Data Management in Large-Scale Education Research

Crystal Lewis

2023-06-29

Contents

Chapter 1

Preamble

This is the in-progress version of *Data Management in Large-Scale Education Research*. To see a previous version of this material, please visit this website.

The results of educational research studies are only as accurate as the data used to produce them. - Aleata Hubbard (2017)

1.1 Introduction

In 2013, without knowing that the term research data management existed, I accepted a position with a prevention science research center. My job was to coordinate the collection and management of data for federally funded randomized controlled trial efficacy studies taking place in K-12 schools, along with a team of investigators, other research staff, part-time data collectors, and graduate students. While I had some experience analyzing and working with education data, i.e. ECLS-K, I had no experience running research grants, collecting original data, or managing research data, but I was excited to learn.

In my time in that position I learned to plan, schedule, and track data collection activities, create data capture tools, organize and document data inputs, and produce usable data outputs; but I didn't learn to do those things through any formal training. There were no books, courses, or workshops that I learned from. I learned from colleagues and a large amount of trial and error. Since then, as I have met more investigators, data managers, and project coordinators in education research, I realize that this is a common method for learning data management—mentoring and "winging it". And while learning data management through these informal methods helps us get by, the ramifications of this unstandardized system are felt by both the project team and future data users.

1.2 Why this book

Research data management is becoming more complicated. We are collecting more data, in sometimes very novel ways, and using more complex technologies, all while increasing the visibility of our work with the push for data sharing and open science practices (Briney 2015; Nelson 2022). Ad hoc data management practices may have worked for us in the past, but now others need to understand our processes as well, requiring researchers to be more thoughtful in planning their data management routines.

1.2.1 Lack of training, resources, and standards

In order to implement thoughtful and standardized data management practices, researchers need training. Yet there is a clear lack of data management training in higher education. In a survey of 274 psychology researchers, Borghi and Van Gulick (2021) found that only 33% of respondents learned data management from college level coursework, while 64% learned from collaborators, and 52% learned from self-education. In their survey of 202 education researchers (PIs and Co-PIs), Ceviren and Logan (2022) found that over 60% of respondents reported having no formal training in data management, yet across eight different data management practices, respondents were responsible for data management activities anywhere from 25-50% of the time. Similarly, in a survey of 150 graduate students in a school of education, when asked if they needed more training in research data management, the average overall score on a scale from 1 to 100 was 80, while the overall confidence in managing data score was only 40 (Zhou, Xu, and Kogut 2023). Furthermore, of the training that does exist, usually provided through university library systems, most material is either discipline agnostic or STEM focused, leaving a gap in training on how to apply skills to the field of education which has unique issues, particularly around working with human subjects data (Nichols Hess and Thielen 2017).

Without training, resources and formal support systems are the next best option for learning best practices. Within university systems, in addition to providing periodic training, research data librarians provide data management planning consultation for researchers and their teams. There is also a wealth of existing research data management books and manuals written for broad audiences which I will link to in this book. However, while education researchers are starting to put out some excellent resources (Neild, Robinson, and Agufa 2022; Reynolds, Schatschneider, and Logan 2022), I still find there is a dearth of practical guides for researchers to refer to when building a data management workflow in the field of education, especially those working on large-scale longitudinal research grants where there are many moving pieces. Researchers are often collecting data in real-world environments, such as school systems, and keeping that data secure and reliable in a deliberate and orderly way can be overwhelming.

Last, unfortunately, while other fields of research, such as psychology, appear to be banding together to develop standards around how to structure and doc-

ument data (Kline 2018), the field of education has yet to develop agreed upon rules for things such as data documentation or data formats. This lack of standards leads to inconsistencies in the quality and usability of data products across the field (Borghi and Van Gulick 2022).

1.2.2 Consequences

A lack of training in data management practices and an absence of agreed upon standards in the field of education leads to consequences. Implementing subpar and inconsistent data management practices, while typically only resulting in frustration and time lost, also has the potential to be devastating, resulting in analyzing erroneous data or even unusable or lost data. In a review of 1,082 retracted publications from the journal PubMed from 2013-2016, authors found that 32% of retractions were due to data management errors (Campos-Varela and Ruano-Raviña 2019). In a 2013 study surveying 360 graduate students about their data management practices, 14% of students indicated they had to recollect data that had been previously collected because they could not find a file or the file had been corrupted, while 17% of students said they had lost a file and been unable to recollect it (Doucette and Fyfe 2013). In their study of 488 researchers who had published in a psychology journal between 2010 and 2018, Kovacs, et al. (2021) asked respondents about their data management mistakes and found that the most serious data management mistakes reported led to a range of consequences including time loss, frustration, and even erroneous conclusions.

Poor data management can even prevent researchers from implementing other good open science practices. In waves 1 and 2 of the Open Scholarship Survey being collected by the Center for Open Science (2022), the team has found that of the education researchers surveyed who are currently not publicly sharing their research data, about 10% mentioned "being nervous about mistakes" as a reason for not sharing. Similarly, when surveying 780 researchers in the field of psychology, researchers found that 38% of respondents agreed that a "fear of discovery of errors in the data" posed a barrier to data sharing (Houtkoop et al. 2018).

The well-known replication crisis is another reason to be concerned with data management. Failure to implement practices such as quality documentation or standardization of practices (among many other reasons), resulted in one study finding that across 1,500 researchers surveyed, more than 70% had tried and failed to reproduce another researcher's study (Baker 2016).

1.3 About this book

While the field as a whole may not have agreed upon guidelines for data management, there are still practices that are proven to result in more secure, reproducible, and reliable data. My hope is that this book can be a foundation to

help researchers think through how to build a quality, standardized data management workflow that works for their team and their projects. As suggested in the title of this book, this content is designed to specifically help teams navigate the complicated workflows associated with large-scale research, such as randomized controlled trial studies, but ultimately these practices are applicable to any research project, no matter the scale.

This book should be viewed as a handbook to be referenced regularly and is not necessarily meant to be read in its entirety in one sitting. While perusing through the entire book to better understand the entire research data life cycle is very helpful, this book is also intended to have chapters referenced as needed when you are ready to start planning a specific phase of your project.

1.3.1 What this book will cover

This book begins, like many other books in this subject area, by describing the research life cycle and how data management fits within the larger picture. The remaining chapters are then organized by each phase of the life cycle, with examples of best practices provided for each phase. Considerations on whether you should implement, and how to integrate those practices into your workflow will be discussed.

1.3.2 What this book will not cover

It is important to also point out what this book will not cover. This book is intended to be tool agnostic and provide suggestions that anyone can use, no matter what tools you work with, especially when it comes to data cleaning. Therefore, while I might mention options of tools you can use for different tasks, I will not advocate for any specific tools.

There are also no specific coding practices or actual syntax included in this book. To be honest, in many ways I feel that the actual "data cleaning" phase of data management is the *easiest* phase to implement, as long as you implement good practices up until that point. Because of that, this book introduces practices in all phases leading up to data cleaning that will prepare your data for minimal cleaning. With that said, I do provide examples of what I would expect to see in a data cleaning process, I just do not provide steps for any specific software system. That is beyond the scope of this book.

This book will also not talk about analysis or preparing data for analysis through means such as data imputation, removal of legitimate outliers, or calculating analysis specific variables. Written from the perspective of a data manager, the end goal of data management is to build datasets for general data sharing. This means we will cover practices that keep data in its most complete and true, but usable form, for any future researcher to analyze in a way that works best for them.

1.4 Who this book is for

This book is for anyone involved in a research study involving original data collection. In particular, this book focuses on quantitative, observational data collection, although I do think that many of the practices covered can also apply to qualitative data collection as well. This book also applies to any team member, ranging from PIs, to data managers, to project staff, to students, to contractual data collectors. The contents of this book are useful for anyone who may have a part in planning, collecting, or organizing research study data.

1.5 Final note

Planning and implementing new data management practices on top of planning the implementation of your entire research grant can feel overwhelming. However, the idea of this book is to find the practices that work for you and your team and implement them consistently. For some teams that may look like implementing just a few of the suggestions mentioned and for others it may involve implementing all of the suggestions. Improving your data management workflow is a process and it becomes easier over time as those practices become part of your normal routine. At some point you may even find that you enjoy working on data management processes as you start to see the benefits of their implementation!

1.6 Acknowledgements

This book is a compilation of lessons I have learned in my personal experiences as a data manager, knowledge collected from existing books and papers (many written by librarians or those involved in the open science movement), as well as advice and stories collected through interviews with other researchers who work with data. I want to be clear that I did not formally study research data management, unlike research data librarians who are experts in this content. Much of this book will be based off of lessons learned from firsthand experience and this book is my attempt to hopefully save others from making the same mistakes I have personally made or seen others make. I can not emphasize enough that if you work for a university and you have the opportunity to consult with a librarian for your project, you absolutely should!

With that said, there is a long list of people I would like to acknowledge for their contributions to this book and for supporting me in this process.

There were many people who graciously allowed me to interview them about their current data management practices. They are Mary McCraken, Ryan Estrellado, Kim Manturuk, Beth Chance, Jessica Logan, Rebecca Schmidt, Sara Hart, and Kerry Shea. These interviews were integral to supplementing my personal knowledge with the broader experience of others in the field. Yet, they affirmed that yes, data management is hard, especially in the context of some

of the complicated study designs we work with in education research, and that everyone who works in this field wishes that better training, support systems, and standards existed. Thank you to everyone who gave me an hour of their time to share their experiences and knowledge! I also have to give a special thank you to Jessica Logan for being the first person I met who appreciates all things data management as much as I do, and since having our interview, has provided invaluable support while working on this book.

I also want to thank everyone who took the time to read and provide feedback on chapters of this book for me. This includes Meghan Harris, Alexis Swanz, and Allyson Hanson. Your revisions and insight helped make this a more cohesive and useful book!

A special thank you to Keith Herman as well. Many years ago he suggested I write a book titled Data Management in Large-Scale Education Research, based on everything I've learned in my experience as a data manager. At the time I considered his suggestion a fun but impossible idea. Yet after sitting with that idea in the back of my head for several years, I realized his idea was actually not so far-fetched. Thank you to Keith for believing I could do something I didn't even know was possible.

Much appreciation to Wendy Reinke as well. Although she may not know it, she is the first person I learned research data management practices from. Joining a project where she had already created documentation and tracking systems was my first glimpse into building tools that help you manage data and my love of research data management grew out of this experience.

I want to say thank you to the POWER Data Management Issues in Education Research Hub. Regularly meeting with this group of data managers, researchers, students, and professors over the last two years has been an amazing source of both support and learning and has greatly increased my understanding of data management.

Last, thank you to Josh for fully supporting me in the decision to write this book and to Fox for being the reason I remember to step away from my computer from time to time and have fun.

Chapter 2

Research Data Management Overview

2.1 What is research data management?

Research data management (RDM) involves the organization, storage, preservation, and dissemination of research study data (Bordelon n.d.b). Research study data includes materials generated or collected throughout a research process (National Endowment for the Humanities 2018). As you can imagine, this broad definition includes much more than just the management of digital datasets. It also includes physical files, documentation, artifacts, recordings, and more. RDM is a substantial undertaking that begins long before data is ever collected, during the planning phase, and continues well after a research project ends during the archiving phase.

2.2 Standards

Data management standards refer to rules for how data should be collected, formatted, described, and shared (Borghi and Van Gulick 2022; Koos n.d.). Implementing standards for things such as how variables should be collected and named, which items from common measures should be shared, and how data should be formatted and documented, leads to more findable and usable data within fields and provides the added benefit of allowing researchers to integrate datasets without painstaking work to normalize the data.

Some fields have adopted standards across the research life cycle, such as CDISC standards used by clinical researchers (CDISC n.d.), other fields have adopted standards specifically around metadata, such as the TEI standards used in digital humanities (Burnard 2014) or the ISO 19115 used for geospatial

data(Michener 2015), and through grassroots efforts, other fields such as psychology are developing their own standards for things such as data formatting and documentation (Kline 2018) based on the FAIR principles and inspired by the BIDS standard (BIDS-Contributors 2022). Yet, it is common knowledge that there are currently no agreed-upon norms in the field of education research (Institute of Education Sciences n.d.a; Logan and Hart 2023). The rules for how to collect, format, and document data is often left up to each individual team, as long as external compliance requirements are met (Tenopir et al. 2016). However, with a growing interest in open science practices and expanding requirements for federally funded research to make data publicly available (Holdren 2013), data repositories will most likely begin to play a stronger role in promoting standards around data formats and documentation (Borghi and Van Gulick 2022).

2.3 Why care about research data management?

Without current agreed-upon standards in the field, it is important for research teams to develop their own data management standards that apply within and across all of their projects. Developing internal standards, implemented in a reproducible data management workflow, allows practices to be implemented consistently and with fidelity. There are both external pressures and personal reasons to care about developing research data management standards for your projects.

2.3.1 External Reasons

1. **Funder compliance**: Any researcher applying for federal funding will be required to submit a data management plan (see Chapter 4) along with their grant proposal (Holdren 2013; Nelson 2022). The contents of these plans may vary slightly across agencies but the shared purpose of these documents is to facilitate good data management practices and to mandate open sharing of data to maximize scientific outputs and benefits to society. Along with this mandatory data sharing policy, comes the incentive to manage your data for the purposes of data sharing (Borghi and Van Gulick 2022).

2. **Journal compliance**: Depending on what journal you publish with, providing open access to the data associated with your publication may be a requirement (see PLOS ONE (https://journals.plos.org/plosone/) and AMPPS (https://www.psychologicalscience.org/publications/ampps) as examples). Again, along with data sharing, comes the incentive to manage your data in a thoughtful, responsible, and organized way.

3. **Compliance with mandates**: If you are required to submit your research project to the Institutional Review Board (see Chapter 10), the board will review and monitor your data management practices. Con-

cerned with the welfare, rights, and privacy of research participants, your IRB will have rules for how data is managed and stored securely (Filip, n.d.). Data sharing and other legal agreements with research partners will also need to be monitored and honored. Additionally, your organization may have their own institutional data policies that mandate how data must be cared for and secured.

4. **Open science practices**: With a growing interest in open science practices, sharing well-managed and documented data helps to build trust in the research process (Renbarger et al. 2022). Sharing data that is curated in a reproducible way is "a strong indicator to fellow researchers of rigor, trustworthiness, and transparency in scientific research" (Alston and Rick (2021), p.2). It also allows others to replicate and learn from your work, validate your results to strengthen evidence, as well as potentially catch errors in your work, preventing decisions being made based on incorrect data (Alston and Rick 2021). Sharing your data with sufficient documentation and standardized metadata can also lead to more collaboration and greater impact as collaborators are able to access and understand your data with ease (Borghi and Van Gulick 2022; Cowles n.d.; Eaker 2016).

5. **Data management is a matter of ethics**: In education research we are often collecting data from human participants, and as a result, data management is an ethical issue. It is important to consider the environmental, social, cultural, historical, and political context of the data we are collecting (Alexander 2023). When managing data we are often making human decisions about how to collect, organize, and clean data, and in this process we need to be aware of our biases that may affect equitable representation in our datasets. For instance, the order we choose to present categories of race on a survey, the way we choose to collect family relationships, or how we choose to collapse categories of gender during our data cleaning process are just a few examples of ways we may potentially bias our datasets based on our personal lenses (Mathematica n.d.). The process of de-identifying data for data sharing also becomes an ethical issue. Protection of participant identities is not always equally distributed across a dataset. While it may be easy to scrub identifying information for the majority of participants, individuals that represent smaller numbers in the data may still be identifiable and it is important to consider security implications for all participants (McKay Bowen and Snoke 2023). Last, collecting data from human participants means people are giving their time and energy and entrusting us with their information. Implementing poor data management that leads to irrelevant, unusable, or lost data is a huge disservice to research participants and erodes trust in the research process. It is our responsibility to have well-designed research studies with quality data management and data sharing practices that ensure that participant data remains secure, usable, and true so that their efforts lead to maximum societal benefits (Feeney, Kopper, and Sautmann 2022).

2.3.2 Personal reasons

Even if you never plan to share your data outside of your research group, there are still many compelling reasons to manage your data in a reproducible and standardized way.

1. **Reduces data curation debt**: Taking the time to plan and implement quality data management through the entire research study reduces data curation debt caused by suboptimal data management practices (Butters, Wilson, and Burton 2020). Having poorly collected, managed, or documented data may make your data unusable, either permanently or until errors are corrected. Decreasing or removing this debt reduces the time, energy, and resources spent possibly recollecting data or scrambling at the end of your study to get your data up to acceptable standards.

2. **Facilitates use of your data**: Every member of your research team being able to find and understand your project data and documentation is a huge benefit. It allows for the easy use and re-use of your data, and hastens efforts like the publication process (Markowetz 2015). Not having to search around for numbers of consented participants or asking which version of the data they should use allows your team to spend more time analyzing and less time playing detective.

3. **Encourages validation**: Implementing reproducible data management practices encourages and allows your team to internally replicate and validate your processes to ensure your outputs are accurate.

4. **Improves continuity**: Data management practices such as documentation ensures fidelity of implementation during your project. This includes implementing practices consistently during a longitudinal project, or consistently across sites. It also improves project continuity through staff turnover. Having thoroughly documented procedures allows new staff to pick up right where the former staff member left off and implement the project with fidelity (Borghi and Van Gulick 2021; Cowles n.d.). Furthermore, good data management enables continuity when handing off projects to collaborators or when picking up your own projects after a long hiatus (Markowetz 2015).

