at most five researchers who reused the same study. In a number of cases, there is only on researcher who reused a study. In these cases I simply use "D" without a number for the first part of the reference, followed by a reference number. For instance, D-05 references a fifth researcher who reused a data study that no other researcher I interviewed reused. These references go from D-01 to D-13.

An issue I have not addressed in this section that I raised in the beginning of it is how I came to think that satisficing as a concept did not provide the most accurate characterization of the behavior researchers discussed in the interviews. I address this question as part of a larger analysis of satisficing in section 4.2.2.

### 3.5 Integration of Survey and Interview Results

There were several ways I integrated analysis and results from the quantitative and qualitative phases. First, I used responses from the survey to finalize my interview protocol and select researchers to interview. For example, I used results about the proportion of researchers who lacked desired knowledge about data and the associated impact on their research to determine how to balance my interests in understanding satisficing and knowledge bounding in the interviews. I additionally used responses from the survey to choose particular interview

respondents so my investigation of satisficing and knowledge bounding would cover as wide a range of researcher characteristics as possible.

I also triangulated results from the interview and survey and used responses from the interview to explain or enhance my analysis of results from the survey. For instance, the survey showed the rate of what I initially believed to be satisficing in my sample and details about the amounts and types of knowledge researchers desired but could not obtain. The interviews provided a frame of comparison that helped me understand that satisficing did not accurately describe researchers' behavior and that instances of lacking desired knowledge were underreported. The interviews also provided a possible explanation for survey findings that researchers tended to reuse data with which they were already familiar (namely, the large amount of time and difficulty involved in reusing data and the pressure for reuse to result in a successful publication). I was additionally able to put findings from the survey and interviews together to learn that researchers made decisions about reusing data at multiple levels (influenced by dimensions of both personal and social reuse equilibrium) and determine more particularly under what circumstances lacking desired knowledge had a negative impact on the outcomes of research.

Finally, results from the survey served to amplify some findings from the interviews. For instance, my findings that early career researchers were particularly affected by social considerations and pressures and used data reuse as a means of advancing their academic careers were put in greater relief by evidence that a majority of respondents in the random survey sample were early career researchers.

## 3.6 Research Inference (Validity)

### *3.6.1 Inference Evaluation Framework*

In their first chapter in the *Handbook of Mixed Methods in Social & Behavioral Research*, Tashakkori and Teddlie (2003) introduce a new terminology for concepts related to research validity and integrity specific to mixed methods. They do so for two reasons. First, they believe "mixed methods researchers should adopt a common nomenclature that transcends the QUAL and QUAN orientations when the described processes are highly similar and when appropriate terminology exists" (Tashakkori and Teddlie, 2003, p. 36). Second, they argue that "a common nomenclature is necessary when the existing QUAL and QUAN terms have been overly used or misused" (Tashakkori and Teddlie, 2003, p. 36). As support for their proposed terminology they point to nearly three dozen terms that are used in qualitative and quantitative research to refer to validity, noting their belief that, due to its multiple uses, the term "has become a catchall term that is increasingly losing its ability to connote anything" (Tashakkori and Teddlie, 2003, p. 36).

Tashakkori and Teddlie propose a framework of validity and credibility for mixed methods research that is based on the concept of inference. They choose inference because the term is flexible enough to take a variety of meanings across both qualitative and quantitative research. They note, for instance, that, "the dictionary definitions for infer include making conclusions, as well as both the term "cause," which is associated with the quantitative orientation, and the term "induce" (the root word for induction), which is associated with the qualitative orientation" (Tashakkori and Teddlie, 2003, p. 35).

Tashakkori and Teddlie introduce four other terms to fill out a range of concepts historically related to research validity. These include first, inference quality, which encompasses concepts of internal validity in quantitative research and credibility in qualitative research (i.e., "the degree to which a researcher believes that his or her conclusions accurately describe what

really happened in the study" (Tashakkori and Teddlie, p. 36)). Inference quality in their view comprises two additional terms: design quality and interpretive rigor. Design quality denotes, "the standards for the evaluation of the methodological rigor of the mixed methods research" and interpretive rigor denotes, "the standards for the evaluation of the accuracy or authenticity of the conclusions" (Tashakkori and Teddlie, p. 37). The final term is inference transferability, which refers to the generalizability of results, discussed in quantitative research as external validity and in qualitative research as transferability.

Within these four terms, Tashakkori and Teddlie propose four dimensions of evaluation. These are the following:

- within-design consistency: "consistency of the procedures/design of the study;
- conceptual consistency: "the degree to which the inferences are consistent with each other and with the known state of knowledge and theory";
- interpretive agreement: "consistency of interpretations across people (e.g., consistency among scholars, consistency with participants' construction of reality)";
- interpretive distinctiveness: "the degree to which the inferences are distinctively different from other possible interpretations of the results and the rival explanations are ruled out (eliminated)" (Tashakkori and Teddlie, 2003, p. 40).

Tashakkori and Teddlie (2003) accompany each of these dimensions with a set of questions, which they note are not exhaustive, that might be posed in order to evaluate the mixed methods research. I represent the questions in relation to the Tashakkori and Teddlie's vocabulary of inference in Table 3.13. I include in this table the strategies I used to achieve inference quality, which I discuss in section 3.6.2 below.

*Table 3.13 Representation of Tashakkori and Teddlie's (2003) Framework for Inference Quality*

| Term | Sub-term | Evaluation Dimension | Evaluation questions | Inference quality strategies | |
|---|---|---|---|---|---|
| | | | | Survey | Interviews |
| Inference quality | Design Quality | Within-design consistency | Is the design consistent with the research questions/purpose? | Peer review and debriefing<br>Survey testing<br>Survey pilot | Peer review and debriefing<br>Interview testing<br>Interview pilot |
| | | | Do the observations/measures have demonstrated quality? | | |
| | | | Are data analysis techniques sufficient/appropriate for providing answers to research questions? | Research consultation<br>Logistic regression | Research consultation |
| | | | Do the results happen the way the investigator claims they did? Do they have the necessary magnitude/strength or frequency to warrant the conclusions (demonstrated results in QUAL research, magnitude of effect in QUAN research)? | Inter-rater reliability testing (for coding of open-ended responses) | Inter-rater reliability testing (for coding of qualitative data) |
| | | | Are the inferences (e.g., emergent theory or explanations) consistent with the results of data analysis? Do they strongly "follow" the findings? | Triangulation (with interview results) | Triangulation (with survey results)<br>Member checking |
| | | | Are the inferences consistent with the research questions/purposes? Are the inferences obtained in each of the two strands (QUAL and QUAN) consistent with the corresponding research questions/purposes? | Peer review and debriefing<br>Research consultation | Peer review and debriefing<br>Research consultation |
| | | Conceptual consistency | Are answers to different aspects of the research question/purpose consistent with each other? | Peer review and debriefing<br>Survey testing<br>Survey pilot | Peer review and debriefing<br>Interview testing<br>Interview pilot |
| | | | Is the final (global) inference consistent with the ones obtained on the basis of QUAL and QUAN strands of the study? | Logistic regression | Triangulation<br>Comparison of phase findings |
| | | | Do the inferences take the current literature into consideration? | Literature review<br>Research consultation | Research consultation |
| | | | Are the inferences consistent with the state of knowledge? If not, do the inferences offer explanations (theory) for the inconsistency? | Triangulation (with interview results) | Triangulation (with survey results) |
| | Interpretive Rigor | Interpretive agreement | Do other scholars agree that the inferences are the most defensible interpretation of the results? | Peer review and debriefing<br>Literature review | Peer review and debriefing<br>Literature review |
| | | | If participants' construction of the events/relationships is important to the | Peer review and debriefing | Peer review and debriefing |

| | | | | | |
|---|---|---|---|---|---|
| | | | researcher, do the interpretations make sense to participants of the study? | | |
| | | Interpretive distinctiveness | Are the inferences distinctively superior to other interpretations of the same findings? | Peer review and debriefing<br>Literature review | Peer review and debriefing<br>Literature review |
| | | | Are there other plausible explanations for the findings? | | |

*Note*. The table includes two aspects of inference quality (design quality and interpretive rigor) and four dimensions associated with these aspects (within-design consistency, conceptual consistency, interpretive agreement, and interpretive distinctiveness).

The dimensions above do not include inference transferability. Tashakkori and Teddlie (2003) take the position that inference transferability is relative. They believe any inference has *some* degree of transferability (e.g., to other settings, populations or times). The key determination in their view is about the specific range of transferability (Tashakkori and Teddlie, 2003, p. 42). I discuss the strategies I use to achieve inference quality and determine the range of transferability below.

### *3.6.2 Inference Quality Strategy*

I took several steps to maximize the inference quality of my study. As represented in Table 3.13 and discussed above, inference quality refers to both design quality and interpretive rigor. Design quality, in turn, encompasses within-design consistency and conceptual consistency and embodies "the standards for the evaluation of the methodological rigor of the mixed methods research" (Tashakkori and Teddlie, p. 37). Interpretive rigor represents "the standards for the evaluation of the accuracy or authenticity of the conclusions" (Tashakkori and Teddlie, p. 37).

The first strategy for inference quality I used was peer review and debriefing (Creswell, 2007). Throughout the prelim and proposal phases of the dissertation, as I reviewed literature, discovered gaps and developed a research design, my advisor played a primary role in "asking hard questions about methods, meanings, and interpretations" (Creswell, 2007, p. 208). Members of my prelim and proposal committees, and fellow students played a key role in ensuring the inference quality of my study as well.

The second strategy I used was testing of my research protocols. Before finalizing the survey, I solicited feedback on the survey instrument from eleven researchers: five in the social sciences and four in other fields). I offered $15 gift cards in exchange and received extensive

feedback from most researchers, the majority of which I incorporated into the wording and structure of the survey questions. I next conducted a pilot survey including 114 researchers to test the performance of my questions and whether the questions elicited the kind of data I expected. My analysis of the pilot survey data resulted in the addition of two questions to the survey and the modification of several others that enhanced my ability to interpret the survey results. From the pilot I also gained valuable experience drawing samples and searching for researcher emails, which I used to improve the quality of inferences and inference transferability in the full study.

I performed similar testing of the protocol I used in my interview research. As reported in section 3.4.4, I did not make many adjustments to the protocol after these tests, but they allowed me to test the flow of questions in different scenarios and gain confidence that my questions could yield the kinds of responses I expected. Testing of the research protocols relates to within-design and conceptual consistency, which I sought to improve through my reliance on the survey results in developing my interview protocol (see section 3.4.3) and though adjustments I made to my research protocol even after I started conducting my research (see section 3.4.4).

With regard to sampling and data analysis both, I took advantage of research consultations offered by Consulting for Statistics, Computing &Analytics Research (CSCAR) at the University of Michigan.

In the analysis phase of my research, I used interrater reliability testing (Cho, 2008), several modes of triangulation, and member checking to maximize inference quality. Interrater reliability testing is a strategy for evaluating "the extent to which two or more independent coders agree on the coding of the content of interest with an application of the same coding scheme" (Cho, 2008). I conducted interrater reliability testing with the help of a research

assistant on selected interview transcripts, using Scott's pi and Holsti's coefficient in the testing (see section 3.4.8.3). Interrater reliability testing applies primarily to within-design consistency.

Triangulation (comparing data collected through multiple means) is a strategy for achieving inference quality that mixed methods are particularly well suited to provide (Morse, 1991; Tashakkori and Teddlie, 2003, Creswell and Creswell, 2018). I triangulated results of the quantitative and qualitative phases of my study in compiling and presenting my findings (see section 3.5). Triangulation applies primarily to within-design and conceptual consistency.

Finally, I developed each phase and component of my research in close consultation with the literature on data reuse. I used the literature to identify the gaps in research that I investigated in my study, determine my specific research questions, and identify concepts to include in my background research and interviews. I intentionally selected my study population (researchers who have reused data from ICPSR) because of the large body of literature that exists on data reuse in the social sciences and the potential to draw on this literature to compare and explain my results. My use of literature applies to within-design consistency, interpretive agreement, and interpretive distinctiveness.

### *3.6.3 Inference Transferability Strategy*

Inference transferability refers to the generalizability of research results or the ability to extrapolate from conclusions "beyond the particular conditions of the research study" (Tashakkori and Teddlie, 2003, pp. 37-38). Tashakkori and Teddlie note that quantitative researchers have referred to the concept as external validity but they propose the term "transferability" to avoid confusion surrounding the term "validity." In order to determine the range of transferability of my results I relied on statistical analysis and research consultations. The statistical validity of the results of the survey in relation to the survey population are

dependent on the samples I drew, the number of respondents I received, and the appropriate application of statistical methods. I was guided in these primarily by peer review, debriefing, and research consultations, in combination with my own understanding and application of statistical analysis. I obtained enough responses in the survey samples selected by type of data (quantitative or qualitative) to generalize many of my results with a high level of confidence (99%) and low error (5%) to the population of reusers of quantitative and qualitative data from ICSPR. I did not obtain sufficient responses to make inferences related to some variables or researcher subgroups (e.g., the subgroup of researchers who lacked knowledge they desired about data).

Wider transferability (outside the population of users of ICPSR data) depends largely on the literature and on the determinations of those reading the results of the research. To the degree to which my survey results (supported by results from my interviews) align with findings in other data reuse settings there will be some transferability. In terms of the interviews themselves, the degree of transferability will be determined largely by the degree of resonance the results have with the experiences of those who read my research. I have noted above the similarity of the reuse instances in my study to cases in a case study. As discussed in section 3.6, a significant source of transferability of my results will depend on the accuracy and reliability with which I have been able to present and analyze individual instances of reuse and make comparisons across reuse instances. Cogent analyses could enable repository curators and administrators, for example, to see meaningful connections between my findings and their own experiences that lead to insights about or have implications for operations in their own contexts.

In whichever way it may occur, my hope is that my study design—using statistical methods on one hand and a large number of individual reuse instances on the other—will

facilitate the transferal of the results into policies and practices that support scholarly reuse of research data.

### 3.7 Relation of Data Collected to Research Questions and Theoretical Framework

The strategies I developed for data gathering, analysis, and integration were designed to answer specific research questions having to do with knowledge bounding and knowledge satisficing in data reuse. My methods were also designed to produce findings that could contribute to a larger discussion about data, and information about data, that we collect and make available from data archives. This discussion involves consideration of overarching policies for data deposit as well as the goals and assumptions that drive them. It also involves consideration of the ways we go about making decisions, including how we design and build archives, and the data artifacts we collect and make available from them.

In my dissertation research, I drew on theories of Donna Haraway and Herbert Simon and placed them in conversation with one another in order to wrestle with critical issues in the design of archives in a context where technical solutions are inadequate to meet social needs. I developed a framework that offered ways of thinking about data boundaries, partial perspectives, and adaptive systems that I believed would allow me to gather data that could help to address important gaps in data reuse research.

In the tables below, for each of the gaps I identified in the literature, I present the concepts I investigated and the specific research questions and strategies I used (i.e, survey, interviews, or background research) to collect relevant data. For instance, Table 3.14 shows the concepts I investigated through the survey and semi-structured interviews related to how researchers bound knowledge about data. The final column of the table indicates to which research question(s) the concepts relate. In addition to gap areas identified in section 2.8, I have

added a table showing the concepts, sources of data, and research questions I draw on in my discussion of how the ongoing conversations Haraway imagines might take place in the context of data "artifacts" and repository design.

The purpose of the tables is to illustrate, in concert information in section 2.8, the correspondence of information collected from each data source with a) my areas of strategic data collection, b) my specific research questions, and c) my theoretical framework. These correspondences will enable me to place my minute findings in a larger context and suggest ways that findings in the specific area of data reuse could contribute to or enhance our understanding of broader theories.

*Table 3.14 Relation of Research Concepts, Questions, and Sources of Data to Gap 1*

<table>
<tr><th colspan="3">Gap 1: How researchers bound the knowledge they obtain in order to decide to reuse data (including whether or not researchers satisfice in obtaining desired knowledge)</th></tr>
<tr><td>Concept</td><td>Source of data</td><td>Correspondence to Research Questions</td></tr>
<tr><td>Bounding knowledge about the data<ul><li>How the researcher determined they had enough information about these areas to decide to reuse the data</li><li>What knowledge was desired when making the decision but not available or limited</li></ul></td><td>Interview</td><td rowspan="2">1. How do researchers determine the boundaries of the knowledge they obtain about data in order to reuse them in their research?<br><br>1b. How do researchers determine how much of what kinds of knowledge is enough?<br><br>2. Do researchers satisfice?<br><br>2a. If so, how can knowledge satisficing be characterized?</td></tr>
<tr><td>Satisficing</td><td>Survey</td></tr>
</table>

*Table 3.15 Relation of Research Concepts, Questions, and Sources of Data to Gap 2*

| **Gap 2: Factors that affect knowledge bounding and knowledge satisficing** | | |
|---|---|---|
| *Concept* | *Source of data* | *Correspondence to Research Questions* |
| Background about the research | Interview / Background research | 1c. What do researchers report influences these determinations [about how much of what kinds of knowledge are enough]?<br><br>2b. What factors influence knowledge satisficing?<br><br>2c. What reasons do researchers give for why they satisfice? |
| Deciding to reuse the data | Interview | |
| Information about researcher's perspective | Survey | |
| Factors that may have influenced knowledge bounding | Interview / Background research | |
| Researcher "distance" from the data | Survey | |
| Experience with data reuse | Survey | |
| Reuse ability | Survey | |
| Reuse motivation | Survey | |
| Research process | Survey | |
| Reused data quantitative or qualitive | ICPSR | |

*Table 3.16 Relation of Research Concepts, Questions, and Sources of Data to Gap 3*

| **Gap 3: The relative priority that researchers assign to different types of knowledge about data in particular reuse instances, and why** | | |
|---|---|---|
| *Concept* | *Source of data* | *Correspondence to Research Questions* |
| Bounding knowledge about the data<br>What was most important to know about the data?<br>Why this knowledge was important? | Interview | 1a. What knowledge about data is most important to researchers to reach reuse equilibrium and why? |

*Table 3.17 Relation of Research Concepts, Questions, and Sources of Data to Gap 4*

| **Gap 4: The impact of knowledge satisficing on the outcomes of research and researchers' attainment of their researcher goals** | | |
|---|---|---|
| *Concept* | *Source of data* | *Correspondence to Research Questions* |
| Impact of knowledge satisficing | Interview | 2d. What is the perceived impact of knowledge satisficing? |
| Outcomes of research | Survey | |
| Goal attainment | Survey | |

*Table 3.18 Relation of Research Concepts, Questions, and Sources of Data to "Conversations"*

| **Discussion of how Haraway's "conversations" can take place in the context of data repository design** | | |
|---|---|---|
| *Concept* | *Source of data* | *Correspondence to Research Questions* |
| Bounding knowledge about data<br>How the researcher learned what was most important about the data<br>What strategies were taken to mitigate effects of lack of knowledge | Interview | 1d. How do researchers obtain the knowledge they desire? |
| Factors that might mitigate knowledge satisficing and its effects | Interview | |

*Note*. How research concepts, research questions, and sources of data related to how the "conversations" Haraway imagines can take place in the context of data repository design.

## 3.8 Conclusion

My overarching goal through my research design was to explore the analytical power that could be gained by applying Haraway's and Simon's theories to the investigation of how and why researchers bound the knowledge they obtain to make decisions about reusing data. The care I took in each step of the process, from drawing samples to developing research instruments, conducting analysis, integrating results and drawing conclusions directly impacted the quality of my inferences and ability to accurately assess what more can be learned through the application of these theories. I built mechanisms into the design to ensure that care was taken (e.g., peer review, debriefing, and interrater reliability). The mixed methods approach and the ability it offered to compare and analyze data from multiple sources also served to maximize the integrity and applicability of my research.

# Chapter 4 Findings

## 4.1 Introduction to Findings

This chapter is separated into seven sections. In this first section 4.1, I introduce the findings by providing an overview of the samples of researchers I surveyed and interviewed and a description of the control variables I used in my statistical analysis. In section 4.2, I present findings related to one of my main research questions: do researchers satisfice? Then, in sections 4.3 through 4.6, I present findings according to the four gap areas I identified in the literature (slightly re-ordered):

- 4.3 Factors that affect knowledge bounding and knowledge satisficing
- 4.4 How researchers bound the knowledge they obtain in order to decide to reuse data
- 4.5 The relative priority that researchers assign to different types of knowledge about data in particular reuse instances, and why
- 4.6 The impact of knowledge satisficing on the outcomes of research and researchers' attainment of their researcher goals

Since findings from my hypothesis tests are distributed throughout these sections, a table summarizing the results of these tests together is included in Appendix F.

In section 4.7, I describe findings related to the idea of "conversation" introduced in my theoretical framework: Donna Haraway's notion that it is through the interactions and conversation between situated perspectives that rational knowledge is produced. As I quoted in the introduction:

> Rational knowledge is a process of ongoing critical interpretation among 'fields' of interpreters and decoders. Rational knowledge is power-sensitive conversation (King, 1987a) (Haraway, 1991, p. 196)

I begin with a description of the survey and interview samples.

### *4.1.1 Characterization of Survey Sample*

I selected participants for the survey using three stratified sampling strategies: a full sample of researchers who reused qualitative data, a random sample of researchers who reused qualitative data, and an additional sample stratified by the number of citations to data by researchers in the first two samples. In the inferential analyses below, I include the first two samples (those who used qualitative data and those who used quantitative—a total of 767 researchers). There was not a large enough set of responses from users of qualitative data (only 28) to treat them as a sub-group in separate analyses, e.g., to compare years of experience among reusers of qualitative data to years of experience among reusers of quantitative data. However, I included both groups in my inferential analyses in order to represent repository users generally (who include both users of qualitative and quantitative data). A more complete characterization of the samples is as follows:

- seven hundred thirty-nine from a sample of 2,405 researchers randomly selected from a population of 6,951 who reported reuse of data—listed in the ICPSR Bibliography and identified by ICPSR as quantitative—between January 2014 and September 2019
- twenty-eight from a full sample of 107 researchers who reported reuse of data—listed in the ICPSR Bibliography and identified by ICPSR as having a qualitative component in the same time frame

I obtained the combined 767 responses from these samples out of 2,512 researchers contacted (30.5%). This is well over the sample size of approximately 400 required according to Yamane's formula (Israel, 1992) to provide results with a 95% confidence level and 5% error for a dichotomous variable given a population of 6,951 individuals. Israel notes that some researchers increase the sample size by 10% to compensate for people the researchers are not able to contact, and some increase the sample size by 30% to compensate for non-response. My sample size exceeded both of these, as well as the sample size required to achieve a 99% confidence level with 5% error (which, according to Cochran, would be approximately 633) (Cochran, 1963, cited in Israel, 1992).

I did not include the third sample I drew (stratified by number of citations to reused data) in the inferential analyses since that sample was drawn for a different purpose (to facilitate the selection of interviewees in my qualitative analysis) and because I found differences in the significance of some results when I included this group with the other two. I did include researchers from this third sample, however, in the analysis of more qualitative aspects of the survey analysis, including the types of knowledge that researchers reported were limited or lacking. I also included researchers from the third sample in descriptive statistics about the subgroup of researchers who reported lacking knowledge they desired about data, such as those surrounding the impact that lacking knowledge had on their research.

Whenever the third sample is included, I indicate this, noting also that the number of respondents out of all three samples who lacked desired knowledge was 223 researchers. In addition, for ease of reference, from here forward I will refer to knowledge that is limited or lacking simply as "knowledge lacking" or "lack of desired knowledge" using the words "lacking" or "lack" to encompass situations where the desired knowledge was both completely and partially lacking.

A table detailing the characteristics of the survey respondents according to the questions I asked and data from the ICPSR archive is given in Appendix G. Characteristics of the interview respondents are given as well for comparison. Unless otherwise, noted, the numbers given are out of the 767 researchers in the combined samples of researchers who reused quantitative and qualitative data.

Overwhelmingly, the format in which most researchers reported reuse of the data I asked them about was a journal article (612). Theses were a distant second (60), followed by

conference proceedings (30) and book sections (16). Similarly, most of the data studies reused were survey data (630).

Most researchers were very involved in (310) or solely responsible for (372) the decision to reuse the data and in most cases, there was no one else (377) or only one other person (156) involved in that decision. Most researchers were very involved in (359) or solely responsible for (334) determining the goals of the research. Close to even amounts of researchers reused data from ICPSR (390) and from a different source (309). A large majority of respondents played a leading role in their research (i.e., PI or co-PI) (571), and the largest proportions of researchers were PhD students (260), assistant professors (121), researchers (111), assistant or associate researchers (82), and associate professors (79).

As tabulated from my qualitative coding of research domains, most researchers worked in areas of health (222) or justice (150). The majority of data studies reused had less than 25 citations by researchers in my sample (though the number of citations in my samples and in the population were proportional, fewer citations were represented in my samples). Most researchers (591) believed that the data they reused were created to be reused. While most (546) reported they were not at all involved in the original research, a wide majority of researchers reported having good (157), significant (243) or in-depth knowledge (238) either themselves or on their research team about the original collection of the reused data.

Most researchers were conducting research in their primary domain of research (670) and the majority of researchers (600) had less than 10 years of experience doing research in their primary domain. Most researchers had less than 10 years of experience reusing data in their primary domain of research (449); more than a fifth of these (101) had no experience reusing

data. 381 researchers had less than 10 years of experience reusing data in any domain; 77 had none.

The two most common purposes for reusing data were answering a new question (608) and testing a theory (519) (researchers could choose as many purposes as were applicable). The third most common purpose was to combine the data with other data (237). In general across reuse purposes, researchers most frequently reported that the data met their reuse goals as expected. Ninety-four percent or more of researchers in every category reported the data met their reuse goals as, better than, or much better than expected. Most researchers reported that their research questions either changed after they began to reuse the data or developed as they reused them (589). One hundred seventy-eight researchers reported their research questions were fixed prior to conducting their research.

One hundred ninety-five out of the 767 researchers in the first two samples reported lacking knowledge they desired about data. Out of the full group of researchers surveyed (883 including the third sample drawn by citation count), 223 reported lacking desired knowledge. The most common type of knowledge researchers reported lacking was a type I categorized as data supplement, describing knowledge that was either more detailed or broader in scope than knowledge available about the data or the population being studied (see section 3.4.8.5 for fuller discussion).

Across all types of knowledge for which researchers reported lacking desired knowledge (a total of 297 knowledge areas indicated by the 223 researchers) 52% of the time, researchers obtained half or less of the knowledge they desired. In areas where researchers did not obtain all the knowledge they desired about data, the most frequent sources from which they obtained at least some knowledge were data documentation (64), the original data creators (39) and the data

themselves (38). In this question, researchers could choose multiple sources, and 62 did not respond.

Most researchers in the group of 223 reported that a lack of knowledge negatively impacted the outcomes of their research slightly (48), moderately (54), or very much (27). Within the same group, 90 reported a lack of desired knowledge negatively impacting their research by limiting reuse in some way (e.g., the scope of analysis, the way analysis could proceed, or the depth of analysis). Other researchers compensated for the lack of knowledge in some way (46) or described the negative impact in terms of an opportunity that was lost to do more with the data (28).

### *4.1.2 Characterization of Interview Sample*

Detailed characteristics of the 26 researchers I interviewed are given along with characteristics of survey respondents in the table in Appendix G. I provide a summary of characteristics below. My primary aim in selection for the interviews was to achieve maximum variation among the interviewees in characteristics that could have a bearing on the way researchers bounded the knowledge they obtained about data. I was by and large successful in this, as can be seen from the table in Appendix G. The numbers of researchers I selected under each criteria is in general both diverse, and reflective of the distribution of responses received in the survey. The table does not show characteristics of the pool of researchers from which I selected interview candidates. In a number of cases where there appears to be a large bias there was simply a lack of diversity to begin with. For example, I intended to interview researchers who believed the data they reused were created to be reused and researchers who believed they were not. However, only six researchers out of the 284 who agreed to be contacted were sure the data had not been created to be reused. I contacted four of these but none responded to be

interviewed. Therefore, 22 researchers I interviewed were sure the data had been created to be reused and none were not (three were not sure).

Most of the researchers' publications were journal articles (17), followed by theses (four), though many of the journal articles came out of theses. Most of the data were survey data (20) and most researchers (16) indicated there were no others involved in the decision to reuse the data. However, 16 researchers said they were very involved in the decision to reuse the data and only eight said they were solely responsible, indicating that while some researchers may have been solely responsible for the decision to reuse the data, at least someone of these felt influenced by others (otherwise at least 16 would have replied they were solely responsible).

Most interviewees (24) indicated that they were either very involved in (11) or solely responsible for (13) determining the goals of the research. Most (15) obtained their data from ICPSR and most of the researchers (22) were in a leading role in their research (e.g., PI). The most frequent domains I coded researchers working in were justice and health. Twenty-five of the researchers reused quantitative data and just one reused qualitative data. Ten of the data studies reused had more than 100 citations in my sample, four had between 25 and 50, six had between five and 24, and six had less than five.

Twenty-two of the researchers believed the data they reused were created to be reused. Most (20) were not at all involved in the original research while the majority (17) had either in-depth (10) or significant (seven) knowledge themselves or on their research team about how the original data were created. The largest proportion of researchers were PhD students (11) when they conducted their research; five were assistant professors, four were researchers, three were full professors, two were associate professors, and one was a postdoc.

For nearly all researchers (24), the research they conducted with the data was in their primary domain of research. Most researchers had between one and five (10) or five and 10 (seven) years of experience in their primary domain. At the same time, 11 researchers reported having one to five years of experience reusing data produced in the same domain as these data and there was a more even spread of researchers with other durations of experience (three researchers each with none, zero to one, five to 10, and 15 to 25 years of experience, two with 10 to 15 and one with 25 to 35). Numbers of researchers with experience reusing data in *any* domain were: zero years (4), zero to one years (two), one to five years (eight), five to 10 years (five), 10 to 15 years (one), 15 to 25 years (three), and 25 to 35 years (three). These differences suggest that some researchers were new to reusing data and some had been reusing data longer in other domains than this domain.

Most researchers reused the data to answer a new question (21) or test a theory (14) (this was a multiple response question). Only two researchers reported meeting their reuse goals somewhat worse than expected (one each for the purposes to answer a new question and to compare data with other data). All others met their goals as or better than expected.

Most researchers (21) had a fluid approach to research question development (i.e., their research questions changed in some way after they started reusing the data). Eleven reported they lacked knowledge they desired about the data (15 did not). For those who lacked knowledge they desired, only one researcher reported the knowledge they lacked was not at all important (though they also reported the lack of knowledge had a negative impact on their research). Aside from one other where the response was missing, all other researchers reported the knowledge that was lacking as important (4), very important (1), or essential (4).

The most common type of knowledge lacking was data supplement (16) followed by data collection (seven), data (three), data analysis (one), and data reuse (one). Researchers who obtained any of the knowledge they desired that they did not have obtained it from advisor(s) (one), data documentation (two), knowledge of others on the team (two), literature (one), original data creators (three), the data themselves (six), and other sources (one). Most of these researchers were impacted slightly (two), moderately (five), or very much (two) by the lack of knowledge and reasons for impact included that it limited the outcomes of their research (11), they compensated for the lack of knowledge in some way (eight), they adjusted their research objectives or analysis techniques (four), they obtained the desired knowledge (two), an opportunity was lost to do something more (two) or they determined that the knowledge they had was satisfactory (one).

### *4.1.3 Control Variables*

In section 2.9.4 I introduced the three control variables I used when testing my hypotheses on the sample of 767 researchers. These were number of citations to data, whether researchers believed the data were created to be reused, and whether the researcher obtained the data from ICPSR or from a different source. Here, I provide descriptive and other information about each of these variables for additional context in interpreting the hypothesis tests.

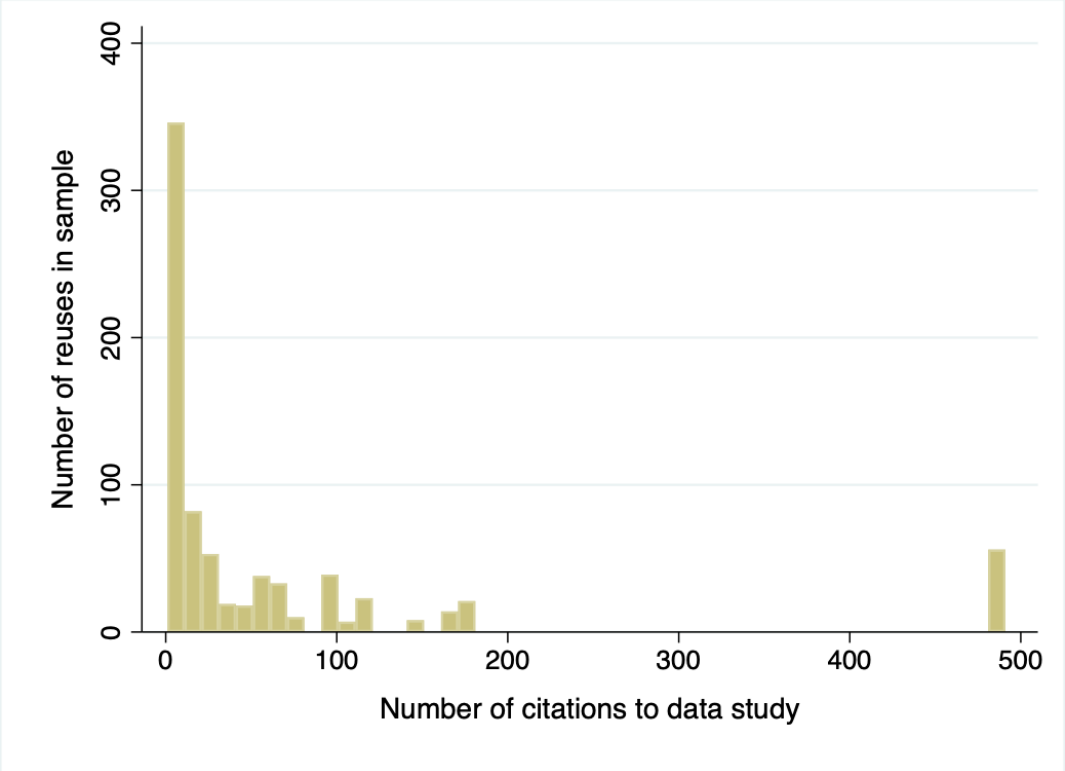
In addition to the control variables, I used general estimating equations to account for clustering effects arising from multiple researchers in my sample reusing the same data. More specifically, I used the Stata "xtgee" function (https://www.stata.com/manuals/xtxtgee.pdf ) to fit population-averaged panel data models. Unless otherwise noted, I include the control variables and account for clustering effects in all my regression analyses below.

#### *4.1.3.1 Number of Citations*

I included number of citations as a control variable based on the reasoning that if the data were well cited, there would be more information about them, or an accepted threshold of the amount of knowledge about the data that was needed to reuse them. More available information and norms about the thresholds of knowledge sufficient to reuse would make it less likely for researchers to lack knowledge about these data than data that were less cited and less generally known. To better understand how and in what form to include the number of citations as a control variable, I calculated some descriptive and inferential statistics.

Figure 4.1 shows the frequency of reuse of data studies with different numbers of citations in my sample. In all, 383 data studies were reused at least once. The bars in the histogram have a width of 10 citations and indicate that the vast majority of reuse instances are of data studies with fewer than 50 citations (accounting for 33 studies and 505 instances of reuse). There are spikes in the frequency of reuse at different points for data studies with between 90 and 500 citations. The spike from 90-100 involves reuse of just two data studies (i.e., citations to two studies together account for about 53 instances of reuse). The spike from 110-120 also involves two data studies. All of the spikes after that involve reuse of just one data study (i.e., one data study cited between 480-490 times in my sample was reused by 56 researchers). The figure reveals—to the extent that my sample represents the larger population of researchers who reuse data from ICPSR—that while some individual data studies receive a large amount of reuse, the majority of reuses were conducted with data that were less frequently cited.

*Figure 4.1 Histogram of the Number of Citations of Reused Data Studies*




I conducted a test to determine whether a log transformed variable for the number of citations was more appropriate to use in hypothesis testing as a control than an untransformed variable. As I would be performing logistic regressions and the number of citations was continuous, it made sense to log-transform the number of citations so the relationship being examined was based on the log of both sides of the equation for the logistic regression. For an empirical test, I ran a logistic regression without log-transforming the number of citations (Table 4.1) and a logistic regression log-transforming the number of citations (Table 4.2).

The results of the regressions show that the data are better modeled using the log-transformed variable for number of citations. The chi-square for the model using the log-transformed variable is 0.036 (see Table 4.2), compared with 0.305 for untransformed variable (see Table 23), and the model shows a significant effect of a change in the log-transformed number of citations on lacking desired knowledge (0.906). In particular, the log-transformed

model indicates that a one percent increase in the number of citations is associated with close to a 90.6% increase (i.e., a 9.4% decrease) in the odds of lacking knowledge. In subsequent testing, therefore, I used the log-transformed number of citations.

*Table 4.1 Estimated Association between a Change in the Number of Citations and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Number of Citations | 0.999 | 0.324 | 0.997 | 1.001 | |
| Constant | 0.362 | 0.001 | 0.300 | 0.436 | *** |
| Number of obs | 767 | | | | |

*** *p<.01, ** p<.05*

*Table 4.2 Estimated Association between a Change in the Natural Log of Number of Citations and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Number of Citations | 0.905 | 0.041 | 0.822 | 0.996 | ** |
| Constant | 0.447 | 0.001 | 0.332 | 0.602 | |
| Number of obs | 767 | | | | |

*** *p<.01, ** p<.05*

#### *4.1.3.2 Data Created to be Reused*

As I noted in section 3.3.1.3, I included the researcher's assessment of whether the data they reused were created with an intention to be reused (represented in the table as "Researcher's perception of original data creation purpose") based on the Zimmerman's (2008) notion that standardization in the curation of data could act to overcome "distance" between a researcher and reused data. I reasoned that if data were created with the intention to be reused, the data might be more likely to have undergone a higher degree of curation than data that were created without an intention to be reused. In the survey results, 591 (77.1%) researchers were sure the data were intended to be reused, 81 (10.6%) were sure the data were not intended to be reused, and 87 (11.3%) were not sure (eight researchers (1.0%) did not respond).

Table 4.3 shows a cross-tabulation of knowledge lacking and whether the data were created to be reused. A chi-square test did not reveal a significant association between the two variables.

*Table 4.3 Cross-Tabulation of Lacking Desired Knowledge and Whether Researchers' Believed the Original Data Were Created to be Reused*

| Lack of desired knowledge | Original data created to be reused | | | |
|---|---|---|---|---|
| | Yes | No | Not sure | Total |
| Yes | 143 | 28 | 22 | 193 |
| No | 448 | 53 | 65 | 566 |
| Total | 591 | 81 | 87 | 759 |

Pearson Chi2 = 4.04 Prob = 0.1325

#### *4.1.3.3 Data Source*

I included the source of the data (whether the researcher obtained the data to reuse from ICPSR or a different source) as a control to account for differences in researcher knowledge about data that might be attributed to the way data were curated or made available from different sources. Table 4.4 shows a cross tabulation of knowledge lacking and data source, Overall, 390 of those who responded (55.8%) said they reused data from ICPSR and 309 (44.2%) said they reused data from a different source. There were 68 missing responses. A chi-square test did not show a significant association between the variables.

*Table 4.4 Cross-Tabulation of Lacking Desired Knowledge and Source of Reused Data*

| Lack of desired knowledge | Source of reused data | | |
|---|---|---|---|
| | ICPSR | Different source | Total |
| Yes | 106 | 66 | 172 |
| No | 284 | 243 | 527 |
| Total | 390 | 309 | 699 |

Pearson Chi2 = 3.15 Prob = 0.0760

#### *4.1.3.4 Definition of "Lacking knowledge"*

I asked researchers a number of questions in the survey to better understand their responses about knowledge they desired but lacked about data and to better interpret whether I

would define them as lacking knowledge they desired about data for the purposes of my study. These included questions about:

- whether knowledge was lacking at the time the researcher was considering reusing the data
- whether knowledge was lacking at the time the researcher decided to reuse the data
- the importance of the purpose for which the researcher was reusing the data (e.g., a background purpose, to answer a new research question, to validate data, etc.)
- the amount (percent) of desired knowledge the researcher was able to obtain
- the importance of the knowledge the researcher reported lacking
- whether the lack of knowledge had a negative impact on the outcomes of the research

I asked whether researchers lacked knowledge at the time they were considering using data and at the time they decided to use the data to differentiate between researchers who 1) knew everything they desired to know before they even started reusing the data (likely indicating a prior familiarity with the data), and 2) researchers who lacked desired knowledge when considering reusing data. In the second group, I sought to distinguish researchers who either a) found out everything they desired before deciding to reuse the data, or b) decided to reuse the data even though there were things they still desired to know. I was interested in any researchers who reported lacking knowledge about data, but I wanted to be able to compare the ratios of those who lacked knowledge to those who did not in both of these groups (those who lacked knowledge when considering and those who lacked knowledge when deciding).

Ultimately, 306 (39.9%) respondents said there was additional knowledge about the data they desired when considering using the data; 429 (55.9%) said there was not—i.e., 55.9% likely had some prior knowledge of the data or were recommended to reuse the data by a trusted source. 31 were not sure. There was one missing response but based on other responses in the survey I concluded this respondent did not lack desired knowledge.

Of the 306 who said they lacked knowledge when considering whether to reuse the data, 140 (45.6%) also said they lacked knowledge they desired about data when deciding to reuse the

data (9 of these were not sure). Among the 429 who did not lack knowledge when considering whether to reuse the data, 47 (11%) said that they lacked knowledge when deciding to reuse the data (six were not sure – I perform further analysis of these groups in section 4.2 below). Of the 31 who were not sure whether they lacked knowledge when considering whether to reuse the data, four (12.9%) said they lacked knowledge when deciding to reuse the data (nine were not sure).

Between the different groups, then, there were 191 instances where researchers reported lacking desired knowledge (140 in the group of 306, 47 in the group of 429, and four in the group of 31 who were not sure whether there was additional knowledge they desired when considering reusing the data). Of all of those in either group who were not sure whether they lacked desired knowledge when deciding to reuse the data, nine indicated in further survey questions that they did in fact lack knowledge in at least one of the three knowledge areas I asked about. This brought the total to 200.

In relation to the remaining criteria, if the researcher reported that the importance of the reuse purpose was not important, that the knowledge lacking was not important, and the lack of knowledge had no negative impact on their research, I did not count the researcher as lacking knowledge (if any negative impact resulted from the lack of knowledge, I counted the researcher as lacking knowledge of the other two). Moreover, if the researcher reported obtaining 100% of the knowledge they indicated was lacking at the time the decision to reuse the data was made, I did not count them as lacking knowledge unless there was other evidence of lacking knowledge (e.g., from the free text responses about the reasons knowledge was lacking).

With these adjustments, I identified 195 researchers (26%) out of 767 in the combined quantitative and qualitative samples who lacked knowledge they desired about data when they

decided to reuse the data. I created a single variable that included all of these adjustments for easier use in my analyses. Unless otherwise specified, this variable for "Lack of desired knowledge" is the one I use in the analyses I present.

## 4.2 Do Researchers Satisfice?

### *4.2.1 Research Question 2. Do Researchers Satisfice? Findings From the Survey (H1)*

My findings in this section relate to my second research question: Do researchers satisfice in the knowledge they obtain about the data they reuse? I investigated this through my first hypothesis, which was that researchers do satisfice in the knowledge they obtain, and through researcher responses to questions in the interviews.

My first hypothesis was that researchers satisficed in the knowledge they obtained when determining whether to reuse data. As I discussed above and show in Table 4.5, 195 researchers (25.4%) out of 767 from the combined samples of quantitative and qualitative data reusers reported that there was knowledge about the data they desired but were not able to obtain at the time they made the decision to reuse the data. This hypothesis was thus supported by my findings.

*Table 4.5 Cross-Tabulation of the variable for knowledge limited when deciding to reuse data*

| Lack of desired knowledge | Freq. | Percent | Cum. |
|---|---|---|---|
| No | 572 | 74.58 | 74.58 |
| Yes | 195 | 25.42 | 100.00 |
| Total | 767 | 100.00 | |

In addition to asking researchers if they lacked desired knowledge at the time they decided to reuse the data, I asked researchers if they lacked knowledge they desired when considering to reuse the data. As noted in section 2.9.1.1, this was important because if I had not asked, I would not know if researchers who said they did not lack desired knowledge about data ever lacked desired knowledge.

Of the 546 (166+380) people who were sure they did not lack knowledge at the time of decision (see Table 4.6), 166 (30.4%) were sure they lacked knowledge at the time of consideration. So 380 (69.6%) did not lack knowledge at the time of consideration or time of decision, suggesting that a large proportion chose data they knew from the outset would fulfill their research aims. I discuss further evidence of this finding in sections 4.3, 4.4, 4.5, and 4.6.

*Table 4.6 Cross-Tabulation of Knowledge Lacking at the Time the Researcher Considered Reusing the data and the Time they Decided to Reuse the Data*

| Knowledge lacking when deciding | Knowledge lacking when considering | | | |
|---|---|---|---|---|
| | Yes | No | Not sure | Total |
| Yes | 140 | 49 | 6 | 195 |
| No | 166 | 380 | 25 | 571 |
| Total | 306 | 429 | 31 | 766 |

### *4.2.2 Research Question 2. Do Researchers Satisfice? Findings From the Interviews*

During the interviews, I began to get the sense that satisficing was not a good characterization of the experience researchers had reusing data. I was surprised by this at first, since I found during the qualitative portion of my study that even more researchers lacked desired knowledge than had indicated so in the survey. (That is, while 15 out of 26 researchers reported they did not lack knowledge in the survey, I found from reading researchers' articles and conducting interviews that all but six lacked some kind of knowledge they desired about the data—i.e., 20 out of 26 lacked knowledge they desired). If researchers lacked knowledge they desired about the data, it seemed they must have "settled" for acceptable data. How could satisficing be a poor fit?

I introduced findings about the inadequacy of the satisficing paradigm during the discussion of the development of my interview protocol in section 3.4.4, but a key theme that repeated itself as I conducted the interviews was that researchers did not see themselves operating in an environment where it was possible to achieve an optimal state of knowledge

about the data. There was no "best" option when using secondary data. Researchers weighed the pros and cons of using data that were less than ideal from the start and negotiated "tradeoffs" that would allow them to get the most they could out of the data. As one researcher explained,

> So when you're doing your calculations, you compare what you don't have to other data sets. They don't have even this much, so I should be kind of happy with what I have. This is the best. This is not the best, this is better than others in terms of addressing my bigger research question…
>
> You don't go with the best because there's usually, in terms of secondary data, there is no best. That's my experience at least. You go with the be[s]t- ish. The better means the one that answers your bigger question. You can miss some certain things, certain variables, but at least, if you're interested in [topic], you need that [dataset] that I knew. And so I knew that there pros and cons and I kind of weigh them and [data study] was able to provide me with more pros when it came to addressing my research questions. (D1-02)

In speaking of the shortcomings of data and translating those into future investigation, another said,

> you pose what would have been better and might be addressed in future studies…Because if you wait to have a perfect data, you will not do any research. (D1-04)

The sentiment that lacking knowledge was just the nature of reusing data was widespread. One researcher (D-01) described their experience this way:

> So, you kind of use these data sets and they cover a range of information that are generally useful for this kind of analysis. And then they will be here and there a few information that will be missing. There will be some variables where you don't get a response from all the individuals, so they will be missing. So, I mean, these are the problem that you face. (D-01)

Several researchers pointed to limitations that were understandable because of the expense involved in gathering data:

> But I think once you get to understand the study- think well, yeah, this is very expensive. So there will be always limitations and that's your compromise, right? (D2-04)

This was echoed by others:

> And that happens almost in any data that you're looking at because [the data creators] can't ask everything. They have to make budgetary decisions about how much information is going to be collected on each topic, and so forth. And so, in these large surveys there are trade-offs between covering a number of areas and getting some information versus going very, very in-depth on one area or two or three. (D1-04)

> Being that these data were collected across seven years and because they're so extensive, and this was millions of dollars in grant money to collect this data, it really made projects that I was interested in possible, even if it wasn't the exact specific project that I wanted to do. (D4-02)

One researcher spoke about the nature of imperfect data in the context of data reuse in general, and the value of data being available to reuse rather than being "locked up in a box":

> But to think that you're the only person that can answer important questions with the data that you collect, or that somehow because you did the interviews, you're the only people who can write about [type of data]—I understand how, yes, you might be the best-case scenario, but all data is flawed. No data is perfect. There's mistakes in everything. There's issues with everything. And if our job as scientists is to use to the best of our ability what is available to us to discern what is happening in the world, then let us use that somewhat imperfect data. It's better than having it locked up in a box. And science is about, you know, progress. And if somebody finds something different then that's great. So I guess for me, the data fit better than I thought it could, and that there's a lot of potential out there for other [type of data] to be helpful to scientists. Especially when they're asking different kinds of questions. (D-03)

A consequence of working with data they believed to be sub-optimal, as the preceding quotes attest, was the need for researchers to make compromises or tradeoffs. However, within stories of compromise, researchers expressed feelings of excitement (as in the quote above) that came with the possibility of making new discoveries. In another example, researcher D-11 discussed the challenges of not knowing everything they would like about data, and the discomfort that came with that, as part of what was involved in working on the "frontiers of knowledge":

> And so that known unknown does cause some anxiety, but it's also, I think, it just comes with the territory, right? We're on the frontiers of knowledge…You do what you can, you acknowledge it, and…if there's ever any evidence that what I'm doing is wrong, I can correct myself. But it's just a known unknown. I know that I can't get it. (D-11)

The tension I felt in the interviews between researchers' struggling on one hand to work with data that were, by definition, less than ideal, and the excitement on the other of making new discoveries gradually caused a change in the way I understood satisficing in data reuse. I heard the frustrations, challenges, and disappointments researchers expressed. But I also began to recognize in researchers' descriptions of reusing data, the complexity, creativity, and novelty of the work they did to investigate new questions and problems in spite of limitations. This was often expressed in an understated way as doing "the best they could" with the data, as in the quotes below.

> you go with what you have, because up until this point, we did not have these questions asked in a [description of sample]. So yeah, that's [an inadequacy in the data that is] pretty common…And so what you do is you do the best you can for what it is that you want to study. And then you make the best case for it to the reviewers that are reading your paper for peer review. (D1-04)

> And if we could've indicated [phenomenon of interest], it would have been a richer paper and a better paper, but we ended up just say, you know, "This is as good as we can go." (D2-01)

One researcher described their approach to making the most of data in a way I came to understand to as a "reuse mindset," or a mindset of using what you have to develop something new:

> what I learned is to try to make use of the existing resources to develop something new…And it's a lesson I learned from the reuse of secondary data. Because, you only have such kind of things on your hand, so what you can do is to try to think in different dimensions, to make use of the existing data set and try to develop something new, and try to enrich it by another research method. Such as the qualitative research method…Yes, of course, it is a constraint for you to use the secondary data, but on the other hand, it also helps you to think about how to use up the existing resource to fulfill your objectives.
>
> Nowadays, I do not have any hesitation to make use of secondary data to write something new. Because I have such kind of experience. I know that we only have such kind of data on hand. You have to do something—you have to identify something new. Otherwise, you cannot write your paper. You're not able to keep a job, you are not able [to produce research] and so on. Therefore, the PhD experience- or the experience of using [data

> study] helped me to develop this kind of mindset. Indirectly, obviously. It is something indirect, not direct…my professor did not do such kind of things. He just provides the dataset and asks us, you have to think by yourself. You have to look for something new by yourself. (D3-02)

This approach—having a reuse mindset—differed to me in an important way from a satisficing mindset. Given a satisficing mindset, I would imagine researchers describing how they fell short of an ideal; how they reached an "acceptable" as opposed to an "optimal" equilibrium with the data. What I found, however, was that most of the researchers simply did not understand "optimal" to be a possibility in the context of reusing data.

In an environment where the value of collecting original data was ascendant, researchers were performing creative acts using pre-existing data and obtaining results that might never—or in some cases would never—be possible given constraints on time and resources, or on the possibility of collecting additional data. While a satisficing mindset and reuse mindset might seem to share similarities (i.e., each evokes the achievement of goals in the face of limited resources), a satisficing mindset suggests a limited achievement of aims and disappointment (the researcher might have done better but did not), whereas a reuse mindset suggests the best and inherently creative achievement of aims given the attendant circumstances. This is an important distinction that I return to in the discussion of my findings.

While the idea of satisficing did not well characterize the behavior of the researchers I interviewed, the fact remained that researchers lacked knowledge about data that they desired. Throughout the remainder of my findings, I continue to report on this phenomenon (lacking knowledge) according to my research questions, and I keep the wording of the research questions the same for consistency. However, I report on the phenomenon of lacking desired knowledge about data, rather than satisficing.

### *4.2.3 Research Question 2a. If so, How Can Knowledge Satisficing Be Characterized? Findings From the Survey*

My findings in sections 4.2.1 and 4.2.3 showed that researchers lacked knowledge they desired about data at the time they decided to reuse them. In this section, I report on findings from the survey that characterize the knowledge that was lacking, including the type of knowledge, the importance of that knowledge to the research, and the amount of desired knowledge that researchers were able to obtain.

#### *4.2.3.1 Findings About Type of Knowledge Lacking*

Among the three samples I drew for the survey, 223 researchers reported lacking knowledge they desired about data by my definition (see section 4.1.1). Of these, 159 reported at least one and as many as four types of knowledge they were lacking, for a total of 307 descriptions of knowledge lacking. For 10 of these descriptions, I either determined that knowledge was not actually lacking at the time the decision to reuse the data was made or I was not able to determine definitively whether knowledge was lacking (and there was an indication that it was not). This left 297 descriptions of lacking knowledge for 156 researchers. Table 4.7 shows the number of researchers who reported or whose responses I coded as being more than one type of knowledge that was lacking.

*Table 4.7 Number of Types of Knowledge Coded by Number of Researchers*

| Number of coded types of knowledge | Number of researchers | Percent of researchers |
|---|---|---|
| 1 | 66 | 22.22 |
| 2 | 84 | 28.29 |
| 3 | 135 | 45.46 |
| 4 | 12 | 4.04 |
| Total | 297 | 100 |

Table 4.8 shows the first and second level categories I used to code the different kinds of desired knowledge and the distribution of the categories.

*Table 4.8 First and Second Levels of Coded Categories for Knowledge Researchers Desired That Was Lacking, Count of Each Code, and Code Definition*

| **Level 1 Knowledge area**<br>Level 2 Knowledge area | **Count** | **Definition used for coding** |
|---|---|---|
| **Data supplement** | **130** | |
| Coverage | 96 | Data beyond those that were made available. |
| Detail | 29 | More detail about available data |
| Missing | 5 | Researchers desired the data that were missing |
| **Data collection** | **66** | |
| Limitations | 2 | Limitations of the data |
| Instrument | 11 | The survey instrument |
| Measurement | 5 | Measurement of phenomena |
| Missing data | 7 | Why data were missing |
| Population | 1 | The study population |
| Process | 14 | The data collection process |
| Respondents | 2 | The study respondents |
| Response rates | 2 | The study response rates |
| Sampling | 20 | Sampling methods |
| Time | 2 | When the data were collected |
| **Data analysis** | **33** | |
| General | 1 | How data were analyzed |
| Cleaning | 7 | How the data were cleaned |
| Coding | 15 | How the data were coded |
| Method | 1 | How the data were processed |
| Missing data | 1 | How missing data were handled |
| Scale development | 3 | How scales were developed |
| Transcription | 1 | How data were transcribed |
| Weighting | 4 | How weightings were determined |
| **Data** | **28** | |
| General | 1 | Knowledge about the data |
| Accuracy | 1 | Data accuracy |
| Available | 1 | What data were available to analyze |
| Consistency | 4 | How data compares with other data from the study |
| How used at time | 1 | How the data were used by the data collectors |
| Limitations | 1 | Limitations of data |
| Missing | 2 | Nature of the missing data |
| Quality | 5 | Data quality |
| Relationship between variables | 1 | Relationship between two variables of interest |
| Reliability | 3 | Data reliability |
| Structure | 2 | Data structure |
| Translation | 2 | Translation of the data |
| Validity | 4 | Validity of data |
| **Data documentation** | **13** | |
| General | 6 | Description of the data |
| Definitions | 7 | How measured concepts were defined |
| **Data reuse** | **6** | Strategies/processes for reusing data |
| **People** | **6** | |
| Data analyzers | 3 | Those who analyzed the data |
| Data collectors | 3 | The data collectors |
| **Other** | **5** | |
| Articles (translation) | 1 | Translations of articles pertinent to the study |
| Not sure | 3 | Could not interpret what knowledge was lacking |
| Previous research results | 1 | Previous research results |
| **Data access information** | **4** | How to access data |
| **Data management** | **2** | How data were managed |
| **Data comparability** | **1** | The comparability of the data |
| **Data context** | **1** | Unspecified data context |
| **Data reporting** | **1** | Public reporting of data |
| **Data validation** | **1** | How data were validated |
| Total | 297 | |

It is clear from the table that data supplement was the largest category of knowledge researchers lacked, with 130 entries. This is the knowledge "of" data that I described in section 3.4.8.5 (knowledge of a sample or population itself) and stands in contrast to knowledge "about" data, as knowledge in the other categories could be construed. I identified three kinds of supplemental knowledge: data supplement (coverage), the largest category; data supplement (detail); and data supplement (missing). Data supplement (coverage) was when researchers wanted broader knowledge than was available in the data, such as measurements of a phenomenon in a different population or answers from respondents about an additional topic. Data supplement (detail) was when researchers wanted more detail than was indicated in the data. For instance, if the data indicated patients were on "other" medications, a researcher might want more details about what the "other" medications were. Data supplement (missing) was when researchers desired to have the data that were missing.

The second largest category was data collection with 66 entries. The most frequently desired knowledge about data collection had to do with knowledge about the sample or how sampling for the data was done. This included knowledge about whether there was bias in the sample, whether the sample was inclusive, how decisions about the sample were made, the size of the sample, why information was collected for some populations but not others, and sample exclusion criteria.

Knowledge about the process of data collection was the next most frequently lacking. Researchers in this group indicated they wanted more detailed information, including tacit information about how the data were collected. Often the desire was for more specific knowledge without the specific knowledge being indicated, but details about the survey

methodology, context of data collection that might have affected data quality, and how many times respondents needed to be contacted are some examples.

Knowledge about the survey instrument, missing data, measurement of variables, and other knowledge were also included in this category. Here, the lack of knowledge did not have to do with wanting to obtain the data that were missing but rather with understanding why the data were missing. Here also, data collection (exclusion) had to do with understanding what instances of a historical phenomenon were included or excluded in an assembled index.

The third most frequent category of knowledge researchers reported lacking was knowledge about data analysis (33 entries), especially about how data were coded and cleaned, how variables should be weighted, the scales used for variables in the analysis, how analysis proceeded including handling of missing data, and procedures for data transcription.

Knowledge about data (25 entries) was the next largest category of knowledge that was lacking. Researchers desired knowledge particularly about data quality, consistency, validity, and reliability; also data structure and accuracy, and the limitations of the data. Researchers desired translation of the data from another language, to know what data were available (they did not have access to the full data at the time they decided to reuse them), to the know more about the relationship between variables, and how variables were used by the people who originally collected them. Data (missing) here refers to the nature of the missing data (what kind of data were missing, as opposed to having the data or understanding why the data were missing).

Data documentation was the fourth largest category with 12 entries. Several researchers wanted knowledge about how variables were defined or differed from one another. Other researchers desired some form of documentation for parts of the data they were interested in.

Data reuse, people, "other," data access information, and data management were the next largest categories of desired knowledge with six, six, five, four, and two entries respectively. Data reuse included knowledge about how to handle variable weights or perform statistical functions; people included knowledge about the people who collected or analyzed the data previously; "other" included knowledge I was unsure of how to interpret or categorize, and knowledge not related to the data themselves such as previous research results and translations of articles (not data) related to the study; data access information included knowledge about when data would be released or why data would not be released. Data management had only two entries. These had to do with knowledge about discrepancies between the data retrieved from ICPSR and another source.

There were four more kinds of knowledge that researchers lacked that I only coded once. These were data comparability, data context, data reporting, and data validation. These had to do with, in order, how easy or difficult it was to merge the data with other data, knowledge about data context that was unspecified, public reporting of the data, and how the data had been validated (as opposed to whether the data had been validated with respect to knowledge about data).

One of the main findings these results pointed towards, and which gained support as I proceeded with my analysis, was the finding that to researchers—at least to researchers who reused data held in ICPSR—what went into the archive (the data that were collected or produced) was as much or possibly even more important to facilitating research with secondary data as improving documentation or adding details about the context of data creation.

#### *4.2.3.2 Findings About Importance of Knowledge Lacking*

Table 4.9 shows the distribution of the importance of the knowledge that was desired but lacking to researchers. There were seven missing values, and 24 researchers who reported that the knowledge that was lacking was not at all important to their decision to reuse the data. Aside from these, however, nearly 90% of the areas in which researchers reported lacking knowledge were somewhat important to the researchers' decisions to reuse the data.

*Table 4.9 Distribution of the importance of the knowledge that was lacking to researchers' decisions to reuse data.*

| Knowledge importance | Frequency | Percent | Cumulative Percent |
|---|---|---|---|
| Essential | 33 | 11.11 | 11.11 |
| Very important | 70 | 23.57 | 34.68 |
| Important | 89 | 29.97 | 64.65 |
| Somewhat important | 74 | 24.92 | 89.56 |
| Not at all important | 24 | 8.08 | 97.64 |
| . (missing) | 7 | 2.36 | 100.00 |
| Total | 297 | 100.00 | |

Table 4.10 shows the distribution of the importance of the desired knowledge that was lacking in relation to the different knowledge categories. For each category of knowledge, a relatively high proportion of researchers believed the knowledge was important, very important, or essential. Data, data access information, data collection, data comparability, data documentation, and "other" stood out as being very important or essential to high proportions of researchers. These findings mainly speak to the substance of the data presented in this dissertation about knowledge lacking. The knowledge indicated was, by and large, knowledge that was important to the researchers to conduct the research they would have liked.

*Table 4.10 Distribution of the Importance of Desired Knowledge That Was Lacking*

| Knowledge area Level 1 | Knowledge importance | | | | | | |
|---|---|---|---|---|---|---|---|
| | Not at all important | Somewhat important | Important | Very important | Essential | . (missing) | Total |
| Data | 1 | 4 | 7 | 14 | 2 | 0 | 28 |
| | 3.6 | 14.3 | 25.0 | 50 | 7.1 | 0 | 100 |
| Data access info | 0 | 1 | 0 | 3 | 0 | 0 | 4 |
| | 0 | 25.0 | 0 | 75.0 | 0 | 0 | 100 |
| Data analysis | 5 | 9 | 6 | 10 | 3 | 0 | 33 |
| | 15.2 | 27.3 | 18.2 | 30.3 | 9.1 | 0 | 100 |
| Data collection | 5 | 15 | 20 | 14 | 12 | 0 | 66 |
| | 7.6 | 22.7 | 30.3 | 21.2 | 18.2 | 0 | 100 |
| Data comparability | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| | 0 | 0 | 0 | 100 | 0 | 0 | 100 |
| Data context | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 0 | 0 | 100 | 0 | 0 | 0 | 100 |
| Data documentation | 0 | 2 | 1 | 5 | 4 | 1 | 13 |
| | 0 | 15.4 | 7.7 | 38.5 | 30.8 | 7.7 | 100 |
| Data management | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| | 0 | 50 | 0 | 50 | 0 | 0 | 100 |
| Data reporting | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 0 | 0 | 100 | 0 | 0 | 0 | 100 |
| Data reuse | 2 | 0 | 2 | 0 | 2 | 0 | 6 |
| | 33.3 | 0 | 33.3 | 0 | 33.3 | 0 | 100 |
| Data supplement | 10 | 38 | 48 | 19 | 9 | 6 | 130 |
| | 7.7 | 29.2 | 36.9 | 14.6 | 6.9 | 4.6 | 100 |
| Data validation | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| | 0 | 0 | 0 | 0 | 100 | 0 | 100 |
| Other | 0 | 2 | 1 | 2 | 0 | 0 | 5 |
| | 0 | 40 | 20 | 40 | 0 | 0 | 100 |
| People | 1 | 2 | 2 | 1 | 0 | 0 | 6 |
| | 16.7 | 33.3 | 33.3 | 16.7 | 0 | 0 | 100 |
| Total | 24 | 74 | 89 | 70 | 33 | 7.0 | 297.0 |
| | 8.1 | 24.9 | 30 | 23.6 | 11.1 | 2.4 | 100 |

*Note*. The first row of each category contains frequencies and the second row contains row percentages.

#### *4.2.3.3 Findings About Amounts of Knowledge Obtained*

Table 4.12 gives descriptive statistics about the amount of knowledge researchers obtained out of what they desired in each category of knowledge. The table excludes missing values (i.e., where researchers indicated there was knowledge they were lacking but did not enter how much knowledge of the desired knowledge they were able to obtain), of which there were 27. In the process of coding types of knowledge researcher's desired (reported on in the previous section), in some cases I changed missing values for amount of knowledge to zero and for source

of knowledge to NA (because the amount was zero) based on the researcher's response to other questions. If I had doubts (if the changes were not explicitly merited based on the other responses) I did not make the changes.

The variation in the standard deviation in Table 4.11 demonstrates a wide range among researchers in the amount of desired knowledge they reported obtaining. A limitation of the question I asked in the survey about the amount of obtained knowledge is that it did not specify at what point the knowledge was obtained: i.e., whether it was obtained at the time the decision to reuse the data was made, or at some point after that but before the conclusion of the research. Therefore, I am unable to make this distinction in my results.

My results do indicate, however, that in 14 knowledge areas across nine researchers (data not shown) researchers who reported lacking knowledge at the time the decision to reuse the data was made obtained 100 percent of the knowledge at some point by the end of their research. In an additional 12 knowledge areas across 10 researchers (data not shown), researchers obtained at least 90 percent of their desired knowledge. Overall, then, only 19 out of 156 researchers (12.2%) obtained all or most (90%) of the knowledge they desired in 27 out of 297 (9%) of the knowledge areas I identified. This means that 135 researchers (88.8%) obtained less than 90% of the knowledge they desired.

Like the finding about the importance of the desired knowledge that was lacking, this finding demonstrates the substantive nature of the phenomenon of lacking desired knowledge (i.e., that it is highly prevalent). I present further findings along these lines, focused on the impact of lacking desired knowledge on research outcomes, later in section 4.6.

*Table 4.11 Amount of Desired Knowledge Obtained by Category*

| | N | Min | Max | Mean | STDev | 25th | Median | 75th | 95th |
|---|---|---|---|---|---|---|---|---|---|
| Data supplement | 115 | 0 | 100 | 36.19 | 31.06 | 0 | 35 | 61 | 85 |
| Data collection | 62 | 0 | 100 | 63.53 | 27.12 | 50 | 70 | 81 | 100 |
| Data analysis | 33 | 0 | 100 | 51.64 | 30.79 | 26 | 50 | 79 | 100 |
| Data | 25 | 10 | 100 | 58.44 | 26.99 | 40 | 60 | 80 | 98 |
| Data documentation | 12 | 20 | 96 | 60.75 | 25.41 | 39.5 | 70 | 79 | 96 |
| Other | 5 | 8 | 64 | 44.40 | 22.38 | 40 | 50 | 60 | 64 |
| Data reuse | 5 | 0 | 90 | 57.20 | 36.01 | 50 | 61 | 85 | 90 |
| People | 5 | 50 | 71 | 58.40 | 8.79 | 50 | 60 | 61 | 71 |
| Data access information | 2 | 30 | 31 | 30.50 | 0.71 | 30 | 30.5 | 31 | 31 |
| Data management | 2 | 31 | 80 | 55.50 | 34.65 | 31 | 55.5 | 80 | 80 |
| Data comparability | 1 | 31 | 31 | - | - | - | - | - | - |
| Data context | 1 | 95 | 95 | - | - | - | - | - | - |
| Data reporting | 1 | 49 | 49 | - | - | - | - | - | - |
| Data validation | 1 | 84 | 84 | - | - | - | - | - | - |

*Note*. Missing responses are excluded.

### *4.2.4 Summary*

I found through my survey results that just over 25% of researchers lacked knowledge about data that they desired at the time they decided to reuse them in their research. I was also able to characterize the knowledge that researchers lacked by type of knowledge, importance, and amount of knowledge obtained. Some of my main insights from these findings were high importance to researchers of the knowledge they lacked about data, the prevalence of lacking knowledge, and the notion, from my findings about types of knowledge lacking, that what goes into the archive (the data that are collected or produced) are as much or possibly even more important to facilitating research with secondary data as improving documentation or adding details about the context of data creation.

To summarize, I found by coding descriptions of 297 areas of knowledge that 156 researchers reported lacking about data (out of 223 researchers from all samples) that most researchers lacked either more detailed knowledge about the data than was available, or broader knowledge. Nearly 90% of the time, the knowledge researchers lacked was at least somewhat

important to their decisions to reuse the data. The distribution of areas of knowledge by count were:

- Data supplement (130)
- Data collection (66)
- Data analysis (33)
- Data (28)
- Data documentation (13)
- Data reuse (6)
- People (6)
- Other (5)
- Data access information (4)
- Data management (2)
- Data comparability (1)
- Data context (1)
- Data reporting (1)
- Data validation (1)

Researchers obtained the greatest median amounts of the knowledge related to data collection and data documentation (for each, researchers reported obtaining a median of 70% of the total knowledge they desired). The next highest areas of attainment were knowledge about data analysis, data, other, data reuse, people, and data management. In these areas, researchers obtained around 50-60% of the knowledge they desired. Researchers obtained the lowest median amounts of knowledge in the areas of data supplement (35%) and data access information (30.5%). Note that I received more than 10 responses about the amount of obtained knowledge in only the first five categories in the bulleted list above. 19 out of 156 researchers (12.2%) ultimately obtained all or most (90%) of the knowledge they desired (across 27 of 297 of the knowledge areas I identified). 135 researchers (88.8%) ultimately obtained less than 90% of the knowledge they desired.

While I was able to quantify the number of researchers who lacked knowledge about data they desired and characterize the knowledge they lacked to some degree, I found from my interviews that satisficing did not well characterize the behavior of researchers who lacked

desired knowledge. The interviews revealed that researchers viewed the act of reusing others' data as subordinate to the optimal condition of gathering data themselves to answer their precise research questions (which is not to say that they did not see great value and advantages to reusing data, which they did).

In the absence of an optimal outcome, rather than seeing themselves as satisficing in what they accomplished, researchers viewed themselves as making the most of a non-optimal situation. The absence of an optimal outcome places researchers' behavior outside the realm of satisficing. Rather than researchers falling short of an optimal bar, my findings indicate that researchers leveraged their own and others' skills, knowledge, and experience, to make creative, novel contributions to scientific knowledge. I discuss more in section 5 why this alternate understanding of data reuse is important.

### 4.3 Factors That Affect Knowledge Bounding and Knowledge Satisficing

In this section, I present findings about factors that affected how researchers bounded knowledge or determined the thresholds of knowledge about data that were sufficient to decide to reuse data. I identified a lack of research on this topic as a gap in the literature in section 2.8. I have decided to discuss factors that affect knowledge bounding prior to addressing how knowledge bounding occurs (which I do in section 4.4—another gap I identified in the literature) because discussing relevant factors sets important context for the strategies researchers used employ to bound knowledge.

As in section 4.2, I begin with a presentation of findings from the survey, particularly those related to my 4$^{th}$ through 10$^{th}$ hypotheses, all related to factors that affect knowledge bounding. After presenting the results of the hypothesis tests, I construct a statistical model of

lacking desired knowledge using relevant factors from my analyses. The findings from the survey relate to my research question 2b: What factors influence knowledge satisficing?

After presenting the model for lacking desired knowledge, I present findings from the interviews relating to factors that affect knowledge bounding. However, these findings relate to my research question 1c, which is, more specifically, what researchers report influences their determinations about how much of what kinds of knowledge are sufficient to decide to reuse data.

In presenting these findings, I subsume research question 2c (What reasons do researchers give for why they satisfice?) under research question 1c. I do this because my findings caused me to reconceive this question as "What reasons do researchers give for why they lack knowledge they desire about data?" and I think this question is most appropriately addressed by considering how researchers bounded knowledge rather than why they did not gain all the knowledge they desired. I additionally discuss findings in a separate section (4.7.3) related to what inhibited and facilitated researchers' reuse of data. Those findings do not necessarily represent why researchers did not obtain all the knowledge they desired, but they do speak somewhat to that question.

### *4.3.1 Research Question 2b. What Factors Influence Knowledge Satisficing? Findings From the Survey (H4-H10)*

I reported findings related to my first hypothesis in section 4.2. My second hypothesis had to do with the effect of knowledge satisficing on research outcomes and the third had to do with the effect of knowledge satisficing on the achievement of data reuse goals. I report on these further in section 4.6. The remaining hypotheses (four through ten) had to do with the association between different factors on knowledge satisficing or the probability of satisficing in the presence of different factors and I treat them in turn below. For each of the hypothesis tests, I

substituted the phenomenon of lacking desired knowledge for satisficing, given my findings about the poor fit between satisficing and the circumstances I encountered. In each of the tests, I used a binary variable for lacking desired knowledge (researchers either lacked desired knowledge or did not) as the dependent variable and the individual factor I was testing as the independent variable. Where I performed regression analysis I also included the control variables and accounted for clustering effects.

#### *4.3.1.1 H4. There Is No Association Between Knowledge Satisficing and Reuse of Qualitative or Quantitative data*

I hypothesized that there was no relationship between lacking desired knowledge and reuse of qualitative or quantitative data and indeed did not find one. Table 4.12 shows a cross-tabulation of lacking desired knowledge and type of data. A chi-square test of independence yielded a test statistic of 0.002, which was not large enough ($p = 0.965$) to reject the null hypothesis that there is no relationship between the variables.

*Table 4.12 Cross-Tabulation of Knowledge Lacking and Whether the Researcher Reused Quantitative or Qualitative Data*

| Lack of desired knowledge | Data qualitative or quantitative | | |
|---|---|---|---|
| | Quant | Qual | Total |
| Yes | 187 | 7 | 194 |
| No | 550 | 21 | 571 |
| Total | 737 | 28 | 765 |

Pearson Chi2 = 0.002 Prob = 0.9645

I also tested the null hypothesis that reusing qualitative or quantitative data does not predict a lack of knowledge including the control variables in a logistic regression. Here, the regression did not show a significant difference in the odds of lacking knowledge when the researcher reused qualitative or quantitative data ($p = .844$; see Table 4.13). I thus did not find any evidence to suggest a significant association.

*Table 4.13 Estimated Association between Reused Data Being Qualitative or Quantitative and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Qual quant | 1.093 | 0.844 | 0.449 | 2.665 | |
| Number of Citations (ln) | 0.916 | 0.081 | 0.830 | 1.011 | |
| Creation purpose | 0.777 | 0.178 | 0.539 | 1.121 | |
| Data source | 0.718 | 0.071 | 0.501 | 1.029 | |
| Constant | 0.774 | 0.653 | 0.253 | 2.365 | |
| Number of obs | 691 | | | | |

*** p<.01, ** p<.05

*Note:* Quantitative data is the reference category.

#### *4.3.1.2 H5. The Greater a Researcher's "Distance" From Data, the Higher the Probability of Knowledge Satisficing*

In my survey, I operationalized the concept of "distance" through two questions: one that asked about the researcher's involvement in the original research and another that asked about the knowledge of the respondent's research team (including the respondent) about aspects of the original data collection that were relevant to reusing the data. These questions both had five-point Likert-scale response options ranging from not involved (or no knowledge) to solely responsible (or in-depth knowledge). I used responses to these two questions to test the hypothesis that a greater "distance" from the data increased the probability of lacking desired knowledge.

##### *4.3.1.2.1 Involvement in Original Research*

I performed a logistic regression, including the control variables, to test the null hypothesis that there was not a significant association between a researcher's involvement in the original research and the odds of lacking desired knowledge about reused data. I used lacking desired knowledge (a binary variable) as the dependent variable and involvement in the original research As Table 4.14 shows, I did not find a significant relationship between involvement in the original research and the odds of lacking desired knowledge (p=0.440). I therefore could not

reject the null hypothesis and concluded that a researcher's involvement in the original research did not affect the odds of lacking desired knowledge.

*Table 4.14 Estimated Association between Original Involvement in the Research and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Original involvement | 0.947 | 0.440 | 0.824 | 1.088 | |
| Number of Citations (ln) | 0.913 | 0.069 | 0.827 | 1.007 | |
| Data purpose | 0.776 | 0.177 | 0.537 | 1.121 | |
| Data source | 0.751 | 0.121 | 0.524 | 1.078 | |
| Constant | 0.891 | 0.759 | 0.427 | 1.859 | |
| Number of obs | 692 | | | | |

**** p<.01, ** p<.05*

#### *4.3.1.2.2 Knowledge On a Researcher's Team*

I used a logistic regression (including the control variables) to test the null hypothesis that there was not a significant association between researchers lacking desired knowledge about data and the amount of knowledge present on a researcher's team (including the researcher themselves) about the original creation of the reused data. The levels of original knowledge in the survey were no knowledge, some knowledge, good knowledge, very good knowledge and in-depth knowledge. The regression in Table 4.15 shows that having knowledge about the creation of the original data had a significant association with the odds of a researcher lacking desired knowledge. I therefore rejected the null hypothesis. In particular, I concluded that the odds of lacking knowledge increased 0.724 times (i.e., decreased by 0.276 times) for every additional level of knowledge held about the original data creation by the data reuse research team. Put in other terms, the odds of not lacking knowledge increased by about 1.4 times for each greater level of knowledge).

*Table 4.15 Estimated Association Between the Amount of Knowledge About Original Data Creation and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Original knowledge | 0.724 | 0.001 | 0.619 | 0.848 | *** |
| Number of Citations (ln) | 0.909 | 0.062 | 0.822 | 1.005 | |
| Data purpose | 0.767 | 0.153 | 0.533 | 1.000 | |
| Data source | 0.893 | 0.552 | 0.615 | 1.297 | |
| Constant | 2.108 | 0.081 | 0.912 | 4.874 | |
| Number of observations | 692 | | | | |

**** p<.01, ** p<.05*

#### *4.3.1.3 H6. The More Experience a Researcher Has in Their Primary Domain of Research, the Lower the Probability of Knowledge Satisficing*

I measured experience through questions about the researcher's departmental affiliation, primary domain of research, and years of experience in their primary domain of research. In the end, I decided to exclude an analysis of departmental affiliation because I felt the researcher's domain provided a more granular and accurate representation of their field of research (at least, I reasoned, department affiliation would not be more accurate) and to reduce the time I would need to conduct my analysis.

##### *4.3.1.3.1 Primary domain of research*

I coded information about primary domain at a broad level (level 1) and a more granular level (level 2). Many researchers listed more than one domain and I coded these where I felt I could reasonably distinguish between them (up to five domains each for levels 1 and 2). Table 4.16 shows the distribution of researchers who I coded as having zero, one, two, three, four, or five primary domains.

*Table 4.16 Distribution of Number of Primary Domains*

| Number of domains | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 (missing) | 60 | 7.82 | 7.82 |
| 1 | 502 | 65.45 | 73.27 |
| 2 | 174 | 22.69 | 95.96 |
| 3 | 22 | 2.87 | 98.83 |
| 4 | 8 | 1.04 | 99.87 |
| 5 | 1 | 0.13 | 100.00 |
| Total | 767 | 100.00 | |

In the end I coded 56 level 1 categories. These are listed along with the count of researcher responses I coded both individually and in combination in Table 4.17. For instance, the second column in the table indicates that I coded 222 researchers as having "health" as one of or the only primary domain. Columns 3-10 show the number of codes per researcher that were also coded along with the code in column 1. For instance, I coded 127 researchers as having indicated "health" only. Seven researchers I coded as both "health" and "justice," 43 I coded as both health and life course, etc. It is important to note that columns three to ten are not exclusive. That is, I coded seven researchers as "health" and "justice" and 43 as "health" and "life course," but a single researcher being coded as "health," "justice," and "life course" would add one to each. It is therefore not possible to know from the table, for instance for "organizations," whether the total of five includes one researcher coded as "health" and another coded as "justice", or whether one researcher was coded as "health" and "justice." For the same reason, the columns for an individual domain (e.g., health) will add up to more than the total count (e.g., 222). Nevertheless, the table gives a sense of the co-occurrence of codes by showing the number of times there are co-occurrences (regardless of the number of researchers involved).

The dark gray shading indicates domains that I coded less than 10 times in the sample. I included these in a single category that I refer as "less common domains 1" in my regression analysis below. These domains plus the domains shaded in light gray comprise the "less common domains 2" category I refer to as well below.

*Table 4.17 Primary Research Domains Coded and Count of Researchers*

| Domain | Count | health | justice | life course | sociology | economics | social behavior | education | population | policy |
|---|---|---|---|---|---|---|---|---|---|---|
| health | 222 | 127 | 7 | 43 | 3 | 5 | 1 | 3 | 6 | 7 |
| justice | 150 | 7 | 101 | 10 | 5 | 2 | 2 | 2 | 0 | 2 |
| life_course | 118 | 43 | 10 | 44 | 2 | 0 | 5 | 5 | 0 | 3 |
| sociology | 54 | 3 | 5 | 2 | 36 | 0 | 0 | 2 | 5 | 0 |
| economics | 50 | 5 | 2 | 0 | 0 | 36 | 0 | 0 | 0 | 1 |
| social_behavior | 37 | 1 | 2 | 5 | 0 | 0 | 27 | 1 | 0 | 0 |
| education | 34 | 3 | 2 | 5 | 2 | 0 | 1 | 18 | 0 | 1 |
| population | 33 | 6 | 0 | 0 | 5 | 0 | 0 | 0 | 19 | 0 |
| policy | 30 | 7 | 2 | 3 | 0 | 1 | 0 | 1 | 0 | 14 |
| inequality | 26 | 15 | 1 | 2 | 0 | 3 | 0 | 0 | 0 | 0 |
| health_services | 21 | 6 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 2 |
| relationships | 20 | 3 | 4 | 9 | 0 | 0 | 2 | 1 | 0 | 1 |
| methodology | 19 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| political_science | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| race | 11 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| communities | 10 | 1 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| gender | 9 | 3 | 3 | 6 | 0 | 0 | 1 | 0 | 0 | 0 |
| social_science | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| poverty | 8 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| cognitive_science | 6 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| decision_making | 6 | 2 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| organizations | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| social_biology | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| work | 4 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| history | 4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| theory | 4 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| ethnicity | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| geography | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| politics | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| sexuality | 3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| social_work | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| assessment | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| communication | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ecology | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| engineering | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| religion | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| charity | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| arts | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| atmospheric_science | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| citizenship | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| civic_engagement | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| data_science | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| housing | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| informatics | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| information_studies | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| information_technology | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| intergroup_relations | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| library_information | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| na | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| single_mothers | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| social_cohesion | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| social_networks | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| social_programs | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| technology | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| transportation | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| urban_planning | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 954 | | | | | | | | | |

*Note*. Many researchers indicated more than one domain. The highlighted rows are those included in the "less common domains 2" category.

I tested the significance of the relationship between each primary research domain and the odds of lacking desired knowledge using the level 1 coded domains with logistic regressions that included the control variables. The results are shown in Table 4.18 (the table shows the results of many individual regressions; I did not include all domains as variables a single regression together). All domain areas with 10 or more responses are listed individually. As indicated above, I grouped the remaining domain areas with less than 10 into the "less common domains 1" category. I made a dichotomized variable and assigned zero to researchers who I coded as working in any one of the top 16 domains and one to researchers who I coded as working in a less common domain (with less than 10 responses).

There were no domains for which there was a significant association with the odds of lacking desired knowledge at the $p < .05$ level. In my comparison of more and less common domains, the odds of lacking knowledge were about 2.3 times greater for researchers in the "less common domains 1" category.

When I added additional coded domains to the "less common domains 1" category (i.e., communities, race, political science, methodology, and other domains I coded as having more than 10 responses), the effect on the odds of lacking knowledge remained significant up until the addition of the "policy" domain. Once I added "policy," the association between domain and the odds of lacking knowledge was not significant. The odds ratio and other information for the combined shaded domains (from inequality to urban planning at the very bottom) is included in the "less common domains 2" category in the last row of Table 4.18. There were no instances in

the regression for the domain "communities" where the researcher lacked knowledge, so I omitted it from the analysis.

*Table 4.18 Estimated Association Between Primary Domain of Research and Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | N | Sig P>[z] |
|---|---|---|---|---|---|---|
| health | 0.697 | 0.084 | 0.463 | 1.049 | 693 | |
| justice | 1.033 | 0.887 | 0.661 | 1.615 | 693 | |
| life_course | 1.004 | 0.988 | 0.613 | 1.644 | 693 | |
| sociology | 1.281 | 0.455 | 0.669 | 2.453 | 693 | |
| economics | 1.076 | 0.836 | 0.538 | 2.150 | 693 | |
| social_behavior | 0.714 | 0.471 | 0.286 | 1.783 | 693 | |
| education | 0.971 | 0.946 | 0.409 | 2.306 | 693 | |
| population | 0.358 | 0.096 | 0.107 | 1.200 | 693 | |
| policy | 0.767 | 0.576 | 0.304 | 1.940 | 693 | |
| inequality | 0.253 | 0.064 | 0.059 | 1.085 | 693 | |
| health_services | 0.344 | 0.160 | 0.078 | 1.523 | 693 | |
| relationships | 1.321 | 0.605 | 0.460 | 3.792 | 693 | |
| methodology | 0.539 | 0.336 | 0.153 | 1.900 | 693 | |
| political_science | 1.052 | 0.932 | 0.326 | 3.393 | 693 | |
| race | 1.949 | 0.309 | 0.538 | 7.057 | 693 | |
| communities | (omitted) | | | | 693 | |
| less common domains 1 | 2.296 | 0.001 | 1.427 | 3.696 | 693 | *** |
| less common domains 2 | 1.559 | 0.039 | 1.023 | 2.374 | 693 | ** |

**** p<.01, ** p<.05*

These results should be interpreted with caution because of inconsistencies in the way I coded domains (see section 3.3.2.5). However, I do not believe the inconsistencies were significant enough to cast doubt on the general finding that researchers identifying with research domains that were less commonly represented in my sample were much more likely to lack knowledge they desired about the data they reused.

#### *4.3.1.3.2 Experience in domain*

I tested the null hypothesis that there was no relationship between the number of years of experience researchers had in their primary domain of research and the odds of lacking desired knowledge about data and was not able to reject it. I performed the test both with years of experience (Table 4.19) and the log-transformed years of experience (Table 4.20) and neither indicated a significant relationship ($p = .421$ for years of experience and $p = .381$ for the log of

years of experience). I therefore concluded that there was no relationship between a researcher's years of experience in their domain and lacking desired knowledge about data.

*Table 4.19 Estimated Association Between Years Experience in Primary Domain and Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Years experience domain | 0.992 | 0.421 | 0.972 | 1.012 | |
| Number of Citations (ln) | 0.910 | 0.075 | 0.821 | 1.009 | |
| Data purpose | 0.846 | 0.407 | 0.571 | 1.255 | |
| Data source | 0.715 | 0.085 | 0.487 | 1.048 | |
| Constant | 0.778 | 0.514 | 0.366 | 1.652 | |
| Number of obs | 657 | | | | |

**** p<.01, ** p<.05*

*Table 4.20 Estimated Association Between Log of Years Experience in Primary Domain and Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Years experience domain (ln) | 0.915 | 0.381 | 0.749 | 1.117 | |
| Number of Citations (ln) | 0.910 | 0.085 | 0.817 | 1.013 | |
| Data purpose | 0.839 | 0.384 | 0.564 | 1.246 | |
| Data source | 0.727 | 0.105 | 0.494 | 1.069 | |
| Constant | 0.827 | 0.647 | 0.366 | 1.867 | |
| Number of obs | 647 | | | | |

**** p<.01, ** p<.05*

#### *4.3.1.4 H7. The More Experience a Researcher Has With Data Reuse, the Lower the Probability of Knowledge Satisficing.*

I examined a number of variables related to experience to test this hypothesis. They include the following, which I examine in turn below:

- professional position
- experience reusing data produced in the same domain as the data that were reused, and any domain
- frequency of reuse for the same purpose for which these data were reused
- whether the reused data were produced in the researcher's primary domain of research
- the overall frequency with which the researcher reused data
- the number of projects involving reuse of data that the researcher had worked on
- the number of projects involving reuse of data where considerations surrounding reuse were substantially similar to the ones involved in reuse of the data I asked them about

- the number of papers the researcher authored or co-authored describing research involving reuse of data
- the number of papers the researcher authored or co-authored describing research involving reuse of data where considerations surrounding reuse were substantially similar to the ones involved in reuse of the data I asked them about

#### *4.3.1.4.1 Professional Position*

I did not find a significant relationship between professional position and the odds of lacking desired knowledge about data using multiple different ways of looking at position including:

- normalizing into granular categories (the categories are represented in Table 4.21)
- combining positions into students and non-students

For each of these ways of combining professional position I provide below a cross-tabulation of the way of combining professional position and the results of a logistic regression using the way of combining professional position as the independent variable and lacking desired knowledge as the dependent variable.

The main purposes of the cross-tabulations are to show the categories of professional position used in the regression analysis and provide information about the numbers of people in each position who lacked knowledge they desired about the data, even though statistical analysis did not show a significant relationship between the two variables. Since these are the primary purposes, I do not describe the cross-tabulations in more detail. The first cross-tabulation is Table 4.21 below, which shows the granular, categories of positions that resulted from my initial normalization of the free text survey responses.

*Table 4.21 Cross-Tabulation of Professional Position and Knowledge Lacking*

| Researcher position | Lack of desired knowledge | | |
|---|---|---|---|
| | Yes | No | Total |
| Undergraduate student | 2 | 3 | 5 |
| Masters student | 5 | 15 | 20 |
| PhD student | 63 | 197 | 260 |
| Postdoc | 8 | 45 | 53 |
| Assistant Professor | 27 | 94 | 121 |
| Assistant Researcher | 0 | 3 | 3 |
| Associate Professor | 17 | 62 | 79 |
| Associate Researcher | 0 | 3 | 3 |
| Lecturer | 1 | 10 | 11 |
| Professor | 24 | 87 | 111 |
| Researcher | 13 | 31 | 44 |
| Other | 5 | 8 | 13 |
| Total | 165 | 558 | 723 |

Table 4.22 shows the results of a logistic regression estimating the odds of lacking knowledge over the granular categories of professional positions. The results did not show a significant relationship between professional position and the odds of lacking desired knowledge ($p = .720$) and I thus could not reject the null hypothesis that there was not a significant relationship.

*Table 4.22 Estimated Association Between Professional Position and Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Researcher position | 1.012 | 0.720 | 0.950 | 1.077 | |
| Number of Citations (ln) | 0.918 | 0.100 | 0.828 | 1.017 | |
| Data purpose | 0.891 | 0.564 | 0.603 | 1.318 | |
| Data source | 0.673 | 0.042 | 0.459 | 0.986 | ** |
| Constant | 0.675 | 0.333 | 0.305 | 1.496 | |
| Number of obs | 660 | | | | |

**** p<.01, ** p<.05*

Table 4.23 shows a tabulation of positions combined by whether or not the researcher was a student at the time the research was conducted and lacking desired knowledge. Table 4.24 gives the results of the logistic regression using student status as the independent variable, including the control variables. As in the other two regressions above, the results did not show a significant relationship between professional position (by student status) and the odds of lacking

desired knowledge (p = .768) and did not allow me to reject the null hypothesis that there was not a significant relationship.

*Table 4.23 Tabulation of Professional Position Grouped by Student or Not and Desired Knowledge Lacking*

| Position normalized to student or not | Lack of desired knowledge | | |
|---|---|---|---|
| | Yes | No | Total |
| Student | 70 | 215 | 285 |
| Not student | 95 | 343 | 438 |
| Total | 165 | 558 | 723 |

*Table 4.24 Logistic regression: Knowledge Lacking and Professional Position Grouped by Student or Not*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Student or Not | 0.945 | 0.768 | 0.650 | 1.374 | |
| Number of Citations (ln) | 0.999 | 0.299 | 0.997 | 1.001 | |
| Data purpose | 0.890 | 0.563 | 0.601 | 1.319 | |
| Data source | 0.679 | 0.045 | 0.465 | 0.991 | ** |
| Constant | 0.652 | 0.353 | 0.265 | 1.606 | |
| Number of obs | 660 | | | | |

**** p<.01, ** p<.05*

#### *4.3.1.4.2 Experience Reusing Data in "This" or "Any" Domain*

Along with investigating professional position as a component of researcher experience reusing data, I investigated researchers' experience reusing data produced in the same domain as the data of interest in my research, and data produced in any domain. I investigated the relationship between these aspects of experience and the odds of lacking desired knowledge using logistic regressions. Based on the results, I could not reject the null hypotheses that a researcher's years of experience reusing data in the same domain as the reused data (Table 4.25) or any domain (Table 4.26) did not significantly affect the odds of lacking knowledge about data (p = .663 in the first case and p = .223 in the second). Running the regressions with log-transformed years of experience did not result in a significant relationship in either model (not shown; the significance of the independent variable's effect changed to p = .635 for "this" domain and p = .178 for "any" domain).

*Table 4.25 Estimated Association Between Experience Reusing Data Produced in Same Domain as Reused Data and Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Reuse experience this | 0.994 | 0.663 | 0.969 | 1.020 | |
| Number of Citations (ln) | 0.917 | 0.115 | 0.824 | 1.021 | |
| Data purpose | 0.957 | 0.833 | 0.635 | 1.441 | |
| Data source | 0.741 | 0.140 | 0.498 | 1.103 | |
| Constant | 0.579 | 0.180 | 0.261 | 1.286 | |
| Number of obs | 617 | | | | |

*** p<.01, ** p<.05*

*Table 4.26 Estimated Association Between Experience Reusing Data Produced in Any Domain and Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Reuse experience any | 0.986 | 0.223 | 0.964 | 1.009 | |
| Number of Citations (ln) | 0.892 | 0.026 | 0.807 | 0.987 | |
| Data purpose | 0.921 | 0.695 | 0.611 | 1.389 | |
| Data source | 0.738 | 0.134 | 0.496 | 1.098 | |
| Constant | 0.712 | 0.404 | 0.322 | 1.579 | |
| Number of obs | 611 | | | | |

*** p<.01, ** p<.05*

#### *4.3.1.4.3 Frequency of Reuse for a Purpose for Which These Data Were Reused, and Frequency of Reuse Overall*

I investigated potential associations between researchers' frequency of reuse and lacking desired knowledge, both for the specific purposes for which they reused the data of interest and for any purpose (frequency of reuse overall). I did not find that reusing data for any specific purpose had a significant effect on the odds of lacking knowledge, except in the case of reusing data to compare them with other data. When the researcher reused the data for comparison purposes, the odds of lacking knowledge increased by 1.484 times (p = .001) with each increase in the frequency of reuse for this purpose (with response options "never," "rarely," "sometimes," "often," "always"). The results of the logistic regressions for each reuse purpose are given in Table 4.27. The final row of the table includes the results of the regression for knowledge lacking and overall frequency of reuse (i.e., frequency of reuse for any purpose).

*Table 4.27 Estimated Association Between Frequency of Reusing Data for Particular Purpose and Lacking Desired Knowledge*

| Knowledge Lacking | Odds ratio | P>[z] | [95% Conf | Interval] | N | Sig P>[z] |
|---|---|---|---|---|---|---|
| Background | 0.981 | 0.925 | 1.467 | 2.119 | 143 | |
| Validate | 0.834 | 0.357 | 1.227 | 1.362 | 139 | |
| New question | 0.937 | 0.501 | 1.133 | 1.026 | 517 | |
| Replicate | 2.066 | 0.063 | 4.438 | 1.297 | 41 | |
| Combine | 1.266 | 0.161 | 1.760 | 1.470 | 202 | |
| Compare | 1.484 | 0.049 | 2.199 | 1.456 | 159 | ** |
| Theory | 0.974 | 0.806 | 1.202 | 1.446 | 423 | |
| Tool | 0.640 | 0.277 | 1.432 | 2.351 | 49 | |
| Other | (see note) | | | | | |
| Overall frequency | 1.079 | .479 | -.135 | .288 | 664 | |

*** p<.01, ** p<.05*

*Note*. There were no frequencies for "other" that reused data from ICPSR and the model did not return usable results.

#### *4.3.1.4.4 Reuse of Data in Primary Domain of Research (q30)*

I also investigated the hypothesis that if the data the researcher reused were produced in the researcher's primary domain of research, the odds that the researcher would lack desired knowledge about the data would be less. I did not find a significant relationship with the reused data being in the same domain as the researcher's primary domain, however (p = .537), and was unable to reject the null hypothesis that there is no association between the data being produced in the researcher's primary domain and the odds of the researcher lacking desired knowledge. The results of the logistic regression is shown in Table 4.28.

*Table 4.28 Logistic regression: Knowledge Lacking and Data Produced in Researcher's Primary Domain*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Reuse in primary domain | 0.813 | 0.537 | 0.422 | 1.568 | |
| Number of Citations (ln) | 0.897 | 0.016 | 0.822 | 0.980 | ** |
| Data purpose | 0.871 | 0.486 | 0.591 | 1.284 | |
| Data source | 0.684 | 0.044 | 0.472 | 0.990 | ** |
| Constant | 0.953 | 0.922 | 0.365 | 2.492 | |
| Number of obs | 611 | | | | |

*** p<.01, ** p<.05*

#### *4.3.1.4.5 Number of Reuse Projects and Papers, and Substantially Similar Projects and Papers*

I examined the relationship between lacking desired knowledge and variables relating to the number of reuse projects or publications the researcher had worked on that involved data reuse, and the number of reuse projects or publications where considerations were substantially similar to those in the project and paper of interest in my research. Once again, I did not find any associations that were significant enough for me to reject the null hypothesis that none of these were associated with lacking desired knowledge. The results of the logistic regressions I ran to test the hypotheses is given in Table 4.29. The p-values for the regressions were as follows: number of reuse projects (p = 0.271); number of reuse projects where considerations were substantially similar (p = 0.134); number of publications (p = 0.551); number of publications where considerations were substantially similar (p = 0.363).

*Table 4.29 Estimated Association Between Numbers of Reuse Projects and Publications and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | N | Sig P>[z] |
|---|---|---|---|---|---|---|
| Num reuse projects | 0.943 | 0.271 | 0.850 | 1.047 | 658 | |
| Num similar reuse projects | 0.928 | 0.134 | 0.842 | 1.023 | 632 | |
| Num pubs | 0.971 | 0.551 | 0.882 | 1.069 | 656 | |
| Num similar pubs | 0.958 | 0.363 | 0.872 | 1.051 | 631 | |

*** p<.01, ** p<.05

#### *4.3.1.5 H8. The Greater a Researcher's Ability to Reuse Data, the Lower the Probability of Knowledge Satisficing*

I did not find that reuse ability, either within the researcher's primary domain or outside their primary domain (p = .128 within and p = .984 without) had a significant effect on the odds of lacking desired knowledge about data. The results of logistic regressions are given in Table 4.30.

*Table 4.30 Estimated Association Between Reuse Ability Within Primary Domain of Research and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | N | Sig P>[z] |
|---|---|---|---|---|---|---|
| Reuse ability in domain | 0.894 | 0.128 | 0.774 | 1.033 | 652 | |
| Reuse ability out of domain | 1.002 | 0.984 | 0.861 | 1.165 | 645 | |

**** p<.01, ** p<.05*

#### *4.3.1.6 H9. There Is No Association Between Researcher's Motivation for Reusing Data and Knowledge Satisficing*

Table 4.31 shows the results of logistic regressions I used to test the null hypothesis that there is not a significant association between a researcher's purpose(s) for reusing data and the odds of lacking desired knowledge about data. For these regressions, I counted only cases where researchers indicated that the reason they reused the data was important to their research. For instance, if they indicated that they reused the data to compare with other data but this purpose was not important to them, I did not include that purpose for that researcher in the regression.

*Table 4.31 Estimated Association Between Purpose of Reuse and the Odds of Lacking Desired Knowledge*

| Knowledge Lacking | Odds ratio | P>[z] | [95% Conf | Interval] | N | Sig P>[z] |
|---|---|---|---|---|---|---|
| Background | 1.406 | 0.098 | 0.939 | 2.106 | 693 | |
| Validate | 1.382 | 0.120 | 0.919 | 2.079 | 693 | |
| New question | 1.264 | 0.300 | 0.811 | 1.970 | 693 | |
| Replicate | 1.930 | 0.048 | 1.006 | 3.701 | 693 | ** |
| Combine | 1.456 | 0.049 | 1.002 | 2.118 | 693 | ** |
| Compare | 1.440 | 0.067 | 0.975 | 2.129 | 693 | |
| Theory | 1.531 | 0.033 | 1.035 | 2.264 | 693 | ** |
| Tool | 1.054 | 0.871 | 0.557 | 1.994 | 693 | |
| Other | 0.499 | 0.157 | 0.190 | 1.307 | 693 | |

**** p<.01, ** p<.05*

There were three purposes where I found a significant association between the purpose of reuse and the odds of lacking desired knowledge about data and was able to reject the null hypothesis. These were purposes to replicate or reproduce research results, combine the reused data with other data, and test a theory. In particular:

- The odds of lacking desired knowledge if using data for replication purposes was 1.930 times the odds of lacking knowledge if not.
- The odds of lacking desired knowledge if using data to combine with other data was 1.456 times the odds of lacking knowledge if not.
- The odds of lacking desired knowledge if using data to test a theory was 1.531 times the odds of lacking knowledge if not.

For the other purposes, my data did show any evidence of an association.

#### *4.3.1.7 H10. When Research Questions Change or Develop Over Time There is a Lower Probability of Knowledge Satisficing.*

I included two elements in my operationalization of the researcher's research process in my survey. One was how researchers arrive at their research questions and the other was the source(s) from which they obtain knowledge about data. I only formulated a hypothesis related to research question development, but include results related to the source of knowledge in this section as well.

##### *4.3.1.7.1 Research Question Development*

I tested the null hypotheses that there is not a significant association between a researcher's process for developing research questions and the researcher lacking desired knowledge about data. The test is based on questions where I asked researchers their level of agreement with four conditions when deciding to use the data: already having their research questions in mind, changing their questions as a result of knowledge obtained about the data, determining their questions as they explored the data, or having an "other" method of developing their research questions. I did not code the "other" responses but many of them had to do with the researcher having prior knowledge of the data study when formulating questions or their process being a combination of the other three conditions. From responses to these questions, I also generated a binary variable describing whether the researcher's questions were "fixed" or "fluid."

Questions being "fixed" meant the researcher completely agreed with having their questions in mind and completely disagreed with the other conditions. Being "fluid" encompassed the researcher having a degree of agreement less than "Completely agree" for having research questions in mind or any degree of agreement with having their research questions in mind and any levels of agreement other than "Completely disagree" for the other conditions (e.g., if a researcher completely agreed that that they had their researcher questions in mind, but somewhat agreed that their questions developed as they reused data, I counted their research question development process as "fluid"). Table 4.32 illustrates these rules. I did not incorporate "other" responses into the new "fixed" variable. The results of the regressions are presented in Table 4.33.

*Table 4.32 Rules for Determining Whether Research Questions Were Fixed or Fluid*

| | |
|---|---|
| Fixed | RQ in mind = Completely agree AND<br>As explore = Completely disagree AND<br>RQ change = Completely disagree |
| Fluid | RQ in mind = Completely agree &<br>As explore > Completely disagree OR<br>RQ change > Completely disagree<br><br>OR<br><br>RQ in mind < Completely agree |

As Table 4.33 shows, there were significant associations between a person changing their research questions and developing their research questions as they explored the data and lacking desired knowledge. In particular, the odds of lacking knowledge increased 1.397 times ($p < 0.001$) for each higher level of agreement with changing research questions (e.g., somewhat disagree compared to completely disagree, somewhat agree compared to somewhat disagree, etc.). The odds of lacking knowledge increased 1.207 times ($p = .002$) for each higher level of agreement with developing the research question as the researcher explored the data.

In addition, there was a significant association between a person's research questions being fixed or fluid and lacking desired knowledge: the odds of lacking knowledge were .332 times higher (that is, 0.668 times lower, $p < .001$) when a person had fixed research questions as compared to when their research questions were fluid. Put another way, the odds of not lacking knowledge were about three times higher if the person's research question was fixed prior to deciding to reuse the data.

*Table 4.33 Estimated Association Between Process for Developing Research Questions and the Odds of Lacking Desired Knowledge*

| Knowledge Lacking | Odds ratio | P>[z] | [95% Conf | Interval] | N | Sig P>[z] |
|---|---|---|---|---|---|---|
| RQ in mind | 0.892 | 0.272 | 0.728 | 1.094 | 687 | |
| RQ changed | 1.397 | 0.001 | 1.231 | 1.585 | 684 | *** |
| RQ dev as explore | 1.207 | 0.002 | 1.073 | 1.358 | 686 | *** |
| RQ other | 1.014 | 0.930 | 0.736 | 1.398 | 71 | |
| RQ fixed | 0.332 | 0.001 | 0.197 | 0.557 | 693 | *** |

*** *p<.01, ** p<.05*

#### *4.3.1.7.2 Source of Knowledge*

I only asked the source of the knowledge that was lacking from researchers who indicated they lacked desired knowledge. This sub-group was not large enough to conduct inferential statistics. However, I provide descriptive statistics about knowledge lacking and sources of knowledge in section 4.2.3.4.

### *4.3.2 Model of Satisficing*

In the course of testing the preceding hypotheses, I performed univariate analysis of individual variables (using the control variables) to test for significance. I subsequently attempted to build a model for lacking desired knowledge about data based on these results. Hosmer and Lemeshow (1989) recommend selecting a subset of variables for inclusion in an overall model based on the results of the univariate analyses. They indicate that a more parsimonious model is more numerically stable and easier to generalize (Hosmer and Lemeshow,

1989, pp. 82-83). They also recommend including, at least initially, variables that have a significance of at least 0.25 to observe how their significance is expressed in a model with the more significant variables (Hosmer and Lemeshow, 1989, p. 86). Below, I test models for knowledge lacking including variables with significance of $p < .25$ and $p < .05$.

Table 4.34 replicates Table 3.9 but adds columns for whether the variable was significant in the statistical analysis and whether I selected the variable to include in the statistical model, either at the $p < .25$ or $P < .05$ level. I included in the model the three control variables: (a) number of citations (ln) (log of number of citations to reused data), (b) data purpose (whether the data were created to be reused), and (c) data source (whether the reused data were obtained from ICPSR). As in my analyses above, in each regression I also accounted for clustering effects resulting from multiple researchers reusing the same data.

The dependent variable was the binarized variable I created to indicate whether researchers lacked desired knowledge or did not according to my definition. The independent variables I included at the $p < .25$ level were

- original knowledge (knowledge on the research team about the creation of the original data)
- other 1 (whether the researcher identified with one of the less common domains, defined as those I coded 10 or more times; I coded 16 domains this way, and 40 domains fewer than 10 times)
- years of experience reusing data produced in any domain
- number of reuse projects with reuse considerations substantially similar to the ones involved in reusing these data
- perceived ability to reuse data produced in the researcher's primary domain of research
- reuse purpose
    - background
    - validate
    - replicate
    - compare
    - combine
    - theory
    - other

- RQ fixed (whether or not the research questions were fixed when the researcher decided to reuse the data)
- whether researchers lacked knowledge when considering reusing data (as opposed to when deciding to reuse data)

The inclusion of this last deserves special mention. I asked in my survey whether researchers lacked knowledge at the time they were considering reusing data to better understand the responses of researchers who indicated they did not lack knowledge at the time the decided to reuse data (see section 4.2.1). I obtained evidence from this that supported a conclusion that researchers work hard to choose data ahead of time that will fulfill their reuse aims.

However, I found a significant relationship between lacking desired knowledge at the time when researchers were considering reusing data and the odds of lacking desired knowledge at the time of decision. In particular, in a regression including the control variables (see Table 4.35), I found that the odds of lacking desired knowledge at the time of decision were 3.3 times higher ($p < .001$) if the researcher lacked knowledge at the time they considered using the data. Lacking desired knowledge at the time of consideration was not one of the main predictors I planned to investigate and I had not formulated a hypothesis about it. However, since the relationship was significant, I included lacking desired knowledge at the time of consideration in the final models both to control for its effect and to construct a model that accounted for as much of the variance in lacking knowledge at the time of decision as possible.

The first three shaded rows in Table 4.34 indicate the control variables, which I included in both models. The next five shaded rows indicate the variables that were significant at levels of both $p < 0.25$ and $p < 0.05$, and which I included in the final model ($p < 0.05$).

*Table 4.34 Concept, concept operationalization, significance, and model inclusion status for variables investigated in the survey*

| **Independent Variables** | | | | |
|---|---|---|---|---|
| **Concept** | **Concept operationalization** | **Significant (p<.25)** | **Significant (p<.05)** | **Inclusion in model (p<.05)** |
| Information about researcher's perspective | Whether data were reused | NA | NA | Prerequisite |
| | Researcher's role | No | No | Do not include |
| | Number of people involved in determining whether the data were sufficient to reuse | (Not investigated) | (Not investigated) | Do not include |
| | Researcher's involvement in determining the goals of the research | No | No | Do not include |
| | Data obtained from ICPSR or another source | Yes | No | Include (control) |
| Data quantitative or qualitative | Information obtained from ICPSR Bibliography | No | No | Do not include |
| Number of citations | Information obtained from ICPSR Bibliography | Yes | No | Include (control) |
| Researcher "distance" from the data | Researcher's perception of original data creation purpose | Yes | No | Include (control) |
| | Researcher's involvement in the original research | No | No | Do not include |
| | Knowledge of research team about the way the original data were collected or produced | Yes | Yes | Include<br>Odds of lacking knowledge decrease as knowledge of research team increases |
| Experience with data reuse | Researcher's unit or departmental affiliation | Not investigated | Not investigated | Do not include |
| | Researcher's professional position | No | No | Do not include |
| | Researcher's primary domain of research | Yes | Yes | Include<br>Odds of lacking knowledge increased when researchers identified with a domain outside of those that were most common |
| | Whether the research conducted was in the researcher's primary domain of research | No | No | Do not include |
| | Researcher's years of experience in primary domain of research | No | No | Do not include |
| | Domain of research in which the reused data were produced | Not investigated | Not investigated | Do not include |
| | Researcher's years of experience reusing data produced in this domain | No | No | Do not include |

| | | | | |
|---|---|---|---|---|
| | Researcher's years of experience reusing data produced in any domain | Yes | No | Do not include |
| | How often researcher's work involved reuse | No | No | Do not include |
| | How often researcher's work involved reuse for the same purpose as these data | Only significant for compare<br>Odds of lacking knowledge increase with frequency of reuse to compare (159 cases) | Only significant for replication | Do not include (significantly reduces the sample size in the regression) |
| | Number of reuse projects researcher worked on | No | No | Do not include |
| | Number of reuse projects where reuse considerations were substantially similar to ones involved in reusing these data | Yes | No | Do not include |
| | Number of authored or coauthored published papers describing research involving data reuse | No | No | Do not include |
| | Number of authored or coauthored published papers describing research involving data reuse where conditions were substantially similar to those involved in reusing these data | No | No | Do not include |
| Reuse ability | Researcher's perceived ability reusing data produced in primary domain of research | Yes | No | Do not include |
| | Researcher's perceived ability reusing data produced outside primary domain of research | No | No | Do not include |
| Reuse motivation | Reuse purpose(s) | Yes<br>Replicate<br>Combine<br>Theory<br>Other | Yes<br>Replicate<br>Combine<br>Theory | Include Replicate, Combine, and Theory. Odds of lacking knowledge increases when used for these purposes. |
| | Relative importance of each reuse purpose | NA | NA | Prerequisite (Important) |
| Research process | How research questions were developed | Yes | Yes | Include<br>Odds of lacking knowledge decrease with fixed RQ |
| | Information source(s) most important in obtaining desired knowledge | Descriptive statistics only | Descriptive statistics only | Do not include |
| Satisficing (Satisficing is also an independent variable) | Whether, when considering reusing the data, there was additional knowledge about data that was desired | Yes | Yes | Include (I included it in the survey to better understand responses about lacking knowledge |

| | | | | |
|---|---|---|---|---|
| | | | | at time of decision. However, the relationship was significant so I include it to control for its effect and explain a greater proportion of variance.) |
| | Whether, when the decision to reuse the data was made, there was knowledge about the data that was desired but not obtained or obtained only in part | NA | NA | Dependent variable here |
| | The three most important things the researcher would have liked to know about the data but was not able to obtain, or obtain to the desired degree. | Descriptive statistics only | Descriptive statistics only | Do not include |
| | How much of each type of desired knowledge was obtained | Descriptive statistics only | Descriptive statistics only | Do not include |
| | The importance of the knowledge that was not obtained or obtained only in part to deciding to reuse the data | Descriptive statistics only | Descriptive statistics only | Do not include |

In Table 4.35, I show all of the variables with significance < .25 when included individually in a logistic regression with the control variables and with desired knowledge lacking as the dependent variable. I include the control variables in the table, with the odds ratio, significance, and other values when all of the controls are included in the same regression (i.e., a logistic regression using all of the controls at the same time). I use only the variable for whether or not the research question was fixed (RQ fixed) in place of the multiple variables I measured related to fixed or fluid research questions to simplify the model and because the generated variable seems to carry the significance of the association with lacking desired knowledge.

*Table 4.35 Variables With P-value of 0.25 or Lower in Univariate Analyses Conducted with Control Variables*

| | Odds ratio | P>[z] | [95% Conf | Interval] | N | Sig |
|---|---|---|---|---|---|---|
| Number of Citations (ln) | 0.919 | 0.085 | 2.651 | 8.406 | 693 | |
| Data purpose | 0.779 | 0.180 | 2.815 | 9.621 | 693 | |
| Data source | 0.729 | 0.080 | 1.209 | 3.695 | 693 | |
| Original knowledge | 0.724 | 0.001 | 0.619 | 0.848 | 692 | *** |
| Lack of desired knowledge (considering) | 3.312 | 0.001 | 2.428 | 4.518 | 692 | *** |
| less common domains 1 | 2.296 | 0.001 | 1.427 | 3.696 | 693 | *** |
| Reuse experience any | 0.986 | 0.223 | 0.972 | 1.012 | 611 | |
| Num similar reuse projects | 0.928 | 0.134 | 0.842 | 1.023 | 632 | |
| Reuse ability in | 0.894 | 0.128 | 0.774 | 1.033 | 652 | |
| Background | 1.406 | 0.098 | 0.939 | 2.106 | 693 | |
| Validate | 1.382 | 0.120 | 0.919 | 2.079 | 693 | |
| Replicate | 1.930 | 0.048 | 1.006 | 3.701 | 693 | ** |
| Combine | 1.456 | 0.049 | 1.002 | 2.118 | 693 | ** |
| Compare | 1.440 | 0.067 | 0.975 | 2.129 | 693 | |
| Theory | 1.531 | 0.033 | 1.035 | 2.264 | 693 | ** |
| Other | 0.499 | 0.157 | 0.190 | 1.307 | 693 | |
| RQ fixed | 0.332 | 0.001 | 0.197 | 0.557 | 693 | *** |

**** p<.01, ** p<.05*

The results of the regression with all variables from Table 4.35 included are shown in Table 4.36. The number of observations in this regression was 579. When included in the same model, the only variables significantly associated with lacking desired knowledge at a level of $p < .05$ were the amount of knowledge on the research team about the original research ($p = .001$), whether researchers lacked knowledge about data at the time they considered reusing data ($p < .001$), and whether the research question was fixed at the time the researcher decided to reuse the data ($p = .004$).

*Table 4.36 Estimated Associations Between Variables from Univariate Analysis of Variables With p < .25 and Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Number of Citations (ln) | 0.912 | 0.095 | 0.818 | 1.016 | |
| Data purpose | 0.935 | 0.773 | 0.591 | 1.479 | |
| Data source | 0.896 | 0.639 | 0.567 | 1.417 | |
| Original knowledge | 0.707 | 0.001 | 0.575 | 0.869 | *** |
| Lack of desired knowledge (considering) | 3.733 | 0.001 | 2.522 | 5.525 | *** |
| less common domains 1 | 0.643 | 0.317 | 0.271 | 1.527 | |
| Reuse experience any | 0.974 | 0.105 | 0.943 | 1.006 | |
| Num similar reuse projects | 0.966 | 0.614 | 0.844 | 1.106 | |
| Reuse ability in | 1.029 | 0.801 | 0.825 | 1.283 | |
| Background | 1.186 | 0.542 | 0.685 | 2.052 | |
| Validate | 0.845 | 0.579 | 0.466 | 1.532 | |
| Replicate | 2.313 | 0.069 | 0.938 | 5.702 | |
| Combine | 1.324 | 0.273 | 0.801 | 2.190 | |
| Compare | 1.317 | 0.321 | 0.765 | 2.268 | |
| Theory | 1.434 | 0.151 | 0.877 | 2.345 | |
| Other | 0.837 | 0.768 | 0.256 | 2.739 | |
| RQ fixed | 0.396 | 0.004 | 0.212 | 0.740 | *** |
| Constant | 0.751 | 0.663 | 0.207 | 2.726 | |
| Number of obs | 579 | | | | |

**** p<.01, ** p<.05*

I next performed a logistic regression with lacking desired knowledge as the dependent variable and variables with a significance level of $p < .05$ from the univariate analysis as independent variables (i.e., the variables shaded in gray in Table 4.34) and the control variables. The results of the regression are shown in Table 4.37. This model included 691 observations. As the table indicates, in this model, whether the researcher worked in a less common domain and reuse for particular purposes (replication/reproduction and theory testing) were significant at the $p < .05$ level. The amount of knowledge on the research team about the original research, whether the researcher lacked desired at the time they were considering reusing data, and whether the research question was fixed at the time the researcher decided to reuse the data were significant at the $p < .01$ level.

*Table 4.37 Estimated Association Between All Variables With a P-value of 0.05 or Lower and the Odds of Lacking Desired Knowledge*

| Lack of desired knowledge | Odds ratio | P>[z] | [95% Conf | Interval] | Sig |
|---|---|---|---|---|---|
| Number of Citations (ln) | 0.929 | 0.150 | 0.840 | 1.027 | |
| Data purpose | 0.732 | 0.120 | 0.494 | 1.084 | |
| Data source | 0.833 | 0.372 | 0.558 | 1.244 | |
| Original knowledge | 0.740 | 0.001 | 0.623 | 0.880 | *** |
| Lack of desired knowledge (considering) | 2.921 | 0.001 | 2.120 | 4.025 | *** |
| less common domains 1 | 1.879 | 0.017 | 1.118 | 3.157 | ** |
| Replicate | 2.162 | 0.036 | 1.050 | 4.450 | ** |
| Combine | 1.418 | 0.097 | 0.939 | 2.139 | |
| Theory | 1.534 | 0.048 | 1.004 | 2.344 | ** |
| RQ fixed | 0.385 | 0.001 | 0.222 | 0.666 | *** |
| Constant | 0.775 | 0.615 | 0.288 | 2.088 | |
| Number of obs | 691 | | | | |

**** p<.01, ** p<.05*

I used the Quasi-likelihood under Independence Model Criterion (QIC) to compare the two models. I found that the model including variables significant at the $p < .25$ level (Table 4.36) had a lower QIC value than the model with variables significant at the $p < .05$ level (636.86 compared to 772.94) and was thus the best working correlation structure or preferred model (Cui, 2007).

In this model, when researchers lacked knowledge at the time they were considering reusing data, they were almost four times more likely (3.73) to lack knowledge at the time they decided to reuse data. Each additional level of knowledge about the original research was associated with a 0.71 times increase (i.e., 0.29 times decrease) in the odds of lacking desired knowledge. Having research questions fixed at the time of the decision to reuse the data was associated with a 0.40 times increase (i.e., 0.60 times decrease) in the odds of lacking desired knowledge.

Working in a research domain outside of the 16 most common and reusing data for either purposes of replication or reproduction, or to test a theory were not significant in the model of variables significant at $p < .25$. However, their significance at the $p < .05$ level makes them of

possible interest in future research, particularly research investigating interactions between variables.

### *4.3.3 Research Question 1c. What Do Researchers Report Influences Determinations About How Much of What Kinds of Knowledge Are Sufficient? Findings From the Interviews*

Through the survey results, I found that those that had a significant effect on the odds of lacking desired knowledge about data were (a) the knowledge that researchers or others on their research team had about the original creation of the data, (b) whether researchers lacked knowledge they desired at the time of considering to reuse data, and (c) the process of developing research questions.

In this section I report findings from the interviews about the factors that influenced researchers in determining how much and what kinds of knowledge were sufficient to decide to reuse data. This is not the question of whether researchers lacked knowledge they desired (a question that originates more from Simon's perspective on adaptation and satisficing), but rather what influenced researchers to bound the knowledge they obtained about data the way they did (a question that originates from Haraway's perspective of situated knowledges and understanding objects of scientific study as boundary projects).

The findings presented in this section speak to the broader context of researchers' decisions to reuse data (including when there was knowledge they wanted about data but could not obtain). In particular, I report findings about what motivated researchers to use the data they selected, how data reuse is perceived in their immediate community of research, and what pressures or considerations influenced how the research proceeded. These findings provide important background and context for understanding the specific strategies researchers reported for determining that they had sufficient knowledge about the data to reuse them in their research, which I report on in section 4.4. In section 4.3.3.1, before proceeding to the findings, I provide

additional details about the relevance of this background and context to understanding researchers' achievement of reuse equilibrium and discuss potential implications of findings in this area for the design of environments that support secondary reuse.

#### *4.3.3.1 Relevance of Research Background and Context to Reuse Equilibrium and Support of Secondary Reuse*

In section 1.4.6, I described the enterprise of research in the context of Simon's framework for adaptive systems. In particular, I proposed that research is a strategy we use in light of limitations to our attention, memory, and other features of our inner environments, to understand and adapt to outer environments characterized by uncertainty and complexity (e.g., we might conduct research to better understand our history or how to work toward more fair and functional societies). In this understanding, it is possible to conceive of situations where research that is performed is not well adapted to a researcher's goals. This could occur, for instance, if instruments or methodology were not formulated to measure phenomena accurately or consistently. It could also occur if aspects of the systems designed to enable and support research did not sufficiently insulate the inner environment (the researcher or researcher team) from pressures of the outer environment (for instance, pressures of time, funding, access, perceptions of data reuse) that inhibit or limit research.

In these cases, the methodology or instruments may be well-adapted, but imperfect measurement could occur because researchers did not have access to all the data they would like (including because the data were not collected), all of the support they needed (for instance, support in statistical methods and analysis), or all of the knowledge and connections they required to influence the way primary data were collected. In other words, something about the outer environment—whose effects the systems that support research were not well-designed to handle—limited or inhibited the attainment of their research goals.

When discussing the application of ideas of "configuration" (Grint and Woolgar, 1997) to reuse equilibrium, I proposed that examples of elements of an inner environment in the context of research would be factors internal to the researcher or research team, such as their skills, knowledge, attitudes, beliefs, or assumptions. I proposed that elements of the outer environment would include the research culture or traditions (that influenced the researcher or research team), disciplinary standards for metadata or documentation creation, the status accorded to researchers, publications, or repositories that might be consulted in the process of data reuse, or pressures that affect the time, staffing, or funding available for research

My theoretical framework led me to see the lines between the inner and outer environment as blurring into one another based on an understanding of situated knowledge that is not about dichotomies (i.e., in my case not about articulating specific and distinct properties of inner or outer environments) but about resonances and interactions between environments. In line with this theoretical framework, I set out to develop a research methodology that was capable of measuring aspects of both the inner and outer environments (as opposed to one or the other) and their impact on research with secondary data. I wanted to measure aspects of the inner and outer environments that had an impact on knowledge researchers obtain, in particular. This methodology is reflected in the loose definition of "knowledge" I used in order to learn what researchers wanted to know about data rather than particular knowledge I might have in mind. It is also reflected in my questions to understand what elements of researchers' environments (both inner and outer) might have had a bearing on their attainment of knowledge about data (the subject of this section).