5. **Increases efficiency**: Documenting and automating data management tasks reduces duplication of efforts for repeating tasks, especially in longitudinal studies.

6. **Upholds research integrity**: Errors come in many forms, from both humans and technology(Kovacs, Hoekstra, and Aczel 2021; Strand n.d.). We've seen evidence of this in the papers cited as being retracted for "unreliable data" in the blog Retraction Watch (https://retractionwatch.com/). Implementing quality control procedures reduces the chances of errors occurring and allows you to have confidence in your data. Without implementing these practices, your research findings could include extra

noise, missing data, or erroneous or misleading results.

7. **Improves data security**: Quality data management practices reduce the risk of lost or stolen data, the risk of data becoming corrupted or inaccessible, and the risk of breaking confidentiality agreements.

2.4 Existing Frameworks

Data management does not live in a space all alone. It co-exists with other frameworks that impact how and why data is managed and it is important to be familiar with them as they will provide a foundation for you as you build your data management structures.

2.4.1 FAIR

In 2016, the FAIR Principles (GO FAIR n.d.) were published in Scientific Data, outlining four guiding principles for scientific data management and stewardship. These principles were created to improve and support the reuse of scholarly data, specifically the ability of machines to access and read data, and are the foundation for how all digital data should be publicly shared (Wilkinson et al. 2016). The principles are:

F: Findable

All data should be findable through a persistent identifier and have thorough, searchable metadata. These practices aid in the long-term discovery of information and provide registered citations.

A: Accessible

Users should be able to access your data. This can mean your data is available in a repository or through a request system. At minimum, a user should be able to access the metadata, even if the actual data are not available.

I: Interoperable

Your data and metadata should use standardized vocabularies as well as formats. Both humans and machines should be able to read and interpret your data. Software licenses should not pose a barrier to usage. Data should be available in open formats that can be accessed by any software (e.g., .csv, .txt, .dat).

R: Reusable

In order to provide context for the reuse of your data, your metadata should give insight into data provenance, providing a project description, an overview of the data workflow, as well what authors to cite for appropriate attribution. You should also have clear licensing for data use.

2.4.2 SEER

In addition to the FAIR principles, the SEER principles, developed in 2018 by Institute of Education Sciences (IES), provide Standards for Excellence in Education Research (Institute of Education Sciences n.d.c). While the principles broadly cover the entire life cycle of a research study, they provide context for good data management within an education research study. The SEER principles include:

- Pre-register studies
- Make findings, methods, and data open
- Identify interventions' core components
- Document treatment implementation and contrast
- Analyze interventions' costs
- Focus on meaningful outcomes
- Facilitate generalization of study findings
- Support scaling of promising results

2.4.3 Open Science

The concept of Open Science has pushed quality data management to the forefront, bringing visibility to its cause, as well as advances in practices and urgency to implement them. Open Science aims to make scientific research and dissemination accessible for all, making the need for good data management practices absolutely necessary. Open science advocates for transparent and reproducible practices through means such as open data, open analysis, open materials, preregistration, and open access (Dijk, Schatschneider, and Hart 2021). Organizations such as the Center for Open Science (https://www.cos.io), have become a well-known proponents of open science, offering the open science framework (OSF) (Foster and Deardorff 2017) as a tool to promote open science through the entire research life cycle. Furthermore, many education funders have aligned their fundee requirements with these open science practices, such as openly sharing study data and pre-registration of study methods (Institute of Education Sciences n.d.b).

2.5 Terminology

Before moving forward in this book it is important to have a shared understanding of terminology used. Many concepts in education research have synonymous terms that can be used interchangeably. Across different institutions, researchers may use all or some of these terms. Please review the 17 to gain a better understanding of how various terms will be used throughout this book.

2.6 The Research Life Cycle

The remainder of this book will be organized into chapters that dive into phases of the research data life cycle. It is imperative to understand this research life cycle in order to see the flow of data through a project, as well as to see how everything in a project is connected. If phases are skipped, the whole project will suffer.

You can see in Figure 2.1, how throughout the project, data management roles and project coordination roles work in parallel and collaboratively. These teams may be made up of the same people or different members, but either way, both workflows must happen and they must work together.

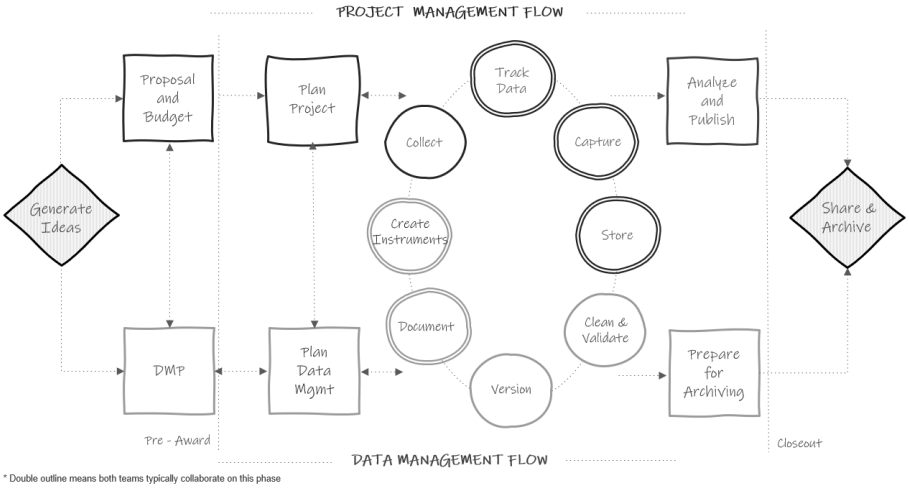


Figure 2.1: The research project life cycle

Let's walk through this chart.

1. In a typical study we first begin by **generating ideas**, deciding what we want to study.
2. Then, most likely, we will look for grant funding to implement that study. This is where the two paths begin to diverge. If the team is applying for federal funding, the proposal and budget are created in the project management track, while the supplemental required data management plan (see Chapter 4) is created in the data track. Again, it may be the same people working on both of these pieces.
3. Next, if the grant is awarded, the project team will begin planning things such as hiring, recruitment, data collection, and how to implement the intervention. At the same time, those working on the data team will begin to **plan** out how to specifically implement the 2-5 page data management plan submitted to their funder and start putting any necessary structures

into place.

4. Once planning is complete, the team moves into the cycle of data collection. It is called a cycle because if your study is longitudinal, every step here will occur cyclically. Once one phase of data collection wraps up, the team re-enters the cycle again for the next phase of data collection, until all data collection is complete for the entire project.
    - The data management and project management team begin the cycle by starting **documentation**. You can see that this phase occurs collaboratively because it is denoted with a double outline. Both teams begin developing documentation such as data dictionaries and standard operating procedures.
    - Once documentation is started, both teams collaboratively begin to create any necessary **data collection instruments**. These instruments will be created with input from the documentation. During this phase the teams may also develop their participant tracking database.
    - Next, the project management team moves into the **data collection** phase. In addition to actual data collection, this may also involve preliminary activities such as recruitment and consenting of participants, as well as hiring and training of data collectors. At this point, the data management team just provides support as needed.
    - As data is collected, the project team will **track data** as it is collected in the participant tracking database. The data management team will collaborate with the project management team to help troubleshoot anything related to the actual tracking database or any issues discovered with the data during tracking.
    - Next, once data is collected, the teams move into the **data capture** phase. This is where teams are actively retrieving or converting data. For electronic data this may look like downloading data from a platform or having data sent to the team via a secure transfer. For physical data, this may look like teams entering paper data into a database. Oftentimes, this again is a collaborative effort between the project management team and the data team.
    - Once the data is captured, it needs to be **stored**. While the data team may be in charge of setting up and monitoring the storage efforts, the project team may be the ones actively retrieving and storing the data.
    - Next the teams move into the **cleaning and validation** phase. At this time the data team is reviewing data cleaning plans, writing data cleaning scripts, and actively cleaning data from the most recent data collection round.
    - And last, the data team will **version** data as it is updated or errors are found.
5. The teams then only move out of the active data collection phase when all data collection for the project is complete. At this time the project team begins analyzing study data and working on publications as well as

any final grant reports. They are able to do this because of the organized processes implemented during the data collection cycle. Since data was managed and cleaned throughout, data is ready for analysis as soon as data collection is complete. Then, while the project team is analyzing data, the data team is doing any additional **preparation to archive** data for public sharing.

6. Last, as the grant is closing out, the team submits data for **public sharing**.

As you work through the remaining chapters of this book, this chart will be a guide to navigating where each phase of practices fits into the larger picture.

Chapter 3

Data Structure

Because data management is made up of just that, data, we need to have a basic understanding of what data looks like. Understanding the basic structure of data helps us write our Data Management Plan, organize our data management process, create our data dictionaries, build our data collection tools, and clean our data, all in ways that allow us to have analyzable data.

3.1 Basics of a dataset

In education research, data is often collected internally by your team using an instrument such as a questionnaire, an observation, an interview, or an assessment. However, data may also be collected from external entities, such as districts, states, or other agencies.

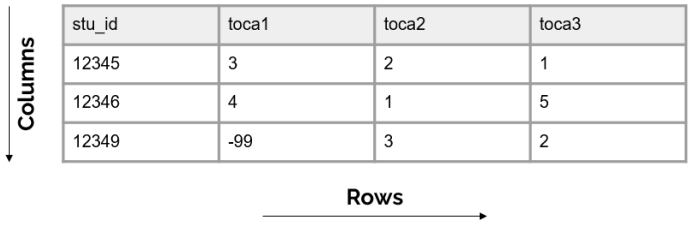
Those data come in many forms (e.g., video, transcripts, documents, files), represented as text, numbers, or multimedia (USGS n.d.b). In the world of quantitative education research, we are often working with digital data in the form of a dataset, a structured collection of data. These datasets are organized in a rectangular format which allow the data to be machine-readable. Even in qualitative research, we are often wrangling data to be in a format that is analyzable and allows categorization.

These rectangular (also called tabular) datasets are made up of columns and rows.

3.1.1 Columns

The columns in your dataset will consist of one or both of the following types of variables:

- Variables you collect (from an instrument or from an external source)



stu_id	toca1	toca2	toca3
12345	3	2	1
12346	4	1	5
12349	-99	3	2

Figure 3.1: Basic format of a dataset

- Variables you create/add (e.g., cohort, intervention, time, derivations)

Unless your data is collected anonymously, every dataset **must** also have the following:

- One or more variables that are **unique identifiers**, sometimes called primary keys. These are variables that uniquely define rows in your dataset (i.e. help you identify duplicate rows), and they also allow you to link data that contain the same identifiers (for example link all student data). It is important to make sure these variables are consistently formatted across files (e.g., always character variables).
- If you plan to link datasets across entities (e.g., link teachers to schools or students to teachers) then you will also need secondary unique identifiers in your dataset (also called foreign keys) that allow you to link across datasets.

We will talk more about creating these identification variables in Chapter 9.

Column attributes

It is important to know that variables have the following attributes:

1. Unique names (no variable name in a dataset can repeat). We will talk more about variable naming when we discuss style guides (see Chapter 8).
2. A measurement type (e.g., numeric, character, date) which can also be more narrowly defined as needed (e.g., continuous, categorical)
3. Acceptable values (e.g., yes/no) or expected ranges (e.g., 1-25 or 2021-08-01 to 2021-12-15). Anything outside of those acceptable values or ranges is considered an error.
4. Labels, descriptions of what the variable represents. This may be a label that you as the variable creator assigns (e.g., "Treatment condition") or they may be the actual wording of an item (e.g., "Do you enjoy pizza?").

3.1.2 Rows

The rows in your dataset are aligned with participants or cases in your data. Participants in your data may be students, teachers, schools, locations, and so

forth. The unique identifier variable mentioned above will denote which row belongs to which participant.

3.1.3 Cells

The cells are the observations associated with each participant. Cells are made up of key/value pairs, created at the intersection of a column and a row. Consider an example where we collect a survey from students. In this dataset, each row is made up of a unique student in our study, each column is an item from the survey, and each cell contains a value/observation that corresponds to that row/column pair (that participant and that question).

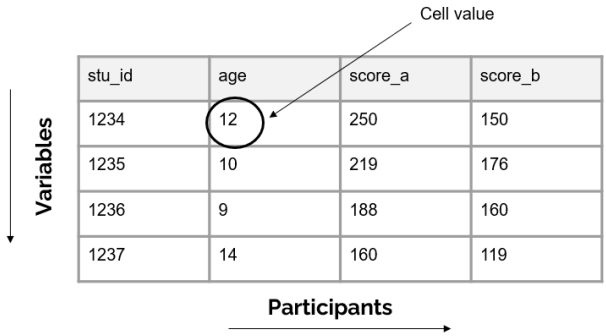


Figure 3.2: Representation of a cell value

3.2 Dataset organization rules

In order for your dataset to be machine-readable and analyzable, it should adhere to a set of structural rules (Broman and Woo 2018; Wickham 2014).

1. The first rule is that your data should make a rectangle. The first row of your data should be your variable names (only use one row for this). The remaining data should be made up of values in cells.

2. Your columns should adhere to your variable type.
   - For example, if you have a numeric variable, such as age, but you add a cell value that is text, your variable no longer adheres to your variable type. Machines will now read this variable as text.

3. A variable should only collect one piece of information. If a variable contains more than one piece of information you may have the following issues:
   - You lose the granularity of the information (e.g., `location` = `Los Angeles, CA` is less granular than having a `city` variable and a `state` variable separately)

not a rectangle

	1234	1235	1236	1237
age	12	10	9	14
	1234	1235	1236	1237
score_a	250	219	188	160
	1234	1235	1236	1237
score_b	150	176	158	119

rectangle

stu_id	age	score_a	score_b
1234	12	250	150
1235	10	219	176
1236	9	188	160
1237	14	160	119

Figure 3.3: A comparison of non-rectangular and rectangular data

text variable

tch_id	age	
12345	22	
12346	24	Space before 24 makes this entry text
12349	49 years old	Text added makes this entry text
12350	36..0	Double decimal point makes this entry text

numeric variable

tch_id	age
12345	22
12346	24
12349	46
12350	36

Figure 3.4: A comparison of variables adhering and not adhering to a data type

- Your variable may become unanalyzable (e.g., a variable with a value `220/335` is not analyzable as a numeric variable). If you are interested in a rate, you can calculate a `rate` variable with a value of `.657`.
- You may lose the variable type (e.g., if you want an `incident_rate` variable to be numeric, and you assign a value of `220/335`, that variable is no longer numeric)

two things in one variable

sch_id	level	incident_rate
235	elementary	55/250
236	elementary	72/303
237	middle	140/410
238	high	219/552

two things in two variables

sch_id	level	incident	enrollment
235	elementary	55	250
236	elementary	72	303
237	middle	140	410
238	high	219	552

Figure 3.5: A comparison of two things being measured in one variable and two things being measured across two variables

4. All cell values should be explicit. This means all cells should be filled in with a physical value.
   - Consider why a cell value is empty
     - If a value is actually missing, you can either leave those cells as blank or fill them with your pre-determined missing values (e.g., -99). See Chapter 8 for ideas.
     - If a cell is left empty because it is "implied" to be the same value as above, the cells should be filled with the actual data
     - If the value for the cell is "implied" to be 0, fill the cells with 0
   - No values should be implied using color coding
     - If you want to indicate information, add an indicator variable to do this rather than cell coloring
5. Your data should not contain duplicate rows. You do not want duplicate rows of a measurement collected **on the same participant**, **at the same time period**. Different types of duplicate rows can occur:
   - A true duplicate row where an entire row is duplicated (the row values are the same for every variable). This may happen if someone enters the same form twice.
   - A unique identifier is duplicated but the row values may or may not be the same across all of the variables. This could happen because one of three reasons:
     1. An instrument is accidentally collected more than once on the same participant in a collection period. This type of duplicate

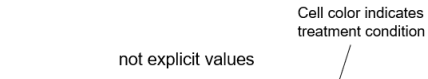
not explicit values

sch_id	year	grade	n_students
204	2020	3	100
		4	80
		5	90
205	2020	3	98
		4	88
		5	91

explicit values

sch_id	year	grade	n_students
204	2020	3	100
204	2020	4	80
204	2020	5	90
205	2020	3	98
205	2020	4	88
205	2020	5	91

Figure 3.6: A comparison of of variables with empty cells and variables with not empty cells



not explicit values

stu_id	date	test_score
12345	2022-04-13	35
12346	2022-04-12	42
12349	2022-04-13	50
12350	2022-04-11	19

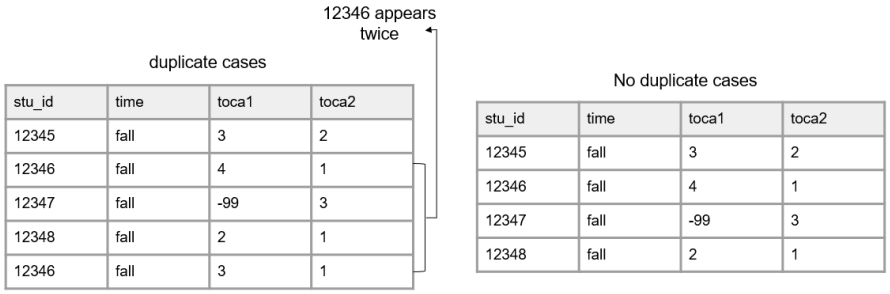
explicit values

stu_id	date	test_score	treatment
12345	2022-04-13	35	0
12346	2022-04-12	42	1
12349	2022-04-13	50	1
12350	2022-04-11	19	0

Figure 3.7: A comparison of variables with implicit values and variables with explicit values

would need to be remedied.

2. A unique identifier was entered incorrectly. In this case you don't actually have a duplicate, you just have an incorrect unique identifier. This error would need to be remedied.
3. More than one variable is used to identify unique participants and the row is not actually a duplicate.
   - Take for example a student id and a class id. Multiple unique identifiers may be used if data is collected on participants in multiple locations and treated as unique data. In this case, the data is not truly duplicate because the combined identifiers are unique.
   - Another example of this is if your data is organized in long format (discussed in Section 3.3.2). In this case unique study identifiers may repeat in the data but they should not repeat for the same form and same time period in your data.



stu_id	time	toca1	toca2
12345	fall	3	2
12346	fall	4	1
12347	fall	-99	3
12348	fall	2	1
12346	fall	3	1

stu_id	time	toca1	toca2
12345	fall	3	2
12346	fall	4	1
12347	fall	-99	3
12348	fall	2	1

Figure 3.8: A comparison of data with duplicate cases and data with no duplicate cases

3.3 Linking data

Up until now we have been talking about one, standalone dataset. However, it is more likely that your research project will be made up of multiple datasets, collected from different participants, from a variety of instruments, and possibly across different time points. And at some point you will most likely need to link those datasets together.

In order to think about how to link data, we need to discuss two things: data structure and database design.

3.3.1 Database design

A database is "an organized collection of data stored as multiple datasets" (USGS n.d.b). Sometimes this database is actually housed in a database soft-

ware system (such as SQLite or FileMaker), and other times we are loosely using the term database to simply define how we are linking disparate datasets together that are stored individually in some file system. No matter the storage system, the general concepts here will be applicable.

In database terminology, each dataset we have is considered a "table". Each table has a primary key that identifies unique entries within a table and each table can be connected through both primary and foreign keys. This linking of tables creates a relational database and we will talk more about this structure when we discuss participant data tracking (see Chapter 9).

Let's take the simplest example, where we only have primary keys in our data. Here we collected two pieces of data from students (a survey and an assessment) in one time period. Figure 3.9 shows what variables were collected from each instrument and how each table can be linked together through a primary key (circled in yellow).

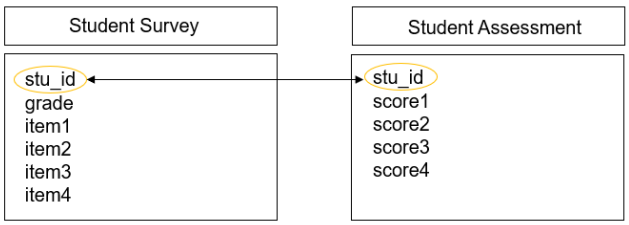


Figure 3.9: Linking data through primary keys

However, we are often not only collecting data across different forms, but we are also collecting nested data across different participants (e.g., students, nested in classrooms, nested in schools, and so on). Let's take another example where we collected data from three instruments, a student assessment, a teacher survey, and a school intake form. Figure 3.10 shows what variables exist in each dataset (with primary keys still being circled in yellow) and how each table can be linked together through a foreign key (circled in blue).

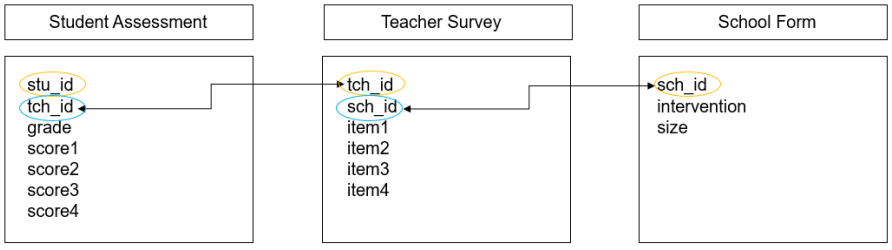


Figure 3.10: Linking data through foreign keys

And as you can imagine, as we add more forms, or begin to collect data across

time, the database structure begins to become even more complex. Figure 3.11 is another example where we collected two forms from students (a survey and an assessment), two forms from teachers (a survey and an observation), and one form from schools (an intake form). While the linking structure begins to look more complex, we see that we can still link all of our data through primary and foreign keys. Forms within participants can be linked by primary keys, and forms across participants can be linked by foreign keys.

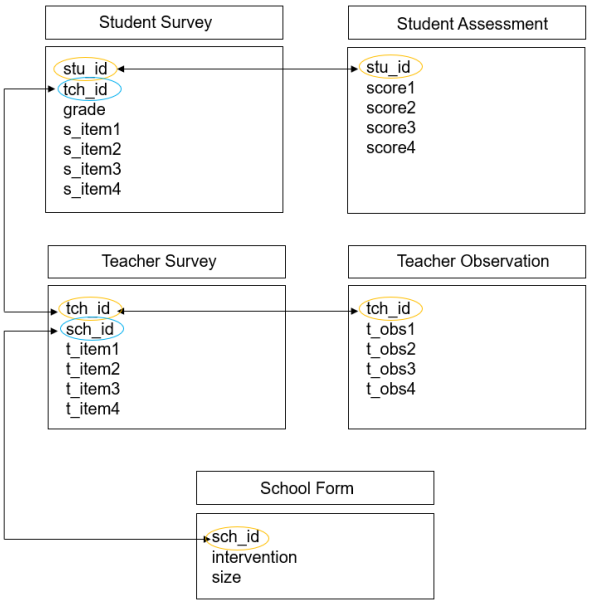


Figure 3.11: Linking data through primary and foreign keys

3.3.2 Data structure

When it comes time to link our data, there are two ways we often think about linking or structuring our data, wide or long.

3.3.2.1 Wide format

When we structure our data in a wide format, all data collected on a unique participant will be in one row. Participants should **not** be duplicated in your data in this format.

This type of format can be used for the following situations:

- To link forms within time
- To link forms across time

- To link forms across participants

The easiest scenario to think about this format is with repeated measure data. If we collect a survey on participants in both wave 1 and 2, those waves of data will all be in the same row (joined together on a unique ID) and each wave of data collection will be appended to a variable name to create unique variable names. We will dive deeper into different types of joins in Chapter 13.

> **Note** It is important to note here, that if your data do not have unique identifiers (primary and/or foreign keys), as is in the case of anonymous data, you will be unable to merge data in a wide format.

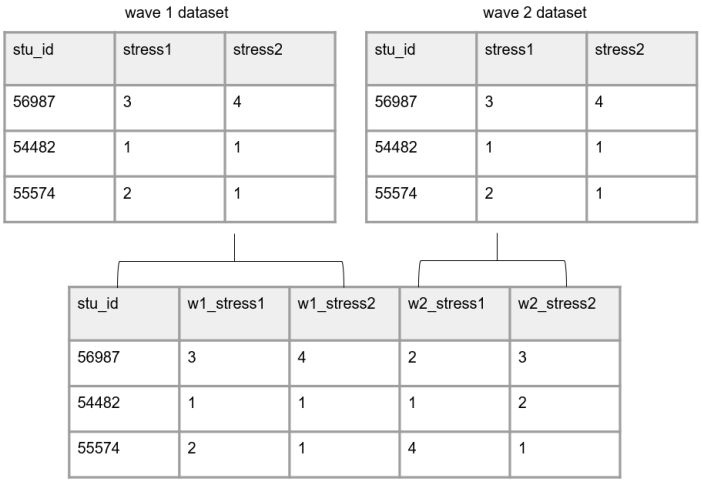


Figure 3.12: Data structured in wide format

3.3.2.2 Long format

In education research, long data is mostly used as a specific way to structure data that is collected over time. In long data a participant can and will repeat in your dataset.

Again, the most straight forward way to think about this is with repeated measure data, where each row will be a new time point for a participant. Here instead of merging forms on a unique id, we stack forms on top of each other, often called appending data. Rows are stacked on top of one another and variables are aligned by variable name. Now instead of linking data by an id, data is now "linked" by variable names. It is important here that variable names and types stay identical over time in order for this structure to work.

In this scenario, we no longer add the data collection wave to variable names. However, we would need to add a time period variable to denote the wave associated with each row of data.

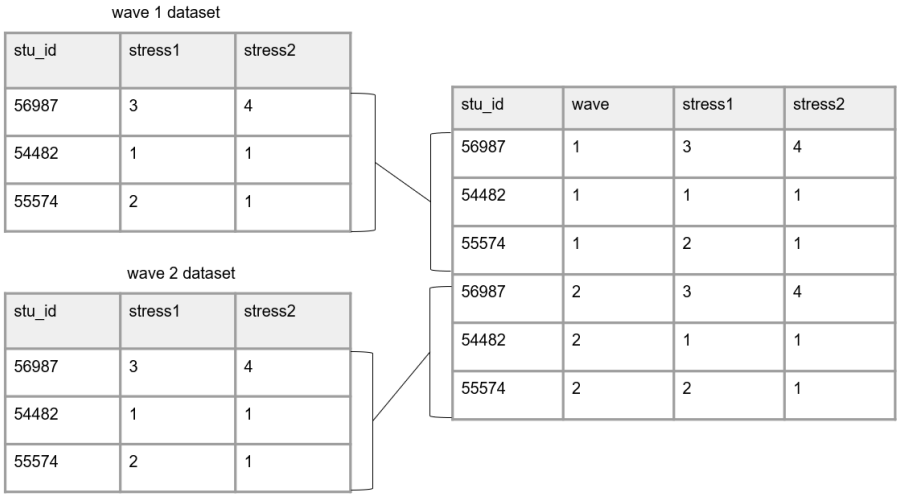


Figure 3.13: Data structured in long format

3.3.2.3 Choosing wide vs long

There are different reasons for constructing your data one way or another. And it may be that you store or share your data in one format, and then restructure data into another format when it comes time for analysis.

Storing data in long format is usually considered to be more efficient, potentially requiring less memory. However, when it comes time for analysis, specific data structures may be required. For example, repeated measure procedures typically require data to be in wide format, where the unit of analysis is the subject. While mixed model procedures typically required data to be in long format, where the unit of analysis is each measurement for the subject (Grace-Martin 2013). We will further review decision making around data structure in Chapter 13.

3.4 File types

These rectangular datasets can be saved in a variety of file types. Some common file types in education research include interoperable formats such as .csv, .txt, .dat, or .tsv, or proprietary formats such as .xlsx, .sav, or .dta.

When you save your files, they will have a file size. Both the number of columns as well as the number of rows in your dataset will contribute to your file size. Just to get a feel for what size your files might be, small datasets (for example 5 columns and <100 rows) may be less than 100 KB. Datasets with several hundred variables and several thousand cases may start to be in the 1,000-5,000 KB range. The type of file you use also changes the size of your data. Saving data in a format that contains embedded metadata (such as variable and value

labels), such as a .sav file, will greatly increase your file size. We will talk about the pros and cons to different file formats in Chapter 14.

Chapter 4

Data Management Plan

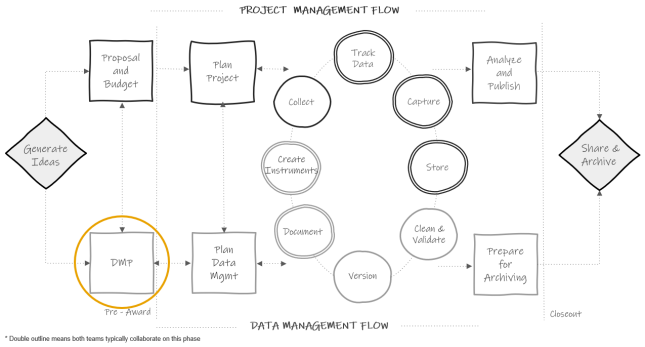


Figure 4.1: Data management plan in the research project life cycle

4.1 History and purpose

Since 2013, even earlier for the National Science Foundation, most federal agencies that education researchers work with have required a data management plan (DMP) or data management and sharing plan (DMSPs) as part of their funding application (Holdren 2013). While the focus of these plans is mostly on the future outcome of data sharing, the data management plan is a means of ensuring that researchers will thoughtfully plan for a research study that will result in data that can be shared with confidence, and free from errors, uncertainty, or violations of confidentiality. President Obama's May 2013 Executive Order declared that "the default state of new and modernized government information resources shall be open and machine readable" (The White House 2013). In August of 2022, the Office of Science and Technology Policy (OSTP) doubled down on their data sharing policy and issued a memorandum stating

that all federal agencies must update their public access policies no later than December 31, 2025, to make federally funded publications and their supporting data accessible to the public with no embargo on their release (Nelson 2022). Even sooner than this, organizations like the National Institutes of Health mandated that grant applicants, beginning January 2023, must submit a plan for both managing and sharing project data(National Institutes of Health n.d.c). The National Science Foundation (NSF) also released version 2.0 of their public access plan in February of 2023, describing how agency plans to ensure that all scientific data associated with peer-reviewed publications that was funded by NSF is publicly shared (National Science Foundation 2023).

4.1.1 Why are DMPs important?

Funding agencies see DMPs as important in maximizing scientific outputs from investments and increasing transparency. Mandating data sharing for federally funded projects leads to many benefits including accelerating discovery, greater collaboration, and building trust among data creators and users. In addition to the benefits viewed by funders, there are intrinsic benefits that come from having to write a data management plan. Having to thoughtfully plan and having transparency in that plan leads to better data management. Knowing that you will eventually be sharing your data and documentation with others outside of your team can motivate researchers to think hard about how to organize their data management practices in a way that will produce data that they trust to share with the outside world (Center for Open Science n.d.).

4.2 What is it?

Generally, a data management plan is a supplemental 2-5 page document, submitted with your grant application, that contains details about how you plan to store, manage, and share your research data products. For most funders these DMPs are not part of the scoring process, but they are reviewed by a panel or program officer. Some funders may provide feedback or ask for revisions if they believe your plan and/or your budget and associated costs are not adequate.

4.2.1 What to include?

What to include in a DMP varies some across funding agencies and the landscape of requirements is currently evolving. You should check each funding agency's site for their specific DMP requirements when submitting a proposal. With that said there are typically 10 common categories covered in a data management plan (Center for Open Science n.d.; Gonzales, Carson, and Holmes 2022; ICPSR n.d.a; Michener 2015) which we will review below.

1. Roles and responsibilities
    - What are the staff roles in management and long-term preservation of data?

- Who ensures accessibility, reliability, and quality of data?
- Is there a plan if a core team member leaves the project or institution?

2. Types and amount of data
   - How is data captured? (e.g., surveys, assessments, observations)
   - Will data be item-level? Summary scores? Metadata only?
     - Datasets from a project may need to be shared in different ways due to legal, ethical, or technical reasons.
   - Will you share raw data and clean data?
   - What are the expected number of files? Expected number of rows/cases in each file?
3. Format of data
   - Will data be in an electronic format?
   - Will it be provided in a non-proprietary format? (e.g., .csv)
   - Will more than one format be provided? (e.g., .sav and .csv)
   - Are there any tools needed to manipulate shared data?
4. Documentation
   - What documentation will you share? (Consider project-level, dataset-level, and variable-level documentation)
   - What metadata will you create?
   - What format will your documentation be in? (e.g., .xml, .csv, .pdf)
   - What supplemental documents do you plan to include when sharing data? (e.g., consort diagrams, data collection instruments, consent forms)
5. Standards
   - Do you plan to use any standards for things such as metadata, data formatting, terminology, or persistent identifiers (PIDs)?
6. Method of data sharing
   - How will you share your data? (e.g., Institutional archive, data repository, PI website)
   - Will data be restricted and is a data enclave required?
   - Is a data use agreement required?
   - How will you license your data?
   - Will your data have persistent unique identifiers?
7. Circumstances preventing data sharing
   - Do you have any data covered by FERPA/HIPAA that doesn't allow data sharing?
   - Do you work with any partners that do not allow you to share data? (e.g., School districts, tribal regulations)
   - Are you working with proprietary data?
8. Privacy and rights of participants
   - How will you prevent disclosure of personally identifiable information when you share data? How will you anonymize data (if applicable)?
   - Do participants sign informed consent agreements? Does the consent communicate how participant data are expected to be used and shared?
9. Data security

    - How will you maintain participant privacy and confidentiality during your project?
    - How will you prevent unauthorized access of data?
    - Consider IRB requirements here.
10. Schedule for data sharing
    - When will you share your study data and for how long?
11. Pre-registration (less commonly required)
    - Where and when will you pre-register your study?

Again, the specifics of what should be included in each category will vary by funder. Here are sites to visit to learn more about the four most common federal education research funder DMP requirements.