It is difficult, based on the evidence I collected in the interviews, to evaluate the relative force of influence or depth of impact of different considerations and pressures on a researcher's

research. However, it is possible to identify patterns in what the researchers mentioned and described and add those to a larger understanding of the environment in which they conducted their research and that affected the attainment of knowledge they desired about their phenomena of interest. I highlight these patterns below. The findings I report do not describe how researchers reached reuse equilibrium, but rather give an idea of the environment in which reuse equilibrium was achieved.

In the four sub-sections below, I describe findings related to 1) factors that motivated researchers to reuse particular data in their research, 2) perceptions of reuse in researchers' immediate communities, 3) researchers' own perceptions and understandings of reuse, and 4) considerations and pressures that influenced the way researchers' reuse projects developed. These motivations, perceptions, considerations, and pressures reveal aspects of the ecosystem in which researchers made decisions about how much knowledge about data was enough to reuse them. These aspects, in turn, and the effects they had on researchers' decisions, provide insights into ways the design of the research ecosystem might be improved in order to better support research with secondary data. Gaining such insights are the ultimate aim of my research.

#### *4.3.3.2 Motivations for Reusing Data*

Researchers reported a variety of motivations for using the data they did, which I categorized as individual, comparative, strategic, and social. Individual motivations were motivations of the researcher themselves (apart from others) and by far, the most commonly reported of these was prior experience with the data. In many cases, this experience came from work done as a research assistant while doing a PhD (D4-02, D1-02, D1-03, D3-02, D2-04) but researchers also had experience collecting the original data (D-09, D-10), had been using the data in their research over a long period of time (D-04, D2-01) or had a relationship with the study in

some other way, such as providing consulting services (D-08). The quotes below illustrate some of these scenarios:

> This was a question I was really interested in learning about. So [the data study] was good because I knew the data already from my previous assignments as a research and teaching assistant. (D1-02)

> [My reuse of these data] goes back to when I first started publishing in graduate school. I have really leaned on that dataset because it is so extensive and so useful in so many different ways. (D4-02)

> So actually I participated in the collection of the data posted on the repository… [A]fter graduating from my graduate school, I wanted to continue to work on the same data set because they're having lots of things to be dealt into, which couldn't be published in the previous articles. (D-09)

One researcher mentioned being "plugged in" to a dataset from the time of its creation because of their interest in the data, attending workshops surrounding the data and having significant interaction with the project PIs (D2-01). Another researcher described the data as their "home turf" when I asked if they had considered other data to reuse (D-10):

> Well, basically most of my work over the past [several] years has come from either [one data study] or [another]…I work with these data…I wouldn't rule [working with other data] out if I had a very specific question about something that in the future I'd also use other data, but for me it's kind of the home turf. (D-10)

In the category of comparative, I placed motivations that had to do with the data being better in some way than other available options. Some referred to the data as the "clear" or "obvious" choice (D-11, D-13) because no other data studies had a comparable scope or universe of cases. Researcher D-11commented:

> why that data appealed to me relative to any other data that existed is because A, it was already collected: it's the benefits, it's secondary data. B, it's the universe of [the population] and so it's very high quality from my perspective. (D-11)

Others described the data as being the "safest bet" for completing the project given their research questions (D-02) (other data studies would make completion less certain), including

variables that had been understudied (D-09), or having a larger sample than others or supporting more generalizable results (D1-02, D1-04, D-12, D-07). As researcher D1-04 explained:

> [The data study] is one of only a handful of studies that take an in-depth look at [this population] and also include the kinds of things…we're interested in…[A] lot of times you have [data] with [specific populations], but you don't have the kinds of measures that were included in [this study].

One researcher chose their data in part because they were interested in doing a comparative analysis and other potential choices had strong conceptual underpinnings that made them more difficult to use for comparison purposes:

> I think the main issue is that you develop a study with a certain framework and that framework has a conceptual underpinning, that also drive the development of item material…I can tell you some of the background questions would, of course, be comparable, but hardly anything else because they have different topics they address. (D-10)

There were some researchers who were influenced in their choice of data by strategic considerations, in that the use of data fulfilled several objectives (e.g., research, training, etc.). For instance, one researcher selected the data for their research because it would be a good learning opportunity for the graduate student working on the project (D-07). Another chose their data because the choice would show others evaluating their work some variety in the data they reused (D-12). This researcher chose their data also to bolster perceptions of validity in their research. As this researcher noted,

> I felt like using a large data set as my example. Even though it's not necessarily logically related to the validity of what I'm presenting as a heuristic, it helps to bolster that. People believe big data, probably rightly so, more than little data." (D-12)

In the social category, I placed motivations for reusing the data that arose from conditions outside the researcher's control. These included data already being available (a quality of all secondary data some called out explicitly)(D2-04, D-03, D-09, D-11), the data being offered as an option by the researcher's advisor (D3-02); and data being "notoriously available and usable,"

because they were mandated to be made available and there was strong centralized support to do this (D-02).

Overall, the main reasons I found that researchers selected their data were prior experience they had with the data, the data offering advantages over other available data, reuse of the data offering second order benefits (such as showcasing skills or bolstering perceptions of validity), and the convenience of reuse.

#### *4.3.3.3 Perceptions of Data Reuse*

For the majority of researchers I interviewed, reuse was common or accepted in their immediate community. Whether it was common or not, however, there were nuances to the stories they told that are revealing about the place secondary data held in the larger research ecosystem. Many researchers noted a perception in their community of primary data collection as the ideal and secondary reuse as something that occurred in situations where resources to conduct primary research were limited. One researcher said that reuse was the most common mode of research in their field, but researchers would prefer otherwise:

> Well, I think that people would like to gather their own data but in quantitative research, it is expensive and time-consuming. It's not qualitative that you go for 20 people. So, it's not that easy. Otherwise, I think that quantitative researchers would love to collect their own data because then you have the freedom to choose your own questions and according to your hypothesis research question, so you can do that. But the problem is that time, money, and manpower. (D-01)

Researcher D1-03 described a similar situation, noting at the same time that secondary reuse was highly valued in their community:

> So, we see it as a valuable resource because sometimes you may not have the financial resources to conduct primary data collection, but there may be data sets out there that you may be able to utilize to answer parts of your research question, maybe not all, but there might be opportunities to answer parts of your research questions. (D1-03)

Researcher D-09 gave a specific example of how their decision to reuse data came about, illustrating how situations where there are deficits in resources can lead to or necessitate data

reuse. They were a faculty member at a research university but because their department was not one of the "golden children" of the university (departments that were well-funded), they did not have access to money to conduct primary research. They also did not have access to an adequately sized pool of students in their own department:

> I am a faculty member in [department], which is a poorer than golden children of the university, [like engineering and others]. So I couldn't have enough money to recruit the lots of participants at the moment. And then now we have kind of subject pool in our own department, but that time I couldn't get access to a huge size of subject pool consisting of students. So I had to deal with those limitations. (D-09)

Another researcher had observations similar to D-01 and said that they were noticing a split between some researchers who were trying to collect data efficiently online to meet their data collection goals, and others who were using secondary data:

> I think most people that I know of are using secondary data sets. One reason why is it is hard to collect your own data. So my colleagues and friends, they go for ICPSR website to download available data sets or they are working with local agencies if they are in [research area] but it's also secondary. I think there is a growing movement to use and conduct internet-based surveys. Those people collect their own data but it takes money to do that. I think I'm noticing two branches of data use. One is regarding the Internet-based surveys such as MTurk or YouGov and another is just relying on secondary data set for their research purposes. (D4-01)

One researcher who noted that perceptions of data reuse varied in their community described changing attitudes towards reuse in their field and recognition, in particular, of its value for meeting the demands of publishing in academia and achieving career objectives:

> I think [how data reuse is perceived] really depends…there is a heavy reliance on grant funding to go and get new data…I have changed some minds to a certain extent in terms of my perspective on this and the kind of the things we're talking about right now, in that, if you want to- especially among some my non-tenured colleagues, in terms of being able to say, "You guys are doing fine publishing and stuff." But if you want be sure that you're going be able to get tenure and meet your tenure clock, [secondary use] exists as an additional tool that if you want to—You know, if there exists an existing data set out there that can answer your research question to a certain extent and then you can continue on to answering that research question by getting grant funding after you publish an initial study using that existing data set, then you're basically double-dipping in it. It

> helps you because you're not having to wait to get that money released and collect the data yourself, while still progressing towards your tenure clock essentially. (D4-02)

Another researcher described data reuse being common, but existing in a hierarchy of research, with primary data collection as the ideal:

> Yeah, everyone does it on the hierarchy of quality or prestige, I guess. That there's always, I think, an emphasis on, in my field, primary data collection. So if you can create your own proprietary data, great, but it's not necessary in that it's completely acceptable to use secondary data. We do it all the time. And as long as its quality secondary data, and you know what you're doing, no issues, right? No barrier to publication, but clearly, if I single-handedly was able to create a database of the universe of [type of data], I would get a lot of credit for that because it's virtually impossible for one person to do. (D-11)

This idea of "credit" for research was described by another researcher who had done work in two related fields (D2-03). In one, they said, reusing data was very common and accepted as "standard practice." In the other, there was a premium placed on putting one's own data set together. It could even contribute to getting a "pass" on shortcomings in the research that might otherwise be barriers to publication. However, in this field, they said, it could take years to put together a set of data, only a few researchers might be interested in the publication, and those who were could be reluctant to reuse them. The researcher describes this and the negative (in their view) effect this has on data reuse in particular and the scientific process more generally:

> A [researcher] or team of [researchers] will spend all this time collecting a data set that's novel. They'll publish a few papers really well with it. And then, the sex appeal of the data are gone. And others are hesitant to reuse it because they don't think it's going to publish as well. Whereas a re-examination of the original question of interest would simply just be the scientific process. And that would be a good thing. Or maybe there are other questions that could be asked from these data that I think people are often reluctant to pursue simply because the data set is no longer novel…
>
> But too much emphasis and too much reward is given to being first as opposed to fully vetting a particular hypothesis through numerous studies and years of work. There's just this really big premium on if you're the first person to do something, you're going to publish in the best journal, and that's going help your reputation. That's going get you tenure, et cetera, et cetera. And then that trickles down into being the first to collect this novel data set. You get a huge, huge reward for that. And then the marginal returns to

> using that data set a second, third or fourth time, especially by people who were not part of the original collection team is really low. And I think that's not good. (D2-03)

One researcher explained that they would like to say data reuse was increasing in frequency but lamented that not enough professors and mentors were encouraging students to use secondary data to the extent that they could be utilized. They said it was one of the great problems in the discipline:

> I would like to say that it's catching on more. But I think that people just keep thinking, I have to do my own research myself. That is the idea that when people get educated in graduate school, they think I have to do my own study. And I don't think enough professors, mentors are encouraging people to use the available data sets to the degree that they could be utilized. And I think it's one of the great problems in social science education, is they just, you know, encourage people, you gotta make your own study…And if you don't do it by yourself, it's not as good. But people can make very important worthy research contributions from these archival datasets. I know it in my bones, and if I was teaching today, that's what I would advocate for more: utilization of these very greatly valuable underutilized sources. (D2-01)

Another researcher expressed a similar sentiment when speaking about perceptions of reuse in their community in a more corporate context. They understood the value and importance of primary data collection but believed there were benefits to secondary reuse that were being overlooked, including avoiding replication of prior research, cost-savings, and reduction of the burden on studied populations:

> And, if anything, I think that the field I work in needs to get a little bit more comfortable with the idea of data reuse. It might not be asking the specific population you want, but if it was asked to a slightly different population and then you have population data from, you know, [data study] or the census or something, a lot of times you can get to the answers you want without having to spend millions of dollars and hundreds or thousands of hours of burden on survey respondents to get the information you need. It's already out there. (D-06)

This researcher noted that there was a divide in their community about how reuse was seen:

> There's a real divide between people who say, "It's always gotta be my own survey," versus people who are a little bit more comfortable with, "Well, wait, what can we do with sort of what's already out there?" (D-06)

A third researcher, coming from a non-profit context where reuse was rare but the community was supportive, talked about their own experience reusing data and becoming an advocate for reuse to make the most of the data produced in primary research:

> And so I pulled down a secondary data set…That was my first experience with it without having a grant. I said, "I want to write this paper on [topic]." And I was like, "Wow, I could just conceptualize a study, find a data set and grab it." And two months later, I have a draft of the paper. So I don't get any sense that my colleagues are pro or against or want to do it one way or another. But my community is a community of people who do primary data collection. And they totally support secondary data analysis. And I think about that all the time. I want my grad students to use my past projects because I can write, you know, two papers from a grant, and I'm already working on the next grants, and I've got all this data that can be analyzed in another way. And that's what I tell my students. That's the beauty of secondary data. The PIs, if you're talking about the [non-profit] world, they're not publishing tons of papers like academics from universities are. So, they're getting the grant, they're doing an evaluation. But they're not doing any theoretical work with some of that data. And it just opens up the avenue to use that data. And there are other colleagues like me who believe that there's so many opportunities, especially when it's recent and if someone's posting the recent data. So in my mind, all the people I work with are pro-using and reusing data." (D-07)

Interviewees expressed the view on one hand, then, that secondary use was something that occurred in situations and environments where resources were lacking. But they also noted that there was tremendous value to data reuse that researchers were not taking full advantage of, even to the point of being wasteful and over-burdening study populations. D2-01 identified a lack of encouragement and education in data reuse, as well as perceptions that research with secondary is not "as good," as a problem in the social sciences. Other interviewees pointed to pressures surrounding the need to publish and prestige gained from primary data collection as factors that influenced attitudes towards and practices surrounding data reuse.

One researcher relayed ways that their attitude about reuse had changed over time. They said that data reuse was accepted in their community at the time they did the study as a PhD student (in the last seven years), but they felt a certain stigma around it. Now, on the other hand, they collaborate regularly with others to reuse data. They said,

> some of us have just decided secondary data analysis works really well for being able to publish a lot, but without having all the resources for getting big grants and doing primary data collection. (D2-02)

They described how receiving funding to do secondary analysis made an impact on them and helped to overcome negative feelings they had earlier about data reuse:

> So I think it made me more confident in doing it. And before I did it- I thought that it was cheating or something 'cause you weren't collecting your own data, but now I see it more as there's all this data out there that's not being analyzed. And really, what I can collect when I do my own studies. I'm about to start data collection now, it's not a randomized sample. I don't have the resources to make as strong of a dataset that some of these studies have. So I know that there's benefits to both, but I no longer see it as cheating the system of- like, not proving yourself by collecting your own data. 'Cause I think I felt insecure about that in my Ph.D. program. Are people going to think that I'm not really a scholar 'cause I didn't collect my own data. (D2-02)

Shifting attitudes was a theme that came out in other interviews as well. In one case, the researcher described their advocacy work for transparency and replicability. They said that even recently, there was not really a positive or negative attitude toward reuse in their community because reuse did not happen that often. But in the wake of the replication crisis in psychology, they said, which had ripple effects in their domain of research, changes were happening:

> So before I started my [job], I felt that people didn't have either positive or negative ideas about data re-use and secondary analysis 'cause that didn't happen that frequently in the field. But I introduce these concepts and then provoke the kind of debates or discussions about what would be ways to improve the science in the field…And I think that they are now pretty outgoing minded to, and friendly to data re-use and secondary analysis compared with the past. (D-09)

In another case, a researcher talked about a division in their field between those who were committed to openness and transparency and those who viewed research as a matter of trust (i.e., if there were a desire to see data, it indicated a lack of trust in them as a researcher). When I probed a little more deeply into the contours of the divide in their community—whether it was evenly balanced or how it was shifting—the researcher replied: "It breaks along generational lines [referring to researchers older in age and younger in age] and time marches on?" (D-12)

It is important to note that for some researchers, reusing data was not really a choice. Either the research was possible because others collected the data—as, for instance data related to law enforcement might be (D-02, D-13)—or, because of substantial barriers to reaching the target population, reusing data others collected previously represented a viable strategy for conducting research (D3-02). In these cases, data reuse was seen as a standard practice or necessity.

My findings about perceptions of data reuse in researchers' immediate communities reveal that data reuse, while valued in those communities, is not valued to the extent that some researchers believe it deserves and is on the contrary often seen as less desirable than primary data collection (including by researchers who reuse data). Many researchers would like to conduct primary data collection but are inhibited in doing so because they lack the resources primary data collection requires. While there are indications that attitudes about the value of data reuse are changing (including attitudes held previously by researchers I interviewed themselves), the view that data reuse is not "as good" as primary data collection still prevailed in communities where researchers conducted their research.

In the next section, I report on what researchers revealed—perhaps because of what they found was in their power to accomplish with secondary data—was a primary value of data reuse.

#### *4.3.3.4 Researcher's Own Perceptions of Reuse*

In asking researchers about perceptions of data reuse in their immediate communities, I gained perspectives on how they themselves viewed data reuse. Something I observed across many interviews was researchers' view of data reuse as a steppingstone or springboard to conducting primary research in the future. Evidence of this appears in some of the preceding quotes (for instance, the quote by researcher D4-02 who talked about getting a grant after an

initial study with secondary data, but this finding is worth highlighting because while it follows logically from the perception that use of secondary data is not as valuable as collecting primary data, it also serves to perpetuate this lesser value as a norm.

When discussing the limitations of data reuse for answering questions, researcher D1-03 described data reuse as preliminary work that could lead to primary data collection in the future (either by the researcher themselves or someone else):

> So, you know, while it is challenging to answer your exact [research] question, you might be able to answer something that is close to it without having to invest a lot of resources in primary data collection. I see it as kind of an opportunity to engage in preliminary work, right? That's how I see the secondary data analysis and you engage in preliminary work and then use that preliminary work to then make the argument for new studies that collect this information slightly differently, or you conduct your own primary data collection project. (D1-03)

Another researcher described how reuse of the data—though it had not yielded everything they desired in that research project—had in fact, been a catalyst for their research career:

> So [the data] gave me what I thought it was going to give me. I didn't create the questions, so they weren't as detailed [as I would have liked], but it gave me a starting point for spring boarding my research focus, and my career…Because I used the framework that I'm continuing to use, I set the stage for these different [research areas] I could focus in on more. (D2-02)

In a similar vein, researcher D4-02 talked about coming to understand how limitations to the data could spur future research. Speaking of a shortcoming in the data they reused, they said:

> You don't just wave it away magically, necessarily, in the limitation section, but you do to some extent, to essentially say, "This is a limitation. This is impetus for future research. This study is imperfect." Every study has limitations and you kind of have to just accept that to a certain extent. And like I said, I've grown into that position of understanding the limitations of that data set and understanding that even though it's limited in all of these ways, it really spurs the capacity to ask for grant funding to do that study perfectly, replicate that study using better data essentially. (D4-02)

Echoing sentiments of researcher D1-03 above, researcher D1-05 noted how their research with secondary data opened possibilities for themselves or others to build further on the

questions they explored. It is notable that although this researcher was in an established position and did not conduct the specific research to further their career, the quote, and their reuse of the data, nonetheless demonstrates the usefulness of data reuse as a foundation or steppingstone to further research (which likely would involve primary data collection).

> So there's utility in the fact that it was data collected and there's not a lot of data out there about [topic] for these particular populations. So I think what we did, we added to the literature in terms of providing this knowledge but, now, if someone else is going to continue on in that area, then they'll have this as a foundation of information and then they can continue to build it. Or, if I should one day decide to do this again, I could build it myself. (D1-05)

It is not clear whether data reuse gained value over time as a result of its use as a steppingstone to further research, or its value as a steppingstone resulted from limitations researchers encountered in conducting their research. Either way, my findings revealed that reusing data was often not sufficient to answer the questions researchers sought to answer by using them, and researchers who had limited resources conducted data reuse explicitly as a strategy to put themselves in a position in the future where they were able to conduct primary research.

Placed together with findings in the two preceding sections, my interviews revealed that in the environment in which researchers reused data, they (a) were substantially motivated to reuse data they are already familiar with, (b) frequently operated in contexts in which data was valued, but comparatively less valued or desired than conducted primary research, and (c) a significant value that reusing data offered was to investigate a topic in a way that required fewer resources and was capable of opening opportunities for future investigation (in particular, involving primary data collection).

The following section on considerations and pressures that affected the development of their research provides insights that help to explain these findings.

#### *4.3.3.5 Considerations and Pressures*

Researchers reported a wide variety of considerations and pressures that influenced the way their research developed or proceeded, which I placed into categories of individual, data, and social. Similar to my coding of motivations for reusing data, I considered individual considerations and pressures to be those personal to the researcher—not imposed by others. Considerations and pressures in the data category are those that center around reuse of the data themselves, specifically. The category of social includes considerations and pressures that were outside the researcher's immediate control and arose more as a result of their social environment.

##### *4.3.3.5.1 Individual*

Two researchers reported that the only pressures that existed surrounding their research were ones they imposed on themselves. The first, researcher D1-04) conducted the research in response to a call for a special journal issue, which was voluntary and for which they reported there was plenty of time to prepare. The second desired to conduct the research to present at an upcoming meeting. The researcher said, "In my case it was, in a sense, fun, and had nothing to do with my work" (D-13). It is worth noting in the context of the social considerations below that both of these researchers were rather advanced (had more than 10 years of experience) in their research careers.

The other kind of individual consideration or pressure I encountered involved researchers who chose specific data because of their desire to be transparent about the data they reused and the processes they used to analyze them. For one researcher, the choice was about whether to use public use data or restricted data. They chose public use data so their research could be more easily replicated. Speaking of a conversation with their thesis advisor, the researcher related the following:

> something else we discussed is that if it would be a burden for me [to use restricted data], it would also be a burden for anybody wanting to replicate this research or to do something similar. And so it was important to me that whatever I wrote up or worked with, I really wanted it to be widely replicable and applicable. And so for me, using public use files was important…the fact that any researcher could access them and could look at my stuff and say, "Okay, I'm going to try to do the same thing or something similar," that was a huge consideration in using the data sets I did. (D-06)

Researcher D-09 considered using their own copy of the data (they had been involved in the original research), or a copy that had been deposited in a repository. They chose to reuse the data in the repository to avoid issues in sharing data with others and so the experience using the data would be consistent for anyone who reused the data:

> The issue is that if I use my own private copy of the data file, then it's just somehow tricky to share that data file with others because of the possible copyright issues or confidentiality issues. And also, that's going to prevent further replication because there are lots of customized variables in my own copy, but if I use the repository version of the data set, then everything is consistent, and every conditions are going to set consistent across different users because they're going to download the same version of the file from the same repository. (D-09)

This decision came from the researcher's deeply held belief in the importance of transparency in science:

> I believe that folks other than the authors should be able to replicate the same analysis, for transparency. And then that is kind of a novel value of secondary data analysis compared with the primary data analysis. Once the researcher conducts secondary analysis and wants to publish the results of the study in a journal, then at the minimum, all the codes, analysis codes like articles or et cetera, which is used in data cleaning pre-processing and analysis, should be shared via open repository so that other readers should be able to repeat the analysis. (D-09)

These findings show that researchers own desires and values are expressed in the selection of data they reuse. This is important in light of considerations and pressures that originate from other sources, as discussed below.

#### *4.3.3.5.2 Data*

Significant among the considerations and pressures researchers described that centered around the data themselves were the time and difficulty involved in data reuse (D1-03, D1-04,

D1-05, D2-01, D2-02, D2-04, D-02, D-05). Researcher D1-05 talked about this while relaying their concern about reusing the data properly:

> I think the biggest concern was making sure I would do [the analysis] correctly and I wouldn't mess it up. But they're large data sets, and so you have to be really specific and know your analysis and to make sure that you're doing it correctly. There's a lot to learn to just getting into the process of using the data, (D1-05)

Researcher D1-03 concurred with this assessment, mentioning, as a result, the need to weigh the pros and cons of reusing a particular dataset:

> [Data reuse] is a very long process and it's not something that you take lightly, especially if you're using it for your dissertation, and you have to weigh the pros and cons. And if you could resolve the cons in other ways, whether that's measuring things differently or…[resolving] issues statistically…then there's always a way to move forward. If not, there's another dataset out there that you could possibly use. (D1-03)

Another researcher explained time to reuse data as a major consideration, pointing out the need to balance the possible benefits of reusing data with the large amount of work involved. Their decision to reuse data was influenced by the broader "publish or perish" framework in which their research took place.

> it's a real investment in time to pick up new data. So, for me, personally, like, when I'm looking at using secondary data, one of the big questions I ask is, "How many projects am I going to get out of this?" We're on a little bit of a publish or perish model. As a tenure track professor, I have to use my time wisely and so there's a question of, I'm going to go through the hassle of getting these data, learning them, recoding, cleaning them, et cetera. It's a big-time investment in order to just get it to be able to use a project that might potentially not work. (D-02)

Another researcher spoke about the factors involved in deciding to reuse data in this same "publish or perish" framework, referencing the "sunk cost fallacy" to illustrate the investment they had in using the data they had originally selected because of the time required to reuse data. In this case, the researcher was pushed toward the "fallacy" (i.e., not switching to other data although that may have been a more "rational" option) by the pressure they experienced to publish:

> There's a term in economics, the "sunk cost fallacy." I have devoted lot of effort and time and energy into this particular dataset that I wasn't willing to just jettison everything, and I really wanted to be able to get a publishable manuscript out of that. And like I said, it's really the measure of a graduate student's success, even at a graduate school level before you actually get any credentials, whether or not you have published papers at that stage…It's not just enough that you're published, but which outlet did you publish in. And, of course, there's this whole hierarchy, like, if you publish in a top-tier, or in a lower-tier publication. Looking back, it's not very healthy. (D-05)

Other researchers, while they did not use the term "sunk cost fallacy" in the description of a decision to stay with data because of the difficulties of switching to new data, discussed this phenomenon as well (D1-03, D2-01, D-02). The researcher D2-01 talked about a process where they might start with a specific interest and data with which they were familiar, but if they were not able to explore their interest to the depth they would like, they would consider using different data. However, they described deliberations that would be involved in making that decision:

> I will have to say, well, gee, if I go there, then I'm going to have to get a clearance to use that data. I'm going to have to learn that terminology that they have for it. And I might have to spend six months of just getting to [know] that data and what can I do with the data I have at hand? And so, you know, I might stay with my [data study] and say, "Okay, well, you know, I'll just go in a different direction with it." And sometimes I may have done that. So, you know, it all depends on just how dominant your study interest is…But, I mean, I'm flexible. And if I see a worthy contribution in [my data study] where I don't have to go barking up a new tree and getting into a whole new set of concerns about getting access, learning the lexicons and all the ins and outs of that database; if I can get some good mileage out of the ones I'm familiar with, I'll stay with them. (D2-01)

This researcher noted they were flexible about what they would study, but the decision to reuse data was negotiated between a "dominant" interest, whether they could make a "contribution," and the difficulty of getting to know new data. Another researcher described a similar "calculus" involved in reusing data. In this instance, the decision was not about switching to another data study but rather about doing a deeper analysis of data. The researcher ultimately decided that the cost to go deeper in time and resources was not worth it given their goals for reuse.

> Once we completed the thesis, we kind of kicked around, "Okay, is there a way to take this project to the next level?" And as I had mentioned, there's this very complex, linking part…And so it was definitely explored whether I should kind of do that and change the [research question]. I think there was agreement and based on the knowledge, that would be a much more substantive, scientific contribution. Those data are just kind of a pain to use. They're not very, like I said, cleaned, they're just kind of clunky. They're not well-suited for research purposes. And so it just became kind of this cost-benefit thing. I knew that this wasn't going to be my primary substantive area. So it just wasn't worth the time and resource commitment to invest in this. (D-02)

An additional consideration brought about by challenges of reusing data was what other expertise they might be able to draw on. The researcher D2-01 described themselves as being the "clutches" of one particular data study but said if things had been different in their career, they might well have "gotten in the clutches of somebody who did work" with a different data study:

> That [different data study] was a masterpiece, longitudinal, they viewed anything and everything, and it was complicated as hell to get to [know]. I never really got to use it much, but had I wanted to go there with it, I probably would have been well advised to find somebody who wrote papers off it and get connected with them. (D2-01)

Data being complicated enough that it was advisable to find knowledgeable collaborators was also discussed by researcher D2-04. They noted:

> Especially if you don't have the experience dealing with datasets, very difficult to use them, very overwhelming, and teamwork, right? Right now, I still partner with people…I am the one who is behind the data. And then I have people that have more ideas on conceptualization. So, if you team up with someone that knows about the dataset, of course, you can use it more. (D2-04).

While other researchers mentioned assembling teams strategically to reuse data (D1-04, D1-05, D2-01, D-01, D-05), a different researcher described how teaming up with someone with methodological expertise but not deep knowledge of the data was insufficient due to the necessity to really know the data:

> because there were so many little details of how to use the weighted data, it was just really complicated. So I had a biostatistician who was going to be a tutor for me during my dissertation. And we met once every couple of weeks, but it ended up being not very helpful. 'Cause you have to be so immersed in the data to really understand what's going on. (D2-02)

These findings reveal that the time and effort required to access, learn about, prepare, and use data was substantial enough that researchers made careful calculations about the benefits reuse of the data would bring. The investment required to reuse data was at times so great that researchers felt pressure to get some kind of publication out of their reuse of the data, even if it was not what they originally desired. They might also continue to reuse data with which they were familiar over data that might better measure their phenomenon of interest or choose not to go as deeply into a topic as they might otherwise because of the time and resources required. A relatively common strategy researchers used to address challenges of reusing data was to collaborate with those who were knowledgeable about the data.

This finding is supported by findings from the survey that the odds of lacking desired knowledge about data decreased 0.60 times if a researchers' research questions remained fixed throughout their study. Apparently, although researchers did not obtain the knowledge they desired, they reached a threshold of knowledge, or reuse equilibrium, that allowed them to proceed with their research nonetheless. To at least a certain extent, then (the extent not being clear from my data), instances where researchers made adjustments to their research questions are indicative of cases where there was research they wanted to perform with the data but were unable because of knowledge they were not able to obtain about the data.

Viewed in light of findings from section 4.2.2 that researchers in general do not seek an optimal equilibrium with data (because they do not believe optimal exists), these findings suggest that even in the general, non-optimal context of data reuse, there are instances where researchers are not able to accomplish all that they desire with data, even though what they are able to accomplish is acceptable by some measure. I come back to this point and address it more fully in sections 4.6 and 4.7.

Looking again at resonances between findings from the interviews and survey, my findings about the time and difficulty involved in reusing data help to explain findings from section 4.3.3 that so many of the researchers I interviewed chose to reuse data with which they had prior experience (i.e., because of the difficulty involved in reusing data). They also support and help to explain findings from the survey, reported on briefly in the description of the survey sample (section 4.1.1) that while most researchers reported they had no or little involvement in the original research, most also reported having good, significant, or in-depth knowledge of the original data creation on their research team. I provide a tabulation of these two variables in Table 4.38. The table shows that 546 of out 767 researchers (71.2%) were not at all involved in the original research and 638 out of 767 researchers (83.2%) had good or better knowledge on their team about the original data creation.

Allowing that some researchers attained knowledge about the original creation on their own (e.g., through use of documentation or by contacting original data creators), the results support a conclusion that a large proportion of researchers were mentored by or collaborated with others who had substantial knowledge of the data, or who themselves had gained significant knowledge of the data through prior experience.

*Table 4.38 Cross-Tabulation of Researchers' Involvement in the Original Research and Knowledge on the Researcher's Team about the Original Data Creation*

| Researchers' involvement in the original research | Knowledge of research team about original data creation | | | | | | |
|---|---|---|---|---|---|---|---|
| | No knowledge | Some knowledge | Good knowledge | Significant knowledge | In-depth knowledge | . (missing) | Total |
| Not at all involved | 23 | 95 | 130 | 179 | 117 | 2 | 546 |
| Somewhat involved | 0 | 3 | 5 | 13 | 18 | 0 | 39 |
| Quite involved | 0 | 0 | 7 | 6 | 17 | 0 | 30 |
| Very involved | 0 | 3 | 10 | 32 | 66 | 0 | 111 |
| Solely responsible | 0 | 3 | 5 | 12 | 20 | 0 | 40 |
| . (missing) | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Total | 23 | 104 | 157 | 243 | 238 | 2 | 767 |

#### *4.3.3.5.3 Social*

My findings in the foregoing sections revealed that researchers frequently reused data they were already familiar with and that this could be explained in part by the time and effort required to learn about and reuse a new set of data. The findings below provide insight into a primary source of time pressures researchers experienced—the pressure to succeed in the academic sphere—and other social pressures they experienced due to considerations about funding and access to data.

I defined social considerations and pressures as ones that resulted from the broader context in which they did their work. Part of this broader context, which pervaded interview after interview, was the pressure to succeed as a professional academic. In many cases, researchers defined success in terms of their research publications. For instance, researcher D-12 explained how pressure to publish caused them to be strategic in the way they thought about their publication pipeline. They adapted their approach to their environment in order put together the most advantageous portfolio for tenure review:

> Oh, man, you'll see. When you're in that position; when the tenure clock's running in your early career. I don't know. And in grad school, you feel it a lot, like disempowered, vulnerable? And then that doesn't change when you get a tenure track job because you're still feelin' like you got a lot to prove. And the metrics that you're proven by are number

> of papers in good journals. And so if I'm thinking about getting to the next milestone that is getting tenure and getting enough papers to be established, sure, I want some papers that are kind of unique little masterpieces…That are heady and technical and impressive, that showcase you flexing your intellectual muscle. Those papers are important, but I felt like I got that in [years past]. I published two really strong, unique methodological papers in high impact journals in my field. And so I felt like I had sort of covered myself in terms of prestige, and basically I just needed to get some good empirical papers in that were easier to write—still good science, but not as ambitious. So that I could get more of them through and ensure that I'd have a full enough CV to get tenure. I imagined that that strategy that I adopted is common across all areas of academia. It's largely a function of the tenure system. For better or worse. I'm not going to opine on whether or not I think there's better ways, there probably are, but that's what we got. And so I've viewed myself as being responsive to the situation I was in, the conditions I was in, and I adapted my approach to publishing to meet that. But now I got tenure, so I write the papers I want to write. (D-12)

Researcher D-09 described a similar circumstance where they needed additional publications as part of their work towards tenure and promotion and explained reusing data "as a kind of insurance or buffer" to maintain a sufficient number of papers in their "publication pipeline" (D-09). Researcher D4-02 provided an additional example of the same kind of pressure that influenced the conduct of their research. They said,

> This project, like all of my projects around that time, was probably motivated by necessity. Realizing that I needed to come up with projects to continue to publish as much as possible to be able to shift over into a faculty role. So really, this came from necessity in that regard. I needed to have a project. (D4-02)

When I asked further about any considerations or pressures, they replied,

> Like I said, the publish or perish climate that we live in academia, it pushed me. If I even want to get a faculty job, I'm going to have to publish above and beyond my day job. So, essentially, that was the pressure there. (D4-02)

Time pressure introduced by the need to publish and achieve academic standing was not unique to junior faculty (as researchers D-09, D-12 and D4-02 described they were at the time). Researcher D-05 described the pressure they experienced to publish as a graduate student as being so intense that researchers might change their research questions as they worked with data in ways that would yield a significant (and therefore publishable) result, even if the significant

result was not something that truly advanced the literature (i.e., publishable findings were more important than meaningful findings).

> Because like I said if you don't have publishable findings, you kind of have to change the research question. It's not politically correct but I think researchers probably face a lot of this especially if they're being told, "Oh, no significant findings." I'm not sure if you heard of the term "salami slicing?" It's basically when they'll have a particular data set they are very well acquainted with. They'll try to find other research questions in there and they will, instead of actually finding something that really moves research forward, they're trying to find something significant to be publishable instead and run with that. So, you know, that's not really moving research forward. I really don't think so. (D-05)

While this researcher did not describe explicit time pressures, such pressure was almost certainly present as it was common among those who conducted their research as graduate students. Researchers D1-03, D2-04, and D3-02 are examples. Pressure to finish in a certain amount of time influenced researcher D1-03 to use data they were more familiar with rather than different data that might have been better for their study:

> In the end, I think [a different data study] would have been probably a better dataset to use. But I think when you're a doctoral student, you're also in a bit of a time crunch to finish within a certain amount of time. I was very familiar and very comfortable with the [data study] because I had already used [the data] in two of my published papers. So I was very familiar with what the variables included and what the sample looked like, and familiar with how to apply the survey weights. So using another dataset would have involved learning the whole data and what it entailed, how to apply those survey data weights, et cetera. (D1-03)

Researcher D2-04 chose to reuse data rather than collect data for their research in order to graduate within a certain period of time.

> I really wanted to graduate in four years. I wanted to graduate sooner than that. But my program was so course-heavy. It just wasn't at all possible. So, I was the first one out of my cohort to graduate. And it still took me the four full four years. But I knew if I collected my own data, it could stretch out for much longer, and I wouldn't have control over it like I would with using secondary data. (D2-04)

This researcher did not give a specific reason for their graduation time frame, but researcher D3-02, who was under similar pressure, cited funding limits:

> The university support my study, but for three years only. If you have to complete your PhD for four years, you have to pay for yourself in the fourth year. If you can complete your study within three years, so okay, you don't have to pay any money. Certain pressure from the financial consideration in terms of—Because I do not come from a rich family. (D3-02)

Among other considerations, including that the researcher's advisor allowed them to concentrate fully on their PhD work (I refer to this later in section 4.7.3.2 on facilitators of researcher's work), choosing to reuse data helped this researcher achieve their aim and complete their degree in three years.

Researcher D1-02 described a similar circumstance regarding both time and funding. They experienced pressure to complete their dissertation on a short timeline with limited resources, both of which influenced how their research proceeded:

> As a quantitative researcher, my resources were limited. I wanted to do a comprehensive study on [topic] and [data study] was almost the most comprehensive data that I could use...I was pressured in terms of time because this is my fourth year and I need to go back to my home country in this summer. I have to defend my dissertation. I have one year, and I'm teaching two classes. (D1-02)

The researcher said they were able to finish in time, thanks in large part to the existence of a data study that included variables in which they were interested. They would have liked to have gained access to the restricted version of the data, but concern about the need to pay for access led them to use the public access version. Researchers D4-01, D-02, and D-07 all reported similar experiences where the need to meet a milestone for their own (or their student's) degree drove their decision to reuse data research.

Another time pressure was the desire pressure to publish findings with a specific data study before others did. Researcher D-02 talked about the pressure this way:

> my overarching concern with this and using really any secondary dataset is you don't have proprietary use of it. So there's, who knows how many, thousands of researchers across the country who are potentially interested in the same topic. So, in order to get it published, it needs to be original research. Who's to say that there's not somebody else

> across the country doing the same project with the same data, who's gonna beat you to publication? (D-02)

Researcher D4-01 talked about using data before it had been too "depleted" by others:

> If the data set is too depleted by previous researchers, if most of the relationships have been tested out of the data set, then that can be another obstacle. So they can produce a kind of pressure. You have to move quicker. (D4-01)

In some cases, researchers described time pressures intersecting with other considerations, particularly around how easy it was to use or access data. For instance, researcher D-09, who needed publications as part of their work towards tenure and promotion, explained that beyond the time saved from reusing data and not needing to collect primary data for their research, the repository where the data were deposited (not ICPSR) was on an exemption list from IRB approval at their institution (meaning that data reused from this repository did not require IRB review). The researcher was therefore able to save additional time in the conduct and publication of their research (D-09).

Another researcher felt under pressure to reuse data before a specified time when they were supposed to destroy the data due to the data's sensitive nature:

> I was under the gun because I had to use the data until I wasn't supposed to use the data anymore. And then I was supposed to like blow it up. (D-03)

They were frustrated in their reuse of the data by the repeated need to renew their access to the data. Furthermore, because they were only able to keep portions of the data they reused, their research was now limited to those portions. These difficulties ultimately caused the researcher to stop reusing the data:

> And so definitely my decision to not use the data and to now be constrained to use only what it is that I have taken from the data if I want to do anything more with it. I don't have access to the full data. That definitely made me less inclined to use it. I stopped using it because it became so time-consuming to worry about the management of the data. (D-03)

Considerations of one kind or another related to access to data were fairly common. For instance, researchers D-03, D-05, and D2-02 all experienced delays to their research because of delays in gaining access to the data. Others described "rigorous" processes involved in applying to access restricted data but saw them on the whole as necessary steps rather than barriers (D1-01, D1-04, D2-01, D2-04). Researcher D-06 considered pursing access to restricted data but did not because of the time and effort that would be involved (including going to a dedicated location to access the data). Researcher D-04, similarly, did not pursue access to restricted data they would have liked to use because of the risk involved in spending time to apply without knowing if they would be able to access the data in the end. After describing the data they would have liked to have and explaining they understood the reasons the data were restricted, they said:

> It requires of a first-year graduate student to take some courage and willing to spend more time on a project—to apply for an IRB proposal—and also, there is no guarantee that you will ultimately get access to official data. So that could have been my obstacle. (D4-01)

In reading the literature, they saw that the names of researchers who had reused the data they desired were all important scholars and they did not believe, as a beginning researcher, they would be allowed to access the data.

> And I was a first year graduate PhD student, and I was thinking, I don't think it'll work out eventually, so I will just stop here just be satisfied with [the data I had]. (D4-01)

Some researchers who described considerations or pressures related to access obtained some kind of additional support that made the data reuse feasible. As a student, for example, researcher D2-02 did not have to pay for access to the restricted data they reused because they were added to IRB-approved research being conducted by a faculty member. They described gaining access would have been a "nightmare" if this had not occurred. While they were able to access the data, they nevertheless experienced delays:

> I just got onto the IRB and had permission to work on the one computer in the office that had it on it when no one else was using it which was really nice cause it takes a long time to get that data. I used it again at another university. And it would have been a lot longer and kind of a nightmare if I had to get it myself back then…I do remember that delayed the start of my dissertation in the fall because we hadn't gotten the data set yet. And I wanted to start in August. And I was a couple months delayed just because they make you jump through all these hoops…And then you can only use it on one computer that is disconnected from the internet and one room that's always locked in all these different steps. So, I had to wait for all of that to get set up to start. (D2-02)

Researcher D2-03 was able to access the data through an institutional agreement that allowed access to a limited number of researchers. Researcher D-11 described a long and detailed application that was required to gain access to data but noted that their university offered support for these kinds of applications and therefore did most of the work. Researcher D-02 was able to learn about the suitability of the restricted data they wished to use ahead of time because their advisor had access to them.

While considerations and pressures due to time, funding, or access were common, not all researchers experienced these. Interestingly, all of those who did not experience such pressures (D1-04, D2-01, D-08, D-10, D-13) reported in the survey having more than 10 years of experience reusing data, and several had more than 15 or 25. Two of the researchers worked outside academia and a responsibility to conduct research fell within the scope of their normal job duties. One of these noted that their main pressure came from finding the time to do research:

> I think the only constraints are usually that you don't have enough time for it [research] and so you just have to try to make time for it…[My work is] all project-driven, as it is in this type of organizations like the one I work for...[it] falls in the category of dissemination of work that we're doing...there are other pressures from work, other work commitments that one may have that make it a bit difficult to find the time for the analysis on the writing of the paper. (D-10)

The other researcher related how,

> in academia you have to worry about grant renewals and all that kind of stuff. In my research, my salary is covered…it's my job to do this research and there's no outside pressure except to publish." (D-08)

A third researcher worked in academia but said there were no pressures in how the research developed; they did not receive grants for most of their work and simply loved doing research:

> No pressure. I mean, we don't have any grants for most of this work. I didn't get any grants ever since I started [reusing data]. We did try to get some initially, but really didn't need them. (D2-01)

A final researcher discussed pressures they experienced during the writing of their dissertation, which was when events occurred that had the biggest impact on the development of their research (the paper was a straightforward publication of their dissertation). However, they said they did not experience pressure during the writing of the paper, which was done during a postdoc.

My findings about the social context out of which considerations and pressures on researchers arose help to explain some of researchers' reuse behavior. Researchers tended to reuse data they were familiar with because the time they had to learn about and reuse data was limited. This occurred because of professional pressures (i.e., to publish as part of tenure and promotion or meet milestones in graduate education), limitations in funding, and pressure to release results before others reusing the same data. Considerations about time also led researchers to use specific data (e.g., to avoid IRB requirements that would delay access to and reuse of data), or not reuse specific data (e.g., because of lack of surety about ultimately being granted access to restricted data or because the time it took to manage the restricted data became too much).

To the extent that social considerations and pressures limited or inhibited researchers' research, they are indicative of ways that the existing research ecosystem may not be well adapted to support research, or as well adapted as it might be. The fact that some researchers

received forms of assistance from institutions or individuals—without whom their research would not have been possible—in itself, points to the scrutiny that social considerations and pressures deserve in enhancing or improving structures that support research with secondary data. It is notable in this regard that many of the considerations and pressures related to time, funding, and access were experienced by researchers early in their careers, and also that the majority of researchers in my sample were in such positions (see Table 4.21). I return to this point in the discussion my findings in chapter 5.

A question remains about what these findings reveal specific to the knowledge researchers obtained about data, and how they speak to possible improvements in design of structures in the research ecosystem that support the attainment of knowledge about data. I leave a fuller discussion of this to chapter 5 but note that my findings about considerations and pressures must be taken together with findings about community perceptions of data reuse and the value placed on reusing data. My findings indicate that there are ways the data reuse ecosystem can be enhanced to support reuse as a steppingstone to further research (i.e., a means of achieving professional goals), and also ways that it can be enhanced to support data reuse and its capability to contribute new and meaningful knowledge that matches and maximizes researchers' aspirations for reusing data, rather than just their professional aims.

#### *4.3.4 Summary*

The factors I found to be significantly associated with lacking desired knowledge—accounting for three control variables (the natural log of the number of citations, whether the data were created to be reused, and the source of the data) and clustering effects—were: (a) knowledge about the original creation of the data, (b) whether researchers lacked desired

knowledge at the time they were considering reusing data, (c) whether research questions were fixed or fluid.

I found that the first two of these factors were associated with a decrease in the odds of lacking knowledge at a significance level of $p < .01$. In particular, the odds of lacking knowledge decreased by .29 times for each greater level of knowledge researchers obtained ($p = .001$) and the odds of lacking knowledge decreased by .60 times when research questions were fixed at the time of the decision to reuse data ($p = .004$). On the other hand, the odds of lacking desired knowledge were close to four times higher (3.73) when researchers lacked desired knowledge at the time they were considering reusing data ($p < .001$).

It makes sense intuitively that researchers with greater knowledge of the original data creation would lack less knowledge about data. That the probability of lacking desired knowledge decreased when research questions were fixed is more surprising, however. I expected that researchers who changed their research questions would exhibit a lower chance of lacking knowledge because they had the opportunity to adapt their study to their level of knowledge. The finding that the greatest predictor of lacking desired knowledge at the time of deciding to reuse data was lacking desired knowledge at the time of consideration is also very interesting. It seems to point to an additional, unmeasured factor that influences the probability of lacking desired knowledge at the time of decision.

Both of these findings align with evidence from the survey and interviews I present in sections 4.4, 4.5, and 4.6 below that researchers make decisions about reusing data at multiple levels. They work hard to select data initially that will help them meet baseline social criteria for success with data reuse (for instance, obtaining a published article), and also push further to see how many of their personal goals for their research they can meet through reuse of the data. It

may be that researchers who are the most sure that data will meet their needs are more able to fix their research questions ahead of time, and that the unmeasured factor is how sure researchers are, before deciding to reuse data, that the data will meet their reuse goals.

To my findings from the survey I added significant information from the interviews about factors in researcher's broader social environments. I learned that researchers were motivated to reuse data that they were familiar with, that offered benefits over other data, and that were strategic and convenient to reuse.

Despite the benefits and value some researchers believed it offered in its own right, secondary reuse took place in a social environment where primary collection of data was preferred and incentivized. Reuse was often seen as a lesser and less desirable option necessitated by a lack of time and resources that could, nevertheless, serve as a steppingstone to conducting (the more desired) primary research.

I also found that researchers reported a variety of considerations and pressures that influenced how their research with secondary data developed. At the individual level, pressures researchers imposed on themselves, and considerations related to maximizing the openness, transparency, and reusability of their own research.

At the level of the data, interviewees reported that research with secondary data could be time-consuming and intimidating and there was a lot to learn about reusing data. As a result, they weighed the costs and benefits of investing resources in reusing data and considered strategies such as collaborating with others who were knowledgeable about the data to lessen or share the burden of reusing data. This finding helped to explain why so many researchers reused data they were familiar with and was supported by findings from the survey that most researchers reported

having good, significant, or in-depth knowledge of the original data creation on their research team even though they had little involvement in the creation of the data themselves.

Another way that researchers dealt with challenges reusing data was to reuse data they were already familiar with, even if other data were a better fit for investigating their phenomenon of interest or if reusing the familiar data necessitated changing something about the scope, depth, or focus of their research. This finding resonated with my finding from the survey that researchers who adjusted their research questions were more likely to lack desired knowledge about data. It provided additional evidence that even in the sub-optimal context of data reuse (where researchers were by definition not accomplishing all of their research aims), when researchers achieved an acceptable reuse equilibrium, they struggled to accomplish all that they desired by reusing the data. I discuss this further in sections 4.6 and 4.7.

Findings at the social level revealed that researchers experienced pressure to complete their research in certain amounts of time due to professional pressures (i.e., to publish as part of tenure and promotion or meet milestones in graduate education), limitations in funding, and pressure to publish results before others reusing the same data. Considerations about time also led researchers to use specific data (e.g., to avoid IRB requirements), or not reuse specific data (e.g., because they were unsure their application to use the data would be successful and the time required to manage restricted data).

I suggested based on my theoretical framework that to the extent that social considerations and pressures limited or inhibited researchers' research, they were indicative of ways that the existing research ecosystem might not be well adapted to support research, or as well adapted as it might be. I noted that the fact that some researchers received forms of assistance from institutions or individuals—without whom their research would not have been

possible—pointed to the importance of examining social considerations and pressures when considering strategies to enhance or improve structures that support data reuse. I pointed out in this regard that considerations and pressures related to time, funding, and access were experienced by researchers early in their careers, and also that the majority of researchers in my sample were in such positions (see section 4.1.1.1 and Table 4.21).

Overall, my findings in this section indicate that there are multiple ways the data reuse ecosystem can be enhanced to support reuse: first, by enhancing its use as a steppingstone to further research (i.e., a means of achieving professional goals), and second (not exclusive to the first) by enhancing its capability to contribute new and meaningful knowledge that matches and maximizes researchers' aspirations for reusing data rather than just their professional aims.

## 4.4 How Researchers Bound the Knowledge They Obtain in Order to Decide to Reuse Data

In this section, I present findings about the strategies researchers used within their particular social environments to bound the knowledge they obtained about data. Because the definition of each of the terms, "knowledge" and "bound," evolved over the course of my study, I provide a brief recap here of how I used them.

In section 3.4.8.5, I introduced difficulties I encountered in defining "knowledge about data," and a distinction I drew been knowledge "about" data (e.g., knowledge about the context of creation or the extent to which the data provided evidence of the phenomenon they were interested in) and knowledge "of" data (e.g., knowledge of what missing values in the data actually were, or additional knowledge about subjects in the study). I concluded that the most important questions for me to focus on were (a) the kind of knowledge the researcher wanted (regardless of whether it was "of" or "about" data or of whether or not the desired knowledge had been collected or included in the scope of the original research) and (b) how they negotiated

or came to terms with—i.e., reached equilibrium with respect to—how much knowledge was enough.

Similarly, in section 3.4.4 I discussed tensions I encountered (researchers appearing to feel I was challenging their research) when I asked researchers how they determined they had "enough" knowledge about data to decide to reuse them. As I reported in that section, researchers' responses led me to change the wording of the question, to ask researchers how they knew the data were sufficient to fulfill their needs, or how they knew they had reached a threshold where their knowledge was enough to reuse the data. This strategy seemed to be effective in removing tension and leading to more open discussions in subsequent interviews.

However, for the purposes of presenting and interpreting my findings, it is necessary to understand that researchers' responses reflected all of these characterizations of knowledge bounding (how researchers determined they had enough knowledge, how they knew the data were sufficient, and knew they reached a satisfactory threshold of knowledge). While there were evidently subtle differences in the way researchers understood my questions (enough to affect their attitudes about my questions) the characterizations bear sufficient similarity to one another to be taken collectively as evidence of how researchers came to reach a state of reuse equilibrium.

### *4.4.1 Research Questions 1 and 1b: 1. How Do Researchers Determine the Boundaries of the Knowledge They Obtain About Data in Order to Reuse Them in Their Research? 1b. How Do Researchers Determine How Much of What Kinds of Knowledge Is Enough?*

Researchers used a variety of strategies and methods to reach reuse equilibrium with respect to the data they reused, either in the face of knowledge they lacked about the data, or in determining that they knew enough about the data to reuse them in their research. I earlier defined reuse equilibrium as something that happened when a researcher determined that a

particular body of data was sufficient to reuse to study their subject of interest and decided to reuse the data to accomplish a specific purpose. My analysis has allowed me to add more nuance to the meaning of "sufficient" in this definition—specifically, to note that what is sufficient depends on the broader context and goals of the research.

For researchers who were conducting their research for the purpose of publication, sufficiency was determined by the researcher's ability to use the data to make a contribution to existing literature. Making a contribution appeared as a fundamental goal and measure of success across the interviews (e.g., D2-01, D2-02, D3-02, D4-01, D-01, D-02, D-05, D-06, D-07, D-08, D-09, D-10, D-11). Researcher D1-05 framed it in terms of finding a gap in the literature:

> I still do secondary data analysis now, like with [data study] and some other data sets, but you really have to be able to go in there and figure out what's the gap, or what's still missing in the literature, and so you really do have to figure out how can you create a new question…based on the data that's available. (D1-05)

Researcher D1-03 discussed making a contribution as a key aim, even if, due to limitations of the data, it was not the full contribution one wanted to make:

> there might be opportunities for you to maybe answer parts of your question in, not all [aspects], but still be able to inform the literature with the limited information that is available or collected from these datasets. (D1-03)

Researcher D3-02 expressed the goal of making a contribution perhaps most succinctly in discussing a particular mindset they developed for reusing data. A portion of what they said, quoted from section 4.2.2 was "you have to identify something new. Otherwise, you cannot write your paper. You're not able to keep a job, you are not able [to produce research] and so on" (D3-02).

As many researchers noted and as I reported on in section 4.3.3.5.3, making a contribution was verified or consummated by the publication of the paper (e.g., D1-05, D-01, D-02, D-05, D-09). To quote from that section, researcher D4-02 explained:

> This project, like all of my projects around that time, was probably motivated by necessity. Realizing that I needed to come up with projects to continue to publish as much as possible to be able to shift over into a faculty role. (D4-02)

And as researcher D-12 put it: "the metrics that you're proven by are number of papers in good journals."

For researchers at the Masters or PhD level, an emphasis on making a contribution existed, but this could be overridden by the goal of demonstrating competence in research. Researcher D1-03 expressed this in talking about a portion of the research that went into their dissertation. They said,

> I just remember feeling like it was not my best work and I was just doing it to get through the motions and to finish. So it was probably the paper that I came to recognize that this one will probably not get published and it's okay. You know, and trying to remember what [our PhD director said at the time], and I'm like, think of this as an exercise. They're not gonna be too concerned about whether there was no finding as long as you did everything in your power to answer the question the best you can and if you justify that you did X, Y and Z, you did all these different approaches to try to respond to your question. Because they're really just examining your skillset rather than, you know, the finding. (D1-03)

Researcher D1-03 reported the director as saying about the dissertation, "This is not going to be the thing that makes or breaks you. It's the things that come afterwards that really are important."

Another researcher D-02 described a similar dynamic in the completion of their dissertation. Speaking of the main consideration in determining that these data would meet their needs for research, they said,

> I mean, the first [influence] was one of my committee members. His big quote was, "A good thesis, is a done thesis." So the biggest thing was, there's going to be bigger fish down the road. Think of this as a milestone you have to get through. And if it gets published, well, that's a great bonus. But most theses don't get published. You're better getting this done in 18 months, and moving on, and having it not published rather than spending three years getting a publication out of it, but now it's going to take you seven years to finish the program. So the priority was, "Identify a project you can get done. Identify a project you can get done quick." (D-02)

A third researcher talked about the lack of pressure to publish they experienced as a student, though they noted they did not think this was "mature":

> In my opinion, it's not good for us to concern about the matter of publication, the matter of how many papers you have, the matter of whether you can publish in the high-ranking journal because I think it's important for us to do something I liked when I was a student. And my thesis supervisor also did not give me any pressure. He just say, "You do something you like first." But it is not so mature, you know. It's not so mature, but he just say, "Okay, you do something you like. No problem…You just try your best and everything will—no problem, okay?" "The best is yet to come. (D3-02)

Most of the strategies I discuss below were not exclusive—they applied to either goal (either making a contribution or demonstrating competence). The biggest difference was who or what body made a final evaluation of the researcher's work. In one case (making a contribution) it was an editorial body. In the other, it was a researcher's thesis committee. Along the way, researchers employed different strategies, methods, and evaluations iteratively in an attempt to reach their reuse goals.

The strategies I report on include:

- conducting preliminary investigation
- adapting research questions
- becoming knowledgeable
- possessing personal qualities
- using data
- interpreting results
- getting feedback

While these individual evaluations were important, they took place in the context and with the knowledge that evaluations of the research would also be conducted by others besides the researcher. A key finding in this regard is that researchers' knowledge that their research would be evaluated in a broader context influenced the ways they reached reuse equilibrium and their determination of how much of what kinds of knowledge were "enough." This finding led me to define a "personal" reuse equilibrium, reached on the basis of a researcher's own

determinations about data in relation to their goals, and a "social" reuse equilibrium, influenced by or reached in relation to broader social norms and requirements (such as publishing or demonstrating competence in research). These were not always distinct, but rather blend into one another, as I describe below.

#### *4.4.1.1 Conducting Preliminary Investigation*

One strategy used by a number of researchers to reach reuse equilibrium was to try to ensure ahead of time (before obtaining the data) that the data would be adequate to use to fulfill their reuse goals. I noted in section 4.3 how common it was in my sample for researchers to reuse data they were already familiar with (e.g., researchers had participated in the collection of the data, been a part of research projects with the data or had reused the data in their research over a long period of time). Researchers selecting data they were familiar with indicates the activation of certain level of prior experience with data that researchers employed to establish their own personal reuse equilibrium. Below I give some examples of researchers who were reusing data earlier in their careers, where, lacking experience, they were more reliant on a social equilibrium.

One researcher, who was a PhD student at the time, had an advisor who had access to the restricted data they wished to reuse, and other advisors who had reused the data. The researcher was able to talk to their advisors to determine ahead of time what their dependent variable would be, whether there were enough cases to conduct their research, and other details (D-02). This preliminary work was important because it gave the researcher basic knowledge that they would be able to do the analysis they wanted. If they were able to find more in the data once they received them, that would be a nice bonus.

Another researcher, with more extensive experience (more than twenty years conducting research), verified that data they reused would fulfill their needs in a couple of different ways.

First, they reused data— and encouraged their students to reuse data—collected by researchers that they knew. This way, they knew something ahead of time about the quality of the data. Second, the researcher had graduate students working with them download the questionnaire and codebook for data they were interested in and ensure that the questions and frequencies of responses were adequate to the research question (D-07). The researcher described their general process with students this way:

> When they come up with an idea and they see a dataset or if they're presenting their dissertation topic to me at the very early stages, I make them pull down the data and run the frequencies on the questions we need. That's the very first thing we do. And I said, "Well, [student name], you're interested in [phenomenon], but 90% of the people didn't answer that question. You wanted that one to be one of your main hypotheses in your paper. You just can't use this data. It's an art. Who knows that that's an artifact of poor conduct of the survey from the PIs or it's just that, I don't know, the behind the scenes, why there are so many missing data." So that's what we look at first. (D-07)

The process this researcher followed with their students, for whom the experience might be their first reuse of data, demonstrates the mentoring and guidance that factor into the social reuse equilibrium researchers obtain, especially early in their careers.

The importance of guidance from advisors and mentors in reaching reuse equilibrium is evident in a quote from researcher D-02 who, like many others (D2-04, D3-02, D4-01, D-02, D-05, D-08), reported finding a gap in the literature as an important part of preliminary investigation:

> I didn't want to spend a year, 18 months on a project that was just going to collect dust. So I did do some intel to make sure, "Is this a novel enough idea that it can get published? Is there a gap in the literature?" Those kinds of things. And you know, "What kind of journals might actually care about this to publish it?" So, I definitely had an eye toward, "Can this be published?" But again, it's pretty early on in my graduate career… My advisors generally said, "Yeah, this would probably get published," but I didn't really have a great pulse on what that process was going to be like. (D-02)

This kind of guidance was commonly reported throughout the interviews and, particularly in the use of literature in advance investigation, shows the continuum of personal and social

influences on reuse equilibrium. One on hand, the researcher consulted literature to identify a novel area to publish in. On the other, they were not yet sure what constituted a publishable product and needed guidance to that end. This tension can be contrasted with the experience of a more advanced researcher (D-08) who themselves determined the specific criteria that data for their research needed to meet. After determining the criteria and spending several days searching for and evaluating datasets, they determined there was only one dataset that satisfied all of their requirements.

In section 4.3, I discussed how important it was for researchers to choose data they knew they could use successfully to meet their reuse goals, especially because of the time and effort involved and external pressures on time and funding. In the face of these pressures, doing advance investigation to confirm the sufficiency of data is an example of a strategy researchers used to reach reuse equilibrium and one in which they might be influenced by a continuum of personal and social factors.