- Institute of Education Sciences [1]
- National Institutes of Health [2]
- National Institute of Justice [3]
- National Science Foundation [4]

4.3 Getting help

Since DMPs are written before a project is funded, and therefore before additional staff members may be hired, oftentimes the investigators developing the grant proposal are the ones who write the DMP. However, when constructing your DMP it is well worth your time to enlist help. If you have an existing data manager or data team, you will most certainly want to consult with them when writing your plan to ensure your decisions are feasible. If you work for a university system, your research data librarians are also excellent resources with a wealth of knowledge about writing comprehensive data management plans. And last, if you plan to share your final data with a repository or institutional archive you will want to contact your repository when writing your plan as well. The repository may have its own requirements for how and when data must be shared and it is helpful to outline those guidelines in your data management plan at the time of submission. You can also specifically write the name of your repository into your data management plan as well. Last, you may want to obtain the help of your colleagues. Your colleagues have likely written DMPs before and many people are willing to share their plans as a way to help others better understand what to include.

Your DMP is a living document and you can always update your plan during or after your project completion. It may be helpful to keep in contact with your program officer regarding any potential changes throughout your project.

[1] https://ies.ed.gov/funding/datasharing_implementation.asp

[2] https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-DMS/writing-a-data-management-and-sharing-plan

[3] https://nij.ojp.gov/funding/data-archiving

[4] https://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf

If you are looking for guidance in writing a DMP, a variety of generic DMP templates for different federal agencies are available, as well as actual copies of submitted DMPs that some researchers graciously make publicly available for example purposes.

Template and Resources

Source	Resource
DMPTool	Templates organized by funding agencies [5]
Figshare	DMP prompts specific to depositing data with Figshare [6]
Hao Ye, et al.	NIH DMS Plan checklist [7]
Harvard Longwood Medical Area RDM Working Group	Annotated DMP template [8]
ICPSR	NIH DMS Plan template with specific recommendations for depositing data with ICPSR [9]
NIH	Sample DMS Plan for human survey data [10]
Sara Hart	A submitted DMP that is publicly available for example purposes [11]
UMN Libraries	Submitted DMP examples from University of Minnesota researchers [12]

4.4 Budgeting

As briefly mention above, funding agencies acknowledge that there are costs associated with implementing your data management plan and allow you to explain these costs in your budget narrative. Costs associated with the entire data life cycle should be considered and may include costs associated with data management personnel, specialized infrastructure, tools needed to collect, enter, organize, document, store, or share study data (UK Data Service 2022), as well as fees associated with data preservation. Make sure to review your funder's documentation for information about allowable costs(Samuel J. Wood Library

[5] https://dmptool.org/public_templates

[6] https://help.figshare.com/article/how-to-write-a-data-management-plan-dmp-and-include-figshare-in-your-data-sharing-plans

[7] https://osf.io/awypt/

[8] https://osf.io/ztjf2

[9] https://www.icpsr.umich.edu/files/ICPSR/nih/FINAL_ICPSR-NIH-DMS-Plan-Template_2023.docx

[10] https://www.nichd.nih.gov/sites/default/files/inline-files/Example_DMS_Plan-Human-Survey-NIH_Format_Page_V2.pdf

[11] https://figshare.com/articles/preprint/Example_of_a_Data_Management_Plan/13218743

[12] https://www.lib.umn.edu/services/data/dmp-examples

n.d.) and time frame for incurring costs. Examples of potential allowable costs include (National Institutes of Health n.d.a):

- Costs associated with curating and de-identifying data
- Costs associated with developing data documentation
- Fees associated with depositing data for long-term sharing in a repository

It can be difficult to estimate the costs of everything that is associated with the vast landscape of managing data. Luckily a few organizations have developed resources to aid in estimating those costs.

Resources

Source	Resource
UK Data Service	Data management costing tool and checklist [13]
University of Twente	Estimating RDM costs review list [14]
Utrecht University	Estimating the costs of data management review list [15]
DataOne	Considerations for providing budget information for a DMP [16]
J-PAL	Research proposal budget considerations [17]

[13] https://ukdataservice.ac.uk//app/uploads/costingtool.pdf
[14] https://www.utwente.nl/en/service-portal/services/lisa/resources/files/library-public/dcc-rdm-costs-estimation.pdf
[15] https://www.utwente.nl/en/service-portal/services/lisa/resources/files/library-public/dcc-rdm-costs-estimation.pdf
[16] https://dataoneorg.github.io/Education/bestpractices/provide-budget-information
[17] https://www.povertyactionlab.org/resource/grant-proposals

Chapter 5

Planning Data Management

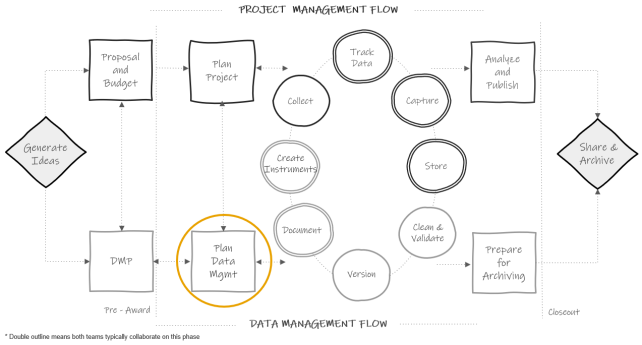


Figure 5.1: Planning in the research project life cycle

Planning data management is distinct from the 2-5 page data management plan (DMP) discussed in the Chapter 4. Here we are spending a few weeks, maybe months, meeting regularly with our team and gathering information to develop detailed instructions for how we plan to manage data according to our DMP. This data management planning happens at the same time that the project team is planning for project implementation (things like how to collect data, how to hire staff, what supplies are needed, how to recruit participants, how to communicate with sites, etc). Team members such as PIs, project coordinators, and data managers, may be assisting in both planning processes.

5.1 Why spend time on planning?

Funder required data management plans are hopeful outlines for future practices. However, the broad theory behind our DMPs do not actually prepare

us for the complex implementation of those plans in practice (Borycz 2021). Therefore, it is important to spend time, before your project begins, planning and preparing for data management. It is an upfront time investment but this sort of slow science leads to better data outcomes. Reproducibility begins in the planning phase. Taking time to create, document, and train staff on data management standards before your project begins helps to ensure that your processes are implemented with fidelity and can be replicated consistently throughout the entire study.

Planning the day to day management of your project data has many other benefits as well. It allows you to anticipate and overcome barriers to managing your data, such as communication issues, training needs, or potential tool issues. This type of planning also saves you time in the long run, removing the last minute scrambling that can occur when trying to organize your data at the end of a project. Last, this type of planning can mitigate errors. Viewing errors as problems created by poorly planned workflows, rather than individual failures, helps us to see how data management planning can lead to better data (Strand n.d.). While data management planning can not remove all chances of errors creeping into your data (Eaker 2016), it can most certainly reduce those errors and prevent them from "compounding over time" (Alston and Rick (2021), p.4).

5.2 Goals of planning

This planning phase should include a series of regular meetings with core decision makers. During this data management planning time there are several goals to keep in mind.

1. Flesh out project goals laid out in a grant proposal (i.e. what data needs to be collected to answer our research questions)
2. Finalize a timeline for goals (i.e. when will data be collected)
3. Lay out specific tasks needed to accomplish goals and come to a consensus regarding all necessary data management decisions (i.e. how will data be collected, stored, managed, and shared)
4. Assign roles and responsibilities (i.e. who will be responsible for tasks)
5. Make decisions around task management and communication (i.e. how will tasks be monitored and communication tracked)

Make sure to come to every meeting with an agenda to stay on track and to take detailed notes. These notes will be the basis for creating all of your documentation in the next phase. All meeting notes can be stored in a central location such as a planning folder with notes ordered by date or in a running document.

At the end of the planning period, the team should have a clear plan for what the project goals are, when goals should be accomplished, how goals will be accomplished, who is in charge of completing tasks associated with goals, and

what additional resources are needed to accomplish goals.

5.3 Planning checklists

Along with your existing data management plan, checklists are great tools to help guide your discussions as you work through this planning process with your team. Below are some sample checklists, one for each phase of the research cycle. These checklists can be added to or amended and brought to your planning meetings to help your team think through the various data management decisions that need to be made at each phase of your research project.

Planning checklists

- Roles and Responsibilities [1]
- Task Management [2]
- Documentation [3]
- Data Collection [4]
- Data Tracking [5]
- Data Capture [6]
- Data Storage and Security [7]
- Data Cleaning [8]
- Data Sharing [9]

5.3.1 Decision-making process

As you move through the remaining chapters of this book, you will begin to learn recommended practices for each phase of the research cycle. Going through each checklist above, you can start to fill in the practices that work for your project for each phase of the study.

This decision-making process is personalized. Borghi and Van Gulick (Borghi and Van Gulick 2022) view this process as a series of steps that a research team chooses, out of a the many possibilities not chosen. Maybe you won't always be able to implement the "best practices" but you can decide what is good enough for your team based on motivations, incentives, needs, resources, skill set, and rules and regulations.

For example, one team may collect survey data on paper because their participants are young children, hand enter it into Excel because that is the only tool

[1] https://docs.google.com/document/d/1o__QsM9N492XgMhRE4ef9GaGVNzvfO4sR
[2] https://docs.google.com/document/d/131cHp9-__NET3futvKH7ECV39rTSTEULE
[3] https://docs.google.com/document/d/1M372uOtVutLxt7VZgCZnxPVDUSTQmm15
[4] https://docs.google.com/document/d/1nvjMHeDmJkQtTT4CoLpUcroYknSDAQyj
[5] https://docs.google.com/document/d/1YM3q0aNEpQAalorr3fs4dXH2aCuompNk
[6] https://docs.google.com/document/d/18fL9M4TKi0k6cC2ubK0VD5Res9yXsS92
[7] https://docs.google.com/document/d/1mxxGaDvFPlQaR7M3wSmwWTgkHa5yiT4t
[8] https://docs.google.com/document/d/12Jx4soafWiZF-1y-ESu1n37aDa-Pa4ZS
[9] https://docs.google.com/document/d/1Bsbjx9aCZlsr8XbLRp3llhJxDkIi56iD

they have access to, and double enter 20% because they don't have the capacity to enter more than that. Another team may collect paper data because they are collecting data in the field, hand enter the data into FileMaker because that is the tool their team is familiar with, and double enter 100% because they have the budget and capacity to do that.

Figure 5.2 is a very simplified example of the decision making process, based on the (Borghi and Van Gulick 2022) flow chart. Of course in real life we are often choosing between many more than just two options!

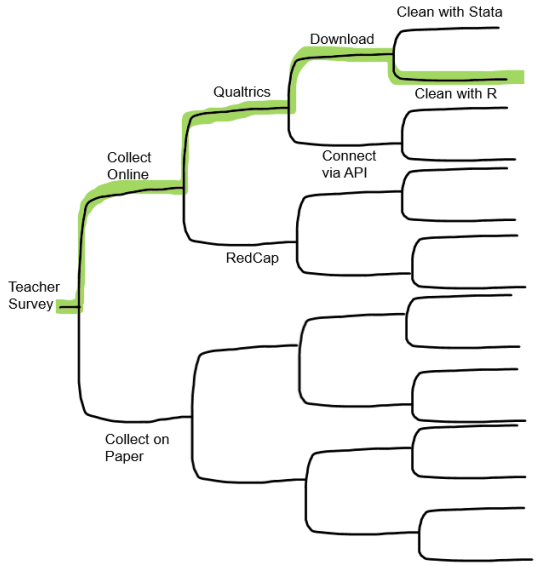


Figure 5.2: A simplified decision-making process

5.3.2 Checklist considerations

It's important to consider how each team and project are unique as you work through these planning checklists. A technique that might work well for one team, may not work out so well for another. Make sure to consider the following:

1. All external requirements
    - Do your practices align with the plan laid out in your DMP? If no, you may need to revise your DMP to match your new decisions - remember your DMP is a living document.
    - Do your practices meet all other external compliance requirements such as those from your Institutional Review Board, your institutional policies, project partner requirements, or government mandates?
2. The skill set of your team

- How does the skill set of your team align with the practices you plan to implement? Will additional training be required?

3. Your available tools
    - What tools are available to your team?
    - Does your organization only allow you to use certain platforms for data storage?
    - What is the complexity of your tools? Will additional training be needed?
4. Your budget
    - Do you have the budget to implement all of the practices you want to implement or will you need to plan something more feasible?
5. Complexity of your project
    - The size of your project, the amount and types of data you are collecting, the number of participants or the populations you are collecting data from, the sensitivity level of the data you are collecting, the number of sites you are collecting data at, and the number of partners and decision makers you are working with, all factor into your data management planning
6. Shared investment
    - Is your entire team invested in quality data management?
    - Is the entire team motivated to adhere to the standards and instructions laid out in your data management planning? If no, what safeguards can you implement to help prevent errors from creeping into your data?

5.4 Data management workflow

The last step of this planning phase is to build your workflows. Workflows allow data management to be seamlessly integrated into your data collection process. Often illustrated with a flow diagram, a workflow is a series of repeatable tasks that help you move through the stages of the research life cycle in an "organized and efficient manner" (CSP Library Research n.d.). As you walk through your checklists, you can begin to enter your decisions into a workflow diagram that show actionable steps in your data management process. The order of your steps should follow the general order of the data management life cycle (specifically the data collection cycle). You will want to have a workflow diagram for every piece of data that you collect. So for example, if you collect the following three items below, you will have three workflow diagrams.

- Student online survey
- Student paper assessment
- Student district-level administrative data

Your diagrams should include the who, what, where, and when of each task in the process. Adding these details are what make the process actionable (Borycz 2021). Your diagram can be displayed in any format that works for you and

it can be as simple or as detailed as you want it to be. A template like the one in Figure 5.3 works very well for thinking through high level workflows. Remember, this is a repeatable process. So while this diagram is linear (steps laid out in the chronological order in which we expect them to happen), this process will be repeated every time we collect this same piece of data.

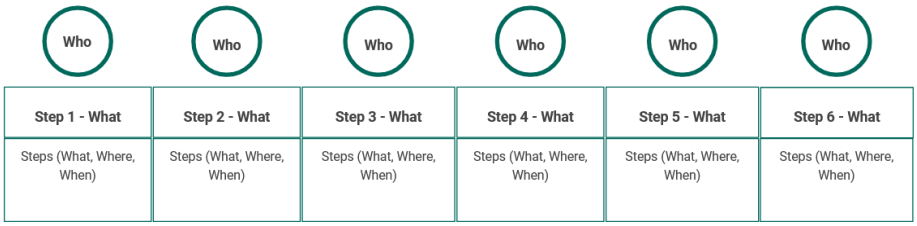


Figure 5.3: A simple workflow template

Here is how we might complete this diagram for a student survey.

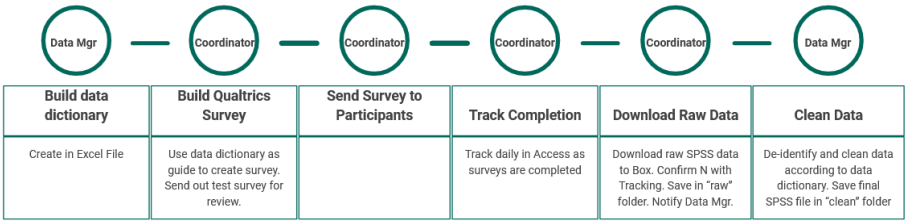


Figure 5.4: Example student survey workflow

But the format truly does not matter. Figure 5.5 is a diagram of the same student survey workflow as above, with more detailed added, and this time using a swimlane template instead, where each lane displays the tasks associated with that individual and the iterative processes that occur within and across lanes.

If you have a working data collection timeline (see Chapter 7) already created, you can even build time into your workflow. Figure 5.6 is another example of the same survey workflow again, this time displayed using a Gantt chart (Duru and Kopper, n.d.) in order to better capture the expected timeline.

While these workflow diagrams are excellent for high level views of what the process will be, we can see that we are unable to put fine details into this visual. So the last step of creating a workflow is to put all tasks (and all final decisions associated with those tasks) into a standard operating procedure (SOP). In your SOP you will add all necessary details of the process. You can also attach your diagram as an addendum or link your SOPs and diagrams in other ways for reference. We will talk more about creating SOPs in Chapter 7.