#### *4.4.1.2 Adapting Research Questions*

Adapting research questions after selecting data was another strategy researchers used to reach reuse equilibrium (e.g., D4-01, D4-02, D-10). This could be considered opposite to what one researcher referred to as a "top-to-bottom" approach (D1-02), where researchers developed their questions first and subsequently identified data to investigate them (this approach was employed by others in my sample as well, e.g., D-02, D-08, D-12). However, I found that even researchers who developed their research questions first sometimes needed to make adjustments based on the data. I cited a quote from D1-05 earlier as evidence of this. They said,

> I always start out with the research question and then you do have to tweak that, too, if you find that it's repetitive or something that's already been done. (D1-05)

Another researcher described performing analysis based on their research question, but being advised, upon not finding significant results, to change their research question to achieve "publishable" findings (D-05).

In an alternative to both of these cases, one researcher described identifying their research question and the data to investigate as a dynamic process:

> So can we take these two data sources and put them together in some sort of meaningful way? So it wasn't necessarily that the research question came first or the data source came first. (D-06)

Abandoning a research question after finding it could not be answered using the data was also reported (D-11), as was reusing a data study starting with one set of research questions and identifying other questions to investigate in the process (either in addition or instead) (D-03, D-13).

Similar to the strategy of conducting prior investigation, in adapting research questions, researchers at times had their own personal motivations and at times were influenced by others' (social) comments and feedback. I discuss the role of this feedback, particularly in the context of goals to publish or demonstrate competence, further in section 4.4.1.7 below.

#### *4.4.1.3 Becoming Knowledgeable*

Many researchers discussed the importance of getting to know the data intimately to become comfortable reusing them. Researchers talked about "sitting" with data (D1-05), "immersing" themselves in data (D1-04, D2-02), getting to know data "frontwards and backwards" (D4-02), "educating" themselves (D-12) about the data, "understanding" data and how to use them (D1-05), paying their "dues" with the data (D2-01), and provided examples illustrating the meticulous attention they paid to all aspects of the data and documentation (D2-03, D2-04). Here is one researcher's description:

> You have to pay your dues with the data. It doesn't jump into your head. You have to learn the labeling of the variables so that you become facile and finding whatever thing you're interested in. And you sit down with the code book and it's like a Bible, and you have to learn the chapter and verse before you can really meaningfully use it. So, you know, that's essential. (D2-01)

Researchers who did this work to gain knowledge about data reported becoming "familiar" (D2-01, D2-04, D3-02, D-03, D-04, D-08, D-10, D-12), "comfortable" (D1-01, D1-03, D2-02, D2-04, D-03), or gaining "competence" (D-03) and later expertise (D1-03, D1-04, D1-04, D2-01, D-03, D-12) with the data. This expertise, then, could translate into "intuition" in using and evaluating data. One researcher described the progression this way,

> And that is what's going to be helpful when you face one of these datasets. It's just gigantic. I mentioned you have to find out your way, you have to read about this, you have to understand and if you don't understand you have to talk, but it's mainly experience. And with experience, comes intuition right, so you develop the intuition of how to evaluate a dataset and you keep working. And it's in the details, I would say. Everything is in the details. (D2-04)

Other researchers described, similarly, how their familiarity with a specific topic, specific data, or reusing data, made it easy for them to evaluate data for its usefulness for their purpose (D-03, D-07, D-12). When I asked one researcher to talk more about how they evaluated whether the data were "good enough" for their purposes, they replied,

> I want to give you a satisfying answer, but I just feel like there was raw data available. I'm a statistician. That's what I deal in, you know? It's my food. So, if you put it in front of me, I know what to do with it, right? (D-12)

Another researcher described knowing they had obtained "enough" familiarity with or knowledge about the data to reuse them when they did not have any more questions for their advisor about the data (they were, rather, explaining things to their advisor), and when they were pointing out errors in the data to the data creators (D-06).

Becoming knowledgeable was an important way that researchers reached a personal reuse equilibrium (in contrast to one influenced by external parties or factors). Like the strategy

of conducting prior investigation—though with important differences—researchers reported a progression of experience that enabled them to be more confident reusing data over time. The differences are that becoming knowledgeable had to do with using a data study rather than selecting one (as conducting prior investigation did), and achieving reuse equilibrium remained more in the scope of the personal (individual) rather than social sphere.

#### *4.4.1.4 Possessing Personal Qualities*

Researchers mentioned several qualities that facilitated becoming knowledgeable about and gaining expertise with data. These generally involved a propensity and like or love for spending time digging through data, and also a desire for personal fulfillment. As related in the cases above, researcher D1-04 identified themselves as a "data geek" and described the large amounts of time they would spend trying different things with the data. Researcher D2-01 referred to themselves as a "data junkie." Researcher D2-03 recalled "how long I could just sit down and I could just nerd out in an ICPSR search engine just looking for data sets." As qualities, these were not strategies researchers used, but rather attributes that gave them motivation and persistence to reach reuse equilibrium.

Researcher D1-05 talked about large surveys as something you had to "commit yourself to" and noted "you have to like data analysis in a way; you have to maybe like the idea of [an] ambiguous process where you're not exactly sure what's going to happen at the end of it." That a tolerance for such ambiguity is needed for research was underscored by another researcher:

> You went and spend maybe two days or three days and something else on it and then oh, the dataset does not able to provide you such kind of things. That is what I mean, the waste of time. However, I still think it is an important process for doing research. Because if you do not have such kind of process, you do not know whether the data set is suitable for you or whether you can find something new. In a certain extent, it's like gambling. (D3-02)

Other researchers described how they "liked" working with data and "loved doing research." In describing their decision to reuse data they had never used before, one researcher explained, "I like data, I like exploring new datasets." Another described their implicit motivation to find time to work with data:

> You know, once you have the data, it's just time to do the analysis, and you find the time. I love doing research. I don't need to get paid for it to do it, so I'm happy to do it. It's a good thing to do. More knowledge. (D2-01)

Having a personal or natural interest (D1-02, D1-05, D3-02, D4-01, D-01, D-11) in their research topic and having a desire to prove something to themselves are two other characteristics researchers mentioned that catalyzed their research. Researcher D-11 emphasized the importance of interest in the development of their research:

> it started with a natural interest in this area. I think you're going to get nowhere as a researcher if you don't just have that natural interest in it. (D-11)

Researchers D-11 and D1-02 both spoke about the passion they had for their research topic as an important motivator for their research.

Another researcher reflected on the fact that they had chosen a project that required deep knowledge of statistics even though statistics was not their passion. They said they now partnered with others who performed the statistical analysis (perhaps partially supporting the point made by D-11) and had been motivated by the desire to prove themselves:

> that might have influenced things 'cause that's not my passion. I don't know why I picked a project that was all statistics for my dissertation when I don't like statistics…[Now] I'm partnering with other people who like to run the analyses, and I'll do the rest. But I think I'm glad I did it 'cause I wanted to prove to myself that I can do it, so I didn't feel like a fraud or something. (D2-02)

A final characteristic that a researcher mentioned was the ability to critically evaluate their own work. One researcher expressed the importance of this as follows:

> and be very critical, not only to what you read, but to what you produce and what you think, it's another important thing, right? So, you have to put several checks and, give you time too, so your ideas can grow, you have to keep distance between what you produce and when you go back to what you produce…[S]o it's not that you write something and then you keep writing on top of it, you write something, you leave it one week to just sit, then you come back, right? (D2-04)

I discuss critical evaluation of work more in the context of interpreting results further below.

The above findings underscore the difficulty of achieving reuse equilibrium (since such a degree of natural interest and even passion are required) and how essential the personal component is to achieving reuse equilibrium. My findings suggest that if data reusers did not obtain their own sense of enjoyment and fulfillment from reusing data, then social norms and requirements—for instance, for achieving a publishable standard of research—would not be enough to motivate research with secondary data.

To express this in slightly different terms and place it in the context of my findings that researchers used secondary data as a steppingstone to further their careers (section 4.3): researchers used secondary data as a steppingstone, but they also appeared to really like what they did. Researchers had their own personal motivations for reusing data, which translated into strategies for reaching a personal reuse equilibrium (such as becoming knowledgeable about data), on top of which social norms and requirements for publishing or demonstrating competence were laid.

#### *4.4.1.5 Using Data*

In addition to obtaining knowledge and expertise about data, and possessing qualities enabling those, researchers reached reuse equilibrium through the approach they took to reusing the data. Some researchers talked about taking an "incremental approach," illustrated by the following quotes, solving each problem or dealing with each issue reusing the data as it came up.

> So I just, every time I would go in—'cause I could only work in one room that had the data on it, and I had a whole notepad of what small issue to get to the next step each day…So I broke down everything I needed to do for the analysis based on the user guide and based on what I wanted to find out. And then I put those steps into my calendar of what I needed to do each week. And I would just constantly double-check myself and make sure the numbers lined up; that they made sense after I ran different things 'cause I was really worried that I was going to make a little like coding area error and mess up the whole thing. (D2-02)

> I was just trying to make sure to, at every step of the way- is this possible? And, you know, if that wasn't possible, is there an alternate approach? And if there was an alternate approach, okay, could I do that approach? And I never hit a point where an alternative approach didn't work. (D1-03)

Other researchers discussed testing the limits of the usefulness of the data for their research, which they talked about in terms of "pushing the limits" of a study, "stretching" a study too much, or "going in too deep." Reusing data was a matter of exploring the limits of what a data study could offer for their purposes and making compromises where necessary to stay within those limits:

> Reading the literature behind, or all the literature published using this kind of study, is very helpful to understand "What are the purposes of the studying [sic]? How you could use it?" So you don't feel you're pushing the limits of the study, right. That's important…And not stretching the study too much: it's a compromise between what you want and what they got to offer to you. (D2-04)

> The data had a lot of things, as I said. It's just when you start to explore more and more, going too deep, then you look for some variable, some questions that maybe you don't have in the data. But that doesn't mean that the data has nothing. So, the data has many things…[Y]ou keep exploring, you keep exploring, you go in too deep, and then you don't find many things. And then you just try to find alternative way to best utilize the information that you already have. (D-01)

Some researchers approached data by identifying factors that could confound or bias their results (D2-03, D2-04, D-08, D-11, D-12, D13), and coming to a point where they had addressed their concerns. This involved including the most likely confounders as control variables, gathering additional data to use in comparison, or using data from different parts of the study to triangulate and verify information and results. Researcher D2-03 related their process as follows:

> Thinking about the variables that you think are important potential confounders, going to the literature and seeing what other people have worried about; that seems to take the top priority on the to-do list for most. And then the question becomes, "Am I missing something?" You know, "Do I need to collect data on more potential confounders that could be driving this relationship between [phenomena]?"
>
> And, oftentimes, you'll do sort of what the data tell you. So, if I'm estimating the relationship between [variables], I'm controlling for all of these other variables that are easily available that I've come up with on my own or referenced the literature to come up with. And if my estimates really don't budge much, they're really insensitive to throwing all these things, then my concerns are a bit quelled and I'm not as worried about working down that list of important confounders and trying to kill myself over collecting more information on other things that are probably less likely to matter if, after controlling for things that really do matter don't have an effect—or that I think might really matter don't have an effect—that's sort of the place where you say, "Yeah, I can probably stop," you know. (D2-03)

A researcher reusing a very different type of data study (administrative data as opposed to a large survey) reported a similar process:

> because all this data is publicly available, I can take a sample and do my own grunt work, so to speak. But I can collect data outside of this database to confirm this is true. There are also existing variables within the database that I can triangulate things, even if it's not perfect in terms of my research question, right? As long as those errors aren't correlated with my dependent variable, then that's just noise. It may be sloppy, but it's not gonna bias the answer, right, in the sense that there's no evidence that it's related to that relationship that I'm testing. (D-11)

Some researchers' approach to data involved applying statistical methods (D-01, D1-03, D3-02, D-06, D-10). These included using statistical methods to remove endogeneity, creating synthetic error in simulation data to approximate error that would occur in collected data, and employing techniques to compare similarities across data gathered in different contexts or to account for people dropping out of a research study over time. One issue raised by a researcher in using statistical techniques is that they can make the main argument of the research more difficult to understand. Speaking of using a technique, a researcher said,

> It should sufficiently address the issue. But these things unnecessarily complicate the main argument. Of course, we always try not to complicate the things, even though we use econometric statistics as tools, but at the end, we would like to write a simple paper

> so that everyone can understand it. So, if you have the data that you want and you want to use it, if you have a data source that provides you everything, all the information, everything you want, then of course, you don't have to go for these kind of methods and unnecessarily complicate the paper. (D-01)

The strategies researchers reported related to using data were largely ones I associated with reaching personal reuse equilibrium (as opposed to social equilibrium). I say "largely," because in at least some cases researchers were also consulting literature (e.g., to understand the limits of what a study could be used to learn), which involves a social component.

#### *4.4.1.6 Interpreting Results*

In section 4.4.1.4, I noted the ability to be critical of oneself and one's work as a quality researchers valued in conducting secondary reuse. Researchers talked about this value in the context of interpreting results as well (D1-05, D2-04, D-10, D-13) and used "caution" or "respect" for data as guides in reaching reuse equilibrium. For instance, for researcher D-10 it was important to understand ways that limitations in the data collection affected the way the data could be analyzed and not going "overboard" in the analysis:

> I think if you work on the generation of data or of instruments, what's always good to have a critical distance from what it is, what you would like to do and how much can be done. And I think any kind of empirical research has limits, of course. If you ask someone about how they feel or how capable they feel themselves to engage in [activity] or interpret something, of course, there will always be some sort of bias though it's not pure measurement yet because you can't just go and look how someone can actually do it…So I think I'm quite critically aware of some limitations we have. And so you also have to be very careful of how you interpret your results and that you don't go overboard. (D-10)

Another researcher talked about the need to be cautious in making and communicating about research conclusions, and to combine this circumspection with knowledge about the data:

> You have to be clear when you write those things or at least, you have to understand your limitations and your language has to be always very cautious when you describe these things. And try to avoid deterministic language or strong language that could arrive to conclusions that may not be supported by your data. So that's a challenging part of using this kind of data sets, just you have to be very careful with the interpretation and how you

> will interpret your results and how you will communicate your results. You have to become very knowledgeable about how the dataset became a dataset, right? (D2-04)

Researcher D1-05 spoke about taking care in the analysis of data in terms of "respecting" the data:

> We went back to the literature to see how other people were [using a certain measure], and so it's just making sure that you're willing to tweak what you need to, and that you take the time to really be careful with the data and just respect the data, and its process and respect what you need to do, and just really understand it. (D1-05)

Taking care in interpreting results was a strategy that, similar to using data, was largely used in the achievement of personal reuse equilibrium. The personal intersected with the social when researchers made comparisons between what they were doing and what other researchers had done with the data.

#### *4.4.1.7 Getting Feedback*

As noted in the section above on adapting research questions (section 4.4.1.2), researchers obtained comments and feedback from others on their research. This is a strategy that many researchers used to reach reuse equilibrium (D1-03, D1-05, D2-02, D2-04, D3-02, D-01, D-02, D-05) and one where social influences were highly in evidence. Researchers received feedback from advisors, mentors, and co-authors, from presentations at conferences, and other sources such as journals. Researcher D2-04 described how, after they obtained a certain level of comfort making decisions about reusing data, they obtained feedback from their advisors and others. Their experience illustrates the interplay between personal and social influences in reaching reuse equilibrium:

> by the time I had to get the dataset, I was pretty much comfortable making those kind of decisions in terms of "Hey, this can work. This cannot work." And then, of course, you go and talk to your mentors and advisors and explain your reasoning. And if they agree, you're usually in the right way. Then you get another random people somewhere there and you explain the same thing, and then well, if nothing all comes out then, you're probably in a good way. You can send a conference, right? I didn't do this. But with other

> datasets that's what I do, usually. I go to a conference. And if the conference goes well, then probably in a good way. (D2-04)

Another researcher's advisors provided guidance on appropriate use of the data:

> because this was a widely used dataset and I had advisors who used it, I was able to get guidance on "Which variables do I use? "Am I doing this correctly?" I was able to check in with people to make sure that it was being used as it was intended to being used. (D-02)

Another researcher reported relying on their advisor's guidance when faced with data that did not include certain information they desired for their study:

> [The data study] does not have such questions and I also at that time- I was a PhD student. I only can rely on my PhD supervisor, my professor, and I also discuss with him and he say, "Okay, it's acceptable. It's an acceptable research limitation. And, therefore, I think it's okay, and I'm also able to answer most of the things I want to know. (D3-02)

This researcher's experience mirrors several described above where, since they had not obtained sufficient personal experience, they were guided largely by parameters for achieving social reuse equilibrium.

An additional source of feedback that researchers mentioned was feedback resulting from the journal review process (D1-01, D-02, D-09, D-12). It was not uncommon for researchers to experience one or multiple rejections, but in the process they also obtained valuable feedback about how to improve their studies. Researcher D-12 described one such process:

> [The paper] got good reviews but it lacked cohesiveness. That was some of the feedback I got, and I submitted it to the top journals in my field, so the probability of it getting in was low, right? But I was getting good feedback because good journals typically have more rigorous reviewers that will give you a harder time, and they're, you know, it's just how it goes, at least in [my discipline]. (D-12)

Another researcher, D1-01, described how comments from the review process caused them to expand the variables used to measure one of their concepts. Researcher D-02 described a scenario with multiple rejections:

> So probably around [date] I started sending it out for review. And then, that process takes a while. It takes a couple of months to hear back, you get rejected, a couple more months and you're back, you're rejected. And, you know, probably bounced around a bit, probably took about maybe 18 months or so in order to get it accepted. (D-02)

An interesting aspect of the journal review process that researchers related is that review was not always a good source of feedback. That is, in addition to describing the review process as valuable for this purpose, researchers mentioned submitting papers to journals that they deemed more appropriate than others for the topic or method they were investigating or using (D2-04, D1-04, D1-05, D-05, D-09). Researcher D2-04 described the variation in expertise that can exist at different journals:

> But I guess, if you go to a journal, I don't know, let's say a top journal in [area 1], the reviewers don't have the expertise of [a specific concept], so the reviewers won't give you much of a hard time however you define [the concept]. (D2-04)

Researcher D2-04 went on to say that the concept they mentioned was very well defined in a second area of research and the scrutiny given to different issues can vary a lot, depending on the journal. Another researcher described a similar differential among journals regarding the age of the data that are reused:

> Sociology journals are more understanding about the age of the data because they know it is unique. They also know that the kind of issues we deal with like [issue] don't go away in 20 years. More likely to be challenged about age of data in interdisciplinary journals. (D1-04)

Researcher D2-04, above, concluded, given differences among journals, that it is best to

> just to look at your research first and then look for the journal that may fit your research. Because trying to accommodate your ideas to your journal, it's very difficult. It's better that you define, you develop your own ideas and when you think they are mature for publication, then look for your journal. (D2-04)

Researcher D1-05 described a slightly more accommodating version of this process, considering a variety of possible journal outlets in case a paper receives rejections:

> Usually as I'm getting to the end of the paper, or even before I start writing the paper, I'll make a list of journals that would be appropriate. So we'll know where to go if we're rejected the first time, or however many times. So that's something that I always keep in mind is the type of paper that will go there, how appropriate the fit is with the aim and scope of the journal. (D1-05)

These experiences illustrate the interplay between personal and social factors in reaching reuse equilibrium. On one hand, researchers used journals as sources of feedback to achieve social reuse equilibrium (i.e., to publish a paper, with standards for publication determined by journal reviewers and editors). On the other, researchers selected publication venues that were appropriate to the personal reuse equilibrium they had obtained (e.g., selecting an outlet where what the researcher produced would be understood and valued as a contribution, or where the standards were not as strict). And researchers had different opinions about whether personal influences should come first (e.g., performing research and then looking for publication venues) or the social (e.g., identifying possible publication venues ahead of time).

A final scenario illustrates both personal and social influences towards reuse equilibrium operating at the same time. A researcher found through the journal review process that their data did not support their desired analysis method. They changed their approach in light of this feedback, but also submitted the paper to a different journal.

> I started from super ambitious when I submitted my paper to the first journal, but realize that [I could not perform their desired analysis]. So I ended with cut down some analysis…based on reviewers' comments. And then I ended up with going down the ladder of the impact of factor at the end. Because once I cut down the most ambitious analysis, because of that kind of methodological issue, then at the same time, the potential impact of manuscripts, my paper gets weaker and weaker. I had to send my manuscript to a journal with a lower impact factor at the end, but the [journal] is a top journal in my field. But at the first I wanted to send the manuscript to [discipline], not [other discipline]. Because I thought that [the paper is more in one discipline than the other], but ended up with sending a manuscript to [journal] because I ended up with considering the topic, not methodology. And I found that [journal] is potentially suitable journal in terms of the topic keyword et cetera, theme. (D-09)

A commonality throughout all of these scenarios is that at the final point (the point of publication in the quotes above, the point where research competence is demonstrated in others), personal and social factors were inextricably linked. In particular, since reaching a publication standard or attaining requisite competence in research are determined by social groups, personal reuse equilibrium must encounter and be influenced in some way by social reuse equilibrium. I present a few more strategies for reaching reuse equilibrium before coming back to this point and explaining its significance in this chapter's summary.

#### *4.4.1.8 Other Strategies*

Some additional ways that researchers reached reuse equilibrium included having confidence in the data (a) because of how the data were produced or the reputation or popularity of the data (D2-03, D3-01, D-02, D-03, D-04, D-07, D-08, D-10, D-11, D-12) or (b) a study and its documentation being so transparent that the researcher felt the researchers "basically did the study for me" (D-12). Researcher D-03 described the experience of reusing data from a repository this way:

> It's almost like, "Are you buying your marijuana at a dispensary, or are you buying it from the person on the corner?" Right? I felt like I was going to a fancy place where it had been curated and was intentional. (D-03)

Researchers also described becoming comfortable reusing data because areas where they lacked knowledge were not their research focus or they were able to use the data to answer most of what they wanted to know (D3-02, D4-01, D-04, D-08).

These scenarios once again illustrate some of the synergies between personal and social reuse equilibrium. Researchers were making their own decisions based on their confidence or comfort in the data. However, at least in the case of selecting data based on the reputation of the

data or the repository, considerations about how the research would be perceived socially were likely factors as well.

### *4.4.2 Summary*

Researchers' strategies for reaching reuse equilibrium had personal and social dimensions. Researchers were influenced by their own determinations about data and also by social norms and requirements, particularly in contexts of publishing or demonstrating competence in research. In several cases, the extent of personal or social influence existed on a continuum that was associated sometimes, but not exclusively, with a researchers' experience. For instance, conducting preliminary research and adapting research questions were strategies that were influenced by personal or social factors, or both, depending in large part on the amount of experience the researcher had reusing data. In contrast, becoming knowledgeable, using data, and interpreting results were strategies used in reaching personal reuse equilibriums.

A key finding of my interviews was that when researchers made individual evaluations of the data and of their research, their ultimate personal determinations of what was "sufficient" were influenced by and designed to meet metrics of success from their social environments. The significance of there being a personal and social reuse equilibrium and of the personal being influenced at some point by the social is that it sets up conditions for a possible gap between the data researchers would like to have to achieve personal reuse equilibrium (i.e., to satisfy personal goals for conducting research) and the data they determine are sufficient to reuse to achieve social reuse equilibrium (i.e., to achieve professional success). The quote from researcher D1-03 above about being able to answer some (not all) parts of your question but still be able to inform the literature is one of many demonstrations from my research that this gap is real.

In the next several sections (4.5, 4.6, and 4.7), I present findings that speak to the existence, size, and importance of the gap between what researchers would like to achieve through data reuse and what is sufficient to them for professional purposes. These findings are undergirded by findings about personal qualities above: namely, that researchers used secondary data to advance their careers, but they also really liked reusing data. Researchers had their own predispositions and propensities for reusing data, which translated into strategies for reaching personal reuse equilibriums (such as becoming knowledgeable about data). And it was on top of personal characteristics and personal goals for reusing data that social norms and requirements for publishing or demonstrating competence were laid.

This is significant because it supports a finding I present in the following sections that researchers wanted to do more than they were able with secondary data, but were limited in various ways. In considering ways to support users of secondary data, it is important to consider whether and how our efforts support the achievement of personal reuse equilibrium, social reuse equilibrium, or both. My contentions are that we can best support the discovery of new knowledge from secondary data by helping researchers achieve their personal reuse equilibriums, and that supporting conversations between data creators and data producers is a key way to do this.

## 4.5 The Relative Priority That Researchers Assign to Different Types of Knowledge About Data in Particular Reuse Instances, and Why

### *4.5.1 Research Question 1a: What Knowledge About Data Is Most Important to Researchers to Reach Reuse Equilibrium and Why?*

In section 3.4.4, I described the challenge, in developing my interview protocol, of reconciling a desire to learn more about circumstances where researchers lacked desired knowledge about data with a desire to gather information about what knowledge was most

important to researchers. I ultimately decided to have two protocols: one to ask researchers who indicated they lacked desired knowledge more about that lack and another to ask researchers who did not lack knowledge about the knowledge that was most important to them. This seemed like the best way to learn the most about my research questions from the interviews.

However, as I reported in section 3.4.4, I found from reading the publications where researchers reported their reuse of data and from interviews themselves that most researchers lacked some kind of knowledge about the data they reused. Since the primary purposes of the interviews were to better understand the nature and extent of knowledge satisficing and learn more about how researchers bounded knowledge about data, I prioritized asking questions having to do with knowledge that researchers desired about data but lacked, rather than questions about the importance of types of knowledge about the data. This meant that when a researcher reported no knowledge lacking but I identified an area of knowledge that appeared from the publication to be lacking (or when they mentioned an area of knowledge during the interview), I asked them about it. In some cases, for instance, if I identified only one area of knowledge lacking from the publication or the researcher's response was relatively succinct, I asked them both about the knowledge lacking and about the knowledge about the data that was most important to them. When I asked about what knowledge was most important, I asked about the top two areas of knowledge.

Ultimately, I obtained information about 23 areas of knowledge that were most important to researchers in 11 out of the 26 interviews (11 out of 15 of the cases where researchers indicated in the survey they did not lack any desired knowledge). For comparison purposes, I coded the knowledge the same way I coded knowledge that researchers reported lacking in the

open-ended survey responses (Table 4.8). Table 4.39 shows the results. I coded some new level 2 categories, which are indicated by gray highlighting.

*Table 4.39 Distribution and Definitions of First and Second Levels of Coding for Areas of Knowledge About Data Researchers Reported as Being Most Important to Them When Deciding to Reuse the Data*

| **Level 1 Knowledge area**<br>Level 2 Knowledge area | **Count** | **Definition used for coding** |
|---|---|---|
| **Data collection** | **13** | |
| Data collection (sampling) | 6 | Sampling methods |
| Data collection (measurement) | 3 | Measurement of phenomena |
| Data collection (process) | 2 | The data collection process |
| Data collection (inclusion) | 1 | What data were collected in general in the study |
| **Data** | **6** | |
| Data (of interest) | 5 | Whether the study had data of interest to the researcher |
| Data (bias) | 2 | Whether the data were biased (e.g., due to non-response) |
| Data (reputation) | 1 | The reputation of the data |
| **Data analysis** | **1** | |
| Data analysis (statistical corrections) | 1 | Whether statistical methods could address bias in the data |
| **Data documentation** | **1** | |
| Data documentation (definitions) | 1 | How measured concepts were defined |
| **Data supplement** | **1** | |
| Data supplement (coverage) | 1 | Data beyond those that were made available |
| **Data supplement** | **1** | |
| Data supplement (detail) | 1 | More detail about available data |
| Total | 23 | |

Researchers most frequently reported knowledge about data collection to be important—knowledge about sample size and sampling methods in particular—followed by how phenomena were measured and details about the data collection process. Several researchers desired to know either what data was collected overall or that the data study included data that would be of interest for their research. Some researchers were concerned about potential bias in the data. For one of these, it was important to know further whether the bias could be addressed using statistical methods. One researcher mentioned the reputation of the data as one of the most important things they wanted to know, and another how concepts that were measured were defined. Only two researchers mentioned knowledge that was not captured in the data (i.e., data supplement) as being one of the most important things to know (one desired knowledge beyond what was captured in the data and one desired more detail about what was captured).

This last finding, that only two researchers mentioned knowledge not captured in the data, stands in contrast to my findings about the frequency of types of knowledge that researchers desired but were lacking (reported on in section 4.2.3.1 and Table 4.8). In that analysis, data supplement was the most frequently reported type of desired knowledge that was lacking.

The discrepancy between these two findings may be explained by the reasons the knowledge in each case was important. The majority of the knowledge areas researchers listed as most important (Table 4.39) are necessary for a basic evaluation of whether the data are suitable to answer the researcher's research questions (e.g., what the sample was, what was measured, whether what was measured was of interest to the researcher, whether there was bias, and whether any bias might be addressed during the analysis). It makes sense that these kinds of questions would be paramount for the researcher to answer and that knowledge that was lacking at the time researchers made the decision to reuse data would be of a second order of importance. In other words, determining how to maximize reuse of the data would come second to determining that the data would be minimally sufficient to fulfill their reuse goals.

If knowledge about data is considered in terms of knowledge of different orders of importance—i.e., first, knowledge that is essential to determining whether to reuse data and second, knowledge that maximizes reuse after that determination is made—the knowledge researchers reported as lacking in the survey (from section 4.2.3.1, Table 4.8) represents a prioritization of knowledge of the second order.

### *4.5.2 Summary*

The areas of knowledge that researchers reported were most important to them most frequently related to knowledge about the collection of data and about the data themselves.

When placed together with my findings from section 4.2.3.1 about the kinds of knowledge researchers desired about data but lacked, findings from this section (4.5) suggest that researchers made decisions about data at two levels (not necessarily sequentially). At the first level, they decided whether data were suitable, at a minimum, to reuse to meet their reuse goals. At a second level, they determined to what degree the data would enable them to maximize the exploration of their phenomena of interest. These findings have important implications for supporting the environment for data reuse that I discuss further in chapter 5.

## 4.6 The Impact of Knowledge Satisficing on the Outcomes of Research and Researchers' Attainment of Their Researcher Goals

The notion in the previous section that researchers made decisions about data at two levels is supported by further findings I obtained from investigating how lacking desired knowledge affected research outcomes. These findings pertain to the second level of decision-making, particularly, where researchers determined the degree to which reusing the data could help them meet their research goals. The findings, in sum, support conclusions that (a) researchers made decisions about data at multiple levels, and (b) while what researchers achieved through reuse of data was often sufficient to satisfy social reuse equilibrium (professional goals) in many cases it was not sufficient to satisfy researchers' personal reuse equilibrium (researchers' own goals for what they what they hope to achieve). This section builds on findings from previous sections to further reveal the existence of a gap between what researchers achieved and what they desired to achieve when reusing data. As described above, this gap points to ways that the research ecosystem might be improved to aid researchers in conducting the research they would like to conduct, and ultimately to more new knowledge from secondary reuse and more refined representations of the phenomena researchers seek to study.

### *4.6.1 Research Question 2d) What Is the Perceived Impact of Knowledge Satisficing? Findings From the Survey (Includes 2di and 2dii)*

I measured the impact of lacking desired knowledge through two questions in the survey. One asked researchers who indicated they lacked desired knowledge whether the lack of knowledge had a negative impact on their research outcomes. Responses were on a five-point Likert scale with the options "not at all," "slightly," "moderately," "very much," and "extremely." The second question asked researchers to enter, for each purpose of data reuse they indicated earlier in the survey (for background purposes, to answer a new question, etc.), how well their reuse of the data met the goals they started with. These two questions formed the basis for my second and third hypotheses, corresponding to my research questions 2di and 2dii. I report on them below.

#### *4.6.1.1 H2. Knowledge Satisficing Has a Negative Impact on Research Outcomes*

In Table 4.40, I show frequencies of the maximum impact researchers reported across the areas of knowledge they indicated where they lacked desired knowledge (up to three per researcher). This means if a researcher reported no negative impact for one area of knowledge, did not answer for another, and reported slight negative impact for a third, the table would reflect "slightly" for that researcher.

In my sample, 154 of the 223 people who lacked desired knowledge by my definition went on to answer the question about impact (69 did not respond). As shown in Table 4.40, 135 (87.7%) of these reported at least some negative impact. On a scale where no impact was one and extreme impact was five, the mean impact was 2.6 (between slightly and moderately) with a standard deviation of 1.05. More than 25% reported a very strong or extreme impact and more than 50% reported at least a moderate impact. Although I could not perform inferential statistics to test this hypothesis because I only asked the question of those who reported lacking desired

knowledge, the results clearly indicate a negative impact on the outcomes of research when desired knowledge is lacking.

*Table 4.40 Distribution of Maximum Negative Impact of Lacking Desired Knowledge on Research Outcomes*

| Maximum impact of knowledge that is lacking | Freq. | Percent | Cum. |
|---|---|---|---|
| Extremely | 6 | 3.9 | 3.9 |
| Very much | 27 | 17.5 | 21.4 |
| Moderately | 54 | 35.1 | 56.5 |
| Slightly | 48 | 31.2 | 87.7 |
| Not at all | 19 | 12.3 | 100.0 |
| Total | 154 | 100.00 | |

Table 4.41 shows the same distribution (degree of negative impact) but tabulates impact for every one of the 297 areas of knowledge for which the 223 researchers reported an impact. There were 290 of these (I did not have a reported impact for seven). Researchers could report more than one knowledge area lacking. Nearly 84% reported that the lack of knowledge had a slight or greater impact on their research outcomes. On a scale where no impact was one and extreme impact was five, the mean impact was 2.5 (between slightly and moderately) with a standard deviation of .99.

*Table 4.41 Distribution of Negative Impact of Lacking Desired Knowledge on Research Outcomes*

| Impact of knowledge lacking | Freq. | Percent | Cum. |
|---|---|---|---|
| Extremely | 6 | 2.0 | 2.0 |
| Very much | 40 | 13.8 | 15.9 |
| Moderately | 96 | 33.1 | 49.0 |
| Slightly | 101 | 34.8 | 83.8 |
| Not at all | 47 | 16.2 | 100 |
| Total | 297 | 100.00 | |

A breakdown of the degree of impact by type of desired knowledge lacking is given in Table 4.42. The table shows the most frequent levels of negative impact are moderate, slight, and not at all. For the knowledge areas of data, data access information, data supplement, data validation, and data reporting, the greatest concentration (percentage) of impact was at

"moderately." The greatest concentration was at "slightly" for data analysis, data collection, data management, and people. Data documentation, data reuse, and other had a more even distribution of impact

*Table 4.42 Impact of Desired Knowledge Lacking by Knowledge Area*

| Knowledge area Level 1 | Negative impact | | | | | | |
|---|---|---|---|---|---|---|---|
| | Extremely | Very much | Moderately | Slightly | Not at all | . | Total |
| Data | 1 | 5 | **12** | 7 | 3 | 0 | 28 |
| | 3.6 | 17.9 | **42.9** | 25.0 | 10.7 | 0.0 | 100 |
| Data access information | 0 | 1 | **2** | 1 | 0 | 0 | 4 |
| | 0.0 | 25.0 | **50.0** | 25.0 | 0.0 | 0.0 | 100 |
| Data analysis | 1 | 6 | 7 | **14** | 5 | 0 | 33 |
| | 3.0 | 18.2 | 21.2 | **42.4** | 15.2 | 0.0 | 100 |
| Data collection | 1 | 4 | 15 | **30** | 16 | 0 | 66 |
| | 1.5 | 6.1 | 22.7 | **45.5** | 24.2 | 0.0 | 100 |
| Data comparability | 0 | **1** | 0 | 0 | 0 | 0 | 1 |
| | 0.0 | **100.0** | 0.0 | 0.0 | 0.0 | 0.0 | 100 |
| Data context | 0 | 0 | 0 | 0 | **1** | 0 | 1 |
| | 0.0 | 0.0 | 0.0 | 0.0 | **100.0** | 0.0 | 100 |
| Data documentation | 1 | **3** | **3** | **3** | 2 | 1 | 13 |
| | 7.7 | **23.1** | **23.1** | **23.1** | 15.4 | 7.7 | 100 |
| Data management | 0 | 0 | 0 | **2** | 0 | 0 | 2 |
| | 0.0 | 0.0 | 0.0 | **100.0** | 0.0 | 0.0 | 100 |
| Data reporting | 0 | 0 | **1** | 0 | 0 | 0 | 1 |
| | 0.0 | 0.0 | **100.0** | 0.0 | 0.0 | 0.0 | 100 |
| Data reuse | 1 | 1 | 1 | 1 | **2** | 0 | 6 |
| | 16.7 | 16.7 | 16.7 | 16.7 | **33.3** | 0.0 | 100 |
| Data supplement | 0 | 19 | **52** | 37 | 16 | 6 | 130 |
| | 0.0 | 14.6 | **40.0** | 28.5 | 12.3 | 4.6 | 100 |
| Data validation | 0 | 0 | **1** | 0 | 0 | 0 | 1 |
| | 0.0 | 0.0 | **100.0** | 0.0 | 0.0 | 0.0 | 100 |
| Other | 0 | 0 | **2** | **2** | 1 | 0 | 5 |
| | 0.0 | 0.0 | **40.0** | **40.0** | 20.0 | 0.0 | 100 |
| People | 1 | 0 | 0 | **4** | 1 | 0 | 6 |
| | 16.7 | 0.0 | 0.0 | **66.7** | 16.7 | 0.0 | 100 |
| Total | 6 | 40 | 96 | **101** | 47 | 7 | 297 |
| | 2.0 | 13.5 | 32.3 | **34.0** | 15.8 | 2.4 | 100 |

Note: First row has *frequencies* and second row has *row percentages.* The most frequently occurring level of impact for each knowledge area is bolded.

#### *4.6.1.2 H3. The Probability of Researchers' Attaining Their Goals for Reusing Data Is Lower in the Presence of Knowledge Satisficing.*

For this analysis, I used responses to survey question 23: "For each of your original reuse purposes, how well did reuse of the data meet your reuse goals?" In this question, researchers were only able to rate their goal attainment for the reuse purpose(s) for which they had earlier indicated they had reused the data. For example, if someone indicated earlier in the survey they

reused the data for background purposes but not to validate data, they would only be able to rate the goal attainment related to their reuse for background purposes. I only counted responses to question 23 where researchers indicated that the purpose(s) for which they reused the data was at least "somewhat" important to their research.

I performed logistic regressions (including the control variables) using knowledge lacking as the independent variable and goal attainment as the dependent variable. I used a binarized version of the original ordinal variable resulting from responses to question 23 (i.e., "much worse than expected," "somewhat worse than expected," "as expected," "better than expected," and much better than expected"). In the binarized version, I combined much and somewhat worse than expected into one category and as expected, somewhat expected and much better than expected in the other. My null hypothesis was that there was no difference in a researcher's ability to achieve their reuse goals whether or not they lacked desired knowledge about data.

I did not find a significant relationship between lacking desired knowledge and the odds of attaining research goals for any reuse purpose, meaning that I could not reject the null hypothesis that there was not a relationship. I show in Table 4.43 the results of logistic regressions. In the regressions for background purposes and purposes to replicate or reproduce research, there were no instances where a researcher lacked desired knowledge and reported meeting their goals worse than expected.

*Table 4.43 Estimated Association Between Variables for Goal Attainment and Lacking Desired Knowledge*

| Goal attainment worse or better than expected | Odds ratio | P>[z] | [95% Conf | Interval] | N |
|---|---|---|---|---|---|
| Background | (omitted) | | | | 57 |
| Validate | 0.125 | 0.079 | 1.276 | 1.409 | 143 |
| New Question | 0.473 | 0.177 | 1.403 | 1.006 | 531 |
| Replicate | (omitted) | | | | 13 |
| Combine | 0.672 | 0.541 | 2.405 | 1.691 | 204 |
| Compare | 0.526 | 0.418 | 2.491 | 1.985 | 164 |
| Theory | 0.684 | 0.429 | 1.754 | 2.556 | 451 |
| Tool | 0.659 | 0.742 | 7.895 | 2.583 | 53 |
| Other | | | | | |

Table 4.44 shows the distribution (first row of each pair) and relative percentages (second row of each pair) of goal attainment responses for each purpose of reuse. As can be seen, no researchers indicated their goal attainment was much worse than expected. “As expected” was the most frequent response across the different reuse purposes. The reuse purposes of new question, combine, compare, and theory were seen by the largest number of researchers and received the most responses overall. In each of these categories, a fair percentage of researchers met their reuse goals either better or much better than expected, in addition to as expected. The categories with the least number of responses were background, validate, replicate, and tool (i.e., to test or develop a tool). These all had smaller percentages of researchers who met their goals better or much better than expected.

*Table 4.44 Distribution and Relative Percentages of Attainment of Research Goals for Different Reuse Purposes*

| Reuse Purpose | Goal attainment | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Much worse | Somewhat worse | As expected | Better than | Much better than | . (missing) | Did not see question | Total |
| Background | 0 | 3 | 128 | 33 | 10 | 3 | 590 | 767 |
| | 0 | .39 | 16.7 | 4.3 | 1.3 | .39 | 76.9 | 100.00 |
| Validate | 0 | 5 | 104 | 43 | 15 | 3 | 597 | 767 |
| | 0 | .65 | 13.6 | 5.6 | 2.0 | .39 | 77.84 | 100.00 |
| New question | 0 | 20 | 374 | 139 | 53 | 13 | 168 | 767 |
| | 0 | 2.6 | 48.7 | 18.1 | 6.9 | 1.7 | 21.9 | 100.00 |
| Replicate | 0 | 2 | 28 | 15 | 2 | 3 | 717 | 767 |
| | 0 | .26 | 3.7 | 2.0 | .26 | .39 | 93.5 | 100.00 |
| Combine | 0 | 13 | 135 | 47 | 29 | 4 | 539 | 767 |
| | 0 | 1.7 | 17.6 | 6.1 | 3.8 | .52 | 70.3 | 100.00 |
| Compare | 0 | 8 | 111 | 53 | 17 | 5 | 573 | 767 |
| | 0 | 1.0 | 14.5 | 6.9 | 2.2 | .65 | 74.7 | 100.00 |
| Theory | 0 | 23 | 310 | 114 | 49 | 8 | 263 | 767 |
| | 0 | 3.0 | 40.4 | 14.9 | 6.4 | 1.0 | 34.3 | 100.00 |
| Tool | 0 | 3 | 35 | 13 | 7 | 1 | 708 | 767 |
| | 0 | .39 | 4.6 | 1.7 | .9 | .13 | 92.3 | 100.00 |
| Other | 0 | 0 | 14 | 3 | 5 | 18 | 727 | 767 |
| | 0 | 0 | 1.8 | .39 | .65 | 2.4 | 94.8 | 100.00 |

*Note*. First row has *frequencies* and second row has *row percentages*

#### *4.6.1.3 Findings About Reasons Lacking Desired Knowledge Had (or Did Not Have) a Negative Impact*

In addition to asking about the kinds of knowledge that were lacking and the amount, importance, and negative impact of knowledge lacking, I also asked about the reasons for the negative impact (or lack thereof). The specific question I asked was, "Did a lack of knowledge about the data in these areas negatively affect the achievement of your desired research outcomes?" I coded responses to the question inductively and iteratively, as I did responses for knowledge areas in section 4.2.3.1. Of the 156 researchers who entered types of knowledge they desired, 117 entered reasons the lack of knowledge did or did not negatively impact the outcomes of their research. Researchers had the opportunity to enter one reason for each type of knowledge they reported missing, and individual entries sometimes included multiple reasons, so there was frequently more than one reason reported.

In all, I coded 229 reasons for impact for the 117 researchers. The remaining 39 researchers either did not answer the question, the response they gave was not relevant to the question (for instance, the researcher explained why they had not been able to obtain the knowledge as opposed to the reason the knowledge did or did not have a negative impact), or I was not able to interpret the response. These appear under the code "Other" in tables 4.46, 4.47, and 4.48 below. Table 4.45 shows the number of researchers I coded as having indicated different numbers of reasons. I coded the largest number of researchers as indicating only one reason but there were substantial numbers at two and three reasons as well.

*Table 4.45 Number of Coded Impact Reasons for Different Numbers of Researchers*

| Number of coded reasons | Number of researchers | Percent of researchers |
|---|---|---|
| 1 only | 61 | 39.10 |
| 2 only | 45 | 28.85 |
| 3 only | 42 | 26.92 |
| 4 only | 7 | 4.49 |
| 5 only | 2 | 1.28 |
| 6 only | 1 | .64 |
| Total | 156 | 100 |

As with knowledge, I coded two levels of impact reasons, one broader level with eight categories and one more granular level with 80 categories to preserve the range of reasons lacking knowledge affected researchers' work. In some ways, given the smaller sample size (156) and the likelihood that researchers did not report all of the knowledge they lacked or the reasons the lack affected their research (see the discussion in section 3.4.8.5), it is the range itself that is of interest. The distribution of categories of impact reasons along with the definitions I used to code each category are shown in Table 4.46.

Each of the eight broad categories in the table indicates the kind of impact that lacking knowledge had, and the granular categories provide additional details. For instance, in the "limited" category, the researcher's analysis was limited in some way. The particular way is given in parenthesis (e.g., analysis of the control variable was limited, the depth of inquiry was

limited, etc.). Also coding definitions follow this pattern, I have included the definitions to ensure the clarity of each code.

*Table 4.46 Distributions and Definitions of First and Second Levels of Coding for Areas of Knowledge Researchers Reported as Lacking*

| Code | Count | Coding definition |
|---|---|---|
| **Limited** | **90** | **Lack of desired knowledge** |
| Analysis (general) | 10 | Limited analysis of the data |
| Analysis (control variable) | 9 | Limited analysis of a control variable |
| Analysis (depth of inquiry) | 4 | Limited depth of inquiry |
| Analysis (determination of sample size) | 1 | Limited ability to determine sample size |
| Analysis (identification of patterns) | 1 | Limited the identification of anticipated patterns |
| Analysis (not done) | 13 | Made it impossible to pursue a line of inquiry |
| Analysis (quality) | 2 | Limited the quality of the analysis |
| Analysis (rigor) | 1 | Limited the rigor of the analysis |
| Analysis (temporal order) | 1 | Limited ability to assess temporal order of data |
| Analysis (variables to include) | 1 | Limited ability to determine what variables to include in the analysis |
| Analysis (variations) | 1 | Limited ability to assess variations in data |
| Comparability (data) | 1 | Limited comparability of data from reanalysis with original data |
| Comparability (definitions) | 1 | Limited ability to compare definitions used with original or other prior definitions |
| Comparability (results) | 3 | Limited comparability of results of reanalysis with original research or other prior research |
| Detail (data) | 4 | Limited detail of the data |
| Explanation in research results | 3 | Limited the ability to explain the original research in the research results |
| Explanation of research results | 2 | Limited the ability to explain the research results |
| Interpretation (data) | 1 | Limited interpretation of the data |
| Knowledge (general) | 1 | Necessary knowledge was not available |
| Research results (claims) | 3 | Limited the claims that could be made |
| Research results (generalizability) | 1 | Limited generalizability of the re-analysis results |
| Research results (impact) | 1 | Limited the impact of the results |
| Research results (increased limitations) | 2 | Increased limitations of the study |
| Research results (inferences) | 1 | Limited inferences that could be made from the results |
| Research results (interpretation) | 2 | Limited interpretation of findings or results |
| Research results (prediction) | 1 | Limited predictions that could be made from the results |
| Research results (Research questions) | 5 | Limited research questions or aspect of researchers' questions that could be explored |
| Research results (robustness) | 1 | Limited the robustness of the study |
| Research results (test theory) | 3 | Limited the ability to test a theory |
| Understanding (data) | 4 | Limited the ability to understand the data |
| Understanding (general) | 1 | Limited the ability to understand some aspect of the research |
| Usefulness (data) | 3 | Limited the usefulness of the data |
| Certainty (accuracy) | 1 | Introduced uncertainty about accuracy of the data |
| Certainty (precision) | 1 | Introduced uncertainty about precision of the data |
| **Compensated** | **46** | **The researcher compensated for the lack of knowledge…** |
| Advisor and colleagues | 2 | By obtaining help from an advisor or colleague |
| Collected data | 1 | By collecting own data |
| Data analysis | 7 | Through analysis of the data |
| Data collector | 2 | Through information from the data collector |
| Data | 7 | By using other data in the dataset |
| Documentation | 2 | Using data documentation |
| Imputation | 3 | Through imputation of data |
| Limitations | 5 | By documenting as limitations |
| Literature | 3 | Using knowledge from literature |
| Outside data | 6 | By using data outside of the dataset |
| Own knowledge | 2 | Using own knowledge |

| | | |
|---|---|---|
| Research | 1 | By conducting additional research |
| Counterbalanced (data) | 5 | Lack of knowledge was counterbalanced by other benefits of the data (e.g., size) |
| **Opportunity lost** | **28** | **Knowledge that was lacking…** |
| General | 6 | Would have improved some aspect of reuse |
| Additional insight | 8 | Would have given greater insight |
| Context | 3 | Would have provided more context |
| Definitive information | 1 | Would have improved definitiveness of data |
| Implications | 1 | Would have enhanced understanding of policy implications |
| More granular analysis | 1 | Would have enabled more granular analysis |
| More meaningful conclusions | 2 | Would have enabled more meaningful conclusions |
| Questions investigated | 1 | Would have expanded the scope of the research questions |
| Reliability | 1 | Would have increased reliability of the data |
| Reuse guidance | 2 | Would have provided guidance for reusing the data (this is not knowledge about data but about reuse) |
| Rigor | 1 | Would have enabled a more rigorous test |
| Sophisticated research questions | 1 | About variables would have enabled more sophisticated research questions |
| **Complicated** | **23** | **Lack of desired knowledge…** |
| General | 4 | Added difficulty to an activity (still doable, but more difficult) |
| Generalizability | 1 | Made it more difficult to determine generalizability |
| Implications | 1 | Made it more difficult to provide implications |
| Interpretation | 2 | Made it more difficult to interpret data or research results |
| Labor | 1 | Required re-doing analysis |
| Thought | 1 | Required re-thinking analysis |
| Delay (analysis) | 6 | Delayed analysis |
| Delay (project) | 7 | Delayed the project |
| **Satisfactory** | **15** | **The researcher determined that existing knowledge was satisfactory based on their own determination or an outside standard (i.e., the results were publishable, the database was reputable)** |
| Database reputation | 1 | Despite any deficiencies, the database was accepted as a source of data in the community of researchers) |
| Obtained sufficient knowledge | 9 | Obtained sufficient knowledge of the data |
| Satisfactory for publication | 5 | The outputs of the research achieved a publishable standard |
| **Adjusted** | **12** | **The researcher …** |
| Analysis | 3 | Adjusted analysis method |
| Objectives | 9 | Adjusted research objectives (including research questions) |
| **Obtained** | **11** | **The researcher…** |
| General | 4 | Obtained the desired knowledge |
| Eventually | 1 | Eventually obtained the knowledge, with some cost as to time or effort |
| Eventually, mostly | 5 | Eventually obtained most of the desired knowledge, with some cost as to time or effort |
| Mostly | 1 | Obtained most of the desired knowledge |
| **Knowledge not important** | **4** | **The knowledge…** |
| Knowledge not important | 3 | Was not important |
| Knowledge of limited usefulness | 1 | Was of limited usefulness |
| **Other** | **87** | |
| Did not answer | 68 | |
| Answer not relevant | 18 | |
| Could not interpret | 1 | |
| Total | 316 | |

The most frequent impact that lacking desired knowledge had on researchers was limiting the research they wished to do. This makes sense in light of my finding that the most frequent kind of knowledge researchers lacked was knowledge "of" the data (i.e., knowledge about the

sample being studied, which I categorized as data supplement (see section 4.2.3.1) as opposed to knowledge about the data themselves,). That is, when researchers did not have knowledge they desired about the sample, it makes sense that their research would be limited. Through further analysis (data not shown) I found that in 48 out of the 90 cases where the reason for impact was limited (53.3%), the area of the knowledge desired was data supplement (the next largest areas were data collection (12) and data analysis (11)).

Not having the desired knowledge limited the analysis that researchers could do (the largest "limited" category, with 44 entries), the results the researcher was able to obtain (20 entries), the ability to compare the data or results with other data (5 entries), the ability to understand the data or research (5 entries), the ability to explain the original research (3 entries) or the researcher's results (2 entries). It also limited the ability to interpret the data and the usefulness of the data, and introduced uncertainty about the data.

The next largest category, compensated, had 46 entries and only two groups of reasons of impact: researcher's compensating for the lack of knowledge in some way, or using other characteristics of the data to compensate for or counterbalance the knowledge that was lacking. Researchers compensated for their lack of desired knowledge by seeking help from advisors and colleagues, collecting their own data, using data analysis techniques (such as sensitivity analysis, constructing new variables, using techniques to validate data or deal with missing data), contacting original data creators to obtain knowledge, using other parts or features of the data, getting a good enough idea from data documentation to satisfy questions about the data, and others as defined in Table 4.46.

A number of researchers reported that they lacked knowledge that, if they had been able to obtain it, would have improved some aspect of their research. There were 28 of these entries

that I coded as "opportunity lost." The lost opportunities extended from lost insights to reduced ability to contextualize results, discuss implications, and formulate meaningful conclusions. Some researchers also reported lost opportunities to perform more rigorous or granular analyses, explore broader or more sophisticated research questions, or to base their conclusions on more definitive data.

The next largest number of entries fell into the category of "complicated" (23 entries). These reasons for impact encompassed a variety of scenarios that made the researcher's work more difficult or time consuming. In some cases, the delay came from waiting for access to data. In others, it took researchers time to find the knowledge they needed in literature or data documentation, or to obtain knowledge from the original data creators.

"Satisfactory" is the next category of reasons for impact (15 entries). In these cases, researchers did not acquire all of the knowledge they desired, but reported that the knowledge they did obtain turned out to be sufficient for their purposes.

"Adjusted" (12 entries), "obtained" (11 entries), and "knowledge not needed" (4) are the final three categories. These encompass situations where researchers gave the following as reasons for the reported impact: adjusting their methods or objectives, having obtained all the knowledge they originally desired, or the knowledge they desired not being important or of limited usefulness.

Table 4.47 shows the amount of knowledge researchers obtained of what they desired for each level 1 reason of impact. The table excludes missing values, of which there were 27. Like Table 4.11, the large standard deviation indicates that researchers obtained a wide range of the knowledge they desired (some closer to the amount they desired and some farther away).

Those who reported that the lack of knowledge limited their research in some way have the lowest median knowledge obtained, at 40.5. Those in the categories of opportunity lost, satisfactory, adjusted, and other cluster together with a median around 50, or half their desired knowledge. Those who compensated for the lack of knowledge in some way or for whom the lack of knowledge complicated their research achieved a median amount of knowledge around 70. Those who obtained the desired knowledge at some point between their decision to reuse the data and the conclusion of their research reported a median knowledge attainment of 91.

My question in the survey implied but was not explicit that the amount of knowledge should be interpreted as the amount at the time of the reuse decision (as opposed to the amount by the end of the research) and researchers interpreted the question differently. It is notable that the maximum knowledge obtained was 100 for all except the categories of satisfactory and adjusted, which align with what might be expected. That is, one would expect lower attained amounts from those who explained the impact of the lack of knowledge by the adjustment of their methods or objectives, or by obtaining an amount of what was desired that was satisfactory.

*Table 4.47 Descriptive Statistics on the Amount of Knowledge Obtained for Each Reason for Impact*

| | N | Min | Max | Mean | STDev | 25th | Median | 75th | 95th |
|---|---|---|---|---|---|---|---|---|---|
| Limited | 82 | 0 | 100 | 40.30 | 31.62 | 5 | 40.5 | 70 | 92 |
| Compensated | 45 | 0 | 100 | 62.09 | 25.88 | 50 | 70 | 81 | 95 |
| Opportunity lost | 24 | 0 | 100 | 48.25 | 32.67 | 20 | 50 | 68 | 100 |
| Complicated | 21 | 5 | 100 | 58.33 | 34.90 | 31 | 69 | 90 | 95 |
| Satisfactory | 15 | 0 | 80 | 47.27 | 23.50 | 27 | 50 | 65 | 80 |
| Adjusted | 11 | 0 | 71 | 35.55 | 28.02 | 0 | 47 | 60 | 71 |
| Obtained | 11 | 70 | 100 | 90.45 | 10.90 | 80 | 91 | 100 | 100 |
| Other | 76 | 0 | 100 | 47.76 | 29.39 | 21.5 | 50 | 71 | 100 |

Table 4.48 shows the distribution of level 1 reasons researchers gave for the impact of knowledge lacking and the degree of negative impact the lack of knowledge had on the outcomes of their research. The row percentages are shown in the second row for each reason. The reason the total in Table 4.48 is 316 and the total in the Table 4.42 (showing the distribution of the

impact of lacking knowledge by category of desired knowledge lacking) is 297 is that some researchers described multiple impacts for a single area of knowledge. The total numbers are largely comparable, however, and the most frequent levels of negative impact are once again moderate, slight, and not at all. That is, most researchers experienced some degree of, but not severe, negative impact from lacking knowledge about data.

It can be seen in Table 4.48 that limited and other are the only reasons of impact where the largest proportion of responses for that reason were associated with a moderately negative impact (45.56% and 35.63% respectively). The largest proportion of coded responses for compensated, complicated, opportunity lost, and satisfactory are associated with a slight negative impact; adjusted is equally split between a slight and moderate negative impact. The largest proportion of responses coded as knowledge not important and obtained are associated with no impact on the research.

*Table 4.48 Distribution of the negative impact of the lack of desired knowledge by reason for the impact (N=156)*

| Impact reason Level 1 | Negative impact of lacking desired knowledge | | | | | | |
|---|---|---|---|---|---|---|---|
| | Extremely | Very much | Moderately | Slightly | Not at all | (missing) | Total |
| Adjusted | 0 | 1 | 5 | 5 | 1 | 0 | 12 |
| | 0 | 8.3 | 41.7 | 41.7 | 8.3 | 0 | 100.0 |
| Compensated | 0 | 1 | 11 | 22 | 12 | 0 | 46 |
| | 0 | 2.2 | 23.9 | 47.8 | 26.1 | 0 | 100.0 |
| Complicated | 0 | 2 | 8 | 13 | 0 | 0 | 23 |
| | 0 | 8.7 | 34.8 | 56.5 | 0.0 | 0 | 100.0 |
| Knowledge not important | 0 | 0 | 0 | 1 | 3 | 0 | 4 |
| | 0 | 0.0 | 0.0 | 25.0 | 75.0 | 0 | 100.0 |
| Limited | 4 | 17 | 41 | 23 | 5 | 0 | 90 |
| | 4.4 | 18.9 | 45.6 | 25.6 | 5.6 | 0 | 100.0 |
| Obtained | 0 | 1 | 2 | 3 | 5 | 0 | 11 |
| | 0 | 9.1 | 18.2 | 27.3 | 45.5 | 0 | 100.0 |
| Opportunity lost | 0 | 6 | 7 | 11 | 4 | 0 | 28 |
| | 0 | 21.4 | 25.0 | 39.3 | 14.3 | 0 | 100.0 |
| Other | 2 | 12 | 31 | 22 | 13 | 7 | 87 |
| | 2.3 | 13.8 | 35.6 | 25.3 | 14.9 | 8.1 | 100.0 |
| Satisfactory | 0 | 0 | 2 | 8 | 5 | 0 | 15 |
| | 0 | 0 | 13.3 | 53.3 | 33.3 | 0 | 100.0 |
| Total | 6 | 40 | 107 | 108 | 48 | 7 | 316 |
| | 1.9 | 12.7 | 33.9 | 34.2 | 15.2 | 2.2 | 100.0 |

*Note*. The first row shows frequencies and the second row shows row percentages.

My findings about the reasons a lack of desired knowledge impacted researchers' work align with findings from section 4.5 that there were multiple levels on which researchers engaged with data. On a first level, they made basic decisions about whether the data were suitable to answer their research questions at a minimum degree of sufficiency. At a second level, through engagement with the data, they determined to what exact degree the data could meet their desired aims.

At this second level, researchers were the most impacted by lacks of knowledge that limited their analysis or the scope of findings they could achieve with the data (some of which they expressed in terms of lost opportunities). In other cases, where researchers obtained slightly more knowledge and were less impacted overall, researchers were able to compensate for knowledge they lacked—sometimes by adjusting their research questions—or persevere through complications that increased the time or effort required to conduct their research.

Researchers who reported that the data ultimately met their needs without any compensation—i.e., in the category of satisfactory—fall into this lower level of impact even though they obtained similar amounts of knowledge overall to those who reported losing an opportunity or that a lack of knowledge limited their research. Researchers who adjusted their research in response to a lack of knowledge did not report much more desired knowledge obtained or much lower negative impact, overall, than those whose research was limited by a lack of knowledge. In fact, their statistics were worse in many regards. And even among those who reported obtaining all the knowledge they desired, half reported that there was still a negative impact of the desired knowledge lacking on their research outcomes (though in these cases, the negative impact came from the additional time needed to obtain the knowledge, or the

fact that obtaining the knowledge did not, in fact, allow them to complete the research they desired).

These results indicate that even when researchers made adjustments in order to have satisfactory outcomes, what they achieved through their research was still less than what they hoped initially (considering achievement in terms of what they might have been able to do had they obtained all the knowledge they originally desired in the satisfactory and adjusted categories, and even when obtaining all the knowledge they desired in the obtained category). This finding gives an important window into strategies for enhancing reuse of social science data, which I return to in chapter 5.

### *4.6.2 Research Question 2dii: What Is the Perceived Impact of Knowledge Satisficing on Researchers' Achievement of Their Goals for Reusing Data? Findings From the Interviews*

In the interviews, I asked researchers to explain answers they gave to a survey question asking how well their reuse of the data met their reuse goals. The 26 interviewees reported a total of 61 reuse purposes that were at least somewhat important to their research (in the survey, each researcher could select multiple purposes). There were eight researchers whom I did not ask about any of their survey responses. This happened because the question was toward the end of the interview and I was sometimes running short on time. Nineteen researchers, then, across 40 reuse purposes, gave a total of 53 explanations in the interviews for why they reported their reuse of the data did or did not meet their reuse goals. Researchers sometimes gave more than one explanation of their survey responses. For instance, a researcher might have felt the data met their goals as expected both because the data allowed them to answer their research question and because the data allowed them to further their career.

In Table 4.49, I present these explanations. "Did not ask" represents instances where researchers indicated a level of goal attainment in the survey, but I did not ask about it in the

interviews. No researchers reported that reuse of the data met their goals much worse than expected.

*Table 4.49 Researcher Explanations for Why the Data Met Their Reuse Goals at Different Levels*

| Goal attainment | Explanation of goal attainment | Count of responses | Researcher(s) |
|---|---|---|---|
| **Somewhat worse** | | **2** | |
| | Reuse did not meet the researcher's expectations | 1 | D1-03 |
| | Did not ask | 1 | D-01 |
| **As expected** | | **47** | |
| | Reuse helped the researcher advance their career | 3 | D2-02, D4-02 |
| | Reuse boosted the researcher's confidence in reusing data | 1 | D2-02 |
| | Did not ask | 12 | D1-01, D1-02, D1-03, D2-01, D2-04, D3-01, D-01, D-10 |
| | Reuse allowed the researcher to produce interesting findings | 1 | D2-02 |
| | There was a large proportion of quality, verifiable data | 4 | D-13 |
| | Reuse met the researcher's needs | 4 | D1-01, D4-02, D-12 |
| | The data were the only data available | 3 | D3-02 |
| | The researcher had prior expectations (from advisor) | 5 | D-02, D-05 |
| | The researcher had prior expectations (from experience) | 5 | D2-03, D3-02, D-07 |
| | The researcher had prior expectations (from investigation prior to reuse) | 1 | D-08 |
| | The researcher had prior expectations (from knowledge of data) | 1 | D-07 |
| | The researcher had prior expectations (from literature) | 4 | D-04, D-11 |
| | The researcher had prior expectations (source undefined) | 2 | D-06 |
| | Reuse resulted in a publishable paper | 1 | D1-04, D-05 |
| **Better** | | **22** | |
| | The data had proven to be a good bet over time for producing results | 3 | D1-01, D1-04 |
| | The data opened new lines of inquiry | 1 | D-03 |
| | Did not ask | 9 | D1-02, D3-01, D-01, D4-01 |
| | Reuse exceeded the researcher's expectations | 3 | D-06, D-09 |
| | There was a large proportion of quality, verifiable data | 1 | D-13 |
| | The researcher's lower expectations were exceeded | 1 | D-11 |
| | Reuse met the researcher's needs | 1 | D3-02 |
| | The data were the only data available | 1 | D3-02 |
| | Reuse resulted in a publishable paper | 2 | D1-04, D-05 |
| **Much better** | | **8** | |
| | Did not ask | 4 | D1-02, D3-01, D-01 |
| | Reuse exceeded the researcher's expectations | 3 | D-09 |
| | Reuse met the researcher's needs | 1 | D1-05 |
| Total | | 79 | |

*Note*. "Did not ask" represents instances where researchers indicated a level of goal attainment in the survey but I did not ask about it in the interviews. The total of 79 responses in the table includes instances where I did not ask. There were a total of 53 responses excluding these.

As the table, shows, many of the explanations for why data met reuse goals as expected had to do with researchers' prior expectations about the data, their previous experience with the data or data reuse, guidance they received from advisors or mentors, or what others had found reusing the data in prior literature. Researcher D-04 expressed it this way:

> You know, I've used the data in one way or another over decades, and [collaborator name] has as well. So we didn't go in blind. We knew what the [data] were all about. We had strong suspicions about what we would find in the [data], and those suspicions were borne out. So, it wasn't a surprise. We were very familiar with the data we were using. (D-04)

Researcher D-11 had a similar response but noted also that while the data met their expectations "as" expected, they had high expectations to begin with:

> I was aware of the substance of the data, which is why I think I responded as expected, because I knew other people had done it. I presumed I could do the same thing, and indeed I generally could. I could use that for the purpose, and therefore, that met my expectations. But I will say I guess it depends on, also, I did have- and this is all relative- but I had relatively high expectations and they were met, which is a good. (D-11)

This articulation of goal attainment in relation to high or low expectations was echoed by others. In researcher D-07's case, the data did not meet their needs as they might have wanted, but they understood that this was "par for the course" when reusing data (and doing research in general):

> I had a preconceived notion what the [data] would be in general, even though there was some missing data, or some sub questions we couldn't use…We had to dump those questions, so it didn't meet my expectations. That's life when you're conducting research, you know, even when you're out doing the primary data collection, things don't work out as planned. I don't say worked out worse than expected. That's par for the course. (D-07)

Similarly for researcher D-06, they expected that it would be difficult to get all the information they needed to compare and combine the data the reused with other data, and it was indeed as hard as they expected.

In some cases (D-02, D-11), researchers' explained that their expectations for reuse were set by or in relation to the expectations of advisors. Researcher D-02 said:

> so I would say my expectations were set by my advisors from the start. This wasn't built up of, "Oh, this is going to be groundbreaking"…It was kind of like, "We're not setting the bar super high. We're achieving what we need to achieve as a mid-tier milestone." So it met those goals. It didn't fall short. It didn't not allow me to achieve the bar. But we kind of knew that the purpose of this project was, "Dip your toes in the water of independent research. Show that you can take a project to fruition, from start to finish, on your own…But this isn't the time to strive for gold, right? You're early on your career. Let's, you know, go through the motions and make sure that you can do this. (D-02)

This reliance on advisors echoes findings from section 4.4.1.7 where researchers used feedback and guidance (including from advisors) to negotiate gaps in knowledge in about data.

There were some instances where a researcher reported in the survey that the data met their needs better or much better than expected and their explanation was simply that the data allowed them to do what they wanted to do (D3-02 D1-05). In the case of researcher D3-02, meeting their needs meant contributing something new (their purpose was to answer a new question); researcher D1-05 did not recall why they said the met their needs better than expected. For the most part, however, researchers indicated the data met their goals better or much better than expected when the data exceeded their expectations (however low or high), or there was something else about the data that was particularly noteworthy.

For instance, both researchers D1-01 and D1-04 noted that the data had been advantageous to work with over a period of time (not just the one instance I was interested in). When discussing their goals for investigating a new question, researcher D1-01 said:

> I think I said better than expected because for me in my mind, there's so many questions I could ask of this data set. I mean, I've used it. I have published with this data set...And have a [grant] from [institution]. And so for me, I think I put better than expected because I've been able to use it, not just to publish articles, but to get a grant and so different questions arise and I'm continuing to use it with other people (D1-01).

In other cases, the large amount of quality data available exceeded the researcher expectations (D-13), the data were the only data available to investigate their questions (D3-02), or the data fulfilled the researcher's needs for a publishable paper. This last was discussed by researcher D1-04:

> Well, I guess I define that in terms of were you able to do an analysis that was then accepted by four critical reviewers? Were you able to get published? That's my criteria here in terms of is the data serving you well, okay? Because if you have an idea and you can't make the case for that imperfect variable and you sent this out to several journals, then I don't think it served me well. But this dataset has served so many people well, served me well. It is not perfect, but it is good enough to advance the ideas that you're trying to get after. (D1-04)

What stands out the most in these findings is the low number of researchers (only three) whose explanation about meeting their goals had to do with achieving findings that exceeded their expectations (D-06, D-09, D-11). Most researchers had a pretty good idea of what they would be able to achieve with the data before they reused them and were able to meet those expectations (even when lacking desired knowledge about data, as 14 of the 18 respondents did) but not exceed them.

This finding, taken together with the finding from section 4.6.1.3 that the most common reason that lacking knowledge had a negative impact was the limits it placed on their research, and the finding from section 4.2.3.1 that the most common kind of knowledge researchers lacked about data was supplementary knowledge, suggest that while researchers were able to reuse data to meet professional goals and expectations for reuse, they were more rarely able to reuse data to investigate their topics of interest as fully as they might have liked.

### *4.6.3 Summary*

In the survey, I found that 135 (87.7%) of 154 researchers who responded indicated that a lack of desired knowledge had at least some negative impact on the outcomes of their research. This confirmed my second hypothesis that lacking desired knowledge negatively impacts the outcomes of research.

Of the 135 researchers who reported a negative impact, 117 provided 229 individual explanations for why a specific lack of knowledge did or not have a negative impact on their research. These were, with counts included, as follows:

- limited (90)
- compensated (46)
- opportunity lost (28)
- complicated (23)
- satisfactory (15)
- adjusted (12)
- obtained (11)
- knowledge not important (4)

The lowest median knowledge obtained was in the category limited (40.5), which was also the most frequent category coded. Limited was also the category with the highest incidence of moderate negative impacts on research outcomes (41) and had a fair number of very (17) or extremely (4) negative impacts. Most other categories had their concentration of impact at the slight level, though some others had significant numbers with both slight and moderately negative impacts: complicated (13 and 8), adjusted (5 and 5), and opportunity lost (11 and 7).

I did not find a significant relationship between lacking desired knowledge and researchers' ability to achieve their reuse goals. In terms of how well researchers' reuse of the data met their research goals. However, across reuse purposes, researchers reported meeting their goals as expected most frequently. The reuse purposes of new question, theory, combine, and

compare, had the largest percentages of researchers who met their reuse goals better or much better than expected in addition to as expected.

My findings from the survey about the reasons a lack of desired knowledge impacted the outcomes of research support findings from section 4.5 about which kinds of knowledge about data were most important to researchers. Findings from section 4.5 indicated that researchers made decisions about data on two levels: a first level in which they made basic decisions about whether the data were suitable to answer their research questions, and a second level where, through engagement with the data, they determined to what exact degree the data could meet their desired aims. Findings in section 4.6.1.3 add support this conclusion. On one level, researchers decided the data were sufficient to reuse (since I selected researchers to interview based on successful reuse) even if they felt an opportunity to do more with the data was lost. On a second level, researchers clearly made efforts, though they lacked knowledge they desired when they decided to reuse the data, to compensate for the knowledge they lacked or persevere through complications that required time and/or effort to address. It is possible that some researchers did not decide to reuse the data until they felt they had compensated enough or persevered long enough to know the data would be sufficient. However, this is unlikely in all or even a majority of cases given my findings about the care researchers take to ensure ahead of time that their reuse of data will be successful. The fact that some researchers felt a research opportunity was lost itself indicates that researchers evaluated whether the data were sufficient (since they reused them anyway) and the scope of what they might have been able to achieve.

I conducted further investigation into the reasons a lack of desired knowledge impacted research outcomes in the interviews. My results showed that in cases where researchers eventually determined their existing level of knowledge about the data was sufficient (my

"satisfactory" category), the amount of knowledge they obtained about the data was not much more than in cases where researchers reported their overall research was limited. Furthermore, in cases where researchers adjusted their research questions or analysis techniques, a significant proportion of researchers (and similar to those whose knowledge was limited) reported a moderate impact on their research. These findings indicate that even when researchers made adjustments to achieve satisfactory outcomes, what they achieved through their research was still less than they might have hoped initially.

This conclusion is supported by results from the interviews where I found that only three of 18 researchers indicated that the findings they obtained reusing the data exceeded their initial expectations. Most researchers, though many lacked desired knowledge about data, had a good idea of what they would be able to achieve with the data ahead of time (from their own experience with the data or with data reuse in general, based on guidance from others, or based on what others had found in previous research) and were able to meet those expectations but not exceed them.

Together with findings from section 4.6.1.3 (that the most common reason that lacking knowledge had a negative impact was the limits it placed on their research), and from section 4.2.3.1 that the most common kind of knowledge researchers lacked about data was supplementary knowledge, my results suggest that (a) researchers made initial assessments about whether data would minimally fulfill their goals for research and (b) researchers who lacked desired knowledge about data and were ultimately successful in their reuse of data commonly met these initial goals for reuse as they expected, however (c) these researchers rarely exceeded their expectations or investigated their topics of interest as fully as they would have liked (since

they achieved their goals as expected in the face of lacking knowledge they desired about the data).

These are key findings that I return to in the discussion (chapter 5), but I provide a brief synthesis of findings and preview here. My findings in section 4.2.2 indicated that researchers did not see reusing data as an optimal strategy for conducting their research. Findings from sections 4.2.1 and 4.2.3 indicated that researchers lacked knowledge they desired about data and the kinds, amounts, and importance of the knowledge they lacked. Section 4.3 revealed factors that affected researchers in reaching reuse equilibrium, and findings in section 4.4 further refined the notion of reuse equilibrium. On one hand, my findings showed that reuse equilibrium existed in the context of broader goals in research (e.g., to publish or demonstrate research competence). On another, they showed that different personal and social reuse equilibriums existed, interacting and playing roles to different degrees in researchers' ultimate decisions about whether to reuse data.

Findings in sections 4.6.1 and 4.6.2 added additional support and nuance to these findings, revealing the particular ways that researchers were affected by limitations in the data they reused. In doing so, they give an idea of the nature of the gap that can exist between researchers' personal reuse equilibriums and social reuse equilibriums. While the reuse equilibriums researchers reach might be acceptable from the point of view of standards such as making a contribution or demonstrating competence in research (i.e., social reuse equilibrium), they represent the meeting of a bar rather than achievement of a potential. While that potential may not be "optimal" (because it would be achieved through data reuse rather than original data collection), the existence of a gap between what is achieved and what researchers desired to achieve is indicative of a gap in the research ecosystem and a way the design of the research

ecosystem might be improved so that the research conducted is a more refined representation of phenomena in the world.

In the next section, I present findings about a mechanism I identified and strategies researchers suggested that speak directly to questions of design and ideas for lessening this gap and improving the landscape for data reuse in the social sciences.

### 4.7 Conversations

In section 1.4.4, I introduced Haraway's argument that recognizing the agency of 'objects' of knowledge—referencing women, in particular, but also other 'objects' of study in the sciences—transforms the way we understand the knowledge we gain from research. Rather than knowledge being 'objective,' she argues that faithful accounts of the world depend on "a power-charged social relation of 'conversation'" (Haraway, 1991, p. 198).

I paraphrased Haraway's ideas by saying that if we understand all knowledge to be partial and situated, there is no one perspective from which objects of knowledge ("data," in my case) can be totally known. Such a recognition requires a consideration of other perspectives and the factors that lead to those perspectives. This consideration I understand to be the 'conversation,' among 'fields' of situated "interpreters and decoders" that Haraway references (Haraway, 1991, p. 198), on which she argues faithful accounts of the world depend.

In section 1.4.5, I also described two applications I make to my research of Simon's framework for understanding natural and "artificial" adaptive systems. The second is my use of his notion of homeostasis to describe "reuse equilibrium," which exists when a researcher determines that a particular body of data is sufficient to reuse and decides to reuse the data to accomplish a specific purpose. The first, which I have not discussed until now, is to understand data that exist in an archive as an artifact that people design to achieve certain purposes (in other

words, an "artificial" adaptive system), including supporting new discoveries and improving trust in science. A key characteristic of these systems is that when they are not well-adapted to the purpose for which they are being used, flaws in the adaptation "show through" (for instance, as I extend Simon's points to argue, in the form of inaccurate results or representations if methods of a research project are not well-adapted to measure phenomena in the environment in which the research is conducted).

In this section, I present findings in three areas that do two things. They first reveal that conversations that occur currently among stakeholders in reuse are crucial to determining the composition of data and the way data reuse proceeds—revealing conversations as fundamental to determining the extent of the possibilities for reusing data. Second, the findings speak to ways that the design of data, and the design of the systems that support data reuse, might be improved to facilitate and enhance research conducted with secondary data, including through designing to enable conversations among participants in the data reuse ecosystem.

The first two areas of findings relate to research question 1d: how do researchers obtain the knowledge they desire? The second set relates to research question 2e: what do researchers believe could mitigate knowledge satisficing?

### *4.7.1 Research Question 1d: How Do Researchers Obtain the Knowledge They Desire? Findings From the Survey*

#### *4.7.1.1 Findings about Source of Knowledge*

In the survey, I asked researchers who indicated they lacked some knowledge they desired to select from a list the source of knowledge that was most important to them. Figure 4.2 shows the sources researchers used when seeking to obtain different types of knowledge about data. The types of knowledge are referenced from section 4.2.3.1 (i.e., the areas of knowledge researchers desired about data where they obtained only limited or no knowledge). The source of

the knowledge was missing for 62 of the 297 knowledge areas I originally coded. I did not include missing values in the figure, but the areas with the most missing values were data supplement (38), followed by data collection (10), data (six), other (3), data access information and data documentation (two each) and people (one). There were 30 instances where the response was not applicable (indicated in the figure as NA and colored bright red at the right of the columns) because the researcher did not obtain any knowledge.

Excluding responses that were missing or not applicable, the greatest number and percentage of instances of knowledge were obtained from data documentation (64 or 31.2%), followed by original data creators (39 or 19.0%), the data themselves (38 or 18.5%), literature (25, or 12.2%), knowledge of others on the research team (12, or 5.9%), one or more colleagues (9, or 4.3%), personal knowledge (six or 2.9%), data repository staff (five or 2.4%), one or more advisors (four or 2.0%), or other (3 or 1.5%).

Of note in these results is that of the 30 instances where researchers did not obtain any knowledge, twenty-five of these were ones where the type of knowledge sought was data supplement. This was not only, then, the category in which researchers most often reported knowledge lacking (section 4.2.3.1), but also the category in which researchers most frequently, by far, obtained none of the knowledge they desired.

*Figure 4.2 Sources From Which Researchers Obtained At Least Some Knowledge They Desired But Lacked*

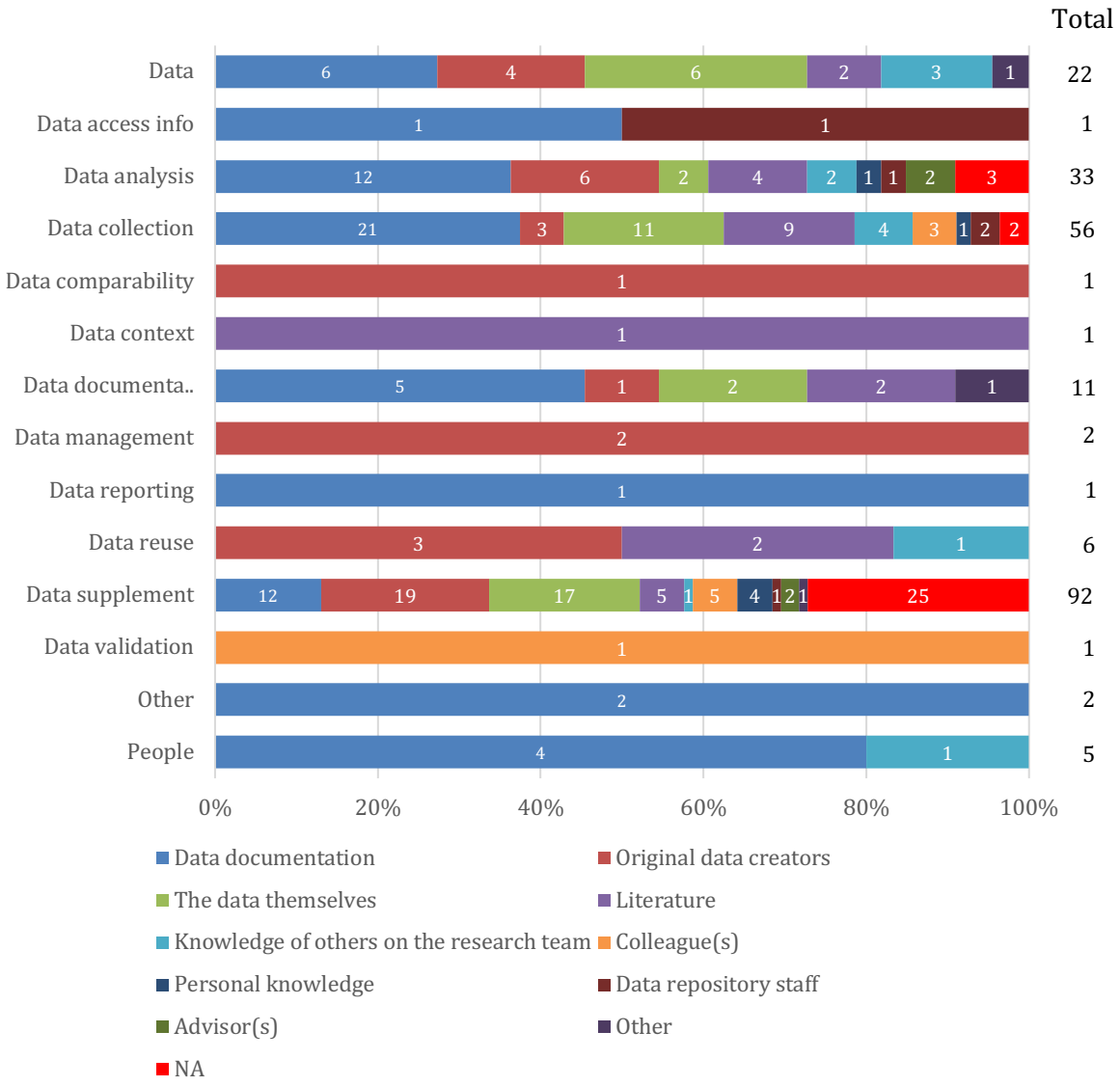



Table 4.50 shows the three most common sources for each area of knowledge, along with a count of the times researchers indicated they obtained knowledge from that source and a column with the total number of times knowledge was obtained from that source for context. It is clear from this table and the one above that data documentation, the original data creators, and The data themselves were the most common sources of knowledge. In fact, in eight of the 14 knowledge categories, data documentation was the most frequent source of information from

which researchers obtained knowledge (data, data access information, data analysis, data collection, data documentation, data reporting, other, and people). The data themselves were the second most frequent source in 4 knowledge categories. The original data creators were the most frequent source for a number of categories of knowledge that were less frequently mentioned (data comparability, data management, and data reuse), and the second most frequent source of knowledge about analyzing the data.

It is notable that in the largest area of knowledge where researchers experienced a lack of knowledge (data supplement), the most common source of knowledge was the original data creators, followed by the data themselves, and finally data documentation. These results indicate that researchers in my sample most often returned to original data creators not when they had difficulty interpreting the data or desired more information about context of the original research, but when there was knowledge they desired that they simply were not able to find in the deposited data.

*Table 4.50 Most Common Sources for Obtaining At Least Some of the Knowledge Desired*

| | Most common | Second most common | Third most common | Total |
|---|---|---|---|---|
| Data | Data documentation (6) | The data themselves (6) | Original data creators (4) | 28 |
| Data access info | Data documentation (1) | Data repository staff (1) | - | 4 |
| Data analysis | Data documentation (12) | Original data creators (6) | Literature (4) | 33 |
| Data collection | Data documentation (21) | The data themselves (11) | Literature (9) | 66 |
| Data comparability | Original data creators (1) | - | - | 2 |
| Data context | Literature (1) | - | - | 2 |
| Data documentation | Data documentation (5) | The data themselves (2) | Literature (2) | 13 |
| Data management | Original data creators (2) | - | - | 2 |
| Data reporting | Data documentation (1) | - | - | 1 |
| Data reuse | Original data creators (3) | Literature (2) | Others on the team (1) | 6 |
| Data supplement | Original data creators (19) | The data themselves (17) | Data documentation (12) | 130 |
| Data validation | Colleagues (1) | - | - | 1 |
| Other | Data documentation (2) | - | - | 5 |
| People | Data documentation (4) | - | - | 6 |

*Note*. The final column shows the total number of instances where knowledge was obtained from that source.

Several findings from this and other sections highlight the importance of interactions with data creators to overcoming limitations data reusers encountered when reusing data. These are

(a) that the area of knowledge where researchers lacked knowledge most frequently was data supplement (section 4.2.3.1), (b) that more than half (53.3%) of the cases where the outcomes of the researcher's work was limited were ones where they lacked knowledge in the category of data supplement (section 4.6.1.3), (c) that researchers obtained none of their desired knowledge most frequently when they lacked knowledge in this category, and 4) that researchers most frequently contacted original data creators when they lacked knowledge in this category. Interestingly, the kinds of limitations data reusers sought to overcome were not always strictly necessary to meet goals of publishing a paper or demonstrating competence in research (as these goals were met at a minimum level in all cases I investigated). However, overcoming the limitations would have helped the researcher attain the potential they desired for investigating their topic or research questions.

The findings I present in the next section speak further to this interaction between data reusers and data creators and what that interaction affords.

### *4.7.2 Research Question 1d: How Do Researchers Obtain the Knowledge They Desire? Findings From the Interviews*

Early in the interviews, a researcher mentioned interacting with the original data creators and influencing them to make changes to their study that would enable the researcher's work. They said,

> When you know the people who were doing the study…and they can add questions that you want added, it's wonderful. I mean, I press them. I don't even know if they would have [made the requested changes]. I bugged them about that for over a decade before they did that finally. And ever since that happened, I've had a lot of good opportunities from it. (D2-01)

This was interesting to me because when I conceived of a project to investigate knowledge about data that researchers desired but were not able to obtain, I was thinking primarily of knowledge surrounding a static body of data. It had not occurred to me to think

about data being updated by the original creators, how that process happens, or how it affects what researchers are able to know about data or what they can do with them (beyond the complications it can introduce for researchers when data are not comparable over time).

The idea that data could be updated over time and that data reusers might influence how that occurred spurred me to think more deeply about the dynamic nature of data in archives and how it might come to be that a researcher could influence what data are collected and made available for them to reuse. I thus began to pay attention when researchers spoke about issues relating to updating data; at times I also asked researchers explicitly about their knowledge of or experience with the data they reused being updated and how conversations surrounding those updates occurred.

I found that researchers' knowledge and experience varied. Some researchers had quite a bit of experience in these conversations. For instance, the researcher above described how they had lobbied the creators of a data study for more than a decade to add a particular measure that would enable research they wanted to do (D2-01). The data creators eventually asked the research to construct a measure that could be added. The researcher did and the measure was added. The same researcher described another instance involving different data where benefactors contributed to a grant that made it possible to add several questions on the researcher's topic of interest.

In another case, a researcher explained how an arrangement existed among bodies in their government. These bodies could submit proposals to have questions added to certain national surveys. From there, survey leaders made determinations about incorporating questions based on the time needed, the cost involved, and whether the proposing body was willing to pay the cost (D-08). For the particular study I interviewed them about, this researcher relayed that certain

questions they were interested in were first added to the survey as part of an international initiative years prior. They said that at the time that they reused the data, there were still additional variables they desired that were not present. However, those variables were added subsequent to their reuse of the data (apparently separately from anything the researcher did) and going forward they would be able to use those variables in their research.

Another researcher, who had reused data they assisted in collecting, described prolonged discussions and interactions among collaborators involved in developing the study and researchers involved in reusing the study (D3-01). These discussions centered on how questions in the study were worded and how constructs in the survey were measured and could thus be analyzed. The researcher noted differences of opinion among collaborators on whether questions should be updated. Some thought that the language needed to be changed to more accurately measure the desired concept in today's society; others thought the language should be kept the same for continuity of the measurement over time.

The researcher (D3-01) also described disagreements between collaborators who had political interests in the study—interests to compare populations in the study based on a single aggregated metric—and those with more scholarly interests who wanted to investigate individual components of the metric. Ultimately, the data that would be valuable for more scholarly interests were not made available.

Researcher D3-01 also described an instance where researchers had approached the data collectors after finding variables that had been removed from the study without documentation, and another where the principle investigators decided to use a particular sampling strategy in a new wave of the study in order to simplify the weighting.

What struck me in all of these scenarios was the impact that decisions about the data—about the construction of the sample, variables included in the study, measurement of concepts, and documentation of data—had on the research that could be conducted. Moreover, the decisions were not a given. In each case, there was some kind of forum for discussion between data creators and users, however arduous or lengthy, to debate practices of data collection that would affect how the data was reused. I considered this a positive and hopeful finding, especially for those I interviewed who seemed further removed from the data collection.

One of these researchers (who had no connections to the original data creators) talked about the benefits that could come if data collectors took secondary users interests into account. They indicated this was something they were trying to advocate for in the paper they wrote and something they believed other researchers would also want, though they did not think data creators were aware of this:

> It would be nice if, for publicly available data, they [data creators] would take into account researchers who were using it for a secondary data analysis—into how they ask the questions, or what they might add in the future. Because I think there would be a lot of people saying the same thing about what they want to see more of. And I don't know if the actual people who are making the surveys ever hear that. (D2-02)

The researcher said that in their experience, secondary reusers were not at all involved in conversations about what data were collected, and they expressed despondence about their own or others' input being sought or taken into account because there was not channel of communication to original data creators:

> That's what I was trying to do by writing a paper on saying, "Hey, none of these surveys ask this and there's a lot of people studying this." But I don't think anyone from those organizations actually is going to read my paper. Maybe that would be nice. But it kind of feels like just, I don't know how you move towards making a change because if enough people are doing secondary data analysis on the same dataset, they're probably going to have some of the same recommendations. And there's never like a feedback loop. I'm just thinking of this now, as I'm saying it. (D2-02)

Other researchers had a vague sense of updates to data collection occurring but were not aware of any efforts to update the data they had reused. When talking about how they became interested in their research, one researcher (D1-04) described how a certain topic was becoming big in the field and a particular measure was being added to surveys. This addition enabled their research.

When I asked about the data being updated, the researcher answered that they weren't aware of any efforts to update variables in the study and that was an issue, but also a reason journals were understanding about accepting studies that reused the data, even though the data might have aged:

> That's one of the problems. And there is not—there is no dataset that has included those things that also has this [population]. So, at the time that we wrote this paper, there weren't very many questions about the timing. (D1-04)

I asked another user of these data a similar question. They responded in general terms about a different data study and imagined that the same would occur for this study if it were ever updated. They were unsure of specifics, but had ideas about the kinds of questions the original PIs would be thinking about if they updated the study:

> But I do think that [updating the survey] would be useful if you're thinking about like with [data study], they do that. They add in additional questions over time. I'm not sure who they're talking to, but I imagine they're looking at the scope of the researchers out there and some of the things that are happening and then they add some information to that. So I imagine that those, the PIs for [data study], if they were to redo their survey, they would probably do the same. They would look and see what new information do we need to collect. What was missing or what's going on now that wasn't happening when you talked about being [span of years], whatever, and those are the things that you would normally do, especially if you're going to continue to do longitudinal studies. (D1-05)

Although researchers had different levels of familiarity with how change occurred in data studies over time, the value to them of including variables that were important to their research

was clear. One researcher spoke about how a lack of the specific variables they desired and their substitution of proxy measures impacted their research:

> But in my heart and, and still to this day, I know that these are proxies, this never replace the original answer to the question [about phenomenon of interest]…because that was very powerful and that was something I really needed. (D1-02)

This researcher talked also about the importance of additional granularity in the data to study their phenomenon of interest, and that not having this granularity could not just limit research but also result in incorrect conclusions from the data:

> If we are not using these variables, our research is always limited. Our questions are not addressed properly. And if datasets limit these [types of information] then that's really not helpful. Then, I can answer basic questions, but see, I didn't have [variable] and I didn't have [variable]…and that limited my understanding. And consequently, our larger understanding of [topic].... if you are not providing the identifiers in terms of [granular] data, that lack of access is just leading to a lot of papers with perhaps a missing conclusions or wrong conclusions even. (D1-02)

This researcher was able to publish their research, and thus, presumably, make the kind of contribution necessary to do so. But they wanted to do more.

The examples above all relate to survey data, but I encountered a similar kind of frustration, and a hint of a remedy, in relation to administrative as well. One researcher talked in-depth about challenges inherent in reusing administrative data.

> but at the end of the day, there's someone in a [organization] entering that data. And you don't know how well that person was trained. You don't know how much this person cares about the data. They've just worked [a long] shift and they're coming in at the end and having to enter [data] into the computer. This is not what they want to be doing. This is not what they signed up for. This is not their job. And you run into that with all sorts of different admin data…And it's a real pitfall as a researcher to say, "Oh, because this is admin data, there's no error in it. Survey data comes with weights. Survey data comes with confidence intervals, it comes with measurements of error. Administrative data does not. So, there's this false sense of security of, "Oh, this is really what's happening on the ground." And it's absolutely not the case. You can't make that assumption and you have to treat it as what it is: the best data that's out there, but it is flawed. It is inaccurate and unfortunately in ways that we usually can't measure. (D-06)

Despite these challenges, another researcher who reused administrative data saw more communication between an institution and researchers who frequently used their data as a way to improve the research that could be done with those data. They ruminated on their experience as follows:

> And this was basically where the data was from, from [institution]…They really weren't providing a lot of information. And so, yeah, that was limiting. And I think, perhaps there really should be more communication or more open lines of communication on what would be exchanged between the researchers and the [institution]...I mean, I'm glad I managed to publish my findings but like I said, I think a lot of things could have been improved in the original data collection. (D-05)

In the previous section, I found that data reusers reached out to data creators, particularly in situations where—though not absolutely necessary to meet professional goals—researchers desired additional data that was not available to them. My findings in this section go beyond revealing that these interactions occur to showing first, that conversations between data reusers and data creators make a difference in at least some circumstances (i.e., toward enabling new and deeper research), and that in circumstances where they don't, researchers believe they could make a difference. In Haraway's terms, these findings expose the existence of "interpreters and decoders" (i.e., researchers) whose situated perspectives are not being considered in a conversation that impacts that scope and depth of the research they are able to conduct. In the context of considerations about how to better support data reuse, this represents an opportunity.

Second, my findings show that not all researchers who reuse data have the same opportunity to have conversations with data creators or ability to influence what data are collected, or how, and what are made available. Making a difference in the data collected in same cases required years of lobbying, the help of benefactors, or membership in government bodies or collaborator groups. In other cases, researchers were not clear how updates or changes occurred or who would be responsible. In all cases, what was at stake for researchers was not

necessarily achieving a publication. It was, rather, investigating their topic at the level they desired, or avoiding coming to faulty conclusions based on inaccurate data. This gap between those who have the ability to influence data collection and those who do not could once again represent an opportunity (from the perspective of data reusers) to make support for data reuse more robust (e.g., by bringing additional perspectives into considerations of data collected and availability).

Overall—to integrate ideas about conversation with Simon's conception of adaptive systems—these findings create an opening for a conception of data as objects of design (whether consciously designed or not) that are not fixed once and for all, but rather can change dynamically over time based on inputs into the design and design priorities. This conception of data allows for the consideration of improvements to design, including those focusing on the processes (conversations) through which it is decided which data are collected and therefore enter the archive.

In the next section, I present findings about what researchers believed impaired their attainment of knowledge about data, and what could improve it. These findings relate also to the conception of data creation and the facilitation of data reuse as objects and processes of design.

### *4.7.3 Research Question 2e: What do Researchers Believe Could Mitigate Knowledge Satisficing?*

During the interviews, I asked researchers about factors that either inhibited or facilitated their attainment of knowledge about the data they reused. I also asked them to think more broadly about the environment for data reuse and comment on shortcomings or barriers to reuse that they saw and what they thought might be done to better support researchers seeking to conduct research with secondary data.

A number of the barriers and facilitators researchers mentioned appear at different points and to different extents in preceding sections. I consolidate them here, however, and give particular attention to factors such as "knowing someone" that I introduce for the first time. I present specific and general barriers to reuse first, followed by specific and general facilitators of reuse. Because of overlap in the specific and general categories of barriers and facilitators, in these sections I have combined the presentation of the specific and general descriptions researchers gave, distinguishing between them in the text as appropriate.

#### *4.7.3.1 Barriers to Data Reuse: Specific and General*

The main barriers to data reuse that researchers described related to accessing data, dealing with shortcomings in the documentation and discovery of data, using data, and challenges having related to the culture surrounding data reuse. There were only two researchers (D2-03 and D-08) who reported that there was nothing in particular that facilitated or hindered their research (these researchers also did not report lacking knowledge about the data they reused). I report on general difficulties D-08 experienced reusing data below, but note that these did not pertain to the research I interviewed them about. In that instance they did not obtain their data from ICPSR.

##### *4.7.3.1.1 Access*

I discussed barriers to accessing data—from complete lack of access, to delays and other inconveniences—as well as some of the factors that lead to such barriers (e.g., sensitive data and incentives around sharing data) in previous sections (see 4.3.3.4.4 and 4.7.2.1). In addition to these issues, some researchers cited the inability to analyze restricted data ahead of time as an impediment to reuse (D-02, D-06, D-11). (ICPSR, in particular, makes it possible for researchers to do limited analysis of some restricted data studies online without having to request access to

restricted data.) Another researcher found the process of requesting access to restricted data confusing and did not understand why the process needed to be repeated multiple times (D-03).

A different researcher reported conducting their research at a time when the data they reused were, on the one hand, publicly available but onerous to assemble from multiple sources, and on the other, available from ICPSR but with restricted access. The researcher was frustrated by the barriers to accessing the data through ICPSR but described the difficulty as a "double-edged sword" since the availability of the data from ICPSR lent the data greater legitimacy:

> Then again, I think the ICPSR thing [i.e., access to data being restricted] is a double-edged sword. On the one hand, it's frustrating. On the other hand, to me, it actually lended legitimacy to [the data] because, to me, there were two curators of data: the [data creator] plus ICPSR. (D-11)

Another had an experience where a perceived lack of transparency about the way the data in ICPSR were curated presented a barrier to reuse:

> The trouble with [ICPSR] dataset is, I don't know what is behind their online engine that they use. So, the choice of sampling weights- For nationally representative surveys, you have to account for any analysis for the stratification process that pick that survey as well as for the weights assigned to those people. And I'm never clear using that online data system, what they're doing to generate their numbers. Probably the prevalence is the percentages are okay but the confidence intervals is what you have to worry about being very different. So that's why I don't tend to use them for final analyses because I don't understand what the programming is behind it. I don't trust it necessarily. (D-08)

Researcher D1-04 described challenges accessing data in general terms but pointed out that the difficulty researchers have gaining access to data is not always the fault of the archive housing the data. It can result from the level of experience an institution has with secondary data analysis—and facilitating secondary analysis for researchers. The difficulty manifests itself on what the researcher referred to as a "continuum of difficulty":

> Sometimes, these datasets have what are called restricted parts. I mean, there's the public release [data], and then if you want more detailed data, you have to go through extra steps to get access to the restricted version. And sometimes doing that can be a little burdensome in terms of the process is not clear sometimes. And sometimes the process

> takes some time. And sometimes, that's not the fault of whoever's housing the data, sometimes it's out of the university that you work at. And sometimes it's universities don't really understand secondary data analysis. Universities have a profile, they're Ag schools, they're medical schools, they're, you know. And sometimes when they're dominated by one of the sciences or agriculture or computer science, they don't understand this. So, they don't have in place people who know immediately what you're talking about when you say restricted datasets. Then there's the you in-between the data archive and the university trying to work out access…and there's a continuum of how difficult this is depending on [your institution]. (D1-04)

They said they understood the ethical concerns surrounding the use of secondary data and the need to be careful about to whom access is granted and under what circumstances and saw the challenges of gaining access as simply the limitations that come with reusing restricted data.

Some additional barriers associated with access that researchers mentioned included an absence of well-curated, timely data (this absence caused them to seek data from local, publicly available sources, which increased concerns about reliability) and an absence of detailed data outside of mainstream areas of research interest (both of these were raised by researcher D-04).

#### *4.7.3.1.2 Data Documentation and Discovery*

Researchers were also inhibited in their reuse of data by barriers arising from poor data documentation (D3-01, D-09) and poor discoverability of data (D-03, D-08, D-13). Researcher D3-01 talked about a deficit in documentation they experienced when trying to reuse the data that still exists today and is particularly problematic for those with less experience reusing data.

> [newcomers] don't really understand it, or they don't really know why you have to have this. They [the data creators] did the report, but then they just don't tell you how you can use the data. I don't know…just a couple of days ago, some students of PhD…or master, they want to use the data. And they said that they have read all this, they still don't know how to use [a particular component of the data]. (D3-01)

Researcher D-09 discussed the importance of well-documented data and connected documentation quality to the quality of the work done in the original research:

> [W]hat I like to emphasize is the transparency and the kind of well-organized variable list or coding scheme, et cetera. All of those kinds of information should be shared with the raw data file. If those points like for the menu or et cetera, are not transparent or crappily organized, then even if we can download a huge raw data containing a list of compressive variables we can never use. Because if there's no coding manual, then there is no way to figure out what number one, number two means et cetera. So in some cases I found that researchers didn't do their job very well. The job, I mean, writing up a good coding scheme, coding manual variable list, explanation, et cetera. So I think those are major barriers [to data reuse]. However, if a transparency can be guaranteed, then I'm going to say there's a possible way to facilitate wider re-use of secondary data sets. (D-09)

There were three main issues researchers raised with data discoverability. The first was the difficulty of searching for data in ICPSR. A researcher related that they were able to find the data they wanted because they knew exactly what they were looking for. However, they felt that finding relevant data was more difficult for their students:

> When I went to ICPSR looking for that data, I knew exactly what I was looking for. I knew I was looking for the [data study]. I knew what to search for. I have since told students they should go to the ICPSR. And you should search by terms and things like that they're interested in. But I don't think it's as easy for them to find. And I don't really know how to fix that. I mean, the more data that there is there—and we all know this that if you don't have like the exact term or something, sometimes it can be difficult to find what it is that you need. But something that let people know about the potential of the data. (D-03)

The other issues had to do with researchers lacking knowledge about what data exist, on one hand, and knowledge about the limitations of the data on the other (i.e., what data do not exist). Researcher D-08 talked about the difficulty of discovering the breadth of relevant data.

> I think the biggest issue in data reuse is that people don't know it's out there. There's no repository you go to that has...every possible dataset, and what's in it, what's not in it, or how you access these datasets. So, like I was saying before, that I found the [data study] and I stopped looking because that was something easy to find. It doesn't mean there weren't other ones that were more difficult to find that I just never got to. Because I didn't know they existed. And that's, I think, the issue for people whose existing datasets, there's no easy way to find out what's really out there. I mean, the only way you can really do it is by reading the literature in the field and hoping in their methods section they've listed where the dataset came from and how you get access. That's really the only way you can do it now. Or, you know, talk to people. So that, to me, is a big disadvantage. (D-08)

In addition to the importance of knowing what data exist, researcher D-13, speaking in the context of what might be done to better facilitate reuse, noted the risk of misuse if researchers do understand additionally what data do not exist:

> I would think the main thing is having [researchers] know that the data sources exist, and having those who have the data sources make it clear what their limitations are...[I]t would probably been better if they were more clear about what they didn't collect. The main thing they probably don't collect is [description], so that anyone who's using [the data] is going to have a massive misunderstanding of [topic]. (D-13)

An additional barrier that one researcher described is what they referred to as "data balkanization" or a separation of data from different fields into separate repositories. The researcher saw breaking down those barriers as key to facilitating reuse. This barrier could be seen as relating to access as well:

> I would maybe describe it as balkanization of data, that I think there are some repositories in specific fields. Economists have their things going on, psychologists have their things going on, and they're silo- balkanized, choose your words. They don't communicate with each other. Though those data can be very complimentary or independently useful across fields. But I think most of the secondary data repositories exist within the field silo. So I think breaking those down can absolutely facilitate the use of secondary. (D-11)

There is an additional dimension to this issue which was raised by another researcher (D-10) who discussed a difficulty associated with copies of data being held by multiple repositories. The difficulty was researchers discovering data in repositories disassociated from relevant documentation. The researcher pointed out the importance in such circumstances of linking to the sites of data producers to ensure that researchers understand the limitations of data and how to use them.

Researchers' responses indicated that poor documentation and poor discoverability of data can result in absolute barriers to data (i.e., not being able to reuse data or not knowing that data exist to be reused) and also incorrect use or interpretation of data (i.e., if appropriate ways of using data or information about is included in the data and not are not documented).

#### *4.7.3.1.3 Using Data*

In addition to barriers to accessing data and shortcomings in data documentation and discoverability, researchers reported difficulty actually using data once it was obtained. I discussed many of these in depth in the context of considerations and pressures that influenced researchers' reuse of data (section 4.3.3.4.3) and researchers becoming knowledgeable about data in order to reach a reuse equilibrium (section 4.4.1.3).

These included first and foremost challenges that researchers experienced due to the time necessary to get to know data and to learn to reuse them. One researcher described data reuse as a "very long process" that was not to be taken lightly (D1-03). Another described getting to know a new dataset as a "real investment" (D-02). Researchers talked about "sitting" with data, "immersing" themselves in the data, and "paying your dues" in terms of spending time with data to develop the requisite competence and eventually expertise and intuition for reusing a given dataset.

In light of the investment needed to reuse data, several researchers noted the "calculus" involved in deciding to reuse data that was new to them as opposed to data they were familiar with (or even doing deeper analysis with familiar data). Some researchers reported opting to stay with familiar data although other data may have been more advantageous because of the work involved in getting to know new data.

Researchers also talked about difficulties that came from a lack of experience reusing data (i.e., it being their first time) and challenges they encountered conducting some forms of analysis, for which it was necessary to seek help, sometimes outside of what was offered in their own departments. About their first time reusing data, researcher D2-02 commented:

> I don't think I realized that there were other datasets that are so much easier to analyze at the time 'cause it was my first time really doing secondary data analysis. So now looking

> back, I think like, wow, that was really complicated for a first time; not something I could just easily do on SPSS or something. (D2-02)

Researcher D1-03 described "complex issues" with their data analysis that required consultation outside their program. Researchers D1-02, D1-05, D-02, and D-05 described similar situations where they sought additional help analyzing data from individuals or groups (e.g., from workshops or trainings).

Two researchers reported a final type of difficulty they encountered, having to do with how the data were curated and made available for reuse. Researcher D-02 described how better "pre-packaging" of data, including the addition information about how the data were intended to be used and better cleaning of data could greatly improve the ease of reuse.

Researcher D-02 had similar thoughts. I quoted them earlier in section 4.3.3.4.3 discussing their calculus for reusing data. In that instance, the researcher decided not to delve deeper into a specific data study because they didn't believe the trouble was worth the benefit they would receive. However, they remarked that they would have reused the data if the data had been better prepared for secondary reuse:

> That said, if the data, the more complex data were more usable, if they were more cleaned, maybe that would change the calculus. Maybe if they invested more resources in them, brought in a team of data scientists to make them more usable for research purposes, that might change the calculus there. (D-02)

All of these posed barriers to reusing data, then: the work involved in getting to know data well enough to reuse them, challenges stemming from the first experience of reuse, difficulty understanding how to analyze data, and obstacles encountered in how the data were made available.

#### *4.7.3.1.4 Culture*

The main cultural barriers researchers described had to do with collaboration, incentives in academia, and guidelines for sharing data. Concerning collaboration, researcher D-05, in whose community data reuse was the norm, noted that researchers often worked by themselves or students who they knew and tended not to reach out to others:

> faculty even in my program, were kind of working in silos. So people weren't really sharing research with each other or with other faculty outside of the department. I mean, there's some collaboration too, but there's always sort of been- well, it's not published anywhere, but something that's been conveyed to me...I noticed people work by themselves and they work with students who they know. They're not really reaching out to other people. So I guess because we have our own areas of interest and if someone is in [field] and I'm in [a different field], then we're working on our respective projects. Why would we be talking with each other as openly? (D-05)

When researchers did collaborate, researcher D-05 described a certain protectiveness of data creators toward data—data which would be reused in the researcher's field—and a conception of reciprocity in their field that was less advantageous to newcomers who perhaps didn't have as much to offer in a collaboration and for whom the stakes of not having a successful outcome were high. The researcher described these, and the siloed nature of research above, as barriers to more transparent discussion about researchers' experiences reusing data that could help researchers identify data that were most appropriate for their research. They said,

> Generally for academics, people don't like admitting there are problems, and that is something that- maybe the culture has to change in order to move forward and address this moving forward. I think some people are very protective of their datasets, especially if it's original data, if they collected it for 20 years or they're working with it for 20 years and only like a privileged few is actually going to get access to it. And then I think there's this whole idea of reciprocity. If I'm collaborating with you, then you would have to offer me something in exchange. And so I think in a way that really speaks to maybe issues of how the culture works and it just doesn't work as well for some people, especially if you're trying to get established and you don't really want to get disappointed with what you find or don't find. (D-05)

In general, researcher D-05 believed that in the culture of data reuse they experienced, there was a mismatch in the incentives that affected scholarship. They believed this was

expressed especially in the reality that researchers might put a lot of time and effort into reusing data for a publication but not get credit for any of that work if somewhere along the publication chain, the research was not deemed publishable (which they experienced):

> It's a lot of effort and I think there's a disconnect between the output of scholars and researchers [in terms of research] and what we're able to achieve with that [i.e., if it is not published]. (D-05)

Other researchers talked about incentives of other kinds as barriers. Researcher D-11 did this while echoing some of what researcher D-05 expressed about researchers being protective of their data:

> At least in my field, there's no benefit to me giving my data to anybody. I actually have an incentive to guard my data- "Well, I can still mine it, but you can't." And I don't know if that's a limitation of secondary data itself, but I do think that even if secondary data is going to be absolutely useful, there's that barrier to accessing it because it's secondary. Someone else owns it, so they have to give it to you. (D-11)

They also talked about a more general lack of incentive for and prejudice against secondary reuse, including replication:

> I will say one shortcoming to using secondary data is I do think that primary data collection is seen as higher in the hierarchy. You know, as more prestigious; it's better research. That culture has its drawbacks. I understand why, because I think particularly in academia, we're seen as academics that are expanding the frontier of knowledge. We're supposed to be doing something new. But on the other hand, particularly in certain fields, you can't replicate. The replicability of studies is horrendous. And so this prejudice against secondary data, I think it does harm as well because there should be reliability there. If X researchers published on this and found X, well, then I should be able to find the same thing too with that data. Yet, there's no incentive to do so. There's no incentive to publish on secondary data. Yes, it's accepted. I'm not saying that there's—but it's seen as better if you do it yourself. (D-11)

Researcher D2-03 gave very similar remarks researchers, believing that too much emphasis was given to being the first to investigate a topic and not enough to vetting conclusions of others' research. They said:

> But too much emphasis and too much reward is given to being first as opposed to fully vetting a particular hypothesis through numerous studies and years of work. There's just this really big premium on if you're the first person to do something. You're going to

> publish in the best journal, and that's going to help your reputation. That's going to get you tenure, et cetera, et cetera. And then that trickles down into being the first to collect this novel data set. You get a huge, huge reward for that. And then, the marginal returns to using that data set a second, third or fourth time, especially by people who were not part of the original collection team is really low. (D2-03)

Researcher D-09 felt strongly that policies promoting the value of secondary analysis were important to dealing more generally with challenges to transparency and replicability in science. They felt that there was a positive trend toward valuing data reuse in academia, but more could be done formally to recognize contributions to open science, such as secondary analysis of data and replication:

> what I felt from my prior research is that a sort of a structure of incentives or like institutional things are really important to promote good practice in science. For instance, as I mentioned, if secondary data analysis is devalued compared to the primary data collection analysis…But at the same time, secondary data analysis should also be well recognized and acknowledged in the field. I think that is one of the ways to deal with the issue of replicability and transparency. So in terms of like a job search or tenure and promotion processes, or journal editorial policies, there should be something about facilitating secondary data analysis at the institutional level or policy level.
>
> In the academia, I think you're getting at least slightly better and better as time goes by. Many journals are now welcoming secondary data analysis and replication studies in general, but some don't. So I'm not going to say there should be radical changes, but this trend should be continued. But in terms of tenure and promotion, I think there should be something more fundamental like providing incentives to open science-related activities. I know some schools consider open science related activities while hiring someone or providing tenure and promotion to faculty members…so I think those kind of things should be considered at the administrative level, as well. (D-09)

Researcher D-12 saw a lack of standardized guidelines for sharing data as a barrier to reuse, suggesting that challenges conducting secondary analysis could persist even if issues related to the status of secondary analysis were addressed. Researcher D-12 pointed to this lack while talking about the benefit a greater number of centralized repositories could have for data reuse (responding to my question about what could improve the environment for data reuse). At the same time, they referenced the incentive not to share data—or the incentive to withhold

data—discussed by researchers D-05 and D-11, and connected a lack of standards to a still-emerging acceptance of open science. They said that although researchers like them were on board with open science, open science practices had not yet filtered into the teaching curriculum:

> I think more just centralized repositories, like for instance open science framework. They've created a way of organizing projects that steers you in the direction of transparency, but it's lacking in terms of providing structure recommendations, basically, on best practices...For instance, my way of practicing open science is when I publish a paper, assuming none of my co-authors object, I extract all the data that was used in the analysis down to the item level. I de-identify it. And I include it on the open science framework page, along with all the analytics scripts, all the outputs, all the supporting information. I put that on open science framework, and it lives there. And my paper, when it's published, is linked to that. But that's just me. Somebody else would do it a different way. Somebody else would potentially provide every data point, every variable that was collected in a dataset. And I don't do that because I might need those variables for the next paper. So I am, maybe I'm only half-hearted about my openness. Because I'm only sharing the parts that are relevant to each paper. I think after a dataset was retired, like I'd gotten everything out of it that I could get out of it, I probably just put the whole thing up there, but now it's already up there in pieces, right? So, I think that's sort of the thing, is that there's no standardized set of guidelines about how we should, as individual researchers, for data that wasn't necessarily funded by a federal grant, how should we go about making that available is still unclear. It hasn't filtered its way into like our training model yet. So I teach methods and statistics to PhD students. And it's not part of the curriculum, practicing open science, because it's still not universally accepted. Like I said, it's a generational shift, and there's field specific levels of embracing it. (D-12)

#### *4.7.3.1.5 Summary*

Shortcomings related to accessing data, data documentation, the ability to discover data (including what data exist and what data do not), the disaggregation of data, and challenges using data were key barriers to reuse that researchers experienced. In addition, researchers encountered difficulties arising from a lack of standards for sharing data and policies supporting data reuse, and incentive structures that undervalued secondary reuse, overlooked scholarly contributions that were not formally published, encouraged siloed research (rather than collaboration) and promoted protection (withholding) of data.

In the context of Simon's framework, we might consider an inner environment made up of researchers seeking to better understand the world and an outer environment of the world which they seek to understand. In this context, we could consider access to data, data documentation and discovery, and standards, policies, and incentives as strategies we have developed over time to enhance the ability of researchers to understand phenomena of the outer environment. If we did all of this, the shortcomings and difficulties that researchers experience, to the extent that they inhibit the achievement of researchers' intended purposes, would indicate aspects of those strategies that are not well adapted to the conduct of research and might stand to be improved through design.

In the next section, I present features of the research environment that facilitated researchers in their reuse of data and ideas they had for how shortcomings might be overcome. In chapter 5, I discuss the implications these barriers and facilitators have, in concert with the conversations I introduced in section 4.7.1, for the design of processes and data themselves that support reuse.

#### *4.7.3.2 Facilitators of Data Reuse: Specific and General*

The primary facilitators of data reuse that researchers described were "knowing someone," having avenues to learn about and become of aware of data, the availability of good documentation, sufficient funding to conduct data reuse, and the ability to access data.

##### *4.7.3.2.1 Knowing Someone*

Through the course of my findings—e.g., in presenting considerations and pressures that affected researchers' reuse of data, the ways researchers negotiated gaps in their knowledge about data, and (briefly above) barriers to reusing data that researchers faced—I have reported on strategies researchers used to overcome challenges they faced in reusing data. These included

collaborating with others and seeking help from advisors, consultants, and venues such as trainings and workshops. Part and parcel of these strategies, which researchers mentioned as facilitators of their reuse of data, was what I came to refer to as "knowing someone." From the time of the decision to reuse data through analysis of data and publishing, researchers talked time and again about the importance of "knowing someone" or having connections that provided crucial input or aid to them at the right time. As research D1-04 put it in the context of analyzing data,

> a lot of times, you run into a problem and your ability to solve that problem depends on knowing someone who is really immersed in the data, who can help you sort it out. (D1-04)

Knowing someone had an impact on researchers from the very beginning of their work. For example, in the environment in which researcher D3-02 conducted their research, their lack of a special connection to an entity in the community they wanted to study meant that it was more advantageous (and more possible) for them to reuse data than to collect their own. Advisors were also a crucial connection for researchers early in their careers. In the case of D3-02 and others, it was a researcher's advisor or mentor who introduced them to the topic of their research (D-06, D3-02), the data they reused (including in many cases facilitating access) (D1-01, D1-02, D1-05, D3-02, D4-01, D4-02, D-06, D-07), and connected them with collaborators and sources of knowledge about the data that were key to their ability to reuse them. Collaborators included other students or researchers who had experience working with the same data (D3-02, D1-03, D1-04). Sources of knowledge included, probably most significantly if they were not connected with them already, the original data creators (D1-01, D1-03, D1-05, D-06, D-07). These connections with original data creators were crucial for some researchers. As researcher D-06 explained,

> And, you know, I had the huge asset of having an advisor who had connections at [organization] so we could go and ask the data creators about it. Not everybody has those connections, and not every data creator is willing or able to answer those questions. (D-06)

Advisors were also important connections because of the control they had over the amount of time students could devote to their research (D3-02) and the advice they gave them on a wide variety of issues and topics. These included what literature existed on their topic and what research was currently being conducted with the data (D-02, D-07), how much knowledge about data was "enough" to reuse the data (D2-04, D-02, D-05, D-06), how to analyze data (D2-02, D4-01), how to revise papers for publication (D-05), and whether or not to pursue access to restricted data (D-06). For all of these reasons, researchers frequently mentioned advisors and mentors as important facilitators of their research.

Some researchers' work was facilitated by access to consultants or tutors, especially for help with statistics (D1-03, D1-05, D2-02, D-05). Others were supported with the provision of office space or assistance gaining access to data, as discussed in section 4.3.3.4.4 (on environment (D2-02, D2-03, D-11). This support from institutions, part of what I considered "knowing someone" was just as crucial as advisors and their connections for researchers' success. Describing issues they encountered using apply weighting in survey analysis, researcher D1-03 related their experience this way:

> I had to then reach out to our stats consultants at [institution] to get further support on how to accurately estimate these models when these estimation issues came up. Which was great at the time because [institution] had incredible resources for doctoral students and faculty to get that consultation support. Something that I realized lacks in other institutions, despite the fact that they're big R1 institutions. I have yet to see something similar when I was at my postdoc at [a different institution]. (D1-01)

The importance of connections to people and institutions is reminiscent of the knowledge and linkages in networks that I reported were necessary for researchers having "conversations"

that influenced the data that were collected and thus made available for research. Like that knowledge and those linkages, the advantage of "knowing someone" (being associated with an institution or individual) who has the ability through their knowledge, experience, and connections, to advance a researcher's work on a particular topic or particular data is not equally distributed. Given the challenges many researchers described to using data and the importance of "knowing someone" to their research, it is hard to imagine those researchers being as successful as they were without the kinds of mentorship and resources that were available to them. The need for mentoring is an aspect of any research but may be particularly crucial for conducting secondary research.

#### *4.7.3.2.2 Learning and Awareness*

Beyond facilitators of research that came through "knowing someone," researchers highlighted workshops and trainings where they had opportunities to learn about data, ask questions of data creators, and meet other researchers working on the same data study as key facilitators of their research (D1-03, D1-04, D1-05, D2-01). Speaking of the importance of workshops generally, researcher D1-03 said,

> So, I think especially for large datasets that are geared towards making it accessible to more people, not just the PIs, having these workshops, whether it's one day, two days a week, in person or virtual, are vital to giving people more information. 'Cause documentation is great. And I think documentation is very important because you can go back to those things and just review information that you may have missed. But sometimes in these in-person workshops [you] are able to get more nuanced information about the thought process about the measurement, or maybe even get one-on-one consultation on using the data and getting some statistical support. (D1-03)

Researcher D1-01 suggested the creation of video mini-trainings to help overcome challenges researchers face in discovering using data if they did not know people or have requisite connections. They talked about this in the context of their own experiences being intimidated with trying to reuse a particular national data set and seeking for data to reuse.

> So there's all these like national data sets that have [topic] focus, but they're in different places. They also don't always include as far as I know- and I shouldn't say that for sure because I think they do have some trainings, but it might be helpful to have like a video training of how best to use this data. Like what do you need to consider?….kind of like a training, but a mini one to go along with the with the data. That might be helpful for reuse, because I think [data study] just scares me because there's so many pieces of it that people say it's really hard to use. So when people say that I just sat back. I'm like, "Well, I'm not using it unless I know somebody that does use it, you know?" (D1-01)

The researcher switched to talking about different data they were able to obtain from the study PIs, still thinking about ways of improving discovery and access. They had difficulty learning how to use the data and had these thoughts:

> I was trying to use [the data] for a grant, but there are all these places you go [to find information]…But it'd be nice to have an overview video or something. And you can't answer every question, but that might be helpful to use it and just publicizing that it exists. I mean, some things you just don't know exists. I knew of this data because someone told me. I don't know if I would have known that if my mentor wasn't at [institution]. I have no idea if I would have found it later on maybe, but it's not like it's going to be a news flash on CNN type of thing, but it would be helpful to maybe put it on Twitter or whatever. I mean, I don't know, use social media to maybe let us know what's out there. (D1-01)

Others also cited awareness about data as key to improving possibilities for reuse. Researcher D2-01 referenced the wonderful job they believed ICPSR was doing to publicize resources and saw the trainings and workshops they offered as key to bolstering data reuse. Researcher D2-03 noted a distinct improvement over time in their students' awareness of data sources and mentioned they now encounter students who have poked around the ICPSR website before they are officially alerted that data are available there. In describing what could improve the environment for data reuse, researcher D-01 pointed to a specific database that brought together data relevant to a particular country. They saw this as an example of what they wished could be available for every country to make it easier for researchers to discover relevant data.