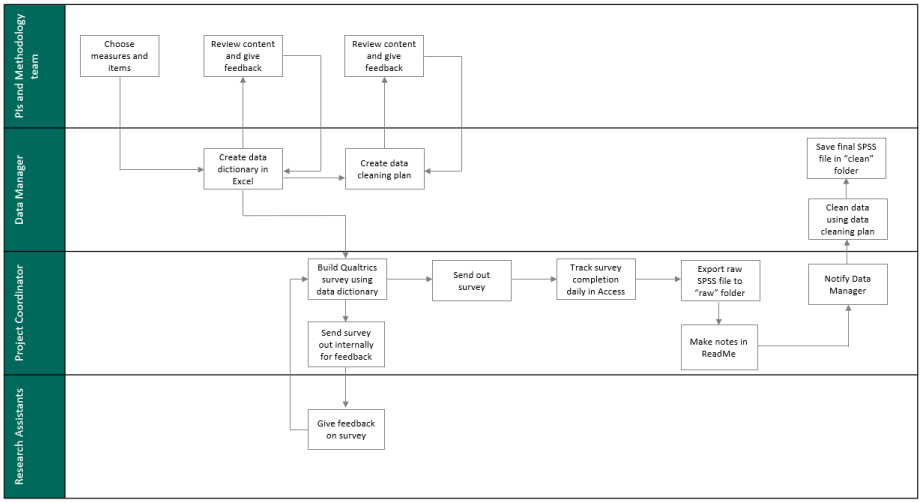


Figure 5.5: Example student survey workflow using a swimlane template

Year 1													
Activity	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July	Responsible
Choose measures													PI and Methodology team
Build data dictionary in Excel													Data Manager
Get feedback on data dictionary and make edits													PI and Methodology Team, Data Manager
Create data cleaning plan													Data Manager
Build Qualtrics survey using data dictionary as guide													Project Coordinator
Review survey and make edits													Research Assistants, Project Coordinator
Collect data with survey													Project Coordinator
Track daily completion in Access													Project Coordinator
Download raw SPSS file and save to Box in "raw" folder													Project Coordinator
De-identify and clean data according to data dictionary													Data Manager
Save clean data file in "clean" folder													Data Manager

Figure 5.6: Example student survey workflow using a Gantt chart

5.4.1 Benefits to visualizing a workflow

Visualizing your decisions in diagram format has many benefits. First, it allows your team to conceptualize their specific tasks in the process, the timing at which their tasks occur, and any dependencies associated with those tasks. It also allows your team to see how their roles and responsibilities fit into the larger research process (Briney, Coates, and Goben 2020). Showing how data management is integrated into the larger research workflow can help team members view data management as part of their daily routine, rather than "extra work" (Borghi and Van Gulick 2022). And last, reviewing workflows as a team and allowing members to provide feedback may help create buy-in for data management processes, potentially leading to better adherence to practices.

5.4.2 Workflow considerations

Similar to the questions you need to consider when reviewing your planning checklists, you also need to evaluate the following things when developing your personalized workflow (Hansen 2017).

- Does your flow preserve the integrity of your data? Is there any point where you might lose or comprise data?
- Is there any point in the flow where data is not being handled securely? Someone gains access to identifiable information that should not have access?
- Is your flow in accordance with all of your compliance requirements (IRB, FERPA, HIPAA, Institutional Data Policies, etc.)?
- Is your flow feasible for your team (based on size, skill level, motivation, etc.)?
- Is your flow feasible for your budget and available resources?
- Is your flow feasible for the amount and types of data you are collecting?
- Are there any bottlenecks in the workflow? Areas where resources or training are needed? Any areas where tasks should be re-directed?

5.5 Task management systems

While tools such as our checklists, workflow diagrams, and SOPs allow us to document and share our processes, it can be tricky to manage the day to day implementation of those processes. The planning phase is a great time to choose a task management system (Gentzkow and Shapiro 2014). Keeping track of various deadlines and communications across scattered sources can be overwhelming and using a task management system may help remove ambiguity about the status of task progress. Rather than having to regularly check in via email for status updates or reading through various meeting notes to learn about decisions made, a task management system allows you to assign tasks to responsible parties, set deadlines based on timelines, track progress, and capture communication and decisions all in one location.

There are many existing tools that allow teams to assign and track tasks, schedule meetings, track project timelines, and document communication. Without endorsing any particular product, some project/task management tools that I know education research teams have used include:

- Trello
- Smartsheet
- Todoist
- Microsoft Planner
- Notion
- Basecamp
- Confluence
- Asana

Of course, as with all processes we've discussed so far, a task management system is only useful if your team is trained to use it, is invested in using it, and actually uses it as part of their daily routine. So make sure to consider this as you choose what tool, if any, is right for you.

Chapter 6

Project Roles and Responsibilities

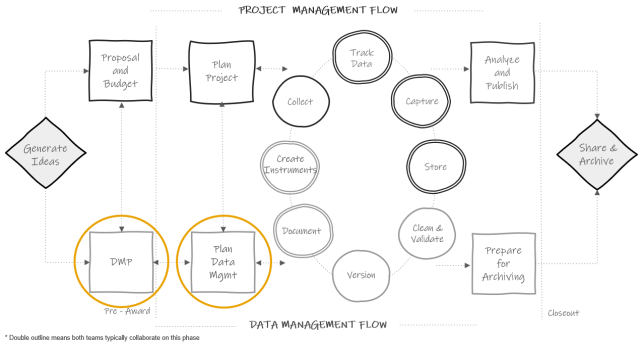


Figure 6.1: Planning in the research project life cycle

Part of the DMP and planning data management phase, as noted in previous chapters, will include assigning roles and responsibilities. In terms of data management, it is important to assign and document roles, not just presume roles, for many reasons including the following (UK Data Service n.d.c):

1. It allows team members to begin standardizing workflows
2. When team members know exactly what is expected of them, it keeps data more secure
3. Creating contingency plans for when staff can no longer fulfill their roles allows for the continuity of practices

6.1 Typical roles in a research project

Before diving in to how to assign and document roles for a project, it is important to get an understanding of typical roles on an education research project team. Your team may be lucky enough to have all of (or multiple of) these roles. Other times, just one person, such as the Principal Investigator (PI), may take on all or multiple of these roles. With that said, if your budget allows it, I highly recommend hiring individuals to fill each of the roles mentioned below to allow team members to specialize and excel in their area of expertise. While learning all aspects of a project is highly recommended to create a cohesive team that works collaboratively, team members that take on too many project roles can be spread too thin and project goals may suffer.

6.1.1 PI and Co-PI

The PIs (or project directors), as well as Co-PIs, are the individuals who prepare and submit the grant proposal and are responsible for the administration of that grant. There are often more than one PI on a project including at least someone with content area knowledge as well as a methodologist. PIs and Co-PIs have varying levels of involvement in research projects and are typically, not always, more hands off in the day to day administration. Even if some tasks are delegated to other research staff, PIs and Co-PIs are ultimately responsible for Institutional Review Board (IRB) submissions and for meeting IRB requirements, as well as for submitting MOUs, budgets, effort reporting, continuing review reports, and any final technical finding reports.

6.1.2 Project Coordinator

The project coordinator (or project manager) is an essential member of the research team. As the name implies, this person typically coordinates all research activities and ensures compliance with agencies such as the Institutional Review Board. Tasks they may oversee include recruitment and consenting of participants, creation of data collection materials, creation of protocols, training data collectors, data collection scheduling, and more. The project coordinator may also supervise many of the other research team roles, such as research assistants.

6.1.3 Data Manager

The data manager is also an essential member of the team. This person is responsible for the organizing, cleaning, documenting, storing, and dissemination of research project data. This team member works very closely with the project coordinator, as well as the PI, to ensure that data management is considered throughout the project life cycle. Tasks a data manager may oversee include data storage, security and access, building data collection and tracking tools, cleaning and validating data, data documentation, and organizing data for sharing purposes.

This role is vital in maintaining the standardization of data practices. If you do not have the budget to hire a full-time data manager, make sure to assign someone on your team to oversee the flow of data, ensuring that throughout the project, data is documented, collected, entered, cleaned, and stored consistently and securely.

6.1.4 Project Team Members

This role refers to any staff hired to help implement a research project which may include full-time staff members, with titles such a research or project assistants for instance, or it may include part-time graduate students. Project team members are typically out in the field, collecting data, or they may also assist in other areas such as preparing data collection materials or assisting with data management. Senior project team members may also assist in implementing training or acting as data collection leads in the field.

6.1.5 Other Roles

The size of a research team and the roles that exist are dependent on factors such as funding, the type of research study, the intervention being studied, or the organization of your specific research institution. Some teams may include additional roles, not mentioned above, such as research director, lab manager, software engineer, database manager, postdoc, analyst, statistician, administrative professional, hourly data collector, outreach coordinator, or coach/interventionist, all who may assist in the research cycle in other ways. Some of these roles will assist in the research data life cycle as seen in the diagram above. Some may be on a path that is hidden from the diagram but still happening, behind the scenes, alongside the process. Take for instance, the role of a coach implementing an intervention that is being studied. Their tasks aren't shown on the original diagram but their work is happening alongside the data collection cycle.

6.2 Assigning roles and responsibilities

Early on in a project you may start to generally assign roles in your data management plan. Remember if you submitted a DMP, you are often required to state who will be responsible for activities such as data integrity and security. Then, once your project is funded and you start to have a better idea of your goals and your budget, you can flesh out the details of your roles. During the planning phase, using tools such as your planning checklists will help you think through more specific responsibilities and tasks associated with each role. When assigning roles and responsibilities, there are several factors to consider (Valentine n.d.).

1. Required skillset

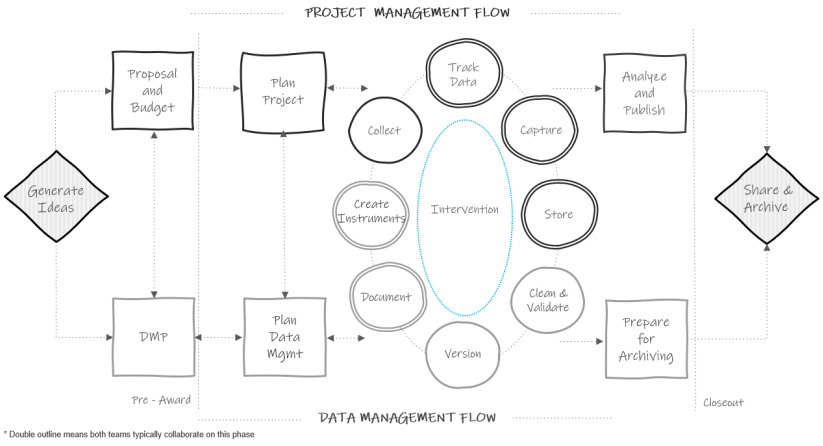


Figure 6.2: Life cycle diagram updated to show hidden processes

In assigning roles and responsibilities, make sure to consider the skills that are needed to be successful in each position. For example, when considering the role of a data manager and the responsibilities associated with that role, you may look for skill sets in the following buckets:

- Interpersonal skills (Detail-oriented, organized, good communicator)
- Domain skills (Experience working with education data, understands data privacy - FERPA, HIPAA)
- Technical skills (Understanding of database structure, experience building data pipelines, coding experience, specific software/tool experience)

The specific skills needed for each role will depend on your project needs as well as the skill sets of the other members of the team.

2. Training needs

In addition to considering skills needed for certain roles, also consider what training is needed to fulfill assigned responsibilities. In roles that work with data, training may include mandated courses from a program like the Collaborative Institutional Training Initiative (CITI) or it may be signing up for training on how to use a specific tool or software. Make sure that your team members are well-equipped to perform their responsibilities before the project begins.

3. Estimated costs

If you are working on roles and responsibilities after your grant has been funded, then your grant budget has already submitted. However, it can still be helpful to thinking through costs associated with overall roles (based on the experience/skillset of the person filling the role) or even broken down by associated

responsibilities (based on things like percent effort or time to complete each task). If discrepancies between the original budget and updated costs are found, often funders will allow PIs to amend budgets.

4. Assess equity in responsibilities

Review how responsibilities are allocated. Consider both the time needed to complete tasks and number of responsibilities assigned to each team member. Make sure you are not overloading any one team member, and reassign tasks as needed.

5. Contingency plans

You should also begin thinking through backup plans should a staff member leave the project or be absent for an extended period of time. This may include cross training staff or a plan for training replacement staff.

6.3 Documenting roles and responsibilities

After assigning roles and responsibilities, those decisions should be documented to avoid any ambiguity about who is doing what. While documentation is a topic that will be covered in the next chapter, I think it is helpful to break the rules and discuss just this one document here while we are covering the topic of assigning roles.

There are many reasons to document staff roles and responsibilities and to store that information in a central, accessible location.

1. It allows your team to easily reference the document to see who is on the project team, what roles they play, and who to contact for questions regarding various project aspects (e.g., who to contact for data storage access).
2. As new tasks arise, team members can refer to the document to see who is best fitted for the assignment.

3. Last, reviewing roles and responsibilities in a document also helps you more clearly see what responsibilities are assigned and how they are assigned. After reviewing the document you can make further revisions if responsibilities need to be added or further redistributed in any way.

This document can be laid out in any format that conveys the information clearly to your team. Figure 6.3 and Figure 6.4 are two example templates. Note that these templates only list overarching responsibilities, not specific steps associated with tasks. Specific actionable steps will be laid out in other process documentation such as standard operating procedures (see Chapter 7) where names are attached to each task.

Since there is no one template for creating a roles and responsibility document,

Project Name:
Date:

The purpose of this document is to clearly articulate the different roles within a project team and the duties each role/person is responsible for.

Title	Role	Name
Project Coordinator	Oversee the completion of research study objectives and research compliance	(Name of Individual)
Responsibilities		
• Hire and train part-time data collectors • Recruit and consent study schools and teachers • Consent study students • Build data collection tools • Organize data collection efforts • Document data collection efforts		

Title	Role	Name
Data Manager	Oversee the design of data collection tools, data documentation as well as the security, management and integrity of study data	(Name of Individual)
Responsibilities		
• Monitor data training compliance for all staff • Build data collection tools • Build data tracking database • Document data management efforts • Clean study data • Oversee data access • Oversee data storage		

Figure 6.3: Roles and responsibilities document organized by role

Project Name:
Date:

The purpose of this document is to clearly articulate the different roles within a project team and the duties each role/person is responsible for.

Phase	Project Coordinator [Name]	Data Manager [Name]	Research Assistant [Name]
Documentation	• Create documentation	• Create documentation	
Create Instruments	• Build data collection instruments • Order supplies	• Build data tracking database	• Test data collection tools and provide feedback
Data Collection	• Hire data collectors • Train data collectors • Schedule data collection	• Oversee integrity of data	• Collect data
Data Capture	• Oversee entry of data	• Build data entry database	• Enter data
Data Storage	• Ensure raw electronic data is stored correctly • Ensure paper data in the field is handled securely • Ensure paper data is stored securely in the office	• Manage data access • Oversee data storage backup and security	

Figure 6.4: Roles and responsibilities document organized by phase

you can really add whatever information helps to most clearly convey the assignments. Some additional columns you may consider adding include:

- Links to related standard operating procedures (e.g., for building a participant tracking database you may link to the specific SOP that lays out steps for building the tool)
- Names of other staff members (if any) that assist with or also contribute to each responsibility
- Timing of each responsibility (e.g., weekly, ongoing, the month of February)

6.4 Data management role

Like I mentioned earlier, I highly recommend hiring a full-time data manager if you are able to budget for this as it allows each team member to have more narrow responsibilities and to implement their tasks with better precision. However, not everyone will have the capacity to do this. If so, it will be vitally important to still assign those data management responsibilities to specific team members. In choosing who to assign these tasks to, you will want to consider several things such as appropriate skill set to manage the data, interest in data management tasks, and time to commit to data management. Oftentimes this responsibility falls to a full-time project coordinator as they are the ones who are intimately familiar with the data, and since they are full-time, they are able to carve out hours for data management tasks. Other times it may be a collaboration between a project coordinator and another staff member, such as a part-time graduate student (who may have more technical skills in terms of data wrangling). No matter who you assign these roles to, just ensure that they are documented and the information is disseminated to the team.

Chapter 7

Documentation

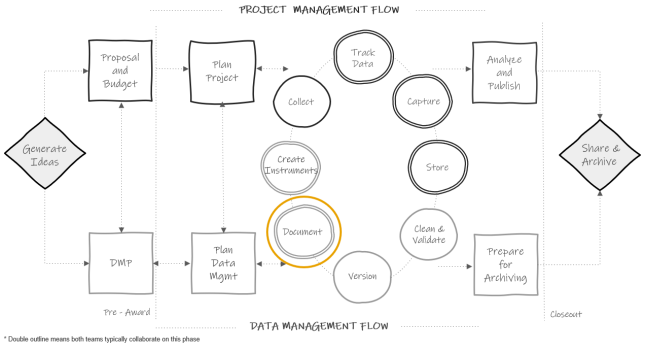


Figure 7.1: Data documentation in the research project life cycle

Documentation is a collection of files that contain procedural and descriptive information about your team, your project, your workflows, and your data. Creating thorough documentation during your study is equally as important as collecting your data. Documentation serves many purposes including:

- Standardizing procedures
- Securing data and protecting confidentiality
- Tracking data provenance
- Discovering errors
- Enabling reproducibility
- Ensuring others use and interpret data accurately
- Providing searchability through metadata

We are going to cover four levels of documentation in this chapter: team-level, project-level, dataset-level, and variable-level. While most of the documentation discussed does fall within its eponymous phase in the research life cycle, some

documents will be created earlier or later and the timing will be discussed in each section. During a project, while you are actively using your documents, the format of these documents does not matter. Choose a human-readable format that works well for your team (e.g., Word, PDF, plain text file, Google Doc, Excel, HTML, OneNote, etc.). When projects are closing out and you are preparing to share your data, you can consider, at that time, how to best make your documents more sustainable, interoperable, and searchable. See Chapter 14 for more information.

The documents below are all recommended and will help you successfully run your project. You can create as many or as few of these documents as you wish. The documents you choose to produce should be based on what is best for your project and your team, as well as what is required by your funder (see Chapter 4) and other governing bodies such as your Institutional Review Board. No matter which documents you choose to implement, it is important to create templates for your documents and implement them consistently within, and even across projects. Implementing documentation using templates, or consistent formats and fields, reduces duplication in efforts (no need to reinvent the wheel) and allows your team to interpret the document more easily. These documents are best created by the team member that directly oversees the process and sometimes that may include a collaborative effort (for example both a project coordinator and a data manager may build documents together).