Two researchers suggested that it would be beneficial to take additional steps not just to facilitate data discovery, but to communicate the value of the data (D-03, D-05). Researcher D-

05 fleshed this idea out a little bit more, describing a hypothetical forum or service similar to Yelp (http://www.yelp.com) but for data where researchers could comment about how well particular data met their research needs. They acknowledged some of the cultural barriers that could be involved, but said that such a service would be something that could save researchers a lot of time and effort.

#### *4.7.3.2.3 Documentation*

Some researchers pointed to good documentation of data as a key facilitator of their work (D2-02, D4-02, D12). These researchers especially appreciated the transparency and ease of use of documentation such as codebooks and having all of the necessary documentation located in one place. Researcher D-12 remarked about how the comprehensiveness and completeness of the documentation of the data they reused gave them the sense that they could have done the study themselves:

> [Measures used in the survey] didn't even have to be fully previously validated because the authors were transparent, in the codebook, and basically literally provided PDFs of the actual survey forms that participants completed, which allowed me to verify the psychometric properties, the reliability of the instruments that were administered. So, it was the authors' essentially commitment to transparency, and the extent to which they were exhaustively comprehensive about all the materials that made it really easy for me to identify and evaluate, "Does this meet my needs?"…I'm a researcher, I could have done that same study. They basically did the study for me. So, that's how I decided I could use the data. (D-12)

Researcher D4-02 noted the importance to them of having comprehensive documentation of the data study in a single location:

> [the creators of the data study] have a nice website that gives all of the information you need. So, when the data was collected, attrition rates, how the variables are actually measured, what they measure, all the conditions of the study, et cetera. That's all really located in one central website place and that really puts this study above other studies and other data sets because it makes it just so easy to be able to go and access that information when you need it…That really gives you a validity check, having that information all in one place, because you can go and see, "All right, there's a relationship here, but am I actually testing the relationship that I think I'm testing?" (D4-02)

One researcher highlighted that there were different styles for learning about data, and different preferred approaches. Some preferred to attend workshops or trainings first to learn about data. Others, like themselves, preferred working with documentation first. They didn't gain much additional information from workshops:

> In general, a couple of workshops that went on was like, I could spend the same time sitting down and reading the information in the webpage—asking a couple of questions, get exactly the same information. I don't think that workshops gives you any key information that is not already written…I'm more the kind of person that prefers to read first, inform myself as much as I can. And when I don't have anything else to read to understand something, then I try to look for workshops or asking people…Some people prefer to be de-briefed first and then they decide to dig in. (D2-04)

Researchers talked about different kinds of documentation as well. Researcher D-06 noted that data creators' inclusion of the code they used to perform specific operations could significantly reduce the amount of time it takes to reuse data:

> So I think one of the huge facilitators for me…is when data creators provide commented basic code for reading in the files, for formatting their variables, and then for maybe performing a simple operation, generating a mean or something like that. That is hugely helpful to me. Especially if they even take it a step further and say, "Okay, you know, here's how to replicate Table 1 in my data." Because that removes literally sometimes months of work, trying to figure out, "Okay, you know, what—how exactly do I have to subset these things or work with these things to get, you know, to get the same outcomes?" (D-06)

Researcher D-07 emphasized the importance to them of having information about the PI of the original study and being able to access all the articles published using particular data. They suggested that data archives collect references to publications the PI is aware of when the data were archived, and noted that they advise their students to use other means (i.e., Google) to search for relevant publications:

> I want to know about the PI…I love it if a data set is archived, when it has a list of every article that's been published; a report that used that data. That's awesome. And I make my students go through that list. You know, "Here are the 20 articles that have been published on that. See what they're doing and actually read them."…I do think having the

> list of all the papers that the PI knows about when they archive that data, that it be there. But that's not always the case, right? Because then other people are using the data and the PIs don't know they've already archived it. So ICPSR doesn't have enough staff to then say, "Here's the other things." But I make my students use Google and Google around and see what's been done since that dataset has been archived. (D-07)

#### *4.7.3.2.4 Funding*

As I reported in section 4.3.3.4.4, several researchers reported funding as a major consideration in how their research developed (D1-02, D2-04, D3-02, D-09, D-11). When thinking about general facilitators of reuse, most researchers spoke about factors outside of funding (e.g., how to facilitate connections, data discovery and use, etc.), but one did mention funding in a general context:

> Secondary data analysis funding may help a lot. Because, although you don't have to go through the data collection process, just sitting down and putting these datasets together, it's considerable amount of work. (D2-04)

#### *4.7.3.2.5 Access*

Similar to funding, and as I have reported, accessing data was a major issue for researchers (though some were able to find ways eliminate or reduce concerns) (D2-01, D2-02, D2-03, D4-01, D4-02, D-03, D-05, D-06, D-11). One researcher pointed to two factors they believed could improve the ability for researchers to conduct research with restricted access data. The first was to enable some level of public analysis of restricted data before making a request.

> is there a way to poke around a head of time? Because, with some of these datasets, I'm going through this now on ICPSR where I'm interested in a data set but it requires an IRB, right? So I have to sit down and think, "Well, I want to use this data set but do I want to sit down for two hours and fill out all this IRB paperwork, send it off, have my IRB kick it back, to say, "You need these revisions, send it back?"" Is it even worth it? It would be great to just say, "Within the ICPSR protocol, could I run some cross-tabs? Could I have a general idea about, you know, is there enough cases? What are the missing data patterns like? Is this going to work out?" before I start investing a bunch of time going down the paperwork rabbit hole. (D-02)

Researcher D-02 argued for allowing more flexible options for accessing sensitive data:

> The third thing I would say, for repository purposes—and COVID has made this clear, right? As academics, we have a lot of flexibility. We can work from home for a year and our job would keep going but also, a lot of times, getting these datasets, there's all these protocols of data storage, "Oh, it needs to be on a hard drive in an office, in a locked drawer. On a non-network computer." It would be great to get more flexibility with this. There's a lot of advances in encryption and data storage and stuff where, like, "Why can't I have this data on an encrypted cloud network or something and access that cloud network from my house?"…You know, a lot of the times, I'll look at secondary data and I'll see another thing that will factor into the judgment of, "Is it worth it?" is, like, "What's the pain on storing this?" Does it need to be in on my office on my crappy non-network--?" My office computer is a super fast Mac that runs great. My non-network computer is like a 10-year-old PC that's in the corner and, when I have to run data on it, it takes an hour cause it's super slow. So, if it's super private data, I have to judge out, like, "Is it worth it?" Now, again, I'm not an expert in data storage but I feel like we're running on a model that's a decade old compared to where we are today in terms of data storage and encryption and stuff. I have the sense we can do better. (D-02)

#### *4.7.3.2.6 Summary*

Throughout their reuse experiences, the connections researchers had with advisors, mentors, collaborators, consultants, tutors, and original data creators were crucial to their success. They provided the experience, guidance, and support that researchers needed to discover, access, obtain and bound knowledge about the data they reused, as well as to understand and analyze them. In the absence of these connections, or sometimes along with or because of them, researchers also benefitted from workshops and trainings about data. They mentioned developing "mini" trainings and videos to accompany data to overcome feelings of intimidation about reusing data and increasing outreach about data to let researchers know what data exists to be reused. Additional strategies related to learning and discovery were aggregating data of interesting around particular topics (e.g., countries) and creating forums for researchers to discuss their experiences reusing specific sets of data.

Researchers also benefitted from comprehensive and complete documentation of the data they reused and having everything available to be able to reuse the data, including relevant code and instructions for how data were processed so researchers spend less time cleaning and trying

to understand data. One researcher noted the value of having access to information about the PI of the original study and articles that have been published that use the data.

Adequate funding for the work involved in reusing data and more—and more flexible—options for accessing restricted data were also seen as key facilitators of reuse. Ideas for access included an option to perform some level of analysis on restricted data prior to making a formal request for access and having options to access restricted data remotely.

The facilitators researchers mentioned map fairly well to the barriers they discussed, including issues related to access, documentation and discovery, and standards (which affect what data and attendant documentation are shared and available for reuse). Some areas that facilitators did not cover relate to policies and incentive structures that affect how data reuse is valued and how it proceeds (encompassing issues surrounding collaboration, or lack thereof, and withholding of data).

While I would argue it is important to advance efforts to improve the design of the research environment along all dimensions, my findings focus attention on design enhancements that address concerns about policies and incentives specifically. Three findings in particular suggest that designing for conversations between data creators and data reusers could have a transformative effect on the depth of researcher that is done, the way data reuse is perceived, and the way contributions made in the context of secondary reuse are valued. These are:

1. in environments where data reuse occurs, primary research is often considered more valuable than secondary reuse and researchers desire to investigate their topics more deeply than what is required to achieve a publication
2. the types of knowledge that researchers lack about data are primary knowledge that is not present in the available data
3. researchers seek to obtain this knowledge through conversations with data creators

I discuss these findings and their implications further in the next chapter.

# Chapter 5 Discussion and Conclusions

In this final chapter, I provide an overview of my research, including my theoretical framework and major findings. I then discuss my findings and their implications in the context of other research. Finally, I present limitations of my research and concluding thoughts.

## 5.1 Overview of Research

### *5.1.1 Background and Theory*

I embarked on my research inspired by possibilities for making new discoveries and obtaining new knowledge in science and scholarship from the reuse of data from prior research. I was skeptical, however, of policy positions that assigned an inherent value to data sharing and wished to place these positions and their underlying assumptions in a critical light. I was concerned, with others (e.g., Mauthner, 2014), that blanket policies advocating for more sharing and wider availability of data would not necessarily serve to advance science and scholarship. My concern was premised on findings about the difficulties involved in reusing data (e.g., Niu, 2009a; Carlson and Anderson, 2007; Andersson and Sørvik, 2013) and findings that data sharing policies could actually serve to harm or detract from research (e.g., Borgman et al., 2007; Vertesi and Dourish, 2011; Mauthner, 2014).

I undertook my research to better understand the types and scope of knowledge about data that researchers sought to obtain when reusing data and how they determined their knowledge was "enough". My goal was to contribute to an understanding of the information important to archive with data, or link to, to best support secondary research. I was also

interested in whether learning about what influences researchers decisions could yield insights more generally into the practice of science in today's research environment.

I took a bottom-up approach in my research. I sought out researchers who had reused data and asked them in a survey if there was knowledge about the data they desired but were not able to obtain at the time they decided to reuse the data. I interviewed some of those that I surveyed to better understand how they determined—regardless of whether they obtained all the knowledge about data they desired or not—how much knowledge was enough to decide to reuse the data. I referred to the process of determining how much knowledge was enough as "knowledge bounding."

Because of challenges prior research has identified in defining "context" (e.g., van den Berg, 2005; Moore, 2007; Carmichael, 2017) and defining "data" separate from "context" (e.g., Moore, 2006; Stvilia et al., 2015) I adopted a theoretical framework that understood data and context together as comprising an "object of knowledge" or "body of evidence" (which I refer to generically as "data") that researchers investigated and used in their research. My framework further understood the boundaries defining these data as fluid and contingent, emerging differently in different reuse scenarios depending on the nature of the data themselves, the perspective the researcher brought to the data, and ongoing interactions between the researcher, data, research process, and other factors (such as the social environment) that might affect the research process.

The forgoing ideas drew upon on Haraway's theory of situated knowledges, which asserts that there is no one perspective from which an object of knowledge can be known. She argues instead that to the extent that we achieve can objectivity, we do so by recognizing the partial perspective that we have on our surroundings, and how that perspective is situated in

relation to others'. Considering the knowledge we have to be both partial and situated makes faithful accounts of our world and objects in the world depend on exchanges or "conversations" between individuals who bear those perspectives. Objects, meanwhile, or "data" in my case, are no longer static or fixed but rather become actors as well in these conversations as their boundaries shift and change, becoming differently perceived, and being productive of different perceptions, in different contexts. This framework allowed me to embrace the complex relationship between data and context and identify a concept, knowledge bounding, that I could operationalize to measure knowledge about data in a research study (whether it might be perceived as knowledge about data or knowledge about context).

Since my project was geared towards understanding the knowledge researchers seek to obtain about data and how that knowledge might be made available, I additionally saw my project in the context of design (i.e., of data and data archives with the goal of maximizing the value of data reuse). Simon's science of design, or "sciences of the artificial" (Simon, 1994) provided a framework for thinking about how people adapt to environments characterized by complexity and uncertainty through the design of "artifacts" (such as archived data and data repositories). I used Grint and Woolgar's (1997) notion of configuration to further conceptualize the way that researchers adapt themselves to reuse data, in particular, by acquiring knowledge about data. I combined elements of Grint and Woolgar 's (1997) and Simon's (1994) ideas to propose that when researchers reach a point where they decide that data are sufficient enough representations of real word phenomena to use in their research (i.e, they have configured themselves sufficiently to reuse the data), they reach a state of "reuse equilibrium."

The final component of my theoretical framework was Simon's theory of satisficing. I used the concept of satisficing to empirically measure knowledge bounding as it was manifested

by researchers. Simon conceives of satisficing in the context of "procedural" rationality (how decisions are made) rather than "substantive" rationality (what decisions are made). As such, it occurs when a person's cognition (a feature of their inner environment) is overwhelmed by the decision possibilities present in their outer environment, causing them to switch to a heuristic strategy to make a decision. In this strategy, they begin with a desired goal and use heuristics to proceed through a problem space, adjusting their expectations downward from their original goal until they reach an "acceptable" outcome.

I modified Simon's conception of satisficing in my operationalization in two ways based on Haraway's assertion that situated knowledge is about tension and resonance rather than dichotomies. First, I did not distinguish between features of a person's inner or outer environment in defining attributes that could affect their decision-making. Second, I allowed that satisficing might occur in the context of substantive rationality. That is, even when a researcher could conceptually understand that different data were better for a research project, they might still "settle" for data that were less optimal to reuse for their purposes. In this way, I did not make assumptions about the factors that could affect decisions researchers made to reuse "acceptable" data rather than data that were optimal for their investigation, or what kinds of decisions might be affected (procedural or substantive).

This approach was similar to the one taken by Agosto (2002) in her study of satisficing in that, while being focused on satisficing behavior (in information retrieval), her methodology was sufficiently open that it was able to detect aspects of both the inner and outer environments that affected her research subjects' searching behavior. It also had similarities at a general level to the naturalistic decision making approach taken by Berryman (2008), in that it allowed for a more holistic measurement of factors in decision-making, rather than being restricted to a certain set of

variables. The key differences in my approach were that I sought to investigate satisficing behavior in particular (a difference from Berryman) and go beyond the reasons users stop their searches (a difference from Agosto). Specifically, I sought to examine the particular phenomena of knowledge acquisition and knowledge bounding in decision-making: i.e., do researchers proceed with "acceptable" rather than "optimal" amounts of knowledge when deciding to reuse data? Whether they do or not, I investigated at what points they determined their knowledge about data was "enough" to reuse them, and what influenced such determinations.

### *5.1.2 Overview of Major Findings*

The major findings from my study were first, that researchers in my sample were not able to reuse data to the degree that they desired; they wanted to do more. Second, I found that in spite of this, satisficing did not well characterize researchers' reuse behavior. Instead, I found that researchers' strategies for reaching reuse equilibrium involved a complex interplay between their personal determinations about the sufficiency of data and social norms and requirements. My findings suggest that facilitating "conversations" between those who create data and those who reuse them could enhance the value of conducting research with secondary data. In the current environment such conversations already occur but opportunities to conduct them or to receive desired outcomes through them are not regularized or evenly distributed.

My findings are important because my evidence indicates that reuse takes place in a research environment where the achievement of social equilibrium (what researchers need to know about data to meet professional aims), is prioritized over personal equilibrium (what researchers would like to know about data to satisfy their desire for discovery and new knowledge). In particular, reuse is conducted in an environment where (a) reusing data is not valued as highly as creating or collecting one's own data and (b) researchers are under

significant pressure to achieve professional aims (i.e., to publish or demonstrate competence in research). Many researchers choose to reuse data to achieve those aims because they do not have the time or financial resources to conduct primary research. As a result, researchers—particularly those early in their careers—may reuse data as a steppingstone to obtain the requisite position or funding to conduct their desired research in the future, as opposed to reusing data as an end in themselves.

I propose that to the extent that a gap exists between the achievement of personal equilibrium (what researchers desire to accomplish through data reuse) and social equilibrium (what is necessary for researchers to accomplish to succeed as academics), opportunities exist to enhance the value and impact of data reuse. This proposition is complicated by findings that researchers who are early in their careers rely on advisors, mentors, and others to help them develop a sense of personal equilibrium, and that, because of the importance of peer review, achieving social equilibrium is required to pursue an academic career. These factors notwithstanding, circumstances where researchers are not able to conduct the research they would like, or to the depth they would like, represent new knowledge that is not obtained or not obtained to the desired degree. Cases where it might be possible and feasible to facilitate deeper research present opportunities to increase the impact and value of secondary research, and, by extension, the value of primary research (since the money invested to create the data would continue to yield gains in knowledge). There are a cascade of advantages to research that could come from increasing the value of secondary research, including increasing the prestige ascribed to conducting secondary research and to sharing data that facilitate valuable secondary research. This could result in incentives to produce and share data that are well-curated for secondary

reuse and bring the vision of making new discoveries and creating new knowledge with secondary data closer to a reality.

It is important to emphasize that none of my findings detract from the value of ensuring that data from prior research are well curated and well documented—quite the opposite. The importance and value of quality data and documentation are evident throughout. What is new in my findings is evidence of the importance to researchers of knowledge of or about data that are not made available and the negative impact that lacking this knowledge has on the outcomes of their research.

## 5.2 Discussion of Major Findings

I now examine findings enumerated above and the evidence that supports them in relation to what has been found in prior research.

### *5.2.1 Researchers Were Not Able to Reuse Data to the Degree That They Desired: They Wanted to do More*

It is well established that researchers struggle with issues of poor quality and completeness of data documentation and of data themselves when reusing data (e.g., Reed, 1992; Bishop, 2007; Corti and Thompson, 2004; Enke et al., 2012, Yoon, 2016). Researchers also contend with issues stemming from not "being there" during the original research—which may actually arise whether or not the researcher was "there" (e.g., Medjedović, 2011)—and of not having a say in the design of the original research. The first problem challenges researchers because they do not have tacit knowledge about the data that original creators have (hindering use and interpretation of data) (e.g., Pasquetto et al., 2019). The second introduces issues of data "fit," where the data do not necessarily support answering the questions a researcher has (e.g., Moriarty et al., 1999, Curty, 2016).

Although Niu (2009b) and Heaton (2004) found that researchers may continue to reuse data despite gaps in their knowledge about the data, we know that this comes at a cost and that lacking knowledge limits the scope of what researchers can accomplish through reuse of data (e.g., Bishop, 2007; Corti and Thompson, 2004). My findings confirm what is known in this regard, as well as the deep importance of data documentation and the data themselves to researchers obtaining the knowledge they desire to successfully reuse data (see, e.g., sections 4.7.1.1 and section 4.7.3).

The main contribution I make to discussions about researchers' ability to fulfill their research aims is my finding that, among all the kinds of knowledge that researchers would like, knowledge that was not present in the data—supplementary knowledge, or what I categorized as data supplement (coverage), data supplement (detail), and data supplement (missing)—was the type of knowledge researchers most frequently desired. This is also the kind of knowledge that researchers most frequently were not able to obtain to any degree, and it accounted for a large proportion (just over half) of the cases where researchers reported that the outcomes of their research were limited in some way.

Secondarily, and on a smaller scale, I found that researchers who lacked desired knowledge about data and were ultimately successful in their reuse of data (i.e., who adjusted their research or reported they obtained knowledge to a satisfactory level) commonly met their initial goals for reuse "as expected." They rarely exceeded their expectations or investigated their topics of interest as fully as they would have liked (evidenced by the fact that they achieved their goals "as expected" knowing that they lacked knowledge they desired about the data).

These findings are important because many data curation efforts today focus on ensuring the quality and usability of data and documentation (see, e.g., Wilkinson et al., 2016; National

Academies of Sciences, 2018 and much related work on FAIR data). My findings revealed that the most common source from which researchers obtained the supplementary knowledge they desired was original data creators. While the benefit data reusers obtain from being in contact with the original creators of data is well documented (e.g., Zimmerman, 2008; Faniel and Zimmerman, 2011; Poole et al., 2016; Pasquetto et al., 2019; Andersson and Sørvik, 2013), researchers who lacked supplemental knowledge about data obtained only 36% of the knowledge they desired, on average, from any source. Along with observations that one-to-one interactions with data creators do not scale (Faniel and Zimmerman, 2011), my findings indicate that something more than a focus on data and documentation quality, or the post-hoc repair of these, could aid researchers in maximizing the attainment of their reuse goals. These are the "conversations" I have referred to and discuss further below.

### *5.2.2 Satisficing Did Not Well Characterize Researchers' Behavior When Reusing Data*

That researchers were not able to do everything they wanted with data appear to contradict my finding that "satisficing" did not provide a good characterization of researchers' behavior in knowledge bounding. A key element of this latter finding, however, is that researchers did not believe it was possible to obtain an optimal amount of knowledge about the data they reused. Thus, instead of adjusting expectations down from an optimal, as satisficing would suggest, they adopted a "reuse mindset" in which they used creativity and skill to make new discoveries and contributions building up from the data that were available.

This may seem like a subtle difference, if a difference at all. After all, in both cases, the reality is that researchers do not meet their desired aims. What I discovered in my research, however, rather than strictly satisficing behavior, was a complex interplay between personal and social goals for reusing data that cannot be explained by satisficing alone. On one hand,

researchers sought to select data that they knew or evaluated ahead of time to be sufficient for a social equilibrium (i.e., to achieve a publication or demonstrate competence). But researchers also had their own research aims that they sought to maximize through their reuse of the data. Their desire for maximization led to frustration with the data (evidenced by the negative impact researchers reported lacking desired knowledge had on their research). At the same time, however, researchers were able to meet their goals for social equilibrium and, frustrations not withstanding, reported that they met their reuse goals "as expected."

The complex behavior I discovered could account for the reason Faniel et al. (2016) did not find a significant relationship between the relevance of data and satisfaction with reuse, and why Borrás' model did not explain the causal relationship she hypothesized between motivations for professional advancement and data reuse. Faniel et al. explained their result by suggesting that their data did not include researchers who reused data that were not relevant. However, it could be that some researchers reused data successfully from a professional point of view (i.e., they were satisfied with their reuse from one point of view), but did so with data that were not entirely relevant to their personal aims (i.e., they were dissatisfied in their reuse). Similarly, a lack of consideration for personal motivations and maximizing behavior could explain why Borrás' model did not explain reuse of data as evidence for scientific claims.

The difference between satisficing behavior and a "reuse mindset" are thus tangible and significant. Moreover, a misperception that satisficing characterizes data reuse behavior could possibly account for the difference between data reuse being seen as form of research on a par with primary research, and something that is "cheating," less valuable, or less prestigious to undertake. I address this possibility in the following paragraphs.

The benefits of reusing data are well documented (e.g., Moriarty et al., 1999; Rew et al., 2000, Borgman et al., 2007; Yoon, 2016), including savings in time and resources and fewer impositions on studied populations, which is especially important in cases where collecting data is difficult for political or other reasons. Despite these benefits, however, according to my findings, while some change is occurring, data reuse continues by and large to be viewed by my participants on the lower end of a hierarchy where greater value is ascribed to primary data collection and primary research.

Much has been written about social dynamics and pressures in academia related to data sharing. These writings propose different cultural models of sharing of scientific ideas through publications and data. Some propose academia as a "gift" culture where researchers share with the expectation of receiving in-kind at a later time (e.g., Hagstrom, 1982; Wallis et al., 2013). Others propose models of sharing based on credibility or reputation (Latour and Woolgar, 1982; Fecher et al., 2015b; see also the discussion of data sharing in Klump, 2017). In discussing some researchers' hesitancy to share data, Mauthner and Odette (2013) note,

> Data sharing activities take place within a social and political context marked by power differentials between respondents, researchers and research agencies, and power relations amongst researchers. These hidden politics of open access may also account for obstacles to data sharing. (p. 60)

While nuanced analyses of data sharing have been conducted, little research investigates cultures of data reuse in academia. Faniel and Yakel (2017), examined disciplinary practices related to data sharing and reuse, but their analysis focused largely on practices of data collection and deposit that enabled reuse, as opposed to motivations and influences that lead to reuse. Other related research has investigated reuse from the perspective of challenges researchers face to share data they have produced (such as lack of time and funding) (e.g., Zimmerman, 2003; Fecher et al., 2015a), but few examine data reuse practices or cultures specifically.

Birnholtz and Bietz (2003) do this to some degree in their examination of social dimensions of data sharing in earthquake engineering, HIV/AIDS research, and space physics. Part of their findings discuss the status data can attract in communities of practice. For instance, some of their informants in earthquake engineering suggested there was a stigma associated with reusing others' data that could "affect one's chances at getting papers accepted in journals and conferences, as well as the respect that is accorded by the community" (Birnholtz and Bietz, 2003, p. 343). They also found that possessing data, or possessing data of quality were markers of status in space physics and HIV/AIDS research, respectively. They noted that in fields where there is less agreement about the problems to be solved and the means of solving them, "much of the creative effort involved in the conduct of research goes into the design of the experiment itself and conceiving of novel ways to collect data." In these instances, "Simply analyzing somebody else's data bypasses this step, which arguably "counts" for less in the battle for scientific reputation" (Birnholtz and Bietz, 2003). My suppositions about perceptions of satisficing in data reuse echo these findings and I return to these ideas further below.

Curty et al.'s (2017) research is some of the most pertinent that focuses primarily on data reuse. They conducted a worldwide survey of scientists in different fields to investigate attitudes and norms that predict data reuse. Most of those surveyed were in the natural (50%) or physical sciences (20%) and 12% were in a grouping of "health, social, and humanities" researchers. Curty et al. examined attitudes about the efficiency and efficacy of data reuse and concerns about the trustworthiness of the data. With respect to norms, they examined factors such as whether researchers felt pressured to collect their own data and whether they only received credit for conducting primary research. They found that the perceived importance of data reuse to

advancing science or a researcher's personal career predicted data reuse and that perceived norms discouraging data reuse did not.

While Curty et al.'s research explored researchers' response to norms, it did not examine the causes or roots of those norms, how norms affected decisions during data reuse (i.e., about bounding knowledge), or the relationship between how reuse was valued and how it was supported. The responses they received furthermore expressed researchers' intentions to reuse data rather than their actual behaviors. Additionally, the factors they used to predict data reuse behavior, including attitudes and norms, and in some analyses experience with data management, ultimately explained an effect of only between 28-36% of the effect on data reuse, meaning that a large portion of the effect on data reuse remained unaccounted for (potentially giving additional credence to the role of complex behavior I described above in data reuse behavior).

One of Curty et al.'s findings that is interesting in light of my research is that they did not find a correspondence between those who shared data and those who reused them. They concluded from their data that behaviors that drive sharing of data are different from those that drive reuse, making these two distinct. Zimmerman (2003) obtained similar results in her study of ecologists, noting, "Social and cultural factors do play a role in data sharing, and they can make the reuse process more difficult, but they do not change the overall approach that ecologists take toward the secondary use of data." (p. 194). These findings contrast with those of Enke et al. (2012), who found a correspondence between those who shared and reused data in a survey of biodiversity scientists. Nevertheless, Curty et al.'s and Zimmerman's findings are important in the context of what I have found because they suggest that the correspondence of power differentials and relations in the context of data sharing and data reuse cannot be taken as

given. Instead, much about what affects data reuse behavior remains to be discovered and merits its own inquiry separate from other behaviors.

Moore's (2007) work examining reuse of qualitative data is also relevant to discussions about cultures of data reuse. Writing about anxieties sociologists may have about archiving data for future reuse—for example, reluctance to expose methodologies and interpretations used in their research in the data they share—she notes that "Research, when written up, is not only a report on research, but is also the construction of an academic career" (p.10). Moore argues that problems of archiving and reuse data are often framed in "terms of the 'professional' academic's concerns about the methodology, epistemology and ethics of research" but that "there is little explicit attention to how the construction of a professional academic career impacts on debates about reusing data" (Moore, 2007, p. 10).

Curty et al.'s and Moore's research both acknowledge the impact that considerations about professional careers have on researchers' behaviors surrounding sharing and reuse of data. These support to some degree my finding and Borrás' assertion that considerations about career advancement play a key role in decision about data reuse. However, neither Curty et al. nor Moore engage in a deeper examination of motivations and impacts surrounding this consideration or how it might contribute to a broader culture or cultures of data reuse.

Borrás' (2020) study is unique in centering the needs to make a contribution (expressed as a publication) or achieve a career milestone as the primary motivators of data reuse. Her research is based on the notion that researchers "make decisions that are nested in larger decisions related to their career goals and influenced by institutional and organizational norms, and power relationships. Their decisions are also influenced by their own personal values, feelings, and personal and familiar situations" (p. 40). Borrás' research subsumes these

assumptions into a causal model. However, she does not focus explicitly on exploring or examining the power relationships, values, or feelings themselves.

There are many possible explanations for why researchers in my sample reported an under-valuing and even de-valuing of data reuse. Epistemological arguments about whether data actually "can" (in a philosophical sense) or should be reused given the problems of not "being there" could make researchers wary of being associated with reuse. The higher prestige accorded to original research could result from the fact that sharing data outside of a publication is a relatively new possibility—traditional publishing outlets may still be unfamiliar with and unsure of how to treat research conducted with secondary data. The fact that part of the work (i.e., data collection) is done by others could make data reuse appear to require fewer skills than primary research.

The creative aspect of data reuse has been observed by researchers in the past (e.g., Rew et al., 2000; Zimmerman, 2003; Specht et al., 2015), but not as often as one would expect given the large amount of literature examining challenges data reusers face in reusing data. Zimmerman (2008) even contrasts secondary research with "creative" primary research when explaining why data sharing takes place in general among close associates in ecological research. She notes some reasons include "few incentives for ecologists to share and a culture that values creative and independent research above secondary use of data" (p. 636). A lower value being ascribed to data reuse in the social sciences could, similar to what Birnholtz and Bietz (2003) reasoned in their study, be due to a perception that it involves less creativity and ingenuity than primary research. With the exception of publishing outlets being unfamiliar with research with secondary data, any of these explanations could be associated with a general perception of data reuse as a "satisficing" or "good enough" strategy for conducting research.

My research makes a contribution to discussions about how to better support reuse by highlighting that (a) researchers' decisions to reuse data involve a complex interplay of personal and social dimensions, (b) primary and secondary research are perceived and valued differently among the community of researchers who reuse social science data, and (c) this difference could have a broader impact on the environment for data reuse. In particular, perceptions of data reuse as satisficing or "less than" may contribute to a lack of interest in conducting secondary research and lessen the perceived value of sharing data (since what prestige may be gained from discoveries made with reused data is assigned to the data reuser rather than the data creator). This is all the more troubling since, according to my findings, satisficing does not accurately characterize what happens in the course of data reuse. Research into the social context and power relations that influence data reuse could thus be a productive area of future research in the social and other sciences given heightened interest in supporting and enhancing the value of reuse.

#### *5.2.2.1 Satisficing as an Analytical Model*

I proposed early in my research that if I were to use satisficing to understand the way researchers bounded knowledge about reused data, I would need to expand the concept of satisficing to allow that factors beyond those defined by people's inner environment as Simon conceived it in relation to satisficing (i.e., their cognitive abilities and constraints) could affect an individual's decision to choose an acceptable rather than an optimal solution. I proposed that this expansion was needed, while allowing that satisficing could be evaluated by assessing differences between initial and terminal aspiration levels when initial aspirations are known (as opposed only to when an optimal solution is not conceivable). I further noted that this conception of satisficing could contribute to a theory of bounded rationality that spanned both procedural and substantive considerations.

I found the model of satisficing productive in its conception of initial and goal states and was able to use it to assess how much of what kinds of knowledge researchers desired about data but lacked. However, I was challenged to measure satisficing in relation to knowledge about data. This was for two reasons. First, I did not include in my survey a measure of whether researchers knew what would constitute optimal knowledge about data (either in kind or amount) at the time their decision was made. Thus, while I was able to measure satisficing in a substantive context (what researchers knew they wanted to know but were not able to obtain), I was not able to measure satisficing in a procedural context (which would occur if researchers were not able to conceive of the knowledge they desired and instead used heuristics to determine how much knowledge was enough).

Second, as discussed in section 4.2.2 above, I found through my interviews that researchers did not believe an optimal amount of knowledge existed because of the nature of reusing secondary data. In the particular context of my research, then, I did not find satisficing as understood in its original (procedural) context to contribute to my understanding of data reusers' behavior. I also did not gather evidence to support a re-conception or expansion of the notion of satisficing. My primary finding was that satisficing was not a productive concept for understanding decisions related to knowledge attainment about data in secondary reuse. The most important takeaway in this regard may be that satisficing does not always apply in situations where individuals appear to select "adequate" solutions to problems or make "satisfactory" choices.

### *5.2.3 Facilitating "Conversations" Between Those Who Create Data and Those Who Reuse Them Could Enhance the Value of Conducting Research With Secondary Data*

In Haraway's scholarship, "conversations" are a means by which stakeholders bearing situated perspectives bring those perspectives together to produce more faithful representations

of the world. As Haraway quotes from King: "Rational knowledge is power-sensitive conversation" (King, 1987a) (Haraway, 1991, p. 196). In the context of data and data repositories, I have proposed that data resulting from scientific research are themselves representations of the world, and that researchers evaluate those representations to determine whether they can reuse them to conduct their own studies (thus achieving "reuse equilibrium" if they decide they can). A discussion about conversations as I have framed it, then, is a discussion about constructing the rational knowledge that is available to be used—i.e., data that are deposited in an archive or made available in some other way—to produce more faithful accounts of the world.

A special feature of this data in the framework of situated knowledges is that as an "object of knowledge," data have the possibility to regain their agentive status. That is, the data created in a research project may no longer be taken as given (fixed) from one perspective but rather as an agent that can change over time and affect and contribute to subsequent knowledge production from multiple perspectives. When the agency of data is taken into account, what the data are and how they are comprised becomes important not only from the perspective of data creators, but of data reusers as well, and the data become part of a larger conversation about the evidence we gather and analyze to produce scientific knowledge.

Challenges of data sharing and deposit are well documented, including concerns about loss of control or misappropriation of data (e.g., Enke et al., 2012), moral and ethical concerns related to confidentiality, privacy, and consent (e.g., Bishop, 2009, Kwek and Kogut, 2015), as well as practical issues cited above related to time and resources to prepare and deposit data (e.g., Tenopir, 2011, Kwek and Kogut, 2015). There are also a significant number of standards and guidelines that exist for researchers about data sharing and deposit, including at ICPSR (e.g.,

ICPSR, 2022). Despite these, however, as I and others have found (e.g., Zimmerman, 2003), researchers remain unsure about what data to share (e.g., only data relevant to a given paper, or all data) or are selective about what they do. There is thus an opportunity to facilitate conversations between data creators and data depositors about what data that currently exist can be shared or documented to enhance support for data reuse. These conversations currently take place facilitated by "intermediaries" (Zimmerman, 2003) such as data repositories and centers for analysis and synthesis of data from disparate sources (Smioski, 2015).

My research highlights the value of facilitating and enhancing conversations at a different point in time, however: before the data are created. While such conversations between data creators and reusers take place, as I have found, I did not find examples of research investigating how they take place, how often, who is involved, or other details. I do know from my research that such conversations can be adhoc and different projects go to different lengths to gather information about who might have a stake in their data or how the data collected might be updated.

My contribution in this arena is to call attention to evidence suggesting that transparent and more intentionally structured channels facilitating conversations between data creators and data reusers about what data are collected could serve to maximize the value and impact of both primary and secondary research.

#### *5.2.3.1 Situated Knowledges as an Analytical Model*

I incorporated Haraway's theory of situated knowledges into my methodology in two ways. First, adopting the premise that boundaries of data are contingent on social factors and situated knowledges, I designed my interviews to investigate a broad range of considerations, pressures, and concerns that might have affected researchers in their attainment of knowledge

about data. Second, I sought to interview clusters of researchers who had reused the same data in order to obtain multiple perspectives (situated knowledges) on reuse of the same data. I thought that doing so would produce more comparable results across investigations of different reuse scenarios—i.e., I believed understanding multiple perspectives on the same data could reveal interesting similarities and differences in how researchers bounded knowledge, both within and across disciplines.

I found the use of situated knowledges to be very productive in the first instance. As I noted in section 5.1.1, leaving interviews open for researchers to discuss a broad range of personal and social factors allowed a holistic assessment of their decision-making and uncovered important aspects of their reuse environments that I may not have learned about otherwise.

In the second instance I did not find situated knowledges as productive. I mentioned in section 3.4.8.5 that I had difficulty determining whether or not similarities or differences in knowledge bounding and influences on knowledge bounding in a cluster of researchers were due to researchers reusing the same data. I attributed this difficulty to researchers giving differing levels of detail in response to my questions (for instance, about considerations, pressures, and concerns surrounding their reuse) and me changing the way I asked questions about knowledge bounding as I proceeded through the interviews. Adding to my difficulty, in a number of cases I was only able to conduct an interview with one person who reused a given data study (rather than multiple reusers as I wished).

While I note the variability in the questions I asked and responses received as a limitation in section 5.4 below, the variability also highlights some challenges in applying the theory of situated knowledges to digital data. One of these is agreement about what the "object" (digital data in my case) is. I asked researchers about the knowledge they desired about data but lacked.

However, researchers had different conceptions of what the data were, evidenced by the fact that some described knowledge "about" data and some knowledge "of" data. The different conceptions were likely influenced by the fact that while a cluster of researchers might have reused the same data study, individual researchers had different research questions and reused different variables or portions of evidence from the study.

Extrapolating to the theory of situated knowledges, a challenge may arise in the ability to reach agreement among a given "field of interpreters," about what the "object" is that individuals have different perspectives on based on their situated perspectives. I was able to observe results and draw conclusions based on an analysis of my entire pool of interviewees. However, a different level of care would have been needed to reliably measure and compare perspectives on the same data.

It is possible that the difficulty I experienced actually exposes and validates a strength of the situated knowledges framework—namely, that it demonstrates that objects of knowledge are truly differently bounded for different individuals. This difference in bounding could pose a challenge, however, when an object becomes the subject of conversation among multiple perspectives. What if perceptions of what an object is are so different that they hinder productive conversation among situated perspectives? There may also be a challenge if the perspectives available, or that speak, represent only a small number of those that might be included. The notion of conversations is powerful, but Haraway provides little guidance as to how conversations might be structured or proceed to achieve their intended purpose (i.e., more rational accounts of our world).

In the following section on implications of my research, I discuss how Simon's conception of design and other ideas give insight into and augment a picture of how conversations about data could be structured and oriented to improve support for data reuse.

## 5.3 Implications of Research

### *5.3.1 Designing for Reuse*

My research has focused from the outset on questions of design—the insights we might gain to improve support for data reuse by bringing together ideas of Simon, who articulated the "science of design" (Simon, 1988) and "sciences of the artificial" (1994), and Haraway, who described situated knowledge and its implications for the ways "objects of knowledge" are bounded and understood. Mediating these perspectives is the notion of the "social-technical gap," or the gap between "what we know we must support socially and what we can support technically" (Ackerman, 2000, p.179).

I proposed that by elucidating data as differently bounded in different situations, situated knowledges provided a framework for investigating and understanding the nature of this gap with respect to data reuse. Alternately, I proposed that Simon's perspective on design might shed light on specific mechanisms by which the conversations Haraway describes might be realized (i.e., by which the perspectives of "fields of interpreters" might be heard and not only the perspective of those who claim "objective" knowledge).

By applying the situated knowledges framework, my research identified the facilitation of conversations between data creators and data reusers as a possible way to enhance the design of data artifacts (and perhaps data archives as well, to the extent they are implicated in the production of data "artifacts"). Here, I consider how the science of design can inform the construction and facilitation of these conversations.

I noted earlier the element of Simon's curriculum for social design that speaks to designing for the future, or "designing without final goals" (Simon, 1994, p. 190). Simon argues that what we are doing when we establish final goals is actually setting initial conditions for our successors. With this in mind, he advocates creating "a world offering as many alternatives as possible to future decision makers, avoiding irreversible commitments that they cannot undo" and leaving "the next generation of decision makers with a better body of knowledge and a greater capacity for experience" (Simon, 1994, p. 187). The purpose of doing these, he says, is not just to enable them to better evaluate possibilities but also to "experience the world in more and richer ways."

Simon's ideas about designing for future generations resonate with remarks given by Russell Ackoff on a similar topic. Ackoff is well-known for the hierarchy of data, information, knowledge, and wisdom he theorized, but is perhaps less well-known for the purpose for which he articulated this hierarchy. The address in which he did so was oriented toward the development of information systems. After describing the hierarchy and his estimate that human minds consist mostly of data, followed in lessening amounts by information, knowledge, and understanding, and, finally, virtually no wisdom, he said:

> Managers of systems are currently drowning in seas of symbols spewed out by mature computer-based management information systems (MIS). More sophisticated computer-based knowledge systems are still young. Younger still are systems that generate understanding. Ones that generate wisdom have yet to be born. *Of what would such a system consist?* It is to this question that this paper is addressed. (Ackoff, 1989, p. 3, emphasis in original)

Ackoff develops an argument in which he describes that information, knowledge, and understanding focus on efficiency, while wisdom is oriented towards effectiveness. He distinguishes the concepts by associating efficiency with growth, which does not necessarily imply an increase in value, and effectiveness with development, which does.

In pursuit of criteria for a system to generate wisdom, Ackoff argues that a new approach to ethics and morality is required that is based not on conformity to rules of conduct but rather on the way decisions are made. In particular, he articulates a principle that "*All those who are directly affected by a decision (the decision's stakeholders) should be involved in making that decision* (Ackoff, 1989, p. 7, emphasis in original). Yet he notes that one stakeholder group, larger than all the rest, is often ignored: that of future generations. He argues that future generations must be allowed to make their own decisions, which requires keeping their options open. While considering long-term interests is something people are not generally disposed to do, Ackoff says that those who have wisdom are recognized as being able to balance these interests with those of the short-term. In short, to paraphrase, wisdom is the ability to see the short and long-range consequences of actions and evaluate them with the goal of preserving peoples' future ability to make their own choices and decisions.

Where do "conversations" fit into this discussion? Ackoff believed that only people were capable of wisdom since wisdom, unlike efficiency, can never be programmed. Moreover, he argued that while information ages rapidly, knowledge ages somewhat more slowly, and understanding is more durable still, wisdom, unless lost, is permanent. Without entering a discussion of the current or future capabilities of artificial intelligence, an important piece of guidance on the structure and function of conversations emerges from this discussion.

This is that conversations be structured to generate and preserve wisdom. On the surface, this seems difficult and intractable. If research involves bounding knowledge, how could it be possible to preserve the ability of all current and future stakeholders to make the most of the data in their research (i.e., to seemingly have access to unbounded knowledge)? This is a valid concern, but I would make a few observations. One is that it is reasonable to start with currently

known stakeholders rather than all potential stakeholders. A second is that not all data have the same function or impact or carry the same expectations. There are data of broad interest that are produced with public money, for example, for which people have greater interest and desires for input, accountability, and return on investment. An initial assessment of research projects parallel to the process of institutional review (as an example), could determine whether select projects (and communities of likely reusers, based on an evaluation of prior research) could benefit from additional review along defined parameters. There are many concerns here, and in the end, the value of additional review to secondary researchers and the public must be weighed against considerations, timelines and interests of the primary researchers.

A third observation is that one of the greatest values that expanding and structuring conversations could contribute is the networks of people of which they are composed and their assets of experience, understanding, knowledge, information, data, resources, and time. If wisdom exists in people, then wisdom can be preserved through conversations involving networks of people. The siloed nature of academia is often the subject of critique. Expanding and strengthening the network of connections between researchers in academia could bring data, knowledge and understanding out of siloed pockets and serve as buffer against their decay. This is especially important in light of the social-technical gap between archived data and reusers (what I have defined as the difficulty of fully representing in a package of deposited data all that occurred in, or all that potential data reusers might like to know about the original research). At some point, connections between primary researchers and their data begin to fray.

Conversations, then, structured with a goal of preserving wisdom in mind, could serve to broaden the depth and scope of research conducted with secondary data, increase the value of conducting research with secondary data (and of depositing and sharing primary data), and

preserve the ability to do both of these over time. Facilitating conversations could also increase the possibility of researchers getting to "know someone" that has the particular knowledge or expertise they need to advance in their research (see section 4.7.3.2.1).

The question might still be asked, especially among those not compelled by Simon's or Ackoff's reasonings about future generations or wisdom, of why data creators and other stakeholders should be motivated to invest time and effort to aid secondary researchers in achieving their personal goals for research. After all, if researchers who reuse data are publishing results and advancing knowledge in their fields and their own careers through those contributions, is there any reason to modify current practice or focus on more than data sharing and data or documentation quality and usability?

There are a few responses to this. One reason is explicit policy aims and expressed desires to maximize innovative potential and return on investment in funded research (e.g., National Research Council, 2003; Holdren, 2013; National Science Foundation, 2018; National Institutes of Health, 2020). Ultimately, if we are to do this, we should be concerned as much with the cost and impact of conducting research as we should about advancing scientific knowledge. Responsible use of research resources means obtaining as much value as possible out of funded research. It also means conducting research without diluting or diminishing the environment for research. If populations are investigated to saturation, or if populations lose trust in the scientific endeavor (whether it is because of perceived poor stewardship of public resources or other reasons), the overall ability to conduct research is diminished, which reduces favorable outcomes for everyone. Researchers know less about populations, reducing the potential to successfully address societal problems, among other concerns. Responsible use of resources is thus in everyone's interest.

Another reason is that researchers are not satisficing—at least not in ways it might be perceived. They are not settling for research that is "good enough." They are maximizing the use of data they determine, given constraints on time and resources, is best to achieve a complex blending of personal and social aims. To the extent that their personal aims outstrip the social, all can benefit by exploring opportunities to meet these in fiscally reasonable and responsible ways. It is possible, as well, that increasing the impact of what can be accomplished with secondary research could increase the prestige ascribed to conducting secondary research and therefore the prestige (for primary researchers) of sharing data that facilitate deeper discovery. This kind of incentivization may be what is needed to maximize "new knowledge from old data" (Zimmerman, 2008) and our progress towards achieving more faithful accounts of ourselves and our world.

As noted above, none of this is to suggest that there is any less value to focusing on existing concerns surrounding data that have already been created. As numerous studies have shown, and as my research confirms, great challenges exist in discovery, access, documentation quality, data quality, development of standards, and numerous other areas. A unique contribution of my research, however, is to highlight the additional benefits that supporting conversations between data creators and reusers could yield.

### *5.3.2 Designing for Reuse Communities*

While my discussion has focused on "conversations" between those who create and reuse data, my research has implications for stewards of data as well (those who preserve and provide access to data for secondary reuse). A perennial challenge of enabling reuse is the impossibility of conceiving of all the possible ways researchers might reuse data and the knowledge they will desire to make those uses. My finding that the strategies researchers use to negotiate gaps in

knowledge about data are situated in broader social settings support an approach to data stewardship that is oriented towards communities rather than individuals alone.

Much work has been done to identify the properties of digital objects that are significant to their preservation and future reuse. Faniel and Yakel (2011) classified prior work on the identification of significant properties of digital data as falling into two categories: studies of significant properties that were data-centric, focusing on the properties of digital objects that needed to be preserved in order to render digital objects over time, and studies that were people-centric, focusing on properties that were important for stakeholders (members of the designated community) to understand and make sense of the data over time. Hedstrom and Lee (2002) developed a model for expressing significant properties of digital materials that could be used to analyze and mitigate risks in preservation. The model included the notion of expressing properties in relation to an entire "designated community." Designated community is a term defined in the Open Archival Information System reference model, which provides a conceptual framework for preserving digital information such that it can be accessed and understood by a "designated community" of users (Consultative Committee for Space Data Systems, 2012).

My research offers a means of thinking about significant properties of digital data in relation to reuse. It suggests that in addition to investigating data- and people-centric properties, significant properties can be identified by seeking to understand how researchers make decisions about reusing data. Since researchers' decisions (at least in an academic context) are situated in and influenced by social settings, this means paying particular attention to how contributions and competence are evaluated among different communities of reusers, what the standards for these are, and what implications these have for how data are curated and collected to facilitate reuse. It might be acceptable in certain communities, for example, to use proxy measures for concepts

related to mental or physical health, where in others it might not. Different communities might have different standards for how or whether measures are validated. Insight into these kinds of norms could provide guides to help data stewards determine the relative suitability of data for use in different communities and tailor stewardship activities accordingly.

There is a significant degree to which, when we think about means of improving the environment for data reuse, we are talking about the needs and choices of individual researchers. However, a defining feature of the scientific enterprise is its reliance on peer review. Individual choices matter, but decision-making in science, to a significant degree, reflects collective values and collective norms of practice. Studies geared towards improving data stewardship and other aspects of the reuse landscape where researcher decision-making is relevant could benefit by taking that collective orientation into account.

### 5.4 Limitations of the Study

My research had several limitations. One was that, because I did not identify specific types of knowledge I was interested in learning about, researchers reported two different kinds of knowledge. These were knowledge "about" data (generally, what might be considered "context," such as how data were created), and knowledge "of" data (which might be considered "data," such as subjects' BMI or socioeconomic status). Although I struggled somewhat with this reporting and with distinguishing kinds of knowledge, in the end I decided that any knowledge that researchers desired but lacked about data was of interest. This is in line with my theoretical framework (which blurs distinctions between data and context). However, it seems to have resulted in some researchers reporting specific kinds of knowledge that others did not, presumably because they were not conceiving of knowledge in the same way. This led, I believe, to an underreporting of desired knowledge lacking, which I discovered when reading the papers

that described my reuse instance of interest prior to my interviews, and in the course of interviews themselves.

A second limitation, similar to the first, was that there was variability in the extent of answers researchers gave to questions in the interviews, and also the specific questions I asked (since my interview questions developed over time and the interviews were tailored to some extent to the researchers personal experiences). I therefore obtained much information from some researchers, for instance, about how they negotiated gaps in their knowledge about data or why a lack of knowledge did or not affect the outcomes of their research in a negative way, but less information about other areas of interest. This variation prevented me from assigning meaning to similarities or differences I observed among researchers I interviewed who had reused the same data. I was not able to determine whether behaviors were due to reusing similar data other factors. Despite this limitation, I did observe patterns across the full set of researchers I interviewed with respect to my primary phenomena of interest in the interviews: how researchers determined how much knowledge was "enough" and what influenced them to do so. These patterns were consistent and strong enough that they yielded meaningful results and meaningful information with respect to my research questions.

A third limitation, that I did not discover until I was preparing to conduct the interviews, was that I had not actually measured procedural dimensions of knowledge satisficing in the survey the way I intended. To have measured knowledge satisficing in Simon's sense, I would have needed to ask researchers whether they knew ahead of time how much knowledge about data they would need to decide to reuse the data, and also how much knowledge they obtained. As it is, I could not tell from the survey responses whether researchers' cognition was overloaded in thinking about how much was knowledge was enough and they used a heuristic to determine

this, or whether they knew in advance and their decision about how much was enough was a substantive decision. Although I incorporated this question into my interview protocol, measuring satisficing in this way turned out not to be as important I as initially thought because of my finding that satisficing did not well characterize researchers' experiences reusing data: they did not seek or expect to obtain an "optimal" amount of knowledge. This lowered the significance of measuring satisficing in my study and put the focus more squarely on behaviors of knowledge bounding.

Several other limitations included bias in the survey and interview samples, and possible self-censoring, social desirability bias, and recall effects in the interviews. I sought to address bias in the survey sample by using random sampling techniques and employing statistical methods to correct for bias that might come from the sample including researchers who reused the same data. I also included control variables in my inferential analyses. One bias I could not address by sampling was the fact that since some institutional members of ICPSR allocate more funds than others to track data reuse in particular areas, the Bibliography does not comprehensively represent reuse of data held in ICPSR (E. Moss, personal communication, 9/6/2019). Reuses of data not being recorded in the Bibliography would introduce error in the precision of my results. Another factor that could affect precision is my choice to limit my study to entries in the Bibliography that cited data "studies" as opposed to data "series" (some entries in the bibliography cited only the broader series of which granular data studies were a part). This excluded a sizable proportion of studies from consideration (837 out of 11,954, or about 7%), which could reduce precision (and inference transferability). While the size of the Bibliography is large (including thousands of works), and my sample size was large enough to achieve 99%

precision (or confidence level), these limitations must be taken into account in interpreting my results.

Self-censoring may have occurred in some interviews, but, as my findings show, most researchers were very open in their discussions with me. I detected little social desirability bias. Aside from annoyance I encountered from some researchers when asking how they knew they had enough knowledge (section 3.4.4), I did not get a sense that researchers regarded my own opinions about the responses. If anything, I felt a lot of desire on my own part for researchers to think well of me so they would be comfortable to express themselves freely.

Recall effects were definitely at play. The most common result of this was that researchers gave general answers about their reuse practices rather than ones that were more specific. This contributed to my hesitancy to draw granular comparisons or make assumptions about sub-groups of researchers based on the data they reused). There were also cases when researchers could not recall why they had responded certain ways in the survey when I asked. I did not see this as much of a problem because while they may not have remembered specifically, most researchers proposed why they answered that way or gave a current answer.

One final limitation related to tacit knowledge. I asked researchers about knowledge they sought and factors that affected how they bounded knowledge, but these would only have encompassed knowledge and factors researchers were explicitly aware of. My study was not able to assess tacit knowledge researchers desired and obtained (or did not) about data or factors that affected knowledge bounding that researchers may not be aware of.

### 5.5 Future Research

My research raises multiple questions that point to areas of future research. Some of these that are the most pertinent relate to conversations, communities, cultures of data reuse, factors

that affect knowledge attainment, theories of decision-making, and the study of knowledge satisficing and knowledge bounding in the reuse of data produced in other domains.

### *5.5.1 Conversations*

My research led me to the conclusion that an impactful and potentially unexplored means of enhancing reuse of secondary data is to support the achievement of researchers' personal aims for reusing data, or personal reuse equilibrium. My research further points to facilitating "conversations" between data creators and data reusers about the data that are created in the first place as an effective strategy for doing so. However, there is much that we do not know about how these conversations take place currently and much that we could learn to understand how best to facilitate them. Some relevant questions include:

- What conversations currently take place about the creation of data that is later deposited in the ICPSR data archive?
- Who is involved? For instance, what role, if any is played by secondary reusers, other data creators, funders, or ICPSR itself?
- What influence do these conversations have on what data are created and why?
- What needs, opportunities, and risks exist for facilitating conversations between data creators and data reusers?

In wide-ranging research on how researchers discover and reuse data, Gregory et al. (2020) note the significant importance of social interactions in researchers' personal networks to enabling their discovery, access, and reuse of data. They propose supporting these interactions in the design of systems and repositories that enable reuse, including by facilitating interactions between data creators and data reusers. They say,

> We could also see a role for an expanded metadata schema as a way to open a conversation between data producers and multiple data consumers within the context of the data themselves at the repository level. In such a system, both data producers and data consumers could contribute information about the data to specific fields. Data producers could describe their own use of the data as well as their concerns about what potential reusers need to know about the data before reusing them. Data users could describe how they used or plan to use the data, as well as pose questions about the data to other users or to the producer. (Gregory et al., 2020, p. 47)

The conversation they propose relates to data that have already been created as opposed to data yet to be created. However, their proposition raises interesting questions about the best ways to facilitate interactions between interested parties. In my interviews, many researchers talked about the importance of workshops to getting to know others, including data creators, and become plugged-in to networks of data reuse. Research is needed, however, into the most effective means of building communities, keeping in mind considerations about fairness in opportunities individuals have to join, participate in, and contribute to them.

### *5.5.2 Communities*

An essential question that is part and parcel of considerations about any kind of social interactions (conversations not least) is in what context, or "community" the interactions take place. In their research, Gregory et al. (2020) considered questions of community as well. They found that the diversity of practices they encountered in data discovery and reuse made it difficult to conceive of designing tools based on individual user profiles. They further noted the difficulty of classifying users according to disciplinary domains, based on the significant number of researchers who identified as belonging to two or more (about 50%) or three or more (about 25%) disciplinary domains. They proposed, instead of individual users, and supplementing considerations about disciplinary domain, to conceptualize communities of reuse in terms of reuse purposes (e.g., to answer a new research question, to integrate data with other data, etc.). They argued that, "communities of use may have similar data needs where specific contextual information supports particular data uses and research stages" (Gregory et al., 2020, p. 253).

My proposal to consider collective values and norms seems to run contrary to Gregory et al.'s findings. However, it is based on the notion that there is something about community norms (including norms in specific disciplines) that merits further investigation. This is the extent to

which norms about what constitutes a contribution or demonstration of competence within a community influence researchers' decisions about what data to reuse and how much they need to know about the data to reuse them.

Taking all of these considerations into account, questions for future research about communities include:

- In what ways and to what extents do researchers' identification with "communities," broadly construed, impact decisions during data reuse?
- Are there communities researchers do not identify with that affect their decisions (e.g., researchers who publish in the same or similar venues)?
- What implications do considerations about community belonging (conscious or not) have for facilitating conversations between data creators and data reusers?
- Do considerations about how to facilitate conversations, or the value of conversations differ by community?
- Do what conversations take place, who is involved, what influence conversations have, or the needs, opportunities, or risk of facilitating conversations vary according to community? If so, how?

There are additional issues to consider having to do with the fact that many data reusers are also data creators, and that perceptions of the value of data reuse vary in different communities. How might belonging to multiple communities or the value of data reuse being perceived differently in different communities influence considerations about facilitating conversations to enhance data reuse?

In pursuing these avenues of research, it is especially important to take an approach that is (a) "situated"—for instance, that acknowledges the range of ways researchers might conceive of "belonging" to a community and incorporates multiple perspectives on what constitute conversations and communities; and (b) oriented toward the future—e.g., that recognizes that conceptions and definitions of conversations, communities, and other terms may change over time.

### *5.5.3 Cultures of Data Reuse*

An area of future research closely related to communities has to do with cultures of reuse. In section 5.2.2, I discussed gaps in research on cultures of data reuse in academia and noted the value of further research into the social context and power relations that influence data reuse—particularly into the reasons for the differential value placed on reusing data compared to conducting primary research. A more nuanced understanding of these dynamics could identify more deeply seated reasons for the challenges researchers face in reusing data and potentially lead to strategies for mitigating those challenges in the future.

### *5.5.4 Factors Affecting a Lack of Desired Knowledge*

Future research could shed light on more specific factors that affect researchers' ability to obtain all the knowledge about data they desire as well. I found when comparing models for lacking knowledge in section 4.3.2 that the model including variables that were significant in univariate analyses at a level of $p < .25$ had a lower QIC value and was thus preferred over the more parsimonious model that included variables significant in univariate analyses at a level of $p < .05$. Importantly, there were three variables that were significant in the $p < .05$ model that were not in the $p < .25$ model. These were (a) working in a research domain outside of the 16 most common and reusing data for either purposes of (b) replication or reproduction, or (c) to test a theory. Future research into factors that affect lacking desired knowledge about data could clarify the associations between these variables and lacking desired knowledge, particularly by examining interactions between these and other variables in the model.

### *5.5.5 Theories of Decision-Making*

An additional area of future research concerns the implications of my findings for the decision-making theories of satisficing and procedural rationality. There are at least three related trajectories.

First, my work complicates a theory of human behavior based on either substantive or procedural rationality because both of these rely on the notion that for a person with a goal there is an optimum way to achieve that goal. My research raises the question: what theory explains behavior when the actor does not believe there to be an optimum outcome?

Second, my work opens exploration into a potentially different kind of procedural rationality. Procedural rationality is concerned with how decisions are made, but procedural rationality does not specify satisficing as its only mechanism (Barros, 2010). Satisficing is a particular theory premised on limitations in a person's ability to compute all possibilities and consequences in a problem space. The fact that researchers make decisions that are not premised on such limitations when making decisions to reuse data (e.g., a decision to continue to reuse one data study over another because they already have experience with it), suggests another mechanism of procedural rationality apart from satisficing. My research uncovered factors that explain why this might occur (e.g., constraints on time in the face of challenges learning to reuse new data and pressures to meet professional goals). However, more research is needed to understand how this decision-making aligns with current theories and to better support researchers in reusing data they feel are most advantageous to study their phenomena of interest.

Third, my research suggests that researchers may satisfice in the selection of data (since they choose data that are minimally suitable to achieve their professional goals) but not in the attainment of knowledge about data. Future research is needed into the first and second-level choices researchers appear to make about reusing data. This research could confirm or better characterize the role of satisficing in first-level decisions, aid in understanding how those decisions differ from second-level decisions involving knowledge bounding, and the implications both of these have for supporting data discovery. For example, is there more that

could be done to make first-level decisions more like second-level ones (i.e., driven by researchers interests as opposed to professional benchmarks) and would that even be desirable?

### *5.5.6 Research Beyond Reuse of Social Science Data*

Finally, my dissertation investigated knowledge satisficing and knowledge bounding in the reuse of data produced in social science research. Since all academic research shares the premise of peer review, there is reason to believe at least some processes of knowledge bounding and associated personal and social considerations are similar in other academic domains. However, there is important work to be done to investigate all of the areas mentioned above—conversations, communities, decision-making, and factors affecting knowledge attainment and knowledge bounding—related to reuse of data produced in other domains as well.

## 5.6 Conclusion

As a PhD student, myself seeking to demonstrate competence in research, this dissertation has been a journey on many dimensions. I have created much data through my survey and interviews. I have reused data from the ICPSR bibliography and by harvesting metadata about data studies from the archive itself. As someone trained in history who has not always seen the distinction between "literature" and "data," I would say I have reused much data from literature as well. I have also struggled to make sense of my own collected data over time as I have sought to reproduce and reinterpret results as my research and analysis skills have evolved. I have not been unaffected by researchers' expressions of the pressure they experienced as graduate students or the struggles they faced to complete their degrees and succeed in academia. Nor have the dynamics of listening to researchers speak about a research hierarchy while their speech provided primary data for my study been lost on me. The research process has thus been eye-opening, therapeutic, and humbling.