Each type of documentation discussed below is a living document to be updated as procedures change or new information is received. As seen in the cyclical section of Figure 7.1 above, team members should revisit documentation each time new data is collected, or more often if needed, to ensure documentation still aligns with actual practices. If changes are made and not added to documentation over long periods of time, you will find that you no longer remember what happened and that information will be lost. It will also be important to version your documents along the way so that staff know that they are working with the most recent version and can see when documents have been updated and why.

> **Note** Creating and maintaining these documents **is an investment**. Make sure to account for this time and expertise in your proposal budget (see Chapter 4). With that said, the return for the investment is well worth the effort.

7.1 Team Level

Team-level data management documentation typically contains data governance rules that apply to the entire team, across all projects. While these documents can be amended any time, they should be started long before you apply for a grant, when your lab, center, or institution is formed.

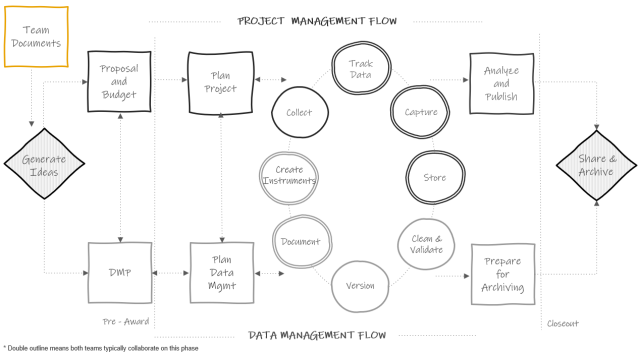


Figure 7.2: Team-level documentation in the research project life cycle

7.1.1 Lab manual

One example of a team-level document is a lab manual, or team handbook. A lab manual creates common knowledge across your team (Mehr n.d.). It provides staff with consistent information about how the team works and why they do the things they do. It also sets expectations, provides guidelines, and can even be a place for passing along career advice (Aczel n.d.; The Turing Way Community 2022). While a lab manual will primarily consist of administrative, procedural, and interpersonal types of information, it can be helpful to include data management content, including general rules about accessing, storing, sharing, and working with data securely and ethically.

Template and Resources

Source	Resource
Balazs Aczel, et al.	Crowdsourced lab manual template [1]
Hao Ye, et al.	Crowdsourced list of public lab manuals [2]
Samuel Mehr	Common Topics in Lab Handbooks [3]

7.1.2 Wiki

A wiki is a webpage that allows users to collaboratively edit and manage content. It can either be created alongside the lab manual or as an alternative to the lab manual is a team wiki. Wikis can be built and housed in many tools such as SharePoint, Teams, Notion, GitHub or Open Science Foundation (OSF). While some lab wikis are public (as you'll see in the examples below), most are not and

[1] https://docs.google.com/document/d/1LqGdtHg0dMbj9lsCnC1QOoWzIsnSNRTSek6i3Kls2Ik

[2] https://docs.google.com/spreadsheets/d/1kn4A0nR4loUOSDn9Qysd3MqFJ9cGU91dCDM6x9aga-8

[3] https://www.rsb.org.uk/images/biologist/2020/Apr__May__2020_67-2/67.2_Handbook.pdf

can be restricted to invited users only. Wikis are a great way to keep disparate documents and pieces of information, for both administrative and data related purposes, organized in a central, accessible location. Your wiki can include links to important documents, or you can also add text directly to the wiki to describe certain procedures. Rather than sending team members to multiple different folders for frequently requested information, you can refer them to your one wiki page.

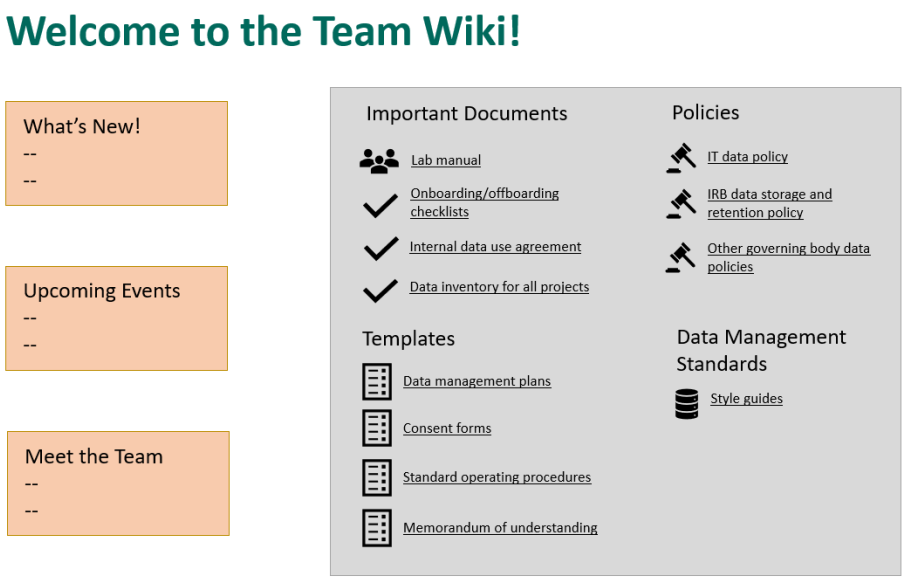


Figure 7.3: Example team wiki with links to frequently requested information

> **Note** Project-level wikis can also be created and be very useful in centralizing frequently referenced information pertaining to specific projects.

Templates and Resources

Source	Resource
Aly Lab	Example public lab wiki [4]
Notion	Company home wiki template [5]
SYNC Lab	Example public lab wiki [6]

[4] https://osf.io/mdh87/wiki/home/
[5] https://www.notion.so/Company-home-240047f7526c4b0091681dc6c95b7e76
[6] https://eur-synclab.github.io/

7.1.3 Onboarding/Offboarding

While **onboarding** checklists will mostly consist of non-data related, administrative information such as how to sign up for an email or how to get set up on your laptop, it should also contain several data specific pieces of information to get all new staff generally acclimated to working with data in their new role.

Similarly, while **offboarding** checklists will contain a lot of procedural information about returning equipment and handing off tasks, it should also contain information specific to data management and documentation that help maintain data integrity and security.

Data related topics to consider adding to your onboarding and offboarding checklists are included in Figure 7.4.

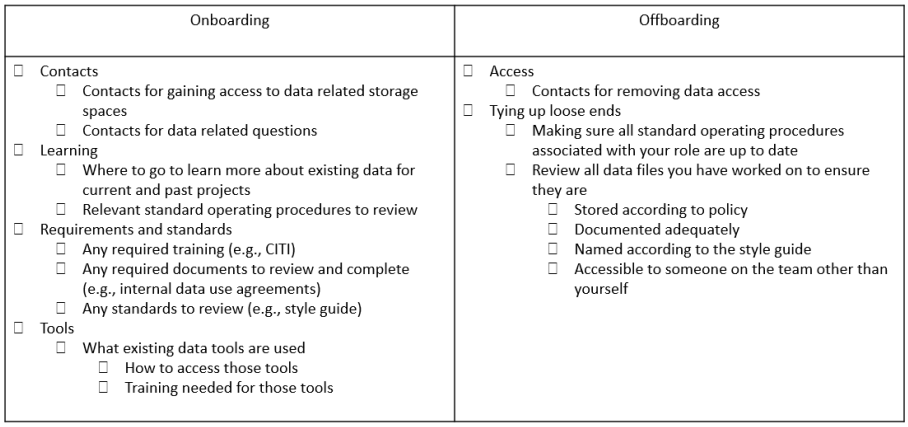


Figure 7.4: Sample data topics to add to onboarding and offboarding checklists

Template and Resources

Source	Resource
Crystal Lewis	Sample data topics to add to an onboarding checklist [7]
Crystal Lewis	Sample data topics to add to an offboarding checklist [8]

7.1.4 Data Inventory

A data inventory maps all data sources collected by the research team (Salfen 2018). As a team grows, the number of data sources typically expands as well.

[7] https://docs.google.com/document/d/1xyU5Q0uUD-PqRKRmMJKpD9lKaGQI6pjs
[8] https://docs.google.com/document/d/1W57cYuYyiqltQNXUITP-jVIf84jao4Ef

It can be very helpful to keep an inventory of what data sources are available for team members to use, as well as details about those data sources.

- Project associated with each data source
- Dates that each data sources was collected
- Storage location of each data source
- Details about each dataset (what the dataset contains, how it is organized, what questions can be answered with the data)
- How data sources are related

7.1.5 Data use agreement

Typically when we think of a data use agreement (DUA) we think of a contractual document typically drafted in conjunction with an external partner to facilitate data sharing. It usually covers the terms for how someone is allowed to use data, considering things like access controls, research participant privacy, and data destruction rules (Geraghty and Feeney n.d.).

However, internally, DUAs can also be very useful in developing a cohesive understanding among team members regarding the terms and conditions of project data use (CESSDA n.d.a). Rules for securely working with data can be added to a lab manual, as many people do, or they can be added to a separate data use agreement where staff members can sign (Filip, n.d.) or check a box acknowledging that they have read and understand the policies.

Ideas of content to include in a DUA are included in Figure 7.5.

Template and Resources

Source	Resource
Crystal Lewis	Example of content to include in an internal data use agreement [9]

7.1.6 Style guide

A style guide is a set of standards for the formatting of information ("Style Guide" 2023). It improves consistency and a shared understanding within and across files and projects. This document includes conventions for procedures such as variable naming, variable value coding, file naming, versioning, file structure, and even coding practices. It can be created in one large document or separate files for each type of procedure. I highly recommend applying your style guide consistently across all projects, hence why this is included in the team documentation. Since style guides are so important, and there are so many recommended practices to cover, I have given this document its own chapter. See Chapter 8 for more information.

[9] https://docs.google.com/document/d/1fCFBULZeCBRyt0v2k4-Jb_9zBrk9En29

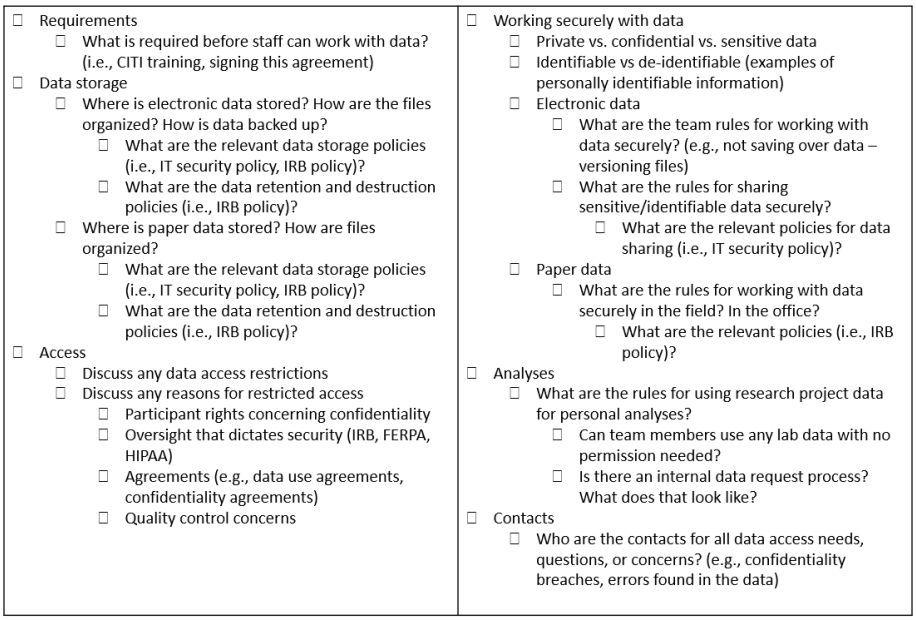


Figure 7.5: Example of content to include in an internal data use agreement

Template and Resources

Source	Resource
Hadley Wickham	Example R coding style guide [10]
Strategic Data Project	Example style guide [11]

7.2 Project level

Project-level documentation is where all descriptive information about your project is contained, as well as any planning decisions and process documentation specifically related to your project. Again, while most of these documents are created in the documentation phase, some documents such as the data management plan (started before your project is funded), checklists and meeting notes (started during the planning phase), or a participant flow diagram (started after data is collected) will begin at other points throughout the cycle.

[10] https://style.tidyverse.org/
[11] https://hwpi.harvard.edu/files/sdp/files/sdp-toolkit-coding-style-guide.pdf

7.2.1 Data management plan

As discussed in Chapter 4, if your project is federally funded it is likely that a data management plan was required. This project-level document is created in the DMP phase, long before a project begins. However, your DMP can continue to be modified throughout your entire study. If any major changes are made, it may be helpful to reach out to your program officer to keep them in the loop as well.

7.2.2 Checklists and meeting notes

Checklists, as discussed in Chapter 5, are documents that are created (or copied from existing sources) and reviewed during the planning phase. Using checklists facilitates discussion and allows your team to build a cohesive understanding for how data will be managed throughout your entire project. As you work through the checklists, all decisions made should be documented in meeting notes. After the planning phase is complete, decisions should be formally documented in applicable team, project, data, or variable-level documents (e.g. research protocol, SOPs, style guide, or roles and responsibilities documents). Even beyond the planning phase though, all meeting decisions and discussions should continue to be documented in meeting notes and used to update formal documentation as needed.

7.2.3 Roles and responsibilities document

Using the checklists reviewed during the planning phase, your team should begin assigning roles and responsibilities for your project. In the planning and documentation phase, those designations should be formally documented and shared with the team. In Chapter 6 we reviewed ways to structure this document. Once this document is created, make sure to store it in a central location for easy referral and update the document as needed.

Templates and Resources

Source	Resource
Crystal Lewis	Three roles and responsibilities templates [12]

7.2.4 Research protocol

The research protocol is a comprehensive project plan document that describes the what, who, when, where, and how of your study. Many of the decisions made in your data management plan and while reviewing your planning checklists will be summarized in this document. If you are submitting your study to your Institutional Review Board, you will most likely be required to submit this

[12] https://drive.google.com/drive/folders/1nhDgOVfESrZLYfvcrTU_I2dnsOtq3TkV

document as part of your application. A research protocol assists the board in determining if your methods provide adequate protection for human subjects. In addition to serving this required purpose, the research protocol is also an excellent document to share along with your data at the time of data sharing, and an excellent resource for you when writing technical reports or manuscripts. This document provides all context needed for you and others to effectively interpret and use your data. Make sure to follow your university's specific template if provided, but common items typically included in a protocol are provided in Figure 7.6.

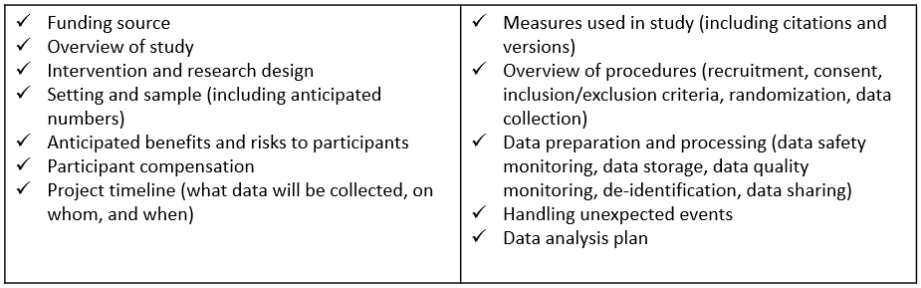


Figure 7.6: Common research protocol elements

When it comes time to deposit your data in a repository, the protocol can be revised to contain information helpful for a data end user, such as known limitations in the data. Content such as risks and benefits to participants might be removed, and numbers such as study sample count should be updated to show your final numbers. Additional supplemental information can also be added as needed.

Template and Resources

Source	Resource
Crystal Lewis	A template to create a project-level summary document for data sharing (based on an IRB research protocol) [13]
Jeffrey Shero, Sara Hart	IRB protocol template with a focus on data sharing [14]
The Ohio State University	Protocol template [15]
University of Missouri	Protocol template [16]
University of Washington	Protocol checklist [17]

[13] https://docs.google.com/document/d/1wOLFFurs0t2rANxyD6rQ7xoFbg5LPmeA
[14] https://figshare.com/articles/preprint/IRB_Protocol_Template/13218797
[15] http://orrp.osu.edu/files/2011/10/GuidelinesforWritingaResearchProtocol.pdf
[16] https://docs.research.missouri.edu/human_subjects/templates/Social_Behavioral_Educational_Protocol_Template.docx
[17] https://depts.washington.edu/wildfire/resources/protckl.pdf

7.2.5 Supplemental documents

There is a series of documents, that while they can absolutely be standalone documents, I am calling supplemental documents here because they can be added to your research protocol as an addendum at any point to further clarify specifics of your project.

1. Timeline

The first supplemental document that I highly recommend creating is a visual representation of your data collection timeline. When you first create these timelines they will be based on your best estimates of the time it will take to complete milestones, but like all documents we've discussed, they can be updated as you learn more about the reality of the workload. This document can be both a helpful planning tool (for both project and data teams) in preparing for times of heavier and lighter workloads, as well as an excellent document to share with future data users to better understand waves of data collection. There is no one format for how to create this document. Figure 7.7 is an example of one way to visualize a data collection timeline.

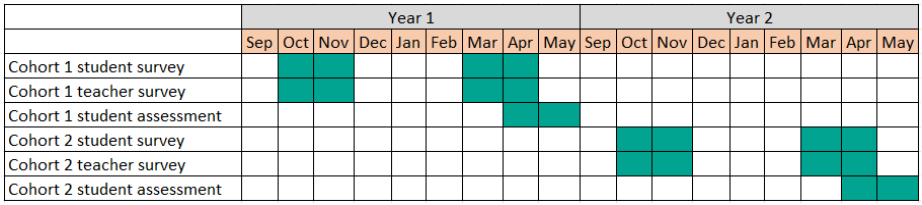


Figure 7.7: Example data collection timeline

2. Participant flow diagram

A participant flow diagram displays the movement of participants through a study, assisting researchers in better understanding milestones such as eligibility, enrollment, and final sample counts. These diagrams are helpful for assessing study attrition and reasons for missing data can be described in the diagram (Nahmias et al. 2022). In randomized controlled trial studies, these visualizations are more formally referred to as CONSORT (Consolidated Standards of Reporting Trials) diagrams, (Schulz et al. 2010) as seen in Figure 7.8 ("CONSORT 2010 Flow Diagram" n.d.). They provide a means to understand how participants are randomized and assigned to treatment groups. As you can imagine though, this diagram cannot be created until at least one wave of data has been collected, and must be updated as more waves are collected. Your participant tracking database, which we will discuss in Chapter 9, will inform the creation of this diagram.

3. Instruments

Actual copies of instruments can be included as supplemental documentation. This includes copies of surveys, assessments, forms, and so forth. It can also

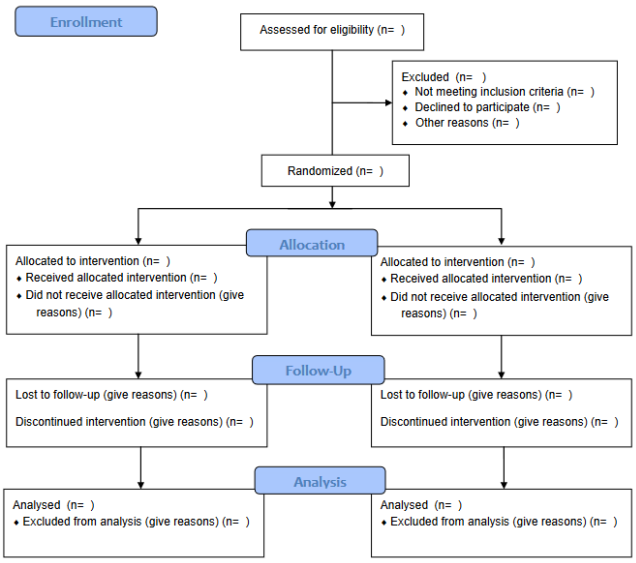


Figure 7.8: 2010 CONSORT flow diagram template

include any technical documents associated with your instruments or measures (i.e. a technical document for an assessment or a publication associated with a measure you used). Sometimes researchers will annotate instruments to show how items were named or coded.

4. Flowchart of data collection instruments/screeners

You can also include flowcharts of how participants were provided or assigned to different instruments or screeners to help users better understand issues such as missing data (Tourangeau 2015).

5. Consent forms

Consent forms(see Chapter 10) can also be added as an addendum to research protocols to give further insight into what information was provided to study participants.

6. Related publications

You may also choose to attach any publications that have come from your data as an addendum to your protocol.

7.2.6 Standard operating procedures

While the research protocol provides summary information for all decisions and procedures associated with a project, we still need documents to inform how the procedures are actually implemented on a daily basis (NUCATS n.d.). Standard

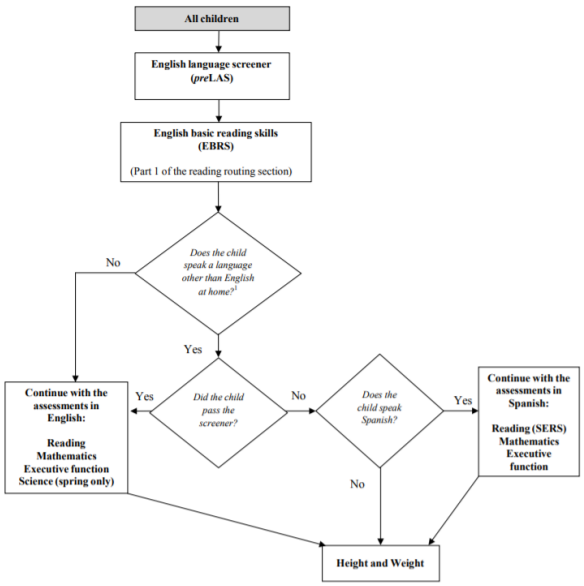


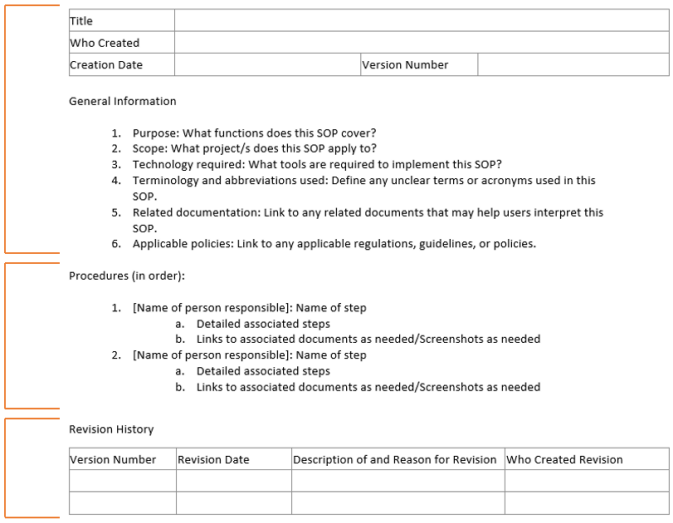
Figure 7.9: Flowchart of an ECLS-K:2011 kindergarten assessment

operating procedures (SOPs) provide a set of detailed instructions for routine tasks and decision making processes. If you recall from Chapter 5, every step that we added to a data collection workflow is then added to an SOP and the details fleshed out. Not only will you have an SOP for each type of data you are collecting (i.e., survey, assessments, observations), you should also have SOPs for any other decisions or processes that need to be repeated in a reproducible manner or followed in a specific way to maintain compliance (Hollmann et al. 2020). Many of the decisions laid out in your protocol will be further detailed in an SOP. Examples of data management procedures to include in an SOP are provided in Figure 7.10. Additional project management tasks such as recruitment procedures, personnel training, data collection scheduling, or in-field data collection routines, should also be documented in SOPs, ensuring fidelity of implementation for all project procedures.

✓ Consent/assent procedures
✓ Inclusion/exclusion criteria
✓ Assigning study IDs
✓ Randomization and blinding
✓ Building tools
  ✓ Data collection tools
  ✓ Data tracking tools
	✓ Data collection workflows
	✓ Data entry procedures and decision rules
	✓ Data scoring procedures and decision rules
	✓ Data cleaning decision rules
	✓ Data storage and transfer procedures and decision rules
	✓ Data archiving procedures and decision rules

Figure 7.10: Examples of data management processes or decisions to develop an SOP for

In addition to giving staff instruction on how to perform tasks, SOPs also create transparency in practices, allow for continuity when staff turnover or go out on leave, create standardization in procedures, and last, because an SOP should include versioning information, they allow you to accurately report changes in procedures throughout the project. You will want to create a template that is used consistently across all procedures, by all staff who build SOPs.



Title			
Who Created			
Creation Date		Version Number	

General Information

1. Purpose: What functions does this SOP cover?
2. Scope: What project/s does this SOP apply to?
3. Technology required: What tools are required to implement this SOP?
4. Terminology and abbreviations used: Define any unclear terms or acronyms used in this SOP.
5. Related documentation: Link to any related documents that may help users interpret this SOP.
6. Applicable policies: Link to any applicable regulations, guidelines, or policies.

Procedures (in order):

1. [Name of person responsible]: Name of step
   a. Detailed associated steps
   b. Links to associated documents as needed/Screenshots as needed
2. [Name of person responsible]: Name of step
   a. Detailed associated steps
   b. Links to associated documents as needed/Screenshots as needed

Revision History

Version Number	Revision Date	Description of and Reason for Revision	Who Created Revision

Figure 7.11: Standard operating procedure minimal template

In developing your SOP template, like the one in Figure 7.11, you should begin with **general information** about the scope and purpose of the procedure, as well as any relevant tools, terminology, or documentation. This provides context for the user and gives them the background to use and interpret the SOP. The next section in the SOP template, **procedures**, lists all steps in order. Each step provides the name of the staff member/s associated with that activity to ensure no ambiguity. Each step should be as detailed as possible so that you could hand your SOP over to any new staff member with no background in this process and be confident they can implement the procedure with little trouble. Specifics such as names of files and links to their locations, names of contacts, methods of communication (e.g., email vs instant message), and so forth should be included. Additions such as screenshots, links to other SOPs or workflow diagrams, or even links to tutorials can also be embedded. Last, any time revisions are made to the SOP, clarifying information about the update is added to the **revision** section and a new version of the SOP is saved. This allows you to keep track of what changes were made over time, including when they were made and who made them.

Template and Resources

Source	Resource
Crystal Lewis	SOP template [18]

7.3 Dataset Level

Our next type of documentation applies solely to your datasets and includes information about what data they contain and how they are related. It also captures things such as planned transformations for the data, potential issues to be aware of, and any alterations to the data. In addition to being helpful descriptive documentation, a huge reason for creating dataset documentation is authenticity. Datasets go through many iterations of processing which can result in multiple versions of a dataset (CESSDA n.d.a; UK Data Service n.d.b). Preserving data lineage by tracking transformations and errors found is key to ensuring that you know where your data come from, what processing has already been completed, and that you are using the correct version of the data.

Not **all** of your dataset-level documentation will be created in the documentation phase and we will talk about the timing as we review each document.

7.3.1 Readme

A Readme is a plain text document that contains information about your files. These stem from the field of computer science but are now prevalent in the research world. These documents are a way to convey pertinent information to collaborators in a simple, no frills manner. Readmes can be used in many different ways but I will cover three ways they are often used in data management.

1. For conveying information to your colleagues
    - An example of this is if a study participant reaches out to a project coordinator to let them know that they entered the incorrect ID in their survey. When the project coordinator downloads the raw data file to be cleaned by the data manager for instance, they also create a file named "readme.txt" that contains this information and is saved alongside the file in the raw data folder. That way when the data manager goes to retrieve the file, they will see that a Readme is included and know to review that document first.

```
- ID 5051 entered incorrectly. Should be 5015.
- ID 5089 completed the survey twice
  - First survey is only partially completed
```

2. For conveying steps in a process (sometimes also called a setup file)
    - There may be times that a specific data pipeline or reporting process requires multiple steps, opening different files and running different

[18] https://docs.google.com/document/d/1q84UCsn_DVL9aaO_n5T_LCjwLy96FPPB

scripts. This information **can** go in an SOP, but if it is a programmatic type process completed using a series of scripts, it might be easiest to put a simple file named "readme_setup.txt" in the same folder as your scripts so that someone can easily open the file to see what they need to run.

```
Step 1: Run the file 01_clean_data.R to clean the data
Step 2: Run the file 02_check_errors.R to check for errors
Step 3: Run the file 03_run_report.R to create report
```

3. For providing information about a set of files in a directory
   - It can be helpful to add Readmes to the top of your directories when both sharing data internally with colleagues, or when sharing files in an external repository. Doing so can provide information about what datasets are available in the directory and pertinent information about those datasets, including how the datasets are related and can be linked, information associated with different versions, definitions of common prefixes or suffixes used in datasets, or instrument response rates. Figure 7.12 is an example readme that can be used to describe all data sources shared in a project repository (Neild, Robinson, and Agufa 2022).

Dataset	File Name	Record Level	N	# of Variables
Analysis Files				
Student Analysis File	student_analysis.dta	Student	21,144	287
Multiply Imputed Files				
	student_imputed_tfa.dta	Student	5,462	211
	student_imputed_tntp.dta	Student	5,313	211
	stu_tchr_imputed_pooled.dta	Student	8,689	722
	stu_tchr_imputed_pooled_xsm.dta	Student	8,689	722
Teacher Analysis File	teacher_analysis.dta	Teacher	323	207
Classroom Analysis File	classroom_analysis.dta	Classroom-Teacher	523	24
School Analysis File	school_analysis.dta	School	63,148	28
Raw Data Files				
Teacher Survey	teacher_survey.dta	Teacher	301	227
Program Interviews	program_interviews.xlsx	Program	20	106

Figure 7.12: Institute of Education Sciences example Readme for conveying information on files in a directory

Template and Resources

Source	Resource
Crystal Lewis	Readme template for sharing information about a set of files in a directory [19]

[19] https://docs.google.com/document/d/1JWeKLDqtuk79beNJBv5xHueMwki0c7xD

Source	Resource
Crystal Lewis	Readme template for sharing project-level information [20]

7.3.2 Changelog

A changelog is a record of all of the versions of your data and code (UK Data Service n.d.d; Wilson et al. 2017). While there are automatic ways to track your data and code through programs such as Git and GitHub, in the education field where researchers often work with human subjects and identifiable data, users are most often not keeping their study data, during an active project, in a remote repository. Instead, data are usually kept in an institution-approved storage location. Even if a storage location has versioning such as Box or SharePoint, unless users are able to add contextual messages about changes made when saving versions (like a commit message with Git), users will still want to keep a changelog.

A changelog provides data lineage, allowing the user to understand where the data originated as well as all transformations made to the data. It also supports data confidence, allowing the user to understand what version of the data they are currently using and to see if more recent versions have been created and why.

In its simplest form a changelog should contain the following:

- The file name (versioned consistently)
- The date the file was created
- A description of the dataset (including what changes were made compared to the previous version)

It could also include additional information such as who made the change and a link to any code used to transform the data.

These changelogs will most likely not be created until the data capture and data cleaning phases of the life cycle when data transformations begin happening, and can be updated at any point as needed.

Template and Resources

Source	Resource
Crystal Lewis	Changelog template [21]

[20] https://docs.google.com/document/d/1rbED1r0fGAk5CREslc8qQ5378EBV5759eSqdQbp4fHc

[21] https://docs.google.com/spreadsheets/d/1fVFv_QOk90NmDW_9R_h9UnvOY79TbcPOub-dL9zqDuo

Original file name	w1_stu_svy_clean_v01.csv
Original syntax name	w1_stu_svy_cleaningv01.R
Description	Wave 1 clean student survey file

File version	Date created	Change	Syntax version
v01	2022-03-21	Cleaned data using original export: w1_stu_svy_raw_v01.csv	v01
v02	2022-04-11	Three students added to the raw data. Data re-cleaned using: w1_stu_svy_raw_v02.csv	v01
v03	2022-04-15	Corrected error found in recoding of stu_gender. Data re-cleaned using: w1_stu_svy_raw_v02.csv	v02

Figure 7.13: Example changelog for a clean student survey data file

7.3.3 Data cleaning plan

A data cleaning plan is a written proposal outlining how you plan to transform your raw data into clean, usable data. This document contains no code and is not technical skills dependent. A data cleaning plan is created for each dataset you plan to collect (e.g., student survey, student assessment, teacher survey, district student demographic data). Because this document lays out your intended transformations for each raw dataset, it allows any team member to provide feedback on the data cleaning process.

This document can be started in the documentation phase, but will most likely continue to be updated throughout the study, especially as you start digging in to your collected raw data and seeing what additional transformations are needed. Typically the person responsible for cleaning the data will write the data cleaning plans, but those documents can then be brought to a planning meeting allowing other team members, such as PIs, to provide input on the plan. This ensures that everyone agrees on the transformations to be performed. Once finalized, this data cleaning plan serves as a guide in the cleaning process. In addition to the changelog, this data cleaning plan (as well as any syntax used) provides all documentation necessary to assess data provenance, a historical record of a data file's journey.

Before writing any data cleaning plans, it can be very helpful for your team to have agreed upon general norms for what constitutes a clean dataset to help ensure that all datasets are cleaned and formatted consistently. These standards can be written down and stored in a central team or project location for referral and then used to guide your process as you write your data cleaning plan. We

will review what types of transformations you should consider adding to this type of norms document in Chapter 13.

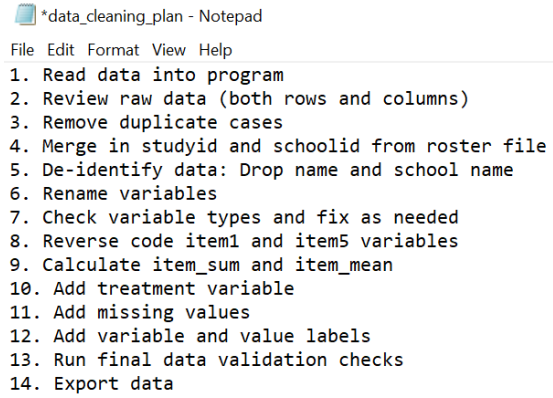


Figure 7.14: A simplistic data cleaning plan

7.4 Variable Level

Our last category of documentation is variable-level documentation. When we think about data management, I think this is most likely the first type of documentation that pops into people's minds. This documentation tells us all pertinent information about the variables in our datasets: variable names, descriptions, types, and allowable values. While variable-level documentation is often used to interpret existing datasets, it can also serve many other vital purposes including guiding the construction of data collection instruments, assisting in data cleaning, or validating the accuracy of data (Lewis 2022a). We will discuss this more throughout the chapters in this book.