My research has also been a chance to “try on” aspects of Haraway’s theory of situated knowledges and consider my own identities in my research. I am a student, a researcher, a father, a husband, a brother, a son, a relative, a friend. I am other things I do not have the desire or courage to divulge. I am a person who is white, who has access to time and resources. I am also vulnerable emotionally, financially, and professionally. What do I make of researching people I may be close to in many ways, and in others very far?

Haraway says that,

> A commitment to mobile positioning and to passionate detachment [characteristics of partial perspective and situated knowledge] is dependent on the impossibility of innocent 'identity' politics and epistemologies as strategies for seeing from the standpoints of the subjugated in order to see well. One cannot 'be' either a cell or molecule - or a woman, colonized person, labourer, and so on - if one intends to see and see from these positions critically. 'Being' is much more problematic and contingent. Also, one cannot relocate in any possible vantage point without being accountable for that movement. Vision is always a question of the power to see - and perhaps of the violence implicit in our visualizing practices. With whose blood were my eyes crafted? These points also apply to testimony from the position of 'oneself'. We are not immediately present to ourselves. Self-knowledge requires a semiotic-material technology linking meanings and bodies. Self-identity is a bad visual system. (Haraway, 1991, p. 192)

She continues:

> The split and contradictory self is the one who can interrogate positionings and be accountable, the one who can construct and join rational conversations and fantastic imaginings that change history. Splitting, not being, is the privileged image for feminist epistemologies of scientific knowledge…The knowing self is partial in all its guises, never finished, whole, simply there and original; it is always constructed and stitched together imperfectly, and therefore able to join with another, to see together without claiming to be another. Here is the promise of objectivity: a scientific knower seeks the subject position not of identity, but of objectivity; that is, partial connection. (Haraway, 1991, p. 193)

> …So location is about vulnerability; location resists the politics of closure, finality, or, to borrow from Althusser, feminist objectivity resists 'simplification in the last instance'. That is because feminist embodiment resists fixation and is insatiably curious about the webs of differential positioning. There is no single feminist standpoint because our maps require too many dimensions for that metaphor to ground our visions. But the feminist standpoint theorists' goal of an epistemology and politics of engaged, accountable

> positioning remains eminently potent. The goal is better accounts of the world, that is, 'science'. (Haraway, 1991, p. 196)

When I began my research, I thought that satisficing was a given and that my project would be finding the patterns of knowledge that were lacking; that future work would involve seeking to assemble this knowledge into archival packages and knowledge networks to enable researchers to make the most of archived data.

I did not consider that an assumption of an "optimal" might have resonances with Haraway's concept of the "god-trick" (an objective view from a single perspective) or that there was any possibility that researchers might have input into the data that were collected. I attribute insights I have had to my attempts to acknowledge my partial perspective and be open to the perspective of others. It is difficult to perpetually exist in a place of partiality with a split and contradictory self. And yet it makes sense to me, as someone in search of better and more faithful accounts of the world (of 'science', as Haraway says), that such a position would lend itself more readily to engaging in conversation with similarly seeking others.

My findings supporting the potential value of facilitating conversations among data creators and data reusers has thus been a personal realization as well, supplemented by insights from Simon's conception of design about how conversations might be structured. As Haraway's and Simon's perspectives themselves are partial, I expect to continue learning and growing in my understanding of their words in the future. In this dissertation, however, they have been excellent conversation partners.

# Appendices

## Appendix A: Survey Protocol

### Introduction

This survey is part of a doctoral research study in the University of Michigan School of Information about the contextual information researchers use when reusing data. The survey is completely optional. You may choose not to answer specific questions or end the survey at any time with no penalty.

Please answer the questions in relation to the data and publication referenced in your survey invitation email and respond as accurately as possible.

In some questions, the numbers 1 to 5 appear in brackets at the end of answer choices. These indicate that the answer choices lie on a scale from 1 to 5.

The survey uses the following definition of Data reuse:

Data reuse means to use data collected in prior research (sometimes referred to as "secondary data") in the conduct of current research. The prior research could have been conducted by yourself or someone else for the same or different purpose(s).

If you combined previously collected data with newly collected data (e.g., in a longitudinal study) that is considered data reuse.

If you only referenced or described data from another study (e.g., in a literature review) that is not considered data reuse.

### Section 1 of 3

1. According to the definition above, did you reuse the data referenced in your invitation email in the research reported in the associated publication?
2. How involved were you in the decision to reuse these data?
3. How many other people aside from yourself were involved in the decision to reuse these data?
4. How involved were you in determining the goals of the research?
5. Were the original data data produced with an explicit purpose to be reused by others?
6. How involved were you in the original research where the data were collected or produced?
7. How would you describe the knowledge of your research team as a whole (including yourself and others involved in conducting the research) about aspects of the original data collection that were relevant to your reuse of the data?
8. Please answer Yes or No to each statement. For what purpose(s) did you reuse these data?

a. To provide background or context for the research (e.g., to develop a questionnaire or obtain calibration information)
b. To validate or corroborate research results
c. To answer a different question from the original research
d. To replicate or reproduce the original research
e. To combine with other data
f. To compare with other data
g. To test a theory
h. To develop or test an algorithm or tool
i. Other (please specify)
j. The data were cited but not reused

9. For each "Yes" response please indicate how important it was to you, at the time you decided to reuse the data, to reuse them for that particular purpose.
10. Please indicate your level of agreement with the following statements.
    a. I already had a research question or questions in mind
    b. I changed my initial research question(s) as a result of knowledge I obtained about the data (e.g., its extent, quality, or representativeness)
    c. I determined the research question(s) as I explored the data that were available
    d. Other (please explain)
11. Some studies listed in the ICPSR Bibliography of Data-related literature obtained the data from a different repository. Did you obtain these data from ICPSR or a different source?

**Section 2 of 3**

12. When you were considering whether to use these data in your research, was there any specific knowledge about the data that was important to fulfilling your reuse purpose(s) that you lacked, or that was limited in some way?
13. At the point when you decided to reuse these data in your research, was there knowledge about the data that you desired but were not able to obtain or only able to obtain in a limited way?
14. Please describe the three most important things you would have liked to know about the data but were not able to learn to the desired degree.
15. On a scale from 0 to 100 where 0 is no knowledge and 100 is optimally desired knowledge, please indicate how much of each type of knowledge you were able to obtain.
16. How important to you was each type of knowledge in determining whether to reuse the data?
17. Did a lack of knowledge about the data in these areas negatively affect the achievement of your desired research outcomes?
18. Please briefly explain why the lack of knowledge did or did not negatively affect the outcomes of your research.
19. Please indicate the information source that was most important to you in obtaining each type of knowledge.
    a. Original data creators
    b. Colleague(s
    c. Advisor(s
    d. Data repository staff
    e. Data documentation

f. The data themselves
g. Personal knowledge
h. Team knowledge
i. Literature
j. Other (please specify)

20. If you entered "Other" as the source of any knowledge above, please specify what information source(s) is applicable for which type of knowledge.
21. For each of your original reuse purposes, how well did reuse of the data meet your reuse goals?

**Section 3 of 3**

Please answer all questions below based on your status at the time the decision to reuse the data was made.

22. Considering roles such as Principal Investigator (PI), Co-PI, data manager, and project manager, please describe your role in the research.
23. With what department(s), center(s), or other organizational unit(s) were you primarily affiliated?
24. What was your professional position (e.g., PhD student, Assistant Professor, etc.)?
25. What was your primary domain of research?
26. Was the research you conducted with these data in your primary domain of research?
27. How many years of research experience did you have in your primary domain of research?
28. In what domain of research were the data produced?
29. About how many years had you been reusing data produced in this domain in your research?
30. About how many years had you been reusing data produced in any domain in your research?
31. How often did any of the research you conducted involve reusing data?
32. How often in any of your research did you reuse data for the same purpose(s) for which you reused these data?
33. Approximately how many projects had you worked on that involved data reuse?
34. In how many of these cases were the considerations involved in reuse substantially similar to the ones involved in reusing these data?
35. Approximately how many published papers had you authored or co-authored describing research involving data reuse?
36. In how many of these papers were the considerations involved in reuse substantially similar to the ones involved in reusing these data?

Please use the following definitions to answer the questions below:

Beginning - Just learning about considerations in conducting research with secondary data.
Intermediate - Comfortable conceiving of and executing research using secondary data
Advanced - Advanced knowledge and experience conceiving of and executing research using secondary data

37. At the time the decision to reuse the data was made, how would you characterize your level of ability reusing data produced inside your primary domain of research?
38. At the time the decision to reuse the data was made, how would you characterize your level of ability reusing data produced outside your primary domain of research?
39. Would you be willing to be contacted with follow-up questions?

## Appendix B: Interview Protocol

### Introduction

Thank you very much for agreeing to participate in this interview. I wanted to start by telling you a little bit about myself and my research. I'm conducting research to understand how to better support researchers such as yourself who reuse data from data archives. I was a librarian for many years in a previous life, working in areas of digital preservation and digital repository management. For a time, I was the assistant director of a large partnership of libraries called HathiTrust that manages a collection of millions of digitized books. I come to my research, then, in part from the perspective a repository manager motivated by providing services to others, and in part from the perspective of a researcher with my own interests in reusing data.

I have questions in my interview about your reuse of data in a specific project, and what I'm really trying to get at is your own personal experience conducting reuse. My study is based on the premise that when we conduct research, we bring to it all of our own experiences, assumptions, physical and mental skills and abilities, and other sensibilities; essentially all of who we are. All of these drive our interests in what we study and, in particular for my study, what we seek to know about the data we use in our research. I'm trying to learn how much and what kinds knowledge we might advocate for to be included in archives of data in order to fulfill our desires for the kinds and depths of research we would like to conduct. So, in answering the questions, I'd like you to reflect broadly on your experiences and motivations to help understand as accurately as possible how we might enhance our archives to better support our work.

### Explanation of Interview Conduct

My goal in the interviews is to set questions up as clearly as possible and listen to researchers talk. There are specific things I would like to know, which will structure the interview, but I want to guide researchers in their specific responses as little as possible. The draft protocol below lists the concepts I would like to ask questions about (e.g., background about the research) and specific questions. My hope is that, through testing, I will be able to make the interviews more conversational, where I obtain the information I would like (including by asking probing questions), but in a way guided by the participants' responses. In other words, I would like to talk about all the areas of interest to me, but do so by pursuing lines of thought as they come up in the conversation, so that not all interviews may have the same exact question order or structure.

### Interview Questions

Background about the research

- I'd like to start with a little bit of background. I wonder if you would talk a little bit about how you became interested in the area of research you see this project as belonging to, and then this project in particular.
- And could you set the scene for this project and give some additional details?

    - I'm wondering things like how the idea for the project came up, how it took shape, and who was involved?
    - [If data not from ICPSR]: I note from the survey that you obtained the data from [source]. Could you talk some more about how that came about?
- What were some of the main considerations or pressures influencing how the research developed and proceeded?
- You mentioned in the survey that members of your research team had [amount of] knowledge of the original data creation. Would you talk a little bit about how that influenced your choice of the dataset and how your research took shape?

Decision to reuse the data

- Would you talk a little bit about the specific data you reused and how you came to the decision to reuse these data in particular?
    - [Follow-up] What led you to choose these data over others?
- Did you have any concerns about reusing the data, either in general or about reusing these data in particular? Would you talk some about those?
- Were there any specific guidelines for data reuse that you followed or consulted, either in terms of methodology or specific instructions for using these data?
- Thinking about those with whom you interact most frequently about your research and who might be involved in commenting or evaluating your research: how is data reuse perceived within this community? Is it something that occurs all the time, something researchers are more reticent to engage in? If so, why?

Bounding knowledge

- I'm interested in learning how researchers know when they've obtained sufficient knowledge about the data to reuse them.
    - [If no knowledge reported lacking] What were some of the most important things you needed to know about the data to reuse them?
    - [If knowledge reported lacking] There seemed to be some places where there was more you would have liked to know about the data. For instance, you mentioned [specific areas of knowledge].
- Would you talk some about, in each of these cases, why the knowledge was important and how you knew these data would fulfill your needs for research? How did you know you had reached a certain threshold where the knowledge you had was enough and what influenced you in making that decision?
- Did you have ideas about how much knowledge would be enough ahead of time or did an idea develop as you reused the data?

Goal attainment

- In terms of how well your reuse of the data met your research goals, in the survey you gave the following responses [repeat responses]. Would you explain a little more why you responded that way?

Sources of knowledge

- What were some of the primary sources you used to obtain the knowledge you desired about the dataset? Would you talk some about how you went about obtaining your

desired knowledge and what that experience was like?

Factors that might mitigate knowledge satisficing and its effects

- Are there particular factors you would say facilitated or got in the way of obtaining the knowledge you desired? What would have helped overcome barriers you experienced?
- Thinking more broadly about your experience reusing data, what are some of the major issues or shortcomings you see for those who seek to reuse data and what changes do you think would be most helpful to support those who are conducting research with secondary data

Closing

- Is there anything else related to obtaining knowledge about data or negotiating situations where desired knowledge is not available that you would like to add?

## Appendix C: Description of Pilot Survey

In November 2019, I distributed a pilot survey to 114 researchers. In conducting the pilot, I followed the sampling procedures and the guidelines for creating the survey instrument described in section 3.3.1. The pilot sample is drawn out of the sample of researchers I used in the full study and the pilot instrument is the instrument that, with a few additional changes, I used in the full study. In brief, the survey sample comprises unique researchers who produced a work in an eligible format (i.e., journal article, thesis, report, conference proceeding, book section, book, or electronic source; see section 3.3.1.1) between January 2014 and September 2019 in which they cited data archived in the ICPSR data repository.

For the pilot study, once I had identified the most recent work of unique authors, I separated the list into researchers who had cited quantitative data and those who had cited qualitative data. I randomly selected 5 of the quantitative data studies that had been cited more than 50 times and produced a randomized list of the researchers who cited each of the 5 data studies. I produced a similar randomized list of researchers who cited qualitative data studies. A research assistant looked up the email addresses of researchers on each list until we had contact information for 20 researchers who cited each quantitative data study and 10 researchers who cited a qualitative data study. I sent an email to each of the 110 researchers with a link to my web-based survey instrument (created using Qualtrics). If an email bounced, I either found a different email address for the researcher or selected a different researcher from one of the randomized lists. In this way, I sent emails to 114 researchers in all.

49 researchers responded and 43 answered enough of the questions about knowledge satisficing that I included their responses in my analysis. The analysis I conducted was primarily to assess the clarity of the questions (whether the questions produced responses that were unambiguous to interpret) and test techniques for producing descriptive statistics and conducting factor analysis and logistic regressions.

As a result of the pilot survey and analysis I added a new question and modified three others in order to improve their precision and therefore the clarity with which I was be able to interpret researchers' responses.

## Appendix D: Coding of Researcher Primary Domains

My final coding of researchers' primary domains is represented in Table 4.17. Researchers' responses, and my coding, reflected topical areas of research. I coded topical areas or domains at two levels: a broad level and a more refined level (the broad level is represented in Table 4.17). In analyzing the domains, I made heavy use of in vivo coding (as will be seen in the examples below). There were some instances where I created codes to group multiple similar areas together. For instance, I used the code "life course" for domains that had to do with the study of people at a particular age in live (i.e., juvenile, adult, older adult).

"Relationships" is another code that I used across multiple domains to group research on some form of relationship specifically (for instance, having to do with intimate partner violence, family relationships, or parent-child relationships). I used multiple codes when appropriate for researcher responses and generally coded as many domains as possible with the overarching goal to capture as many similarities as possible (i.e., as made sense) across researchers. However, as shown below, there were cases where I did not feel comfortable separating researcher's responses into multiple domains. I give specific explanations and examples of my coding below.

Many researchers entered multiple domains or research areas. For instance, "juvenile justice, violence, and mental health", "health and criminal justice" or "health care and public health management." If multiple areas were specified and were easily separable, I separated them. For example, "Art therapy, arts and human development" became

| L1 | L2 |
|---|---|
| Arts | Arts |
| Health | Art Therapy |
| Life course | Behavior (human development) |

If the domains were not easily separable, did not make sense in separated language, or I was not sure whether they should be separated, I kept them together. For example, "health-related judgement and decision-making" became:

| L1 | L2 |
|---|---|
| Decision-making | Decision-making (health-related judgement and decision-making) |
| Health | Health (health-related judgement and decision-making) |

In some cases where there were a combination of factors I combined coding practices. For instance, in the example below ("poverty policy effects on children and families"), I kept the full entry in the codes for Poverty and Policy, but separated "children" and "families" in order to code the entry separately for Life course (related to effects on "children") and Relationships (related to effects on families):

| L1 | L2 |
|---|---|
| Poverty | Poverty (poverty policy effects on children and families) |
| Life course | Behavior (poverty policy effects on children) |
| Policy | Policy (poverty policy effects on children and families) |
| Relationships | Behavior (poverty policy effects on family) |

Coding entries related to psychology was challenging because some researchers indicated their domain explicitly as "Psychology" while others entered a particular area such as "Child development." Moreover, some areas of psychology such as clinical psychology are more specifically related to health whereas others such as social or educational psychology have to do with social phenomena. Still other research areas spanned several of my coding areas, including Health, Justice, Life course, and Relationships. An example is "Child sexual abuse and intimate partner violence." This I coded in the following way:

| L1 | L2 |
|---|---|
| Health | Behavior (child sexual abuse) |
| Justice | Justice (violence — intimate partner) |
| Life course | Behavior (child sexual abuse) |
| Relationships | Behavior (intimate partner violence) |

As in the examples above, I separated two apparently distinct domains (child sexual abuse and intimate partner violence). Then, in this and other instances where the domain had to do with some form of psychological or behavioral research, I indicated this with the prefix "Behavior." In this example, I coded the entry as Behavior (child sexual abuse) in the second level of coding and placed this construct under the appropriate level 1 code (Health) since abuse is related to Health. I also placed Behavior (child sexual abuse) under Life course since the research relates to people in particular stage of life (i.e., children).

I stuck as closely as possible to the language used by researchers, which resulted at times in separate entries for similar items at the level 2 coding: for instance for Behavior (addiction) and Behavior (addictions); Behavior (suicide prevention) and Behavior (adolescent suicide prevention); and Health (behavior) and Health (behavioral). In some cases, I qualified entries to make them easier to group together in aggregate statistics. For instance, I listed epidemiology-related domains in the following way:

| L1 | L2 |
|---|---|
| Health | Health (epidemiology — aging) |
| Health | Health (epidemiology — arthritis) |
| Health | Health (epidemiology — cancer) |
| Health | Health (epidemiology — cardiovascular disease) |
| Health | Health (epidemiology — genetic) |

I was not entirely consistent in this practice because modifying researcher entries did not always serve to help aggregate responses and sometimes obscured meaning I wanted to preserve. For example, I coded "Trauma and behavior in juvenile justice" as follows:

| L1 | L2 |
|---|---|
| Health | Behavior (trauma and behavior in juvenile justice) |
| Justice | Behavior (trauma and behavior in juvenile justice) |
| Life course | Behavior (trauma and behavior in juvenile justice) |

I might have coded this as, e.g.,

| L1 | L2 |
|---|---|

| | |
|---|---|
| Health | Behavior (juvenile justice — trauma and behavior) |
| Justice | Behavior (juvenile justice — trauma and behavior) |
| Life course | Behavior (juvenile justice — trauma and behavior) |

But this would not have served to aggregate the responses with others that were Justice (juvenile) and keeping the full description helped preserve context and meaning of the researcher's response.

There were cases, however, where I coded entries one way under some level 1 categories and another way under another. Some examples were:

| L1 | L2 |
|---|---|
| Health | Health (substance use — youth) |
| Life course | Life course (youth substance use) |

| L1 | L2 |
|---|---|
| Justice | Justice (violence — intimate partner) |
| Relationships | Behavior (intimate partner violence) |

In these cases it was important for me to be able to see aggregations of coded responses together and I did not have concerns about misinterpreting the responses. In general, I formatted codes in such a way as to facilitate my understanding and interpretation of responses for my purposes. These included seeing broad trends at the level 1 level of coding, understanding how the research domains of researchers I selected to interview stood in relation to other researchers, and analyzing the research domains of researchers who reused the same data studies

In most cases, I coded responses into level 1 categories based on categories that were mentioned by the researchers, as opposed to making inferences as to how responses should be coded. I did this so as not to put too much weight on my own interpretations. There are some exceptions. For instance, I coded an entry for "Volunteering and charitable giving" under "Charity":

| L1 | L2 |
|---|---|
| Charity | Charity (charitable giving) |
| Charity | Charity (volunteering) |

In another example, I coded entries of “Elections” and “Political behavior” as Political science. However, some other entries that seemed relevant to Political science I did not code that way. Some examples are “Gender, race and politics” and “Race, ethnicity, politics, policies.” These, I coded as “Politics.” My reason for this, and this practice in other instances, was to avoid over-interpreting entries. In addition, since my main uses of the domain areas were to generally categorize the sample and to help understand where the researchers I selected for interviews fit into the larger group of researchers and not statistical inference, the finer level of granularity in these cases at level 1 did not take away from my analysis.

## Appendix E: Summary of Interviews

In this appendix, I present summaries of several instances of reuse by researchers who reused the same data, and findings from my analysis of these instances. As I stated at the end of section 3.4.8.5, my intention in presenting the instances is to give an idea of how the interviews proceeded and the kinds of similarities and differences I encountered, and also to demonstrate some of the process I went through to analyze interviews and the results of that analysis. My presentation is additionally intended to support the inference transferability of my research results (see section 3.6.3). I have also noted a couple of times the similarity of the reuse instances in my study to cases in a case study. In his discussion of case studies, Flyvbjerg (2006) cites a comment of Lisa Peattie (2001) that "The dense case study…is more useful for the practitioner and more interesting for social theory than either factual "findings" or the high-level generalizations of theory" (Flyvbjerg, 2006, p.238). Accordingly, it is my hope that the summaries of these reuse instances will prove useful to understanding the particular situations of researchers I interviewed and enable readers to more readily relate to and engage with their experiences.

The summary of my findings from analyzing this set of reuse instances represent patterns and themes I identified in my third cycle of analysis. Presenting these instances of reuse and my analysis provides a frame of reference for understanding the interviews findings I present in section 4. In the presentation of those findings, I have incorporated the findings from my analysis of these interviews (see E.2 below) into the larger body of all the interviews I conducted.

I begin my presentation of five instances of reuse with a brief description of the data the researchers reused and an introduction to the researchers using data gathered from the survey. I then provide a summary of my findings from analyzing interviews, followed by summaries of the reuse instances themselves.

## E.1 Introduction

The data study for which I present cases of reuse was a large, nationally representative survey that was well-cited in my sample. I interviewed five researchers who reused the study. Four of the researchers believed the data had been created to be reused and one assumed it had. Four researchers reported they were not at all involved in the original research, while one was very involved. Despite this, most had a very good knowledge about the original collection of the data (two researchers had in-depth knowledge, two had significant knowledge, and one had some knowledge).

With the exception of one researcher who reused data rarely overall, the other four researchers reused data often in their research. One researcher's research questions were fixed prior to deciding to reuse the data, and the others' research questions developed fluidly as they reused the data. Four researchers obtained the data from ICPSR and one from a different source. Only one researcher reported lacking desired knowledge about the data in the survey.

Two researchers were PhD students at the time they reused the data; one was a postdoc, one held an assistant academic position, and one held a full academic position. Four had between one and five years of experience in their primary domain and one had between 15 and 25. Two researchers had between one and five years of experience reusing data in their primary domain, one had five to 10, one had 10 to 15, and one had none. Two researchers had between one and five years of experience reusing data in any primary domain, one had five to 10, one had 15 to

25, and one had none. There was some overlap and some differences in the way the researchers described the domain of research that the reused data were in.

I use the same convention below as in the body of the dissertation to reference interviews. An example of a reference is D1-01. I use the first part of the reference (e.g., D1) to refer to a specific data study and the second part (e.g., 01) to refer to a researcher who reused that study. For instance, D1-03 would indicate a third researcher who reused the D1 data study.

Not all of the instances of reuse I summarize below are the same length. There are two reasons for this. One is that I strove to focus on and highlight portions of the interviews that were particularly relevant to my overall findings and the occurrence of these portions was not equal across all five of these interviews. The second is in taking care to protect the privacy of the interviewees, in some cases it was easier to present general summaries rather than quotes.

### E.2 Summary of findings

Some of the main ideas I drew from this group of interviews included the following:

- Researchers can be influenced to conduct research with secondary data when there is time pressure to complete research or when they do not have the resources to conduct primary research (D1-02, D1-03).
- Secondary data provide imperfect measures of concepts and phenomena researchers wish to study (D1-01, D1-02, D1-03, D1-04, D1-05).
- Since researchers are not able to control how or which questions were asked, the data researchers deemed "best" are often not optimal, but simply better than other options (D1-03).
- Researchers may not choose the data that are the "best" for their research question due to the time and work needed to learn a new data study (D1-03).
- Researchers do their best with available data (D1-03, D1-04), prioritizing the conduct of research that makes some contribution to existing knowledge (D1-03, D1-04, D1-05).
- Researchers frame limitations they encounter in secondary data as preliminary work that can reveal gaps in existing research and form the basis for grant proposals and other future research (D1-01, D1-03, D1-05).
- Research with secondary data is time-consuming; there is a lot to learn about data and analytical techniques. Researchers must "sit with," "commit to" and "immerse" themselves in the data to reach a point where they feel "comfortable" with the data and the results of their analyses; gaining "expertise" with data brings "comfort" (D1-03, D1-04, D1-05).

- Researchers can be intimidated about reusing data, but having access to the original data creators, attending meetings and workshops, and reading about others' reuse of the same data are important ways (in addition to "sitting" with data) that researchers learn about and become comfortable reusing data.
- Researchers are influenced by a diverse range of factors in determining how much knowledge about secondary data is "enough" to achieve their reuse goals. These include peer feedback, conversations with advisors, colleagues, and mentors, and feedback from conference presentations and journal review panels.
- There are different types of data reusers: researchers who reuse data as part of their normal research and researchers who reuse data because of a need or desire for more publications (D1-03).
- Expectations for reuse depend on the goals of reuse. For example, in a dissertation, where there is focus on demonstrating skills, criteria for success are different than a publication (D1-03, D1-04).
- Researchers see publication as a strong measure for successful reuse of data, but journals in different disciplines have different review criteria (e.g., with regard to age of the data) and researchers have choices about where they publish and direct their papers. If one journal is not a good fit, researcher can submit their paper to another publisher (D1-04, D1-05).
- There are personal characteristics such as being a "data geek" or liking or embracing an "ambiguous process" where you are not sure what the outcome will be that can predispose researchers to reuse data, or be advantageous for those who wish to conduct secondary analysis (D1-04, D1-05).
- It is important to have requisite analysis skills, have access to consultation, or ensure that someone on the research team has expertise in the methods to be employed (D1-03, D1-04, D1-05).
- Inability to have full access to data can cause anxiety (D1-01) and gaining access to data can be burdensome and time-consuming for some researchers (D1-02, D1-04). On the other hand, while it means data are restricted for others, having preferential access to data before it is made public can be advantageous (D1-01).
- Despite the difficulties and challenges of reusing data, secondary data offer tremendous possibilities for research.

### E.3 Researcher D1-01

This researcher's interest in their topic stemmed from prior experience as a working professional. They were first introduced to the data study by a mentor as a PhD student. The mentor believed the data would be a good fit for the researcher's interests. The researcher used the data in their dissertation and then subsequently in grant projects and proposals, so by the time they were reusing the data for this research, they already had a substantial amount of experience with the data. In addition, one of their mentors had published on the data in collaboration with

lead researchers on the grant project where the data were originally created. The researcher was able to draw on that mentor's expertise, and others involved in the original research with questions as they worked with the data. They also attended a training on how to use the data and other meetings with researchers working with same data.

The researcher was drawn to the data because of the particular sample population and because the data included variables that aligned with the theoretical construct they wanted to examine. The data could also be used to examine a broad range of outcomes. The researcher had a fixed research question, then, that they wanted to examine using these data.

The paper itself was developed with a colleague for a conference presentation. The researcher felt some pressure to advance the paper since they had presented on it. They did not want it to be a paper that stopped progress after a conference as they mentioned often happens. They also experienced some pressure to publish because the data were not public use at the time. The data were to be made public, however, and the researcher was concerned about getting their research published before the data were more widely available.

The researcher did not have too many concerns about the data themselves. The data were gathered using a complex sample design, which meant the analysis needed to be done a certain way. In addition, even though they had selected the data in part because of the particular variables available, they were also concerned that the variables did not directly measure their theoretical construct and they used some variables as proxy measurements. A final concern had to do with the age of the data.

When speaking about the perceptions of data reuse in the community, the researcher noted that the relevance of the research question and the study were more relevant than whether the data were secondary or primary. The community was aware of the limitations of using

secondary data, i.e., “you can’t always represent exactly what you’re looking to do” (D1-01), and they had never experienced skepticism about secondary analysis itself, other than the limitations of the data. They said that, for themselves, there was always a point in doing secondary analysis where they needed to ask questions that were not asked in the original study—that secondary analysis generated those questions. But in general, their experience presenting on these data was that people immediately saw the relevance and contribution of reusing them for research.

The researcher indicated in the survey they had not lacked desired knowledge about the data, so I asked about the most important things the researcher desired to know about the data. These were the range and type of variables collected, including those outside their area of focus; the sampling design (how the data were collected); the types of measures used, and where the measures came from (specifically to assess their reliability and validity). In terms of the data collected, the researcher knew the data were sufficient for their research questions because of their prior experience with them. In terms of measures, a few different processes were important to their reaching an equilibrium.

One was receiving feedback on the conference paper. After receiving the feedback they talked with the co-author about how to represent their concept and had to make judgement calls about what measures from the data were sufficient to do that. Another was the publishing process. They revised the variables they used to represent the theoretical construct based on reviews from their journal submission. A third was searching literature to assess whether they were representing their theoretical construct in ways that they were comfortable with for a publication and ways that it was consistently represented in the literature. They said that ultimately they reached a threshold of variables that was comfortable to them but the representation of the construct was still an ongoing research problem.

In this case, then, the main challenge the researcher faced was determining whether the data could be used to accurately reflect their construct, and they used community feedback, feedback from a journal, discussion with their collaborator, and searching the literature to reach a point of equilibrium.

The researcher could not recall why they described the data as meeting their reuse goals "as expected" for background purposes. They said in retrospect that they perhaps should have said the data met their goals "as expected" for testing a theory because the data did what they "were supposed to do…I was able to speak to this theory with the application of this data for this question" (D1-01). They responded "better than expected" for using the data to answer a new question because of the success that they and others had had reusing the data: "The ability to be able to generate new questions and ask them all of these data is just endless. It's endless really. I haven't touched some of the variables in the data set yet, and may-may not ever, but--so that's why I think I responded in that way." (D1-03)

Some particular factors that facilitated the researcher's reuse of the data were access to the data, access to researchers involved in the original data collection and a mentor with extensive experience with the data; also access to trainings to use the data and access to early articles written about the data to frame the background and methods of the project, particularly the complex sample. The main barrier they experienced was obtaining the knowledge to conduct the desired analysis on the data.

The main things the researcher thought could facilitate reuse is having a centralized place to access data on their topic of interest and video trainings that gave an introduction to specific data and how to use and analyze them.

### E.4 Researcher D1-02

This researcher's interest in their topic area came from reading and, like D1-01, from personal experience. The researcher was introduced to the data study as a teaching and research assistant as a PhD student and gained significant knowledge about the data through that experience. They were doing readings in their topic and came across a theorist they resonated with; they were inspired to use these data by reading papers of others who had reused them. They chose these data for their dissertation because the data are the most comprehensive available for studying their topic; they had a particular theory they wanted to test and needed the most comprehensive dataset possible to do so. The researcher preferred a model of developing research questions ahead of time and in some ways their research questions were fixed prior to reusing the data (since they were testing a particular theory). However, the data lacked certain measures that limited what the researcher could investigate. It is likely for this reason that the researcher indicated in the survey that their research question changed in the course of the research.

The primary pressure the researcher felt during the project was the time limit they had to finish their dissertation. As a quantitative researcher with limited resources to conduct original research, reusing the data made it possible for them to finish their dissertation on time. Concerns about cost also led the researcher to use the public use version of the data. This led to a final concern about their ability to test the theory using the particular population and in the particular way that they desired, because of measures not available in the data. (D1-02)

The researcher experienced positive perceptions of data reuse in their community. Their dissertation chair was a quantitative researcher who reused data and provided support to the researcher and the researcher had been an assistant for a professor who was reusing these specific data.

The researcher indicated in the survey they had lacked desired knowledge. The main knowledge they lacked were equivalent data about different populations and geographic locations, which I coded as data supplement (coverage). In describing how the researcher negotiated the knowledge they desired about the data but lacked, they described the tradeoff that is always involved in reusing data. Like D1-01, the researcher used proxies for certain measures because there was not direct measurement. “But in my heart and, and still to this day,” they said, “I know that these are proxies” and would never replace a question that asked what the researcher was looking for directly (D1-02). These data were better than other data, however, and the researcher was satisfied with what they were able to put together for the dissertation. They said,

> So when you're doing your calculations, you compare what you don't have to other data sets. They don't have even this much, so I should be kind of happy with what I have. This is the best. This is not the best, this is better than others in terms of addressing my bigger research question. But, [measure] still was a big, I think, thing that I missed, for my research. (D1-02)

The researcher consulted literature to determine what proxy measures would be best, but in the end, was not able to test all components of the theory because of the lack of certain variables.

I did not ask the researcher about achievement of their reuse goals because of the time we spent discussing other topics.

In terms of facilitators and barriers of the researcher’s specific work, and reuse more broadly, they spoke primarily about the kind of data that were lacking and how the lack of those data inhibit researchers generally in her area of research.

### E.5 Researcher D1-03

This researcher reused the data in their dissertation, which developed from their prior research. A particular variable in their area of interest stood out to them and they choose these

data because they wished to investigate the phenomenon among different populations. The researcher was introduced to the data by another researcher while working on a paper together. The researcher reused the data in a collaboration with a different researcher as well and gained familiarity with the data through those experiences. The researcher also attended a workshop on reusing the data. Other sources of knowledge about the data included email updates from ICPSR noting changes in the data, contact with the original researchers (to ask questions about matching the original analysis of data), and reading other studies that reused the data (to answer similar questions).

The primary pressure the researcher mentioned related to reusing the data was time pressure as a doctoral student: being pressured to finish in a certain amount of time. This pressure led the researcher to reuse these data even though there was another data study they were aware of that might have been a better fit. They reused these data because they there would have been a lot involved in learning to reuse the other data. They had already invested that time in reusing these data and could proceed more readily with them.

The main concern the researcher discussed is the same as one mentioned by D1-01 and D1-02: the limitations in measurement that come from secondary use rather than primary collection. Not having control over the way the question that addressed their key variable was phrased, or who was asked, limited the analysis the researcher was able to conduct. Similar to D1-01, the researcher also had concerns about the age of the data.

The researcher described that reuse was highly valued their community. The primary reasons were the ability to perform research—even if limited—in the absence of resources to perform primary research, and the ability to learn about and improve on limitations in prior research. Regarding the first, they said,

> So, we see it as a valuable resource because sometimes you may not have the financial resources to conduct primary data collection, but there may be data sets out there that you may be able to utilize to answer parts of your research question, maybe not all, but there might be opportunities to answer parts of your research questions. (D1-03)

And regarding the second:

> It might not be the PI who improves upon them [the data]. They might be quite happy with how they measure, you know, physical activity let's just say, or how they measure perceptions of need or how they measure service utilization. But for myself, who is interested in perceptions of need from a different angle or a different perspective, it's now my opportunity to, okay, well, I'm gonna take, what's been done before and I'm gonna improve upon that. (D1-03)

The researcher indicated that they did not lack desired knowledge. However, I identified multiple types of knowledge in reading their publication that it appeared they lacked, and that other researchers had identified in the survey as knowledge that they lacked. One of these, that I asked about, involved the inability to distinguish between two measures that had been combined into a single variable. I only asked about one of these because I wanted to take the opportunity (since the researcher reported they had not lacked desired knowledge) to ask the researcher about the most important things they needed to know about the data to reuse them. They responded that they needed to know whether the variables they needed for their research were present (whether the measures would be enough for their purposes), whether there were any methodological issues with the way variables were measured and if so, if they could be resolved through the application of statistical methods. They also talked about a question they would have liked to answer, and attempted to, but that the structure of the data would not allow to their satisfaction.

The researcher reached equilibrium concerning these issues in a variety of ways. First, regarding whether the measures that were present would be enough, they had a lot of conversations, with their dissertation chair, co-chair, and other faculty. These conversations helped the researcher think through statistical methods, theoretical concepts, and strategies of

analysis. Second, the researcher used statistical methods to deal with issues that arose in constructing a model using the variables in that study that was based in the literature. Finally, the researcher pointed to their training, both in statistics and in reuse of these data (during which time they were able to interact with the original project leads), as well as their practical experience working with the data. These were instrumental, they said, in their achieving a sense of comfort with data and confidence that the variables used, statistical tests employed, and overall analysis were, by and large, reasonable and sound.

With regard to the inability to distinguish measures that had been combined in a single variable, they said that they knew it was a limitation going in, but that in the greater scheme of things, the pros of using these data outweighed the cons. Subsequently, the researcher conducted their own study where they collected the measures and ended up combining them (similar to the data they reused) because there was little variation in one of the measures. In retrospect, then, they could understand why the variables might have been combined.

Their way of negotiating their desire to conduct analyses that the data would not support was a little bit different. The researcher themselves was not entirely convinced of their findings, but according to the parameters of their dissertation the analysis they had done was good enough. Their committee was primarily concerned with whether they had done everything they could to answer the research questions, and they thought through alternative means of analysis as thoroughly as possible. The researcher described it this way:

> Obviously, kind of the way we approach this dissertation is, or where I think the director of the PhD program at the time is, this is not your Magnum Opus. It's not supposed to be the best thing you create in your career. The dissertation is an exercise to see if you're able to take a question, answer it the best you can and recognize the limitations that you came across when trying to respond to your question.
>
> And whether or not you find significant findings, a non-finding is a finding. So, the fact that a particular method didn't work out wasn't gonna be the breaker of me not getting my PhD because it was part of the process. So I think the fact that they really emphasize this

> dissertation as a process rather than a product per se, although the product is valuable, right? You wanna be able to publish papers from this. It just- I felt a little more at ease throughout the whole thing--understanding that there were these, estimation limitations [in the dissertation], and that I was doing everything I could given my training and given the expertise on my committee and the available resources at [institution] to respond to the questions…I wish there were better data out there. But, you know, you do what you can with what you have. (D1-03)

The researcher offered a couple of other insights into how they understood or negotiated the tradeoffs and challenges inherent in reusing data. They talked about being able to answer maybe not all of one's research question, but enough to make a contribution to the literature without having to do a lot of "legwork" to collect the data, and also engaging in data reuse as preliminary work to support primary data collection in the future:

> I think the reuse of any dataset, you come across challenges on how to measure your constructs, right? Because you're not involved in the design of the study, you're very limited to how you can assess certain things. However, I think the nice thing about reusing data is the fact that you don't have to spend too much time doing the legwork and there might be opportunities for you to maybe answer parts of your question in, not all, right, but still be able to inform the literature with the limited information that is available or collected from these datasets.
>
> So, you know, while it is challenging to answer your exact question, you might be able to answer something that is close to it without having to invest a lot of resources in primary data collection. I see it as kind of an opportunity to engage in preliminary work, right? That's how I see the secondary data analysis and you engage in preliminary work and then use that preliminary work to then make the argument for new studies that collect this information slightly differently, or you conduct your own primary data collection project. (D1-03)

This was another interview where I did not ask about achievement of reuse goals because of the depth of discussion we had on other issues. I did ask about things that facilitated or were barriers to conducting their research. The researcher mentioned their familiarity with the data, the workshop they attended on reusing the data, and the consultation support available to faculty and doctoral students at their institution as facilitators. A barrier they experienced was the barrier of the time and work it takes to get to know a data study. Given the pressure they were under

when selecting a data study, this barrier caused them to choose a data study they were already familiar with and not explore one that might have been a better fit for their research.

Thinking more broadly about what might facilitate greater reuse, the researcher talked about workshops: whether long, short, in-person, or virtual; the lower the cost the better. They said these were valuable opportunities for researchers who reuse data as a norm or for junior faculty looking to publish to meet the data creators, meet with statisticians and others interested in the data, and have a chance to play with the data. As they related:

> So, I think especially for large datasets that are geared towards making it accessible to more people, not just the PIs, having these workshops, whether it's one day, two days a week, in-person or virtual, are vital to giving people more information. 'Cause documentation is great. And I think documentation is very important because you can go back to those things and just review information that you may have missed. But sometimes in these in-person workshops [you] are able to get more nuanced information about the thought process about the measurement, or maybe even get one-on-one consultation on using the data and getting some statistical support. (D1-03)

**E.6 Researcher D1-04**

This researcher was part of a group of researchers, the members of which often formed subgroups to pursue particular projects. The researcher's area of expertise was emerging as an area of interest more broadly and a recent book advanced a theory relevant to their work. The researcher believed the theory could be used to link two different phenomena and spoke to one of the researchers in the group about putting together a paper for a journal's special issue. That researcher identified others with complementary skills and a group was formed.

There was a clear choice about the data to reuse. Two of the collaborators had in-depth knowledge of the original data collection, and this was one of the only data studies available that included both the variables and the study population the researchers were interested in. Furthermore, these were data that had, as the researcher said, "been through the ringer" (D1-04). The data were archived in ICPSR, which the researcher described as "well-vetted," there was

good documentation, and ICPSR sent updates if they found any mistakes. The researcher said of the data, “I think everything that could have gone wrong or every error would have been corrected” (D1-04).

The documentation and their colleagues were the main sources of the researcher’s knowledge about the data. The researcher did not have any particular training on this data study and described their knowledge as coming through “immersion” in the data and their own expertise. They said,

> Well, with any dataset, you always read the documentation. You read about how the data was collected. You read about the weighting, you know, to make it representative of the nation. You read about the suggested procedures, because when you use data like this- because they're basically using a complex sampling design, which has to be corrected for in your statistical analysis, you read all that documentation. And so by the time you start your data analysis, you have immersed yourself in the technical reports that have come out on the data…So it's just something that you get used to as using secondary data analysis. So, it's when you're trying as I was as a [disciplinary domain], that's what [researchers in this domain] use. And so, you're taking your training from working with [previous] data to working with these surveys. And so they're just steps that you get used to. (D1-04)

The researcher also emphasized the importance of working with the data to get to know them, describing themselves at one point as a “data geek”:

> So when you work with a data set and you're looking at different things, you begin to know what other people have done, what you've done, and you've looked at things in various ways. And a lot of us are data geeks. Before I became [position], I could spend like a day just running models. And, you know, and I’ll- “what if we do this?" You know? What if we do this? So a lot of times you learn things by coding the data differently, you know, taking the categories and making a split between people who didn't finish high school when everybody else- sometimes you need a finer split in [disciplinary domain]. It all depends on what you're looking at. So by the time we got to this, we were- - we kind of had a pretty good idea of how we wanted to approach it. (D1-04)

There were not particular pressures on the group related to the publication. With the paper being written for a special issue, there was no pressure other than to meet the paper deadline. The researcher did not have any particular concerns about the data either, as mentioned

above. The age of the data might have been a concern but the researcher believed the issues relevant to their research had not changed in the time since the data were created.

When I asked about perceptions of data reuse in the researcher's community, they explained that secondary reuse was acceptable because not everyone is able to collect the kind of data needed to address questions in the field. They described a data ecosystem and the role played by ICPSR:

> In sociology generally, except for maybe experimental design people who were into experimental design, secondary data is acceptable. Because of the kind of questions that we ask, because we're interested in questions that revolve around hierarchy, you need large sample sizes to test the kinds of questions that we're interested in and these surveys are expensive. So, you can't just-- not everybody can just go and write a grant and get one. And so, one of the reasons why NIH and NSF invest in these kinds of projects is with the understanding that it will be made available to the larger research community. And that's why they're archived in ICPSR in part. It's a whole system in which member schools have access to these data for a membership fee so that their students and faculty can access them for analysis. So, it's a pretty institutionalized process in say sociology, political science, economics, and to some extent, psychology. (D1-04)

Thus, they said, in their area of specialization, the case for data reuse is already made. In fact, they noted that they train their students in secondary data reuse, including of both quantitative and qualitative data.

The researcher had not identified specific knowledge they desired that was lacking in the survey, but they mentioned a couple of areas in the interview. These had to do with the specific roles of people and types of places relevant to their phenomena of interest.

In negotiating the gap between desired and obtained knowledge related to roles, the researcher said it would have been great to have more detailed information, but there was a lot of value to the study without it. They described their reasoning as follows: "you go with what you have, because up until this point, we did not have these questions asked in a national sample, a nationally representative sample." They went on to explain that this was common practice in

secondary data reuse: “what you do is you do the best you can for what it is that you want to study. And then you make the best case for it to the reviewers that are reading your paper for peer review” (D1-04).

The researcher pointed to the original data collection to explain why this was a common approach to secondary reuse. They said that the original researchers work under budget constraints and have to make decisions about what variables and what granularity of information to collect:

> And that happens almost in any data that you're looking at because they can't ask everything. They have to make budgetary decisions about, you know, how much information is going to be collected on each topic and so forth. And so, in these large surveys, there are trade-offs between covering a number of areas and getting some information versus going very in-depth on one area or two or three. (D1-04)

The way they approached the lack of knowledge about places was similar to the way they approached the lack of knowledge about roles described above:

> So basically, you go with what you have. And then you acknowledge the limitations in the limitations section of the paper…you pose what would have been better and might be addressed in future studies…Because if you wait to have a perfect data, you will not do any research. (D1-04)

Like D1-03, they proposed approaching limitations as areas of future study, and then talked about the importance of making a contribution to existing knowledge. After describing their approach to the research, they said,

> So, this was a contribution for that reason also. So it was kind of like taking this data not only in a national sample but also doing the analysis in a different way that gave us a new information. (D1-04)

While the researcher did not have concerns about the age of the study for their own research, they mentioned that it could be a concern depending on the venue targeted for publication. Age was less important in sociology journals, and more important for interdisciplinary journals. They said,

> So, sociology journals are a bit more understanding about the age of the data and more accepting because they know it's unique. They also know that many of these kinds of relationships we look at, especially if they're dealing with [topic] hasn't changed that much in [time period]. Okay? It's not like, "Oh, we've had this radical decline." Now in the more interdisciplinary professional journals, you are more likely to get challenged on the age of the data. You're more likely to get challenged on it and you can make [an] argument that you really think this has changed. You know, it's not like we're trying to estimate the amount of [phenomenon] as if you were trying to estimate the number of [alternate phenomenon]. What social scientists are interested in is the relationships, what we call the beta coefficients, what impact does this have? How much does it elevate your risk for something rather than estimating the precise amount of risk? We're more interested in hypothesis testing and finding these links between a hierarchy and outcomes. So, it depends on what interdisciplinary journal you go to. Probably the more you venture off into say, public health is where you're more likely to be challenged on the age of the data. (D1-04)

The data met the researcher's goals for reuse "better than expected," the criteria for which they described in stark terms:

> Were you able to do an analysis that was then accepted by four critical reviewers? Were you able to get published? That's my criteria here in terms of is the data serving you well, okay? Because if you have an idea and you can't make the case for that imperfect variable, and you sent this out to several journals, then I don't think it served me well, but this dataset has served so many people well, served me well. It is not perfect, but it is good enough to advance the ideas that you're trying to get after…Other people might define it some other way, but that's how I define it, because you don't want to spend nine months learning the datasets, the ins and out, meeting all the technical support running models, and then there's some kind of fatal flaw or something like that. And this data doesn't have really a fatal flaw. It might not please a particular set of reviewers, but usually you look at what they said, you address as much as you can, and then you send it back to that journal after some time, or you send it to another one. And so, to me, that's what better than expected is to say, you rarely ever strike out using this data. (D1-04)

The researcher did not mention any barriers to their reuse of the data for this project. Factors that facilitated the researcher's reuse of the data, however, included attending trainings and classes while the researcher was completing a postdoc, participating in workshops on data reuse after they obtained a faculty position, and supervising and assisting graduate students with data reuse. These experiences were invaluable for the researcher to gain skills and experience with secondary data.

Thinking about the environment for reuse more broadly, the researcher believed the most helpful thing would be to offer consultation when researchers ran into issues with analysis—a sort of helpdesk for data, especially for large and complex government-funded surveys. In addition, they identified obtaining access to restricted data as a challenge for secondary data reusers—especially graduate students who don't have time to wait for access approval—that could be lessened if universities had a better understanding of secondary data reuse and how to assist researchers in gaining access.

### E.7 Researcher D1-05

This researcher brought interests in their topic from their background and prior research to a postdoc where they reused these data. Gaining access to the data was benefit of the postdoc, in addition to the opportunity to receiving mentoring and training. Another benefit was the ability to use portions of the data that had not yet been made available to the public.

The researcher gained a lot of their knowledge about the data study from the original data creators, from a workshop they participated in on reuse of the data, and from meetings they attended to learn about research being done with the data. Similar to D1-04, they also talked about immersing themselves in the data, or taking the time to really get to know them well and "respecting" them:

> We went back to the literature to see how other people were [using a certain measure], and so it's just making sure that you're willing to tweak what you need to, and that you take the time to really be careful with the data and just respect the data and its process and respect what you need to do, and just really understand it. So I think it's a process that if you get into it and you get used-- I'll say this about the postdoc, my opportunity to do the postdoc actually really taught me how to analyze secondary data, and so that was one of the things that was really helpful. (D1-05)

The researcher's approach to projects in general was to have a research question ahead of time, but to be willing to tweak or change the question based on the data and what others have

already investigated. Like D1-03 and D1-04, they said a key component to a successful project was coming up with research that was new or filled a gap in existing literature (i.e., made a contribution). They explained this in a couple of ways:

> For me, it always starts with the question of, "What is interesting?" Because I think, when you're going to have to spend so much time with these papers, it needs to be a question that you find interesting and then the key is how does it fit within the literature? Is there a gap? Are you just being repetitive? Do you have something that you can actually add? (D1-05)

It was also important to the researcher when developing a research question to think about how the research fit in with their larger career goals:

> The other thing that guides me or has guided my work is also making sure that I'm creating a body of research in an area that I want to develop a level of expertise [in]. So that's also checking the box for me that I was able to add this to the work that I'm doing as a research interest. (D1-05)

The main pressures the researcher mentioned they experienced were pressures of time and pressures to publish.

> One of the things they always told us when we're in the postdoc, you have a whole year to get this many papers out. So you just need to get as many out as you can because after the postdoc, your next step would be an academic position or wherever you're going to go. So you're always thinking about the timing and that's why secondary data was a good idea. You really wouldn't have enough time to collect your own data within a year and then publish it and then get something out… so this is something that I needed to make sure I wrapped up and got out the door. (D1-05)

They said there was pressure also from the funders of the postdoc to publish (funders wanted to be able to report that researchers were able to write a paper and have it publish), and pressure to think about publishing in the context of their future work:

> When you do the postdocs, you do have to submit to your funder that you were able to write a paper and get it published. They'll track your progress as well. My postdoc was through [funding source]. So, you know, for, I think the last five years or so they still track how many papers did you get? What did you produce from the postdoc? So you just want to think about your timing and what can you do? And it really gives you a good sense of how you can build up your publication record as you're thinking about the next paper, or actually, grants in the future. (D1-05)

The researcher's primary concern about the data, due to their complexity, was using them correctly. As part of the postdoc, the researcher worked with a statistician to help ensure this:

> I think the biggest concern was making sure I would do it correctly…they're large data sets and so you really have to be really specific and know your analysis and to make sure that you're doing it correctly. There's a lot to learn to just getting into the process of using the data. (D1-05)

Similar to D1-01, D1-03 and D1-04, the age of the data study is something the researcher also considered, but this was not so much of a concern, as I discuss further below.

When I asked how data reuse is seen in the researcher's community, they said that data reuse was regular practice at the institution where they did they postdoc, but that they were the only researcher who did secondary analysis when they went to a new institution. They said that reuse depends on the purpose of the research. For instance, qualitative researchers were less likely to reuse data, but it is not an either or: researchers might want to do secondary analysis to learn more about something they discovered in earlier research, or compliment current research.

> It depends on your research area and the type of research that you do. So, if you're a qualitative researcher, you probably won't do secondary data analysis. You're gonna go out because you'll have a smaller N, first of all, hopefully, and you'll go out and collect your own data. Or it may be something that you put into a grant application if you're saying, "Well, I want to learn more about this topic," which is something I did previously. So I had a previous grant and I said we wanted to learn more about [topic] and so we proposed to write a paper from [the data], where we can learn more about that because I wouldn't be able to collect that data on my own based on the population, and so I think it's helpful, and I think people use it as they need it, or however it's going to compliment what they're wanting to do, if they're wanting to go and collect their own data as well. So I think it's something that people use. And I think people feel comfortable using it, at least, you know, I do. (D1-05)

They also noted that data are created with the intent of being reused, with an implication that this intent to share can contribute to the value of reuse for secondary researchers:

> When you asked about people's comfort level with the usage of the data, I'll say that there are a lot of people who do create very large data sets, epidemiological data studies that they collect, and I think the intended use of it too, is to be able to share it with other

> researchers, and so, if you have the opportunity to conduct secondary data analysis, it can be an interesting and worthwhile thing, depending on what your purpose for the usage is. (D1-05)

There were several areas of knowledge the researcher desired about the data but was not able to obtain. These mainly had to do with specific details ("what" questions) and reasons ("why" questions) related to their phenomenon of interest. There were a variety of strategies or approaches the researcher used to reach an equilibrium with the data despite not having all the knowledge they desired. The first was finding the gap ahead of time: knowing how they were going to use the data that was different from what others had done. They said this was part of the process when applying to the postdoc:

> And that was part of the process too, in the postdoc, when we actually had to submit an application saying, you know, when you scan the data, you scan the questionnaire to make sure that we weren't replicating papers that were already being done. (D1-05)

The second, similar to D1-03, was using what is available and viewing limitations as a way of setting up future work. They said,

> So I think when you're doing this, it's what's in front of you—what is available to you to analyze and to try to understand. And so it is the, "What information do you have?" and then it leaves open for places, particularly in your discussion, where you can talk about what you want to know more about or what maybe you need to know more about and you always want to situate it within the literature as well. And so we worked with what we had and that's where it really opens up in the area to then further develop a new survey based on what was collected or what you learned and what was missing from there and what you think you need to add there to, you know, even make your research question more robust. So, basically, working with what you have and then figuring out if the next paper or the next project-- So, say, if I were writing this and I was interested in continuing it on and doing a grant off of this, that would be an aim that I could develop. (D1-05)

A third, which relates to the discussion of journals by D1-04, was to look for a publishing venue where the research will be a good fit. The researcher described:

> So I knew I wanted to try for a [topic] journal. Actually, we submitted to the first journal and in some, the journal disbanded, I guess, I don't know. It was unfortunate. And so then I had to resubmit to a different journal. So yes. So as you're getting to the end, usually as

> I'm getting to the end of the paper, or even before I start writing the paper, I'll make a list of journals that would be appropriate. So we'll know where to go if we're rejected the first time, or however many times. So that's something that I always keep in mind is the type of paper that will go there, how appropriate the fit is with the aim and scope of the journal…So I think, probably every paper, you always think about your audience, who you're going to submit the paper to and would it-- How well would it be received, and so that's always a consideration. (D1-05)

Some of the factors they considered with regard to the age of the data (i.e., to evaluate whether the data were still relevant) were 1) whether other people were still writing about the data, 2) whether the issues they were investigating had changed over time, and 3) whether the age of the data was not such a big concern to journals as long as you explain it. They said, "I think a lot of the journals are used to it too; that they have expectation that using it now, may be a little difficult, you could say." (D1-05)

The researcher did not remember so well why they indicated the data met their reuse goals much better than expected for answering a new question and as expected for testing a theory. They posited that the first was because there was a particular aspect of their question the data enabled them to explore. They said they were not sure they were able to investigate a theory so they could not speak to that.

The main facilitators of the researcher's use of the data were access to authors of the original study. They said they were able to get all the information they needed and didn't experience any barriers. When thinking about what could facilitate reuse more broadly, the researcher talked about the need to understand well the "concept and meaning" behind why the data were collected; to understand "not only your research questions, but how you need to use the data—what's the most appropriate statistical package" (D1-05). In discussing this, they reiterated the importance of spending time with the data and "committing" to them, as well as personal characteristics that are beneficial to reusing data:

Just to make sure you really understand, you have to really kind of sit with the survey for a while. The surveys are large and cumbersome, and it's something that you have to commit yourself to and I think you have to like data analysis in a way; you have to maybe like the idea of [an] ambiguous process where you're not exactly sure what's going to happen at the end of it. (D1-05)

## Appendix F: Summary of Hypothesis Tests

Table F.1 gives a summary of my hypotheses, how I tested them, and my findings. Since I found that satisficing did not well characterize researchers' behavior in bounding knowledge about data, I have substituted "lacking desired knowledge" for "satisficing" in the hypotheses below.

*Table F.1 Summary of Hypothesis Tests*

| Hypothesis | How Tested | Findings |
|---|---|---|
| 1. Researchers lacked knowledge they desired when determining whether to reuse data | Descriptive statistics (simple proportion) | 25% of researchers lacked knowledge they desired |
| 2. Lacking desired knowledge has a negative impact on research outcomes | Descriptive statistics (simple proportion) | More than 56% of researchers reported at least a moderate negative impact |
| 3. The probability of researchers' attaining their goals for reusing data is lower when they lacked desired knowledge | Logistic regression | No significant relationship between goal attainment and the odds of lacking desired knowledge |
| 4. There is no association between lacking desired knowledge about data and reuse of qualitative or quantitative data | Logistic regression | No significant relationship between type of data reused and the odds of lacking desired knowledge |
| 5. The greater a researcher's "distance" from data, the higher the probability of lacking desired knowledge.<br><br>Measured as:<br>(a) involvement in the original research<br>(b) knowledge about the original data creation | Logistic regression | (a) No significant relationship between involvement in the original research and the odds of lacking desired knowledge<br><br>(b) Significant relationship between knowledge about original data creation and odds of lacking knowledge (the odds of lacking knowledge decreased by 0.28 times for each greater level of knowledge about original data creation) |
| 6. The more experience a researcher has in their primary domain of research, the lower the probability of lacking desired knowledge<br><br>Measured as:<br>(a) primary domain of research<br>(b) years experience in primary domain | Logistic regression | (a) No significant relationship between primary domain and the odds of lacking desired knowledge. Significant relationship between being in a "less common" domain and the odds of lacking knowledge (the odds of lacking knowledge were 2.3 times greater for researchers in a "less common" domain)<br><br>(b) No significant relationship between years experience in primary domain and the odds of lacking desired knowledge |
| 7. The more experience a researcher has with data reuse, the lower the probability of lacking desired knowledge<br><br>Measures tested: | Logistic regression | No significant relationship between any measure of experience and the odds of lacking desired knowledge except for one measurement under (c) frequency of reuse for the purpose of comparing the data with |

| | | |
|---|---|---|
| (a) professional position<br>(b) experience reusing data produced in the same domain as the data that were reused, and any domain<br>(c) frequency of reuse for the same purpose for which these data were reused<br>(d) whether the reused data were produced in the researcher's primary domain of research<br>(e) the overall frequency with which the researcher reused data<br>(f) the number of projects involving reuse of data that the researcher had worked on<br>(g) the number of projects involving reuse of data where considerations surrounding reuse were substantially similar to the ones involved in reuse of the data I asked them about<br>(h) the number of papers the researcher authored or co-authored describing research involving reuse of data<br>(i) the number of papers the researcher authored or co-authored describing research involving reuse of data where considerations surrounding reuse were substantially similar to the ones involved in reuse of the data I asked them about | | other data (the odds of lacking knowledge increased 1.48 times for each of five levels of increased frequency of reuse from never to always) |
| 8. The greater a researcher's perceived ability to reuse data, the lower the probability of lacking desired knowledge | Logistic regression | No significant relationship between perceived reuse ability and the odds of lacking desired knowledge |
| 9. There is no association between researcher's motivation for reusing data and lacking desired knowledge<br><br>Motivations tested:<br>(a) background<br>(b) validate<br>(c) new question<br>(d) replicate<br>(e) combine<br>(f) compare<br>(g) theory<br>(h) tool<br>(i) other | Logistic regression | Significant relationships for (d) replicate, (e) combine, and (g) theory (odds of lacking knowledge increased by 1.93, 1.45, and 1.05 times, respectively) |
| 10. When research questions change or develop over time there is a lower probability of lacking desired knowledge | Logistic regression | Significant relationship between research questions being fixed (not changing or developing) and the odds of lacking desired knowledge (the odds of lacking knowledge decreased by .67 times when research questions were fixed) |

## Appendix G: Characteristics of Interviewees

Table G.1 shows the final distribution of researcher and data characteristics for the 26 researchers I interviewed and for the samples of researchers I used for inferential statistics from the survey (N=767). I do this to show the relationship between the two—in particular, that the counts of those interviewed correspond fairly well to those that I surveyed. In most cases, the survey sample count reflects the count of researchers in the combined samples of researchers who reused quantitative data and qualitative data (i.e., the samples I used to conduct my inferential statistics, which excludes the sample of researchers drawn by the number of citations to the data). In some cases, where the survey question was asked only researchers who satisficed, I include the counts from the survey of all who responded (i.e., 223; relevant cases below are indicated with a description that includes "(out of 223)"), because that is the population I drew from in order to conduct the interviews and because those are the statistics I give in the body of the dissertation. This population includes the sample of researchers drawn based on the number of citations to the data.

For rows labelled "multiple", it is possible for a dataset to be given multiple designations (e.g., a data study could have a type of both "Medical" and "Experimental." In most cases I was able to achieve a diversity of characteristics of both interviewees and data. In some areas, such as type of reference, dataset type, whether the data were created to be reused or not, whether the reused data were in the researcher's primary domain of research, and others, there was not a lot of diversity to choose from in the respondents who agreed to be contacted.

*Table G.1 Characteristics of Interviewees and Data*

| Survey Concept | Criteria | Interview Sample Count | Survey Sample Count |
|---|---|---|---|
| **Administrative** | | | |
| Type of reference | Diversity | | |
| Book Section | | 0 | 16 |
| Book | | 1 | 11 |
| Conference proceeding | | 2 | 30 |
| Journal article | | 17 | 612 |
| Report | | 2 | 29 |
| Thesis | | 4 | 69 |
| **Type of data study (multiple)** | Diversity | | |
| Observational | | 0 | 96 |
| Census | | 0 | 43 |
| Roll call voting data | | 0 | 1 |
| GIS | | 0 | 1 |
| Administrative | | 1 | 62 |
| Survey | | 20 | 630 |
| Aggregate | | 2 | 44 |
| Clinical | | 3 | 96 |
| Event/Transaction | | 4 | 51 |
| Machine | | 1 | 2 |
| Medical | | 1 | 7 |
| Experimental | | 1 | 14 |
| Text | | 1 | 6 |
| Missing | | 0 | 13 |
| **Researcher's perspective** | | | |
| Researcher’s involvement in the decision to reuse the data | Diversity | | |
| Not at all involved | | 0 | 17 |
| Somewhat involved | | 0 | 21 |
| Quite involved | | 2 | 47 |
| Very involved | | 16 | 310 |
| Solely responsible | | 8 | 372 |
| Number of people involved in the decision to reuse the data | Diversity | | |
| Zero | | 16 | 377 |
| One | | 5 | 156 |
| Two | | 4 | 89 |
| Three | | 1 | 61 |
| More than three | | 0 | 62 |
| Researcher’s involvement in determining the goals of the research | At least somewhat involved | | |
| Not at all involved | | 0 | 8 |
| Somewhat involved | | 0 | 15 |
| Quite involved | | 1 | 45 |
| Very involved | | 11 | 359 |
| Solely responsible | | 13 | 334 |
| Missing response | | 1 | 6 |
| Data obtained from ICPSR or another source | Some ICPSR, some not | | |
| ICPSR | | 15 | 390 |
| Other | | 8 | 309 |

| | | | |
|---|---|---|---|
| Missing response | | 3 | 68 |
| Researcher's role | Leads (diversity after that if possible) | | |
| None | | 0 | 3 |
| Lead | | 22 | 571 |
| Collab | | 1 | 44 |
| Support | | 2 | 90 |
| Missing | | 1 | 59 |
| Researcher's primary domain of research (multiple) | Diversity of domains | | |
| Assessment | | 1 | 2 |
| Communities | | 1 | 10 |
| Economics | | 2 | 50 |
| Education | | 4 | 34 |
| Health | | 7 | 222 |
| Health services | | 1 | 21 |
| Inequality | | 1 | 26 |
| Justice | | 8 | 150 |
| Life course | | 1 | 118 |
| Methodology | | 3 | 19 |
| Population | | 2 | 33 |
| Race | | 1 | 11 |
| Relationships | | 1 | 20 |
| Social behavior | | 2 | 37 |
| Arts | | 0 | 1 |
| Atmospheric science | | 0 | 1 |
| Charity | | 0 | 1 |
| Citizenship | | 0 | 1 |
| Civic engagement | | 0 | 1 |
| Cognitive science | | 0 | 6 |
| Communication | | 0 | 2 |
| Data science | | 0 | 1 |
| Decision-making | | 0 | 6 |
| Ecology | | 0 | 2 |
| Engineering | | 0 | 2 |
| Ethnicity | | 0 | 3 |
| Gender | | 0 | 9 |
| Geography | | 0 | 3 |
| History | | 0 | 4 |
| Housing | | 0 | 1 |
| Informatics | | 0 | 1 |
| Information studies | | 0 | 1 |
| Information technology | | 0 | 1 |
| Intergroup relations | | 0 | 1 |
| Library and information science | | 0 | 1 |
| NA | | 0 | 1 |
| Organizations | | 0 | 5 |
| Policy | | 0 | 30 |
| Political science | | 0 | 16 |
| Politics | | 0 | 3 |
| Poverty | | 0 | 8 |

| | | | |
|---|---|---|---|
| Religion | | 0 | 2 |
| Sexuality | | 0 | 3 |
| Single mothers | | 0 | 1 |
| Social biology | | 0 | 4 |
| Social cohesion | | 0 | 1 |
| Social programs | | 0 | 1 |
| Social science | | 0 | 8 |
| Social work | | 0 | 3 |
| Sociology | | 0 | 54 |
| Technology | | 0 | 1 |
| Theory | | 0 | 4 |
| Transportation | | 0 | 1 |
| Urban planning | | 0 | 1 |
| Work | | 0 | 4 |
| **Data quantitative or qualitative** | Some from qual, some from quant | | |
| Quantitative | | 25 | 737 |
| Qualitative | | 1 | 28 |
| Missing response | | 0 | 2 |
| **Number of citations** | Diversity of high and low citations | | |
| Less than 5 | | 6 | 206 |
| 5-24 | | 6 | 258 |
| 25-50 | | 4 | 54 |
| 51 to 100 | | 0 | 120 |
| More than 100 | | 10 | 129 |
| **Researcher distance** | | | |
| Researcher's perception of original data creation purpose (whether the data were created to be reused) | Diversity of purposes | | |
| Yes | | 22 | 591 |
| No | | 0[1] | 81 |
| Not sure | | 3 | 87 |
| Missing | | 1 | 8 |
| Researcher's involvement in the original research[2] | Diversity of involvement | | |
| Not at all | | 20 | 546 |
| Somewhat | | 2 | 39 |
| Quite | | 1 | 30 |
| Very | | 3 | 111 |
| Solely responsible | | 0 | 40 |
| Missing | | 0 | 1 |
| Knowledge of research team about the way the original data were collected or produced | Diversity of original knowledge | | |
| No knowledge | | 1 | 23 |
| Some | | 5 | 104 |
| Good | | 3 | 157 |

[1] Only 6 out of the 284 researchers who agreed to be contacted were sure that the data had not been created to be reused. I included 4 in the researchers I contacted but none of those responded.
[2] Out of the 284, the distribution of involvement was Not at all (207), Somewhat (16), Quite (9), Very (41), Solely responsible (11). Among the 53 researchers I contacted the distribution was Not at all (41), Somewhat (3), Quite (1), Very (8), Solely (2).

| | | | |
|---|---|---|---|
| Significant | | 7 | 243 |
| In-depth | | 10 | 238 |
| Missing | | 0 | 2 |
| **Experience with data reuse** | | | |
| Researcher's professional position | Diversity of positions | | |
| Undergrad | | 0 | 5 |
| Masters | | 0 | 20 |
| PhD student | | 11 | 260 |
| Postdoc | | 1 | 53 |
| Assistant professor | | 5 | 121 |
| Assistant researcher | | 0 | 3 |
| Associate professor | | 2 | 79 |
| Associate researcher | | 0 | 82 |
| Lecturer | | 0 | 11 |
| Professor | | 3 | 44 |
| Researcher | | 4 | 111 |
| Other | | 0 | 13 |
| Missing | | 0 | 44 |
| Whether the research conducted was in the researcher's primary domain of research | Nearly all were Yes | | |
| Yes | | 24 | 670 |
| No | | 1 | 45 |
| Other | | 1 | 14 |
| Missing response | | 0 | 38 |
| Researcher's years of experience in primary domain of research | Diversity of experience | | |
| None | | 0 | 12 |
| 0<x<=1 | | 1 | 36 |
| 1<x<=5 | | 10 | 242 |
| 5<x<=10 | | 7 | 210 |
| 10<x<=15 | | 2 | 75 |
| 15<x<=25 | | 4 | 87 |
| 25<x<=35 | | 1 | 34 |
| x>=35 | | 1 | 26 |
| Missing | | 0 | 45 |
| Researcher's years of experience reusing data produced in this domain | Diversity of experience | | |
| None | | 3 | 101 |
| 0<x<=1 | | 3 | 66 |
| 1<x<=5 | | 11 | 220 |
| 5<x<=10 | | 3 | 162 |
| 10<x<=15 | | 2 | 56 |
| 15<x<=25 | | 3 | 45 |
| 25<x<=35 | | 1 | 23 |
| x>35 | | 0 | 4 |
| Missing | | 0 | 0 |
| Researcher's years of experience reusing data produced in any domain | Diversity of experience | | |
| None | | 4 | 77 |
| 0<x<=1 | | 2 | 45 |
| 1<x<=5 | | 8 | 198 |

| | | | |
|---|---|---|---|
| 5<x<=10 | | 5 | 161 |
| 10<x<=15 | | 1 | 84 |
| 15<x<=25 | | 3 | 62 |
| 25<x<=35 | | 3 | 29 |
| x>35 | | 0 | 15 |
| Missing response or did not answer the question | | 0 | 96 |
| **Reuse motivation (multiple)** | Diversity of purposes at least somewhat important | | |
| Background | | | |
| Yes | | 8 | 180 |
| No | | 18 | 586 |
| Missing | | 0 | 1 |
| Validate | | | |
| Yes | | 6 | 175 |
| No | | 20 | 591 |
| | | 0 | 1 |
| New question | | | |
| Yes | | 21 | 608 |
| No | | 5 | 158 |
| Missing | | 0 | 1 |
| Replicate or Reproduce | | | |
| Yes | | 1 | 53 |
| No | | 25 | 713 |
| | | 0 | 1 |
| Combine | | | |
| Yes | | 4 | 237 |
| No | | 22 | 529 |
| Missing | | 0 | 1 |
| Compare | | | |
| Yes | | 6 | 198 |
| No | | 20 | 568 |
| | | 0 | 1 |
| Test a theory | | | |
| Yes | | 14 | 519 |
| No | | 12 | 247 |
| Missing | | 0 | 1 |
| Develop an algorithm or tool | | | |
| Yes | | 3 | 62 |
| No | | 23 | 704 |
| | | | 1 |
| Other | | | |
| Yes | | 2 | 41 |
| No | | 24 | 725 |
| Missing | | 0 | 1 |
| **Goal Attainment (multiple)** | | | |

| Background | Diversity of attainment[3] | | |
|---|---|---|---|
| Somewhat worse than expected | | 0 | 3 |
| As expected | | 6 | 128 |
| Better than expected | | 1 | 33 |
| Much better than expected | | 1 | 10 |
| Missing | | 0 | 3 |
| Validate | | | |
| Somewhat worse than expected | | 0 | 5 |
| As expected | | 5 | 104 |
| Better than expected | | 1 | 43 |
| Much better than expected | | 0 | 15 |
| Missing | | 0 | 3 |
| New question | | | |
| Somewhat worse than expected | | 1 | 20 |
| As expected | | 9 | 374 |
| Better than expected | | 8 | 139 |
| Much better than expected | | 3 | 53 |
| Missing | | 0 | 13 |
| Replicate or Reproduce | | | |
| Somewhat worse than expected | | 0 | 2 |
| As expected | | 0 | 28 |
| Better than expected | | 1 | 15 |
| Much better than expected | | 0 | 2 |
| Missing | | 0 | 3 |
| Combine | | | |
| Somewhat worse than expected | | 0 | 13 |
| As expected | | 3 | 135 |
| Better than expected | | 1 | 47 |
| Much better than expected | | 0 | 29 |
| Missing | | 0 | 4 |
| Compare | | | |
| Somewhat worse than expected | | 1 | 8 |
| As expected | | 4 | 111 |
| Better than expected | | 1 | 53 |
| Much better than expected | | 0 | 17 |
| Missing | | 0 | 5 |
| Test a theory | | | |
| Somewhat worse than expected | | 0 | 23 |
| As expected | | 9 | 310 |
| Better than expected | | 2 | 114 |
| Much better than expected | | 3 | 49 |
| Missing | | 0 | 8 |
| Develop an algorithm or tool | | | |
| Somewhat worse than expected | | 0 | 3 |
| As expected | | 1 | 35 |
| Better than expected | | 2 | 13 |
| Much better than expected | | 0 | 7 |

[3] Diversity of goal attainment where researchers lacked desired knowledge or the purpose of reuse was at least somewhat important (for researchers who did not lack desired knowledge); applies to all variables for goal attainment.

| | | | |
|---|---|---|---|
| Missing | | 0 | 1 |
| Other | | | |
| As expected | | 0 | 14 |
| Better than expected | | 0 | 3 |
| Much better than expected | | 0 | 5 |
| Missing | | 0 | 18 |
| **Research process** | | | |
| How research questions were developed | Diversity of fluid and fixed[4] | | |
| Fixed | | 5 | 178 |
| Fluid | | 21 | 589 |
| Sources of knowledge (out of 223)[5] | Diversity of sources | | |
| Advisor(s) | | 1 | 4 |
| Data documentation | | 2 | 64 |
| Knowledge of others on the team | | 2 | 12 |
| Literature | | 1 | 25 |
| Original data creators | | 3 | 39 |
| The data themselves | | 6 | 38 |
| Other | | 1 | 3 |
| Colleagues | | 0 | 9 |
| Data repository | | 0 | 5 |
| Personal knowledge | | 0 | 6 |
| **Outcomes of research** | | | |
| Whether (and how much) a lack of knowledge affected the outcomes of the research | Diversity of impact (based on maximum impact across types of knowledge lacking) | | |
| Not at all | | 0 | 19 |
| Slightly | | 2 | 48 |
| Moderately | | 5 | 54 |
| Very much | | 2 | 27 |
| Extremely | | 1 | 6 |
| Missing | | 1 | 69 |
| Why the lack of knowledge did or did not affect the outcomes of research (out of 223)[6] | Diversity of impact reasons | | |
| Adjusted | | 4 | 12 |
| Compensated | | 8 | 46 |
| Limited | | 11 | 90 |
| Obtained | | 2 | 11 |
| Opportunity lost | | 2 | 28 |
| Satisfactory | | 1 | 15 |
| Complicated | | 0 | 23 |
| Knowledge not important | | 0 | 4 |
| Other (did not answer or answer not relevant) | | 0 | 87 |

[4] For 14 out of the 53 researchers contacted, research questions were "fixed" as opposed to "fluid."
[5] I counted sources of knowledge for distinct mentions of knowledge lacking. For example, if for two mentions of knowledge lacking a researcher obtained knowledge from the data themselves, I added 2 to the count of the data themselves.
[6] I counted distinct mentions of impact reasons, so, e.g., if a researcher mentioned two instances where knowledge was limited, I added both to the count.

| **Satisficing** | | | |
|---|---|---|---|
| Whether, when the decision to reuse the data was made, there was knowledge about the data that was desired but not obtained or obtained only in part | Mix of Yes and No | | |
| Yes | | 11 | 195 |
| No | | 15 | 572 |
| The three most important things the researcher would have liked to know about the data but was not able to obtain, or obtain to the desired degree (multiple) (out of 223)[7] | Diversity of knowledge lacking | | |
| Data | | 3 | 28 |
| Data analysis | | 1 | 33 |
| Data collection | | 7 | 66 |
| Data reuse | | 1 | 6 |
| Data supplement | | 16 | 130 |
| Data access information | | 0 | 4 |
| Data comparability | | 0 | 1 |
| Data context | | 0 | 1 |
| Data documentation | | 0 | 13 |
| Data management | | 0 | 2 |
| Data reporting | | 0 | 1 |
| Data validation | | 0 | 1 |
| Other | | 0 | 5 |
| People | | 0 | 6 |
| How much of each type of desired knowledge was obtained (out of 223)[8] | Diversity of amount obtained | | |
| 0 | | 2 | 36 |
| 1-10 | | 0 | 14 |
| 11-20 | | 0 | 17 |
| 21-30 | | 0 | 20 |
| 31-40 | | 1 | 17 |
| 41-50 | | 4 | 39 |
| 51-60 | | 0 | 17 |
| 61-70 | | 4 | 34 |
| 71-80 | | 3 | 31 |
| 81-90 | | 2 | 21 |
| 91-99 | | 1 | 9 |
| 100 | | 1 | 15 |
| Missing | | 0 | 27 |
| The importance of the knowledge that was not obtained or obtained only in part to deciding to reuse the data (out of 223) | Knowledge lacking must be important (or have an impact) | | |
| Not at all[9] | | 1 | 24 |

[7] I counted distinct mentions of knowledge lacking. For example, if a researcher mentioned they lacked knowledge in three areas that I categorized as data supplement, I added three to the count of data supplement.
[8] I counted amounts of knowledge for distinct mentions of knowledge lacking. For example, if the same researcher obtained between 34 and 66 percent for two distinct areas of knowledge, I added two to the count of 34-66.
[9] In this case, the researcher indicated the knowledge lacking was not important but I included it because they also said the lack of knowledge had a slight impact on their research.

| | | | |
|---|---|---|---|
| Somewhat important | | 0 | 74 |
| Important | | 4 | 89 |
| Very important | | 1 | 70 |
| Essential | | 4 | 33 |
| Missing | | 1 | 7 |

## Bibliography


Ackoff, Russell L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*, 3-9.

Agosto, D. E. (2002). Bounded rationality and satisficing in young people's Web-based decision making. *Journal of the American Society for Information Science and Technology*, *53*(1), 16–27. https://doi.org/10.1002/asi.10024

Allison, L., & Gurney, R. (2015). *A place to stand: E-infrastructures and data management for global change research*. https://www.belmontforum.org/wp-content/uploads/2017/05/A_Place_to_Stand-Belmont_Forum_E-Infrastructures__Data_Management_CSIP.pdf

American Pyschological Association (APA). (n.d.). Gender. APA Style. https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/gender

Andersson, E., & Sørvik, G. O. (2013). Reality lost? Re-use of qualitative data in classroom video studies. *Forum, Qualitative Social Research / Forum, Qualitative Sozialforschung*, *14*(3). http://www.qualitative-research.net/index.php/fqs/article/view/1941

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., & Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, *3*, 135–152. https://doi.org/10.2481/dsj.3.135

Atici, L., Kansa, S. W., Lev-Tov, J., & Kansa, E. C. (2013). Other people's data: A demonstration of the imperative of publishing primary data. *Journal of Archaeological Method and Theory*, *20*(4), 663–681. http://www.jstor.org/stable/43654603

Atkinson, P. (1992). The ethnography of a medical setting: Reading, writing, and rhetoric. *Qualitative Health Research*, *2*(4), 451–474. https://doi.org/10.1177/104973239200200406

Baker, M. (2012). Independent labs to verify high-profile papers. *Nature*. https://doi.org/10.1038/nature.2012.11176

Barbrow, S., Brush, D., & Goldman, J. (2017). Research data management and services: Resources for novice data librarians. *College & Research Libraries News*, *78*(5). https://crln.acrl.org/index.php/crlnews/article/view/16660/18116

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, 11(3), 369. https://doi.org/10.2307/248684

Bernard, H. R. (2012). *Social research methods: Qualitative and quantitative approaches*. SAGE Publications.

Berryman, J. M. (n.d.). *Judgments during information seeking: Emerging themes from an exploratory study into the concept of enough information*.

Berryman, J. M. (2006). What defines "enough" information? How policy workers make judgements and decisions during information seeking: Preliminary results from an exploratory study. *Information Research: An International Electronic Journal*, 11(4). https://eric.ed.gov/?id=EJ1104626

Berryman, J. M. (2008). Judgements during information seeking: A naturalistic approach to understanding the assessment of enough information. *Journal of Information Science*, 34(2), 196–206. https://doi.org/10.1177/0165551507082589

Bilal, D. (1998). Children's search processes in using World Wide Web search engines: An exploratory study. *Proceedings of the ASIS Annual Meeting*, 35, 45–53.

Birnholtz, J. P., & Bietz, M. J. (2003). *Data at work: Supporting sharing in science and engineering*. ACM. https://doi.org/10.1145/958160.958215

Bishop, L. (2006). A proposal for archiving context for secondary analysis. *Methodological Innovations Online*, *1*(2), 10–20. https://doi.org/10.4256/mio.2006.0008

Bishop, L. (2007). A reflexive account of reusing qualitative data: Beyond primary/secondary dualism. *Sociological Research Online*, *12*(3), 1–14. https://doi.org/10.5153/sro.1553

Bishop, L. (2009). Ethical sharing and reuse of qualitative data. *Australian Journal of Social Issues*, *44*(3), 255–272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting qualitative data reuse: A decade on. *SAGE Open*, *7*(1), 2158244016685136. https://doi.org/10.1177/2158244016685136

Blackler, F. (1995). Knowledge, knowledge work and organizations: An overview and interpretation. *Organization Studies*, *16*(6), 1021–1046. https://doi.org/10.1177/017084069501600605

Bloor, D. (1991). *Knowledge and social imagery* (2nd ed.). University of Chicago Press.

Borgman, C. L. (1990). *Scholarly communication and bibliometrics*. Sage.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007a). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*, 17–30. https://cloudfront.escholarship.org/dist/prd/content/qt6fs4559s/qt6fs4559s.pdf

Borgman, C. L., Wallis, J. C., Mayernik, M. S., & Pepe, A. (2007b). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 269–277. https://doi.org/10.1145/1255175.1255228

Bornat, J. (2005). Recycling the evidence: Different approaches to the reanalysis of gerontological data. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869224962/abstract/39450B3EA1AA4C89PQ/1

Brody, A. J. (2011). The archaeology of the extended family: A household compound from Iron II Tell En-Naṣbeh. *Household Archaeology in Ancient Israel and Beyond*, 237–254. https://doi.org/10.1163/ej.9789004206250.i-452.71

Broom, A., Cheshire, L., & Emmison, M. (2009). Qualitative researchers' understandings of their practice and the implications for data archiving and sharing. *Sociology*, *43*(6), 1163–1180. https://doi.org/10.1177/0038038509345704

Carlson, S., & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, *12*, 635–651. https://doi.org/10.1111/j.1083-6101.2007.00342.x

Carmichael, P. (2017). Secondary qualitative analysis using online resources. In N. G. Fielding, R. M. Lee, & G. Blank, *The SAGE Handbook of Online Research Methods* (pp. 509–524). SAGE Publications Ltd. https://doi.org/10.4135/9781473957992.n29

Chao, T. C. (2011). Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. *Proceedings of the Association for Information Science and Technology*, *48*(1), 1–8. https://doi.org/10.1002/meet.2011.14504801125

Chin, G., Jr., & Lansing, C. S. (2004). Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. https://doi.org/10.1145/1031607.1031677

Cho, Y. I. (2008). Intercoder Reliability. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 345–345). SAGE Publications, Inc. https://doi.org/10.4135/9781412963947

Chu, F. T. (1994). Reference service and bounded rationality: Helping students with research (research note). *College and Research Libraries*, *55*(5), 457–467. https://doi.org/10.5860/crl_55_05_457

Church, R. M. (2002). The effective use of secondary data. *Learning and Motivation*, *33*(1), 32–45. https://doi.org/10.1006/lmot.2001.1098

Collins, H. M. (1998). The meaning of data: Open and closed evidential cultures in the search for gravitational waves. *American Journal of Sociology*, *104*(2), 293–338. https://doi.org/10.1086/210040

Collins, H. M., & Pinch, T. (1993). *The golem: what everyone should know about science*. Cambridge University Press.

Coltart, C., Henwood, K., & Shirani, F. (2013). Qualitative secondary analysis in austere times: Ethical, professional and methodological considerations. *Forum: Qualitative Social Research*, *14*(1). http://search.proquest.com/docview/1365213578/abstract/DC5F311C53234E52PQ/1

Consultative Committee for Space Data Systems. (2012). Reference model for an open archival information system. Washington, D.C.: CCSDS Secretariat. https://public.ccsds.org/pubs/650x0m2.pdf

Corti, L. (2000). Progress and problems of preserving and providing access to qualitative data for social research—The international picture of an emerging culture. *Forum: Qualitative Social Researc; Berlin*, *1*(3). https://www.qualitative-research.net/index.php/fqs/article/view/1019

Corti, L. (2005). User support. *Forum: Qualitative Social Research*, *6*(2). http://search.proquest.com/docview/869227642/abstract/B81E2A118AB94DDDPQ/1

Corti, L., & Bishop, L. (2005). Strategies in teaching secondary analysis of qualitative data. *Forum: Qualitative Social Research*, *6*(1). https://doi.org/10.17169/fqs-6.1.509

Corti, L., & Thompson, P. (2004). Secondary analysis of archived data. In C. Seale, G. Gobo, F. Gubrium, & D. Silverman (Eds.), *Qualitative Research Practice* (pp. 297–313). SAGE Publications Ltd. https://dx.doi.org/10.4135/9781848608191

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2nd ed). SAGE Publications.

Creswell, J. W. (2015). Mapping the developing landscape of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of Mixed Methods in Social & Behavioral Research* (pp. 45–68). SAGE Publications, Inc. https://dx.doi.org/10.4135/9781506335193

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, William E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 209–240). SAGE Publications, Inc.

Cui, J. (2007). QIC program and model selection in GEE analyses. *The Stata Journal*, *7*(2), 209–220. https://doi.org/10.1177/1536867X0700700205

Curty, R. G. (2016). Factors influencing research data reuse in the social sciences: An exploratory study. *International Journal of Digital Curation*, *11*(1), 96–117. https://doi.org/10.2218/ijdc.v11i1.401

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PLOS One*, *12*(12), e0189288. https://doi.org/10.1371/journal.pone.0189288

Curty, R. G., & Qin, J. (2014). Towards a model for research data reuse behavior. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–4. https://doi.org/10.1002/meet.2014.14505101072

Curty, R., Yoon, A., Jeng, W., & Qin, J. (2016). Untangling data sharing and reuse in social sciences. *Proceedings of the Association for Information Science and Technology*, 53(1), 1–5. https://doi.org/10.1002/pra2.2016.14505301025

Dale, A. (2004). Secondary analysis of qualitative data. In M. Lewis-Beck, A. E. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods* (Vol. 1). SAGE Publications.

Dallmeier-Tiessen, S., Darby, R., & Gitmans, K. (2014). Enabling sharing and reuse of scientific data. *New Review of Information Networking*. http://www.tandfonline.com/doi/abs/10.1080/13614576.2014.883936

Dargentas, M., & Roux, D. L. (2005). Potentials and limits of secondary analysis in a specific applied context: The case of EDF-Verbatim. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869224968/abstract/D2954108AFA64283PQ/22

Digital Curation Centre. (2018). *Overview of funders' data policies*. http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

De Vocht, L., Van Compernolle, M., Dimou, A., Colpaert, P., Verborgh, R., Mannens, E., Mechant, P., & Van de Walle, R. (2014). *Converging on semantics to ensure local government data reuse* (Vol. 1280). http://ceur-ws.org/Vol-1280/paper4.pdf

Dicks, B., Mason, B., Williams, M., & Coffey, A. (2006). Ethnography and data re-use: Issues of context and hypertext. *Methodological Innovations Online*, *1*(2), 33–46. https://doi.org/10.4256/mio.2006.0010

Donaldson, D. R., & Conway, P. (2015). User conceptions of trustworthiness for digital archival documents. *Journal of the Association for Information Science and Technology*, *66*(12), 2427–2444. https://doi.org/10.1002/asi.23330

Donaldson, D. R., Martin, S., & Proffen, T. (2017). Understanding perspectives on sharing neutron data at Oak Ridge National Laboratory. *Data Science Journal*. https://doi.org/10.5334/dsj-2017-035

Doolan, D. M., & Froelicher, E. S. (2009). Using an existing data set to answer new research questions: A methodological review. *Research and Theory for Nursing Practice: An International Journal*, *23*(3), 203–215. https://doi.org/10.1891/1541-6577.23.3.203

Duff, W. M., & Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behavior of historians in archives. *The Library Quarterly*, *72*(4), 472–496. https://doi.org/10.1086/lq.72.4.40039793

Eisenhardt, K. M. (2002). Building theories from case study research. In A. M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion* (p. 18). Sage Publications.

Elsevier. (2018). *Research data*. https://www.elsevier.com/about/open-science/research-data

Engineering and Physical Sciences Research Council. (2011). *EPSRC policy framework on research data*. https://www.epsrc.ac.uk/about/standards/researchdata/

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —. *Ecological Informatics*, *11*, 25–33. https://doi.org/10.1016/j.ecoinf.2012.03.004

European Commission: Directorate-General for Research and Innovation. (2016). *Guidelines on FAIR data management in Horizon 2020* (ec.europa.eu). Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J., & Yakel, E. (2013). *The challenges of digging data: A study of context in archaeological data reuse*. 295–304. https://doi.org/10.1145/2467696.2467712

Faniel, I. M. (2009). *Unrealized potential: The socio-technical challenges of a large scale cyberinfrastructure initiative* (p. 65). University of Michigan. https://www.researchgate.net/publication/30862491_Unrealized_Potential_The_Socio-Technical_Challenges_of_a_Large_Scale_Cyberinfrastructure_Initiative

Faniel, I. M., & Yakel, E. (2011). Significant properties as contextual metadata. *Journal of Library Metadata*, 11(3-4), 155–165. http://doi.org/10.1080/19386389.2011.629959

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation*, *75*(6), 1274–1297. https://doi.org/10.1108/JD-08-2018-0133

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(3–4), 355–375. https://doi.org/10.1007/s10606-010-9117-8

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. https://doi.org/10.1002/meet.14504901068

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*, 1404–1416. https://doi.org/10.1002/asi.23480

Faniel, I. M., & Yakel, E. (2017). Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. In L. R. Johnston (Ed.), *Curation research data, Volume 1: Practical strategies for your digital repository* (Vol. 1, pp. 103–126). Association of College and Research Libraries. http://www.oclc.org/content/dam/research/publications/2017/faniel-yakel-practices-do-not-make-perfect.pdf

Fear, K., & Donaldson, D. R. (2012). Provenance and credibility in scientific data repositories. *Archival Science*, *12*(3), 319–339. https://doi.org/10.1007/s10502-012-9172-7

Fecher, B., Friesike, S., & Hebing, M. (2015a). What drives academic data sharing? *PLOS One*, *10*(2). https://doi.org/10.1371/journal.pone.0118053

Fecher, B., Friesike, S., Hebing, M., Linek, S., & Sauermann, A. (2015b). *A reputation economy: Results from an empirical survey on academic data sharing. 2015*. DIW Berlin Discussion Paper.

Fidel, R., Davies, R. K., Douglass, M. H., Holder, J. K., Hopkins, C. J., Kushner, E. J., Miyagishima, B. K., & Toney, C. D. (1999). A visit to the information mall: Web searching behavior of high school students. *Journal of the American Society for Information Science*, 50(1), 24–37. https://doi.org/10.1002/(SICI)1097-4571(1999)50:1<24::AID-ASI5>3.0.CO;2-W

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2), 219–245. https://doi.org/10.1177/1077800405284363

Fowler, F. J. (1995). *Improving survey questions: design and evaluation*. SAGE Publications.

Frank, R. D., Chen, Z., Crawford, E., Suzuka, K., & Yakel, E. (2017). Trust in qualitative data repositories. *Proceedings of the Association for Information Science and Technology*, *54*(1), 102–111. https://doi.org/10.1002/pra2.2017.14505401012

Frické, M. (2009). The knowledge pyramid: a critique of the DIKW hierarchy. *Journal of Information Science, 35*(2), 131–142. https://doi.org/10.1177/0165551508094050

Freundlich, A. (2016, May 3). *Feminist standpoint epistemology and objectivity*. https://wordpress.viu.ca/compassrose/feminist-standpoint-epistemology-and-objectivity/

Gillies, V., & Edwards, R. (2005). Secondary analysis in exploring family and social change: Addressing the issue of context. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869227108/abstract/D2954108AFA64283PQ/37

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.2307/1174772

Gregory, K. (2021). Findable and reusable? Data discovery practices in research (S. Wyatt (ed.)) [PhD, Maastrict University]. https://doi.org/10.26481/dis.20210302kg

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, 70(5), 419–432. https://doi.org/10.1002/asi.24165

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or found? Discovering data needed for research. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e38165eb

Grint, K., & Woolgar, S. (1997). Configuring the user: Inventing new technologies. In K. Grint & S. Woolgar (Eds.), *The machine at work: technology, work, and organization* (pp. 65–94). Polity Press.

Hagstrom, W. (1982). Gift giving as an organizing principle in science. In B. Barnes & D. O. Edge (Eds.), *Science in context: Readings in the sociology of science* (pp. 21–34). MIT Press. https://philpapers.org/rec/HAGGGA

Hammersley, M. (1997). Qualitative data archiving: Some reflections on its prospects and problems. *Sociology*, *31*(1), 131–142. https://doi.org/10.1177/0038038597031001010

Haraway, D. (1991). *Simians, cyborgs, and women: The reinvention of nature*. Routledge.

Harding, S. (2001). Feminist standpoint epistemology. In M. Lederman & I. Bartsch (Eds.), *The Gender and Science Reader* (pp. 145–168). London: Routledge.

He, L., & Nahar, V. (2016). Reuse of scientific data in academic publications. *Aslib Journal of Information Management*, *68*(4), 478–494. https://doi.org/10.1108/AJIM-01-2016-0008

Heaton, J. (2004). *Reworking qualitative data*. SAGE Publications.

Heaton, J. (2008). Secondary analysis of qualitative data: An overview. *Historical Social Research*, *33*(3), 33–45. https://doi.org/10.12759/hsr.33.2008.3.33-45

Hedstrom, M., & Lee, C. A. (2002). Significant properties of digital objects: Definitions, applications, implications. *Proceedings of the DLM-Forum*. https://www.ils.unc.edu/callee/sigprops_dlm2002.pdf

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. In *Biometrical Journal* (Vol. 60, Issue 3, pp. 431–449). https://doi.org/10.1002/bimj.201700067

Henderson, S., Holland, J., & Thomson, R. (2006). Making the long view: Perspectives on context from a qualitative longitudinal (QL) study. *Methodological Innovations Online*, *1*(2), 47–63. https://doi.org/10.4256/mio.2006.0011

Hesse-Biber, S. N. (2006). In-depth interview. In S. N. Hesse-Biber & P. Leavy (Eds.), *The Practice of Qualitative Research* (pp. 119–148). Sage.

Higgins, M. (1999). Meta-information, and time: Factors in human decision making. *Journal of the American Society for Information Science*, 50(2), 132–139. https://doi.org/10.1002/(SICI)1097-4571(1999)50:2<132::AID-ASI4>3.0.CO;2-N

Hills, D. J., Downs, R. R., Duerr, R., Goldstein, J. C., Parsons, M. A., & Ramapriyan, H. (2015). The importance of data set provenance for science. *EOS: Earth and Space Science News*, *96*.

Hirsh, S. G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science*, 50(14), 1265–1283. https://doi.org/10.1002/(SICI)1097-4571(1999)50:14<1265::AID-ASI2>3.0.CO;2-E

Holdren, J. P. (2013). *Increasing access to the results of federally funded scientific research*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Holstein, J. A., & Gubrium, J. F. (2004). Context: Working it up, down, and across. In C. Seale, G. Gobo, F. Gubrium, & D. Silverman (Eds.), *Qualitative Research Practice* (pp. 267–281). SAGE Publications Ltd. https://doi.org/10.4135/9781848608191.d24

Horton, L., & DCC. (2014). *Overview of UK institution RDM policies*. http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies

Hosmer, D. W., Jr, & Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Sons.

Huggett, J. (2018). Reuse remix recycle: Repurposing archaeological digital data. *Advances in Archaeological Practice*, *6*(2), 93–104. https://doi.org/10.1017/aap.2018.1

Interagency Working Group on Digital Data. (2009). *Harnessing the power of digital data for science and society* (NITRD). Retrieved from https://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Inter-university Consortium for Political and Social Research. (2019, May 7). *ICPSR recognized as a 2019 recipient of nation's highest museum and library honor*. https://www.icpsr.umich.edu/icpsrweb/about/cms/1687

Inter-University Consortium for Political and Social Research (ICPSR). (2019). *About the bibliography of data-related literature*. https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/citations/methodology.html

Irwin, S. (2013). Qualitative secondary data analysis: Ethics, epistemology and context. *Progress in Development Studies*, *13*(4), 295–306. https://doi.org/10.1177/1464993413490479

Irwin, S., & Winterton, M. (2011a). *Debates in qualitative secondary analysis: Critical reflections* (Timescapes Working Paper Series No. No.4) (p. 24).

Irwin, S., & Winterton, M. (2011b). Qualitative secondary analysis and social explanation. *Sociological Research Online*, *17*(2), 4.

Israel, G. D. (1992). *Determining sample size* (fact sheet PEOD-6). Program Evaluation and Organizational Development, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida. https://www.academia.edu/21353552/Determining_Sample_Size_1

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, 18(1), 3–20. https://doi.org/10.1177/1525822X05282260

Jackson, P., Smith, G., & Olive, S. (2007). *Families remembering food: Reusing secondary data*.

Johns Hopkins Libraries. (2017, January 1). *Funder data-related mandates and public access plans*. http://dms.data.jhu.edu/data-management-resources/plan-research/funders-data-sharing-requirement/funder-data-related-mandates-and-public-access-plans/

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112–133. https://doi.org/10.1177/1558689806298224

Jones, S. (2012). Developments in research funder data policy. *International Journal of Digital Curation*, *7*(1), 114–125. https://doi.org/10.2218/ijdc.v7i1.219

Kafai, Y., & Bates, M. J. (1997). Internet web-searching instruction in the elementary classroom: Building a foundation for information literacy. *School Library Media Quarterly*, *25*(2), 103–111. https://eric.ed.gov/?id=EJ541378

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, *93*(5), 1449–1475. https://www.jstor.org/stable/3132137

Katsanidou, A., Horton, L., & Jensen, U. (2016). Data policies, data management, and the quality of academic writing. *International Studies Perspectives*, *17*(4), 379–391. https://doi.org/10.1093/isp/ekv014

Keene, S. (2005). *Fragments of the world: Uses of museum collections*. Routledge.

Kelder, J.-A. (2005). Using someone else's data: Problems, pragmatics and provisions. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869227115/abstract/CF34FA2F09534637PQ/1

Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the Association for Information Science and Technology*, *59*(3), 363–374. https://doi.org/10.1002/asi.20722

Kemper, E. A., Stringfield, S., & Tashakkori, A. (2003). Mixed methods sampling strategies in social science research. In C. Teddlie (Ed.), *Handbook of Mixed Methods in Social & Behavioral Research* (pp. 273–296). SAGE Publications, Inc.

Kerstholt, J., & Ayton, P. (2001). Should NDM change our understanding of decision making? *Journal of Behavioral Decision Making*, 14(5), 370–371. https://doi.org/10.1002/bdm.390

Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *68*(12), 2709–2719. https://doi.org/10.1002/asi.23892

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., & Scholze, F (2017). The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine*, *23*(3/4). https://doi.org/10.1045/march2017-kindling

King, K. (1987). The passing dreams of choice ... Once before and after: Audre Lorde and the apparatus of literary production (book prospectus). University of Maryland at College Park.

Klump, J. (2017). Data as social capital and the gift culture in research. *Data Science Journal*, *16*. https://doi.org/10.5334/dsj-2017-014

Koesten, L. M., Kacprzak, E., Tennison, J. F. A., & Simperl, E. (2017). The trials and tribulations of working with structured data: -A study on information seeking behaviour. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1277–1289. https://doi.org/10.1145/3025453.3025838

Koesten, L., Gregory, K., Groth, P., & Simperl, E. (2021). Talking datasets – Understanding data sensemaking behaviours. In *International Journal of Human-Computer Studies* (Vol. 146, p. 102562). https://doi.org/10.1016/j.ijhcs.2020.102562

Kriesberg, A., Frank, R. D., Faniel, I. M., & Yakel, E. (2013). The role of data reuse in the apprenticeship process. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–10. https://doi.org/10.1002/meet.14505001051

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. SAGE Publications.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). Emerald.

Kuipers, T., & Hoeven, J. van der. (2009). *PARSE.Insight: Insight into digital preservation of research output in Europe: Survey report*. http://docplayer.net/127428-Parse-insight-deliverable-d3-4-survey-report-of-research-output-europe-title-of-deliverable-survey-report.html

Kwek, D., & Kogut, G. (2015). *Knowledge of prior work and soundness of project: A review of the research on secondary analysis of research data*. National Institute of Education (Singapore). https://repository.nie.edu.sg/bitstream/10497/17598/1/OER_report_Kwek%26Kogut_Nov2015_a.pdf

Langkamp, D. L., Lehman, A., & Lemeshow, S. (2010). Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics*, *10*(3), 205–210. https://doi.org/10.1016/j.acap.2010.01.005

Latour, B. (1990). Drawing things together. In M. Lynch & S. Woolgar (Eds.), *Representation in scientific practice* (pp. 19–68). MIT Press.

Latour, B and Woolgar, S (1982). The cycle of credibility. In Barnes, B & D. O. Edge, (Eds.), *Science in context: Readings in the sociology of science* (pp. 35–43). MIT Press. https://philpapers.org/rec/BARSIC

Latour, B., & Woolgar, S. (1986). *Laboratory life: The social construction of scientific facts*.

Leckenby, D., & Hesse-Biber, S. (2007). Feminist approaches to mixed-methods research. In S. Hesse-Biber & P. Leavy (Eds.), *Feminist research practice*. SAGE Publications, Inc. https://dx.doi.org/10.4135/9781412984270.n9

Lee, C. (2019, October 31). Welcome, singular "they." APA Style. https://apastyle.apa.org/blog/singular-they

Lee, J.-G., & Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Lemieux, V. L., & The ImProvenance Group. (2016). Provenance: Past, present and future in interdisciplinary and multidisciplinary perspective. In *Building Trust in Information* (pp. 3–45). Springer, Cham. https://doi.org/10.1007/978-3-319-40226-0_1

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge University Press. https://doi.org/10.1017/CBO9780511626388

Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology*, *61*(10), 2029–2048. https://doi.org/10.1002/asi.21339

Margaryan, A., & Littlejohn, A. (2008). Repositories and communities at cross-purposes: Issues in sharing and reuse of digital learning resources. *Journal of Computer Assisted Learning*, *24*, 333–347. https://doi.org/10.1111/j.1365-2729.2007.00267.x

Mason, J. (2007). "Re-using" qualitative data: On the merits of an investigative epistemology. *Sociological Research Online*, *12*(3), 3. http://www.socresonline.org.uk/12/3/3.html

Mauthner, N. S. (2014). Digital data sharing: A genealogical and performative perspective. *Studia Socjologiczne,* (3), 177–186. http://search.proquest.com/docview/1722763611/abstract/B497CEAD876842DAPQ/1

Mauthner, N. S., & Doucet, A. (2008). `Knowledge once divided can be hard to put together again': An epistemological critique of collaborative and team-based research practices. *Sociology*, *42*(5), 971–985. https://doi.org/10.1177/0038038508094574

Mauthner, N. S., & Parry, O. (2009). Qualitative data preservation and sharing in the social sciences: On whose philosophical terms? *Australian Journal of Social Issues*, *44*(3), 291–307. https://doi.org/10.1002/j.1839-4655.2009.tb00147.x

Mauthner, N. S., Parry, O., & Backett-Milburn, K. (1998). The data are out there, or are they? Implications for archiving and revisiting qualitative data. *Sociology*, *32*(4), 733–745. https://doi.org/10.1177/0038038598032004006

Mayernik, M. S. (2012). Data citation initiatives and issues. *Bulletin of the Association for Information Science and Technology*, 38(5), 23–28. http://doi.org/10.1002/bult.2012.1720380508

McKay, D. (2014). Bend me, shape me: A practical experience of repurposing research data. In *IEEE/ACM Joint Conference on Digital Libraries* (pp. 399–402). https://doi.org/10.1109/JCDL.2014.6970196

Medjedović, I. (2011). Secondary analysis of qualitative interview data: Objections and experiences. Results of a German feasibility study. *Forum: Qualitative Social Research*, *12*(3). http://search.proquest.com/docview/898516565/abstract/64CE610513034583PQ/1

Medjedović, I., & Witzel, A. (2005). Secondary analysis of interviews: Using codes and theoretical concepts from the primary study. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869224973/abstract/12B1F26EC6C541F3PQ/1

Medjedović, I., & Witzel, A. (2011). Sharing and archiving qualitative and qualitative longitudinal research data in Germany. *IASSIST Quarterly*, *34*(3), 42–42. https://doi.org/10.29173/iq461

Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, *29*, 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. SAGE.

Mooney, H., & Newton, M. (2012). The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication*, 1(1), eP1035. http://doi.org/10.7710/2162-3309.1035

Moore, F. P. L. (2010). Tales from the archive: Methodological and ethical issues in historical geography research. *Area*, *42*(3), 262–270. http://www.jstor.org/stable/40890880

Moore, N. (2006). The contexts of context: Broadening perspectives in the (re)use of qualitative data. *Methodological Innovations Online*, *1*(2), 21–32. https://doi.org/10.4256/mio.2006.0009

Moore, N. (2007). (Re)using qualitative data? *Sociological Research Online*, *12*(3), 1–13. https://doi.org/10.5153/sro.1496

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research*, *40*(2), 120-123. https://doi.org/10.1097/00006199-199103000-00014

National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Research Data and Information, & Committee on Toward an Open Science Enterprise. (2018). *Open science by design: Realizing a vision for 21st century research*. National Academies Press. https://www.nap.edu/catalog/25116/open-science-by-design-realizing-a-vision-for-21st-century

National Academy of Sciences. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age* (p. 368). National Academies Press: Washington, D.C. https://doi.org/10.1126/science.1178927

National Institutes of Health. (2016, August 29). *NIH Sharing policies and related guidance on NIH-funded research resources*. https://grants.nih.gov/policy/sharing.htm

National Institutes of Health. (2020, October 29). *Final NIH policy for data management and sharing*. NIH Grants and Funding. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

National Institutes of Health Office of Behavioral and Social Sciences. (2018). *Best practices for mixed methods research in the health sciences*. National Institutes of Health.

https://www.obssr.od.nih.gov/wp-content/uploads/2018/01/Best-Practices-for-Mixed-Methods-Research-in-the-Health-Sciences-2018-01-25.pdf

National Research Council. (2003). *Sharing publication-related data and materials*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/10613

National Science Foundation. (2007). *Cyberinfrastructure vision for 21st century discovery*. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

National Science Foundation. (2018). *Building the future: Investing in discovery and innovation*. https://www.nsf.gov/pubs/2018/nsf18045/nsf18045.pdf

National Science Foundation. (n.d.). *Dissemination and sharing of research results*. https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Niu, J. (2009a). Overcoming inadequate documentation. *Proceedings of the American Society for Information Science and Technology*, *46*(1), 1–14. https://doi.org/10.1002/meet.2009.145046024

Niu, J. (2009b). *Perceived documentation quality of social science data.* University of Michigan, Ann Arbor, MI.

Notz, P. (2005). Secondary qualitative analysis of interviews. A method used for gaining insight into the work/life balance of middle managers in Germany. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869227056/abstract/D2954108AFA64283PQ/41

O'Connor, H., & Goodwin, J. (2010). Utilizing data from a lost sociological project: Experiences, insights, promises. *Qualitative Research*, *10*(3), 283–298. https://doi.org/10.1177/1468794110362875

Oleksik, G., Milic-Frayling, N., & Jones, R. (2012). Beyond data sharing: Artifact ecology of a collaborative nanophotonics research centre. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 1165–1174). New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145376

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Organization for Economic Co-operation and Development. (2015). *Making open science a reality*. https://www.innovationpolicyplatform.org/content/open-science

Pampel, H., & Dallmeier-Tiessen, S. (2014). Open research data: From vision to practice. In *Opening science* (pp. 213–224). Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-00026-8_14

Parezo, N. J. (1996). The formation of anthropological archival records. In W. D. Kingery (Ed.), *Learning from things: Method and theory of material culture studies* (pp. 145–174). Washington, D.C.: Smithsonian Institution Press. https://trove.nla.gov.au/version/38653395

Parry, O., & Mauthner, N. (2005). Back to basics: Who re-uses qualitative data and why? *Sociology*, *39*(2), 337–342. https://doi.org/10.1177/0038038505050543

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review*, 1(2). https://doi.org/10.1162/99608f92.fc14bf2d

Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. Eos, Transactions American Geophysical Union, 91(34), 297–298. http://doi.org/10.1029/2010EO340001

Pätzold, H. (2005). Secondary analysis of audio data. Technical procedures for virtual anonymization and pseudonymization. *Forum: Qualitative Social Research*, 6(1). http://search.proquest.com/docview/869227193/abstract/D2954108AFA64283PQ/40

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723–744. https://doi.org/10.1007/s11192-016-1887-4

Pienta, A. M., Alter, G. C., & Lyle, J. A. (2010). *The enduring value of social science research: The use and reuse of primary research data*. "The organisation, economics and policy of scientific research" workshop, Torino, Italy. http://deepblue.lib.umich.edu/handle/2027.42/78307

Pine, K. H., Wolf, C., & Mazmanian, M. (2016). The work of reuse: Birth certificate data and healthcare accountability measurements. *iConference 2016 Proceedings.* https://doi.org/10.9776/16320

Piwowar, H. A., Carlson, J. D., & Vision, T. J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the Association for Information Science and Technology*, *48*(1), 1–4. https://doi.org/10.1002/meet.2011.14504801337

Piwowar, H. A., & Chapman, W. W. (2008, June). A review of journal policies for sharing research data. *Proceedings ELPUB 2008 Conference on Electronic Publishing*. http://elpub.scix.net/data/works/att/001_elpub2008.content.pdf

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Podesta, J., Pritzker, P., Moniz, E. J., Holdren, J. P., & Zients, J. (2014). *Big data: Seizing opportunities, preserving values*. Executive Office of the President. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Redman-Maclaren, M., Mills, J., & Tommbe, R. (2014). Interpretive focus groups: A participatory method for interpreting and extending secondary analysis of qualitative data. *Global Health Action; Abingdon*, *7*. http://dx.doi.org.proxy.lib.umich.edu/10.3402/gha.v7.25214

Reed, J. (1992). Secondary data in nursing research. *Journal of Advanced Nursing*, 17(7), 877–883. https://doi.org/10.1111/j.1365-2648.1992.tb02011.x

Rennstam, J., & Ashcraft, K. L. (2014). Knowing work: Cultivating a practice-based epistemology of knowledge in organization studies. *Human Relations*, *67*(1), 3–25. https://doi.org/10.1177/0018726713484182

Rew, L., Koniak-Griffin, D., Lewis, M. A., Miles, M., & O'sullivan, A. (2000). Secondary data analysis: New perspective for adolescent research. *Nursing Outlook*, 48(5), 223–229. https://doi.org/10.1067/mno.2000.104901

Ribes, D., & Jackson, S. J. (2013). Data bite man: The work of sustaining a long-term study. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 147–166). MIT Press. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6462156

Rolland, B., & Lee, C. P. (2013). Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 435–444). New York, NY, USA: ACM. https://doi.org/10.1145/2441776.2441826

Rung, J., & Brazma, A. (2012). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, *14*(2), 89–99. https://doi.org/10.1038/nrg3394

Saldaña, J. (2009). *The coding manual for qualitative researchers*. SAGE.

Savage, M. (2005). Revisiting classic qualitative studies. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869227066/abstract/4030492B6F774FCCPQ/1

Schacter, J., Chung, G. K. W. K., & Dorr, A. (1998). Children's internet searching on complex problems: Performance and process analyses. *Journal of the American Society for Information Science*, *49*(9), 840–849. https://doi.org/10.1002/(SICI)1097-4571(199807)49:9<840::AID-ASI9>3.0.CO;2-D

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, *29*(1), 65–88. https://doi.org/10.1146/annurev.soc.29.110702.110112

Schwartz, C. A. (1989). Book selection, collection development, and bounded rationality. *College and Research Libraries*. https://doi.org/10.5860/crl_50_03_328

Shankar, K. (n.d.). Order from chaos: The poetics and pragmatics of scientific recordkeeping. *Journal of the American Society for Information Science and Technology*, *58*(10), 1457–1466. https://doi.org/10.1002/asi.20625

Shapin, S., Schaffer, S. (1985). *Leviathan and the air-pump : Hobbes, Boyle, and the experimental life*. Princeton University Press.

Sherif, V. (2018). Evaluating preexisting qualitative research data for secondary analysis. *Forum: Qualitative Social Research*, *19*(2). http://search.proquest.com/docview/2022516538/abstract/D7B0F8123F0C4989PQ/1

Simon, H. A. (1994). *The sciences of the artificial* (2nd ed.). MIT Press.

Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society*, *1*(1), 25–39. https://doi.org/10.1007/BF02512227

Smioski, A. (2010). Establishing a qualitative data archive in Austria. *IASSIST Quarterly / International Association for Social Science Information Service and Technology*, 6. https://iassistdata.org/sites/default/files/iqvol34_35_smioski.pdf

Specht, A., Guru, S., Houghton, L., Keniger, L., Driver, P., Ritchie, E. G., Lai, K., & Treloar, A. (2015). Data management challenges in analysis and synthesis in the ecosystem sciences.

*The Science of the Total Environment*, *534*, 144–158. https://doi.org/10.1016/j.scitotenv.2015.03.092

Stake, R. E. (2006). *Multiple case study analysis*. The Guilford Press.

Stanley, M. (2013). Where is that moon, anyway? The problem of interpreting historical solar eclipse observations. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 77–88). MIT Press. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6462165

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLOS One*, *8*(6). https://doi.org/10.1371/journal.pone.0067111

Straughn-Navarro, R. (2016). Provenance in situ: Documenting information of origins across GLAM contexts. *International Information & Library Review*, *48*(4), 287–293. https://doi.org/10.1080/10572317.2016.1243964

Sturges, P., Bamkin, M., Anders, J., & Hussain, A. (2014, January 1). *Access to research data: Addressing the problem through journal data sharing policies*. https://jordproject.wordpress.com/access-to-research-data-addressing-the-problem-through-journal-data-sharing-policies/

Stvilia, B., Hinnant, C. C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2015). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology*, *66*(2), 246–263. https://doi.org/10.1002/asi.23177

Sveinsdottir, T., Wessels, B., Smallwood, R., Linde, P., Kala, V., Tsoukala, V., & Sondervan, J. (2013). *Stakeholder values and relationships within open access and data dissemination and preservation ecosystems*. Policy RECommendations for Open access to research Data in Europe (RECODE). http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research*. SAGE Publications, Inc.

Temple, B., Edwards, R., & Alexander, C. (2006). Grasping at context: Cross language qualitative research as secondary qualitative data analysis. *Forum: Qualitative Social Research*, *7*(4). http://search.proquest.com/docview/869232116/abstract/FE54EF58E6D940EDPQ/1

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLOS One*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS One*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Thompson, P. (2000). Re-using qualitative research data: A personal account. *Forum: Qualitative Social Research*, *1*(3). https://doi.org/10.17169/fqs-1.3.1044

Thomson, R., & Holland, J. (2003). Hindsight, foresight and insight: The challenges of longitudinal qualitative research. *International Journal of Social Research Methodology*, *6*(3), 233–244. https://doi.org/10.1080/1364557032000091833

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, U.K.; New York: Cambridge University Press.

Travers, M. (2009). A not so strange silence: Why qualitative researchers should respond critically to the qualitative data archive. *Australian Journal of Social Issues*, *44*(3), 273–289. https://doi.org/10.1002/j.1839-4655.2009.tb00146.x

Trevelyan, J. (2016). Extending engineering practice research with shared qualitative data. *Advances in Engineering Education*, *5*(2), 1–31. http://proxy.lib.umich.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ofs&AN=116895880&site=ehost-live&scope=site

Tyler, A. R. B., Suzuka, K., & Yakel, E. (2020). Complementary data as metadata: Building context for the reuse of video records of practice. *International Journal of Digital Curation*, *15*(1), 13. https://doi.org/10.2218/ijdc.v15i1.701

UCLA: Statistical Consulting Group. (n.d.a). Logistic regression power analysis. https://stats.idre.ucla.edu/stata/dae/logistic-regression-power-analysis/

UK Data Service. (n.d.). *Reusing qualitative data*. https://www.ukdataservice.ac.uk/use-data/guides/methods-software/qualitative-reuse

University of New Mexico Libraries. (2018). Research guides: Digital data management, curation, and archiving: data sharing policies. http://libguides.unm.edu/c.php?g=232325&p=2402343

van den Berg, H. (2005). Reanalyzing qualitative interviews from different angles: The risk of decontextualization and other problems of sharing qualitative data. *Forum: Qualitative Social Research*, *6*(1). http://search.proquest.com/docview/869227101/abstract/D2954108AFA64283PQ/24

van den Berg, H., Wetherell, M., & Houtkoop-Steenstra, H. (2004). *Analyzing race talk: Multidisciplinary perspectives on the research interview*. Cambridge University Press.

van Weijen, D. (2012). The language of (future) scientific communication. *Research Trends*, (31). https://www.researchtrends.com/issue-31-november-2012/the-language-of-future-scientific-communication/

Vertesi, J., & Dourish, P. (2011). The value of data: Considering the context of production in data economies. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (pp. 533–542). New York, NY, USA: ACM. https://doi.org/10.1145/1958824.1958906

Vogel, G. (2011, October 31). Report: Dutch "Lord of the data" forged dozens of studies (update). *Science Insider*. http://news.sciencemag.org/europe/2011/10/report-dutch-lord-data-forged-dozens-studies-update

Vogt, W. P. (2008). Quantitative versus qualitative is a distraction: Variations on a theme by brewer and hunter (2006). *Methodological Innovations Online*, 3(1), 18–24. https://doi.org/10.4256/mio.2008.0007

Wajcman, J. (2002). Addressing technological change: The challenge to social theory. *Current Sociology*, *50*(3), 347–363. https://doi.org/10.1177/0011392102050003004

Wallis, J. C., Borgman, C. L., Mayernik, M. S., Pepe, A., Ramanathan, N., & Hansen, M. (2007). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. In *Research and advanced technology for digital libraries* (pp. 380–391). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74851-9_32

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLOS One*, *8*(7). http://dx.doi.org/10.1371%2Fjournal.pone.0067332

Walters, P. (2009). Qualitative archiving: Engaging with epistemological misgivings. *Australian Journal of Social Issues*, *44*(3), 309–320. https://doi.org/10.1002/j.1839-4655.2009.tb00148.x

Wan, X., & Pavlidis, P. (2007). Sharing and reusing gene expression profiling data in neuroscience. *Neuroinformatics*, *5*(3), 161–175. https://doi.org/10.1007/s12021-007-0012-5

Warwick, C., Galina, I., Rimmer, J., Terras, M., Blandford, A., Gow, J., & Buchanan, G. (2009). Documentation and the users of digital resources in the humanities. *Journal of Documentation*, *65*(1), 33–57. https://doi.org/10.1108/00220410910926112

Wästerfors, D., Åkerström, M., & Jacobsson, K. (2014). Reanalysis of qualitative data. In U. Flick, *The SAGE handbook of qualitative data analysis* (pp. 467–480). SAGE Publications, Inc. https://doi.org/10.4135/9781446282243.n32

Weber, N., & Chao, T. (2011, June 1). *A multi-disciplinary analysis of data reuse activities*. International Association for Social Science Information Service & Technology, Vancouver, BC. http://www.iassistdata.org/conferences/2011/presentation/2847

Weinberger, D. (2010, February 2). The problem with the data-information-knowledge-wisdom hierarchy. *Harvard Business Review*. https://hbr.org/2010/02/data-is-to-info-as-info-is-not

Wheeler, G. (2020). Bounded rationality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/bounded-rationality/

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS One*, *6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE Publications, Inc.

Winerman, L. (2017). Trend report: Psychologists embrace open science. *Monitor on Psychology*, *48*(10). http://www.apa.org/monitor/2017/11/trends-open-science.aspx

Wolski, M., Howard, L., & Richardson, J. (2017). A trust framework for online research data services. *Publications*, *5*(2), 14. http://dx.doi.org.proxy.lib.umich.edu/10.3390/publications5020014

Wojtkiewicz, R. A. (2016). Control modeling. In *Elementary regression modeling: A discrete approach* (pp. 53–77). SAGE Publications. https://methods.sagepub.com/book/elementary-regression-modeling/i668.xml

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, data use, and scientific inquiry: Two case studies of data practices. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 19–22). New York, NY, USA: ACM. https://doi.org/10.1145/2232817.2232822

Yamane, T. (1967). *Statistics, an introductory analysis*, 2nd Ed., New York: Harper and Row.

Yan, T., Kreuter, F., & Tourangeau, R. (n.d.). Evaluating survey questions: A comparison of methods, 27. https://www.researchgate.net/profile/Ting_Yan3/publication/272349195_Evaluating_Survey_Questions_A_Comparison_of_Methods/links/54e1f2b90cf24d184b11f82b.pdf

Yardley, S. J., Watts, K. M., Pearson, J., & Richardson, J. C. (2014). Ethical issues in the reuse of qualitative data: Perspectives from literature, practice, and participants. *Qualitative Health Research*, *24*(1), 102–113. https://doi.org/10.1177/1049732313518373

Yin, R. K. (2014). *Case study research: design and methods*. SAGE.

Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–6. https://doi.org/10.1002/pra2.2016.14505301126

Yoon, A. (2017a). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956. https://doi.org/10.1002/asi.23730

Yoon, A. (2017b). Role of communication in data reuse. *Proceedings of the Association for Information Science and Technology*, *54*(1), 463–471. https://doi.org/10.1002/pra2.2017.14505401050

York, J., Gutmann, M., & Berman, F. (2016). Will today's data be here tomorrow? Measuring the stewardship gap. In *Proceedings of the 13th International Conference on Digital Curation*. Bern, Switzerland: Swiss National Library.

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data*.

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, 7(1–2), 5–16. https://doi.org/10.1007/s00799-007-0015-8

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values*, *33*(5), 631–652. Retrieved from http://www.jstor.org/stable/29734058

Zsambok, C. E., & Klein, G. A. (Eds.). (1997). *Naturalistic decision making*. L. Erlbaum Associates.