7.4.1 Data dictionary

A data dictionary is a rectangular formatted collection of names, definitions, and attributes about variables in a dataset (Bordelon n.d.a; Gonzales, Carson, and Holmes 2022; UC Merced Library n.d.). This document is both a planning tool and a tool used for interpretation. A data dictionary is most useful if created in the documentation phase, before a project begins, because it is integral to many other phases of a study (Lewis 2022a; Bochove, Alper, and Gu n.d.).

This document should be structured similar to a dataset, with variable names in the first row (Broman and Woo 2018). What tool you use to build your data dictionary is up to you, but there are key pieces of information that should be included, as well as optional fields that can be helpful as well (Johns Hopkins Institute for Clinical and Translational Research n.d.).

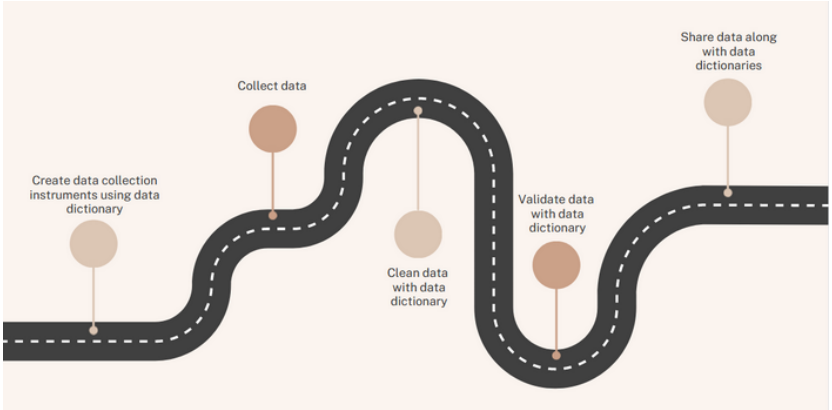


Figure 7.15: The many uses for a data dictionary

7.4.1.1 Creating a data dictionary for an original data source

Before you begin to build these dictionaries you will need to have the following:

1. Your style guide already created
   - We will talk more about style guides in Chapter 8, but this document will provide team or project standards for naming variables and coding response values.

2. Documentation for your measures
   - If you are collecting data using existing measures (i.e. existing scales, existing standardized assessments), you will want to collect any documentation on those measures such as technical documents or copies of instruments. You will want your documentation to provide information such as:
     - What items make up the measures/scales/assessment? What is the exact wording of the items?
     - How are items coded? What are allowable values?
     - Are there any calculations/scoring/reverse coding needed?
     - If items are entered into a scoring program and then exported, what variables are exported?
   - See Figure 7.16 for example of the information that could be pulled from a publication if using the Connor Davidson Resilience Scale (CD-RISC) (Connor and Davidson 2003).
3. Any data element standards that you plan to use
   - See Chapter 10 for an overview of existing data element standards

You will then build one data dictionary for each instrument you plan to collect (e.g., student survey data dictionary, teacher survey data dictionary, student

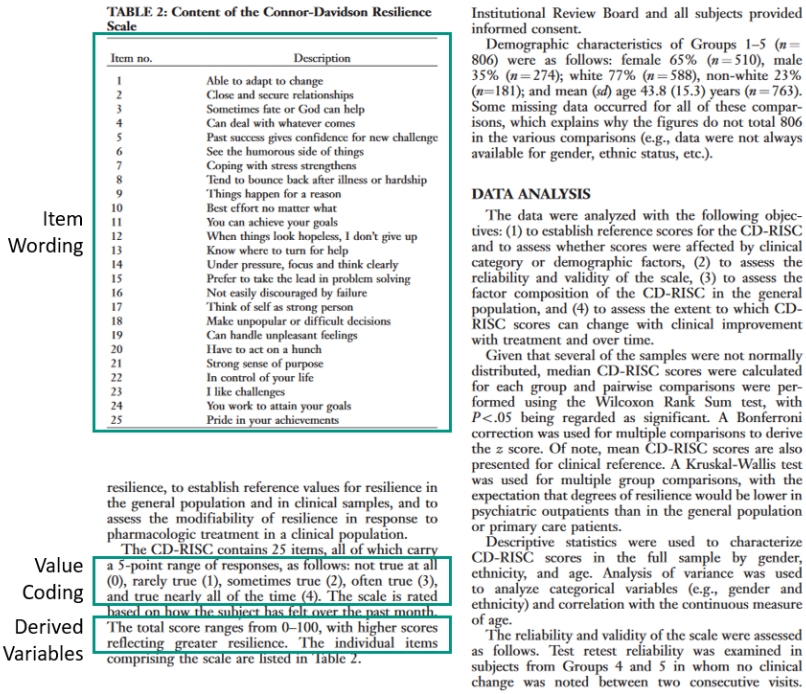


TABLE 2: Content of the Connor-Davidson Resilience Scale

Item no.	Description
1	Able to adapt to change
2	Close and secure relationships
3	Sometimes fate or God can help
4	Can deal with whatever comes
5	Past success gives confidence for new challenge
6	See the humorous side of things
7	Coping with stress strengthens
8	Tend to bounce back after illness or hardship
9	Things happen for a reason
10	Best effort no matter what
11	You can achieve your goals
12	When things look hopeless, I don't give up
13	Know where to turn for help
14	Under pressure, focus and think clearly
15	Prefer to take the lead in problem solving
16	Not easily discouraged by failure
17	Think of self as strong person
18	Make unpopular or difficult decisions
19	Can handle unpleasant feelings
20	Have to act on a hunch
21	Strong sense of purpose
22	In control of your life
23	I like challenges
24	You work to attain your goals
25	Pride in your achievements

resilience, to establish reference values for resilience in the general population and in clinical samples, and to assess the modifiability of resilience in response to pharmacologic treatment in a clinical population.

The CD-RISC contains 25 items, all of which carry a 5-point range of responses, as follows: not true at all (0), rarely true (1), sometimes true (2), often true (3), and true nearly all of the time (4). The scale is rated based on how the subject has felt over the past month. The total score ranges from 0–100, with higher scores reflecting greater resilience. The individual items comprising the scale are listed in Table 2.

Institutional Review Board and all subjects provided informed consent.

Demographic characteristics of Groups 1–5 (n = 806) were as follows: female 65% (n = 510), male 35% (n = 274); white 77% (n = 588), non-white 23% (n=181); and mean (sd) age 43.8 (15.3) years (n = 763). Some missing data occurred for all of these comparisons, which explains why the figures do not total 806 in the various comparisons (e.g., data were not always available for gender, ethnic status, etc.).

DATA ANALYSIS

The data were analyzed with the following objectives: (1) to establish reference scores for the CD-RISC and to assess whether scores were affected by clinical category or demographic factors, (2) to assess the reliability and validity of the scale, (3) to assess the factor composition of the CD-RISC in the general population, and (4) to assess the extent to which CD-RISC scores can change with clinical improvement with treatment and over time.

Given that several of the samples were not normally distributed, median CD-RISC scores were calculated for each group and pairwise comparisons were performed using the Wilcoxon Rank Sum test, with $P < .05$ being regarded as significant. A Bonferroni correction was used for multiple comparisons to derive the z score. Of note, mean CD-RISC scores are also presented for clinical reference. A Kruskal-Wallis test was used for multiple group comparisons, with the expectation that degrees of resilience would be lower in psychiatric outpatients than in the general population or primary care patients.

Descriptive statistics were used to characterize CD-RISC scores in the full sample by gender, ethnicity, and age. Analysis of variance was used to analyze categorical variables (e.g., gender and ethnicity) and correlation with the continuous measure of age.

The reliability and validity of the scale were assessed as follows. Test retest reliability was examined in subjects from Groups 4 and 5 in whom no clinical change was noted between two consecutive visits.

Figure 7.16: Pulling relevant information for the Connor Davidson Resilience Scale (CD-RISC)

assessment data dictionary). All measures/items for each instrument will be included in the data dictionary (see Figure 7.17).

Fields to include	Optional fields to include
Variable name	Skip patterns
Variable label (What is this item?)	Required item (Were participants allowed to skip this item?)
Variable type/format	Variable universe (Who got this item?)
Allowable values/range	Notes (Such as versions/changes to this variable)
Assigned missing values	Associated scale/subscale
Recoding/calculations	Time periods this item is available (if study is longitudinal)
Variable origin (Primary/derived)	Item order
	Remove item (Should this item be removed before publicly sharing data? i.e. PII)

Figure 7.17: Fields to include in a data dictionary

As you build your data dictionary, consider the following:

- Item names
  - Are your variable names meeting the requirements laid out in your style guide?
  - Are there any field standards that dictate how an item should be named?
- Item wording
  - If your items come from an existing scale, does the item wording match the wording in the scale documentation? Do you plan to reword the item?
  - Are there any field standards that dictate how an item should be worded?
- Item value codes for categorical items
  - If your items come from an existing scale, does your value coding (the numeric values assigned to response options) align with the coding laid out in the scale documentation?
  - If your items do not come from an existing scale, does your value coding align with the requirements in your style guide?
  - Are there any field standards that dictate how an items values should be coded?
- Additional Items
  - What additional items will make up your final dataset? Consider items that will be derived, collected through metadata, or added in. All of these should be included in your data dictionary.

    - Identifiers (unique ids)
    - Grouping variables (treatment)
    - Derived variables
        - This includes both variables your team derives (e.g., mean scores, reverse coded variables, variable checks) as well as variables derived from any scoring programs (e.g., percentile ranks, grade equivalent scores)
    - Metadata (Variables that your tool collects such as IPAddress, completion, language)
- What items should be removed before public data sharing (i.e., personally identifiable information)

For demonstration purposes only, the data dictionary in Figure 7.18 uses items from Patterns of Adaptive Learning Scales (PALS) (Midgley 2000). In an actual research study your dictionary would most likely include many more items and a variety of measures.

scale	subscale	var_name	origin	label	values	missing_values	type	transformations	universe
NA	NA	stu_id	primary	Student unique identifier	Range 2000-3000	NA	character	NA	All grades
NA	NA	svy_date	primary	Date of survey	Range 2022-04-01 to 2022-05-20	NA	date (YYYY-MM-DD)	NA	All grades
pals	neighborhood space	pals74	primary	In my neighborhood, I have trouble finding safe places to hang out with my friends.	not at all true = 1, 2 = 2, somewhat true = 3, 4 = 4, very true = 5	-99 = skipped	numeric	NA	Only grades 5-8
pals	neighborhood space	pals79	primary	On the weekends, I can find good and useful things to do in my neighborhood.	not at all true = 1, 2 = 2, somewhat true = 3, 4 = 4, very true = 5	-99 = skipped	numeric	NA	Only grades 5-8
pals	neighborhood space	pals81	primary	After school, I can find many interesting and positive things to do in my neighborhood.	not at all true = 1, 2 = 2, somewhat true = 3, 4 = 4, very true = 5	-99 = skipped	numeric	NA	Only grades 5-8
pals	neighborhood space	pals74_r	derived	In my neighborhood, I have trouble finding safe places to hang out with my friends. – reverse coded	not at all true = 5, 4 = 4, somewhat true = 3, 2 = 2, very true = 1	-99 = skipped	numeric	reverse coded pals74	Only grades 5-8

Figure 7.18: Example student survey data dictionary

The last step of creating your data dictionary, as it should be for every document you create in this documentation phase, is to review the document/s with your team.

- Is everyone in agreement about how variables are named, how values are coded, and our variable types?
- Is everyone in agreement about who gets each item?

- Does the team want to adjust any of the question/item wording?
- Does the data dictionary include everything the team plans to collect? Are any items missing?
    - If additional items are added to instruments at later time points, adding fields to your data dictionary, such as “time periods available”,

can be really helpful to future users in understanding why some items may be missing data at certain time points.

7.4.1.2 Creating a data dictionary from an existing data source

Not all research study data will be gathered through original data collection methods. You may be collecting external data sources from organizations like school districts or state departments of education. In these cases, you will begin building your data dictionaries later in the cycle, when data is received. Rather than the forward-moving flow we discussed before where the dictionary is built first, we will now have to work backwards to answer questions about our data.

The first step in building your data dictionary is to review your existing data. Yet, it turns out that all this tells you is what **does** exist in the data, not what **should** exist in the data. Items could be incorrectly coded, columns could be assigned the incorrect variable type, etc. As you review your data, start to collect questions such as:

1. What do these variables represent?
    - What was the wording of these items?
2. Who received the items?
3. What do these values represent?
    - Am I seeing the full range of values/categorical options for each item? Or was the range larger than what I am seeing?
    - Do I have values in my data that don't make sense for an item?
4. What data types are the items currently? What types should they be?

To answer those questions, you may need to do some additional detective work.

1. Contact the person who originally collected the data to learn more about the instrument and the data.

2. Contact the person who cleaned the data (if cleaned) to see what transformations they completed on the raw data.

3. Request access to the original instruments to review exact question wording, item response options, skip patterns, etc.
4. Request any documentation that the original owners have. For example, do they have their own data dictionaries, codebooks, or syntax that might help you understand what is going on in the data?

Ultimately you should end up with a data dictionary structured similarly to the one above. You may add additional fields that help you keep track of further changes (e.g., a column for the old variable name and a column for your new variable name), and your transformations section may become more verbose as the values assigned previously may not align with the values you prefer based on your style guide or the existing measures. Otherwise, the data dictionary should still be constructed in the same manner mentioned above.

7.4.1.3 Time well spent

The process described in this section is a manual, time consuming process. This is intentional. Building your data dictionary is an information seeking journey where you take time to understand your dataset, create standardization of items, and plan for data transformations. Spending time manually creating this document before collecting data prevents many potential errors and time lost fixing data in the future. While there are absolutely ways you can automate the creation of a data dictionary using an existing dataset, the only time I can imagine that being useful is when you have a clean dataset that you have confidently already verified is accurate and ready to be shared. However, as mentioned before, a data dictionary is so much more than a document to be shared alongside a public dataset. It is a tool for guiding many other processes in your research data life cycle.

Template and Resources

Source	Resource
Crystal Lewis	Data dictionary template [22]

7.4.2 Codebook

Codebooks provide descriptive, variable-level information as well univariate summary statistics which allow users to understand the contents of a dataset without ever opening it. Unlike a data dictionary, a codebook is created **after** your data is collected and cleaned and its value lies in data interpretation and data validation.

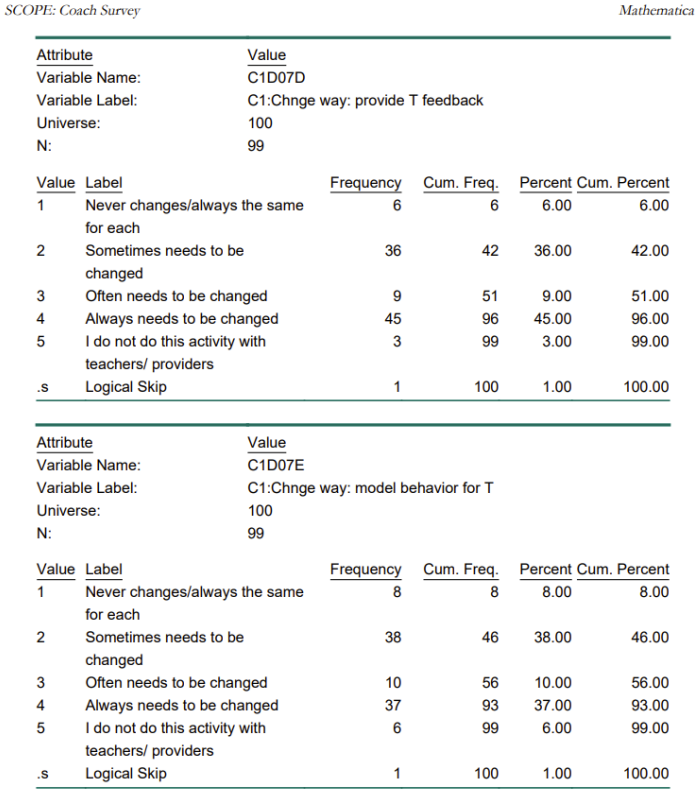
The codebook contains some information that overlaps with a data dictionary, but is more of a summary document of what actually exists in your dataset (ICPSR 2011).

Overlapping information	New information
Variable name	Existing values/ranges
Variable label (What is this item?)	Existing missing values
Variable type	Summary statistics
Value labels	Weighting
	Prefix/Suffix/Acronyms Used (ex: s_ = student survey, t_ = teacher survey, w1_ = wave 1, w2_ = wave 2)

Figure 7.19: Codebook content that overlaps and is unique to a data dictionary

[22] https://docs.google.com/spreadsheets/d/1R-5TIUvAhJRDucVhq4dNg00RR1CG7uQ6MRhje0BBC20

Figure 7.20 is an example codebook from the United States Department of Health and Human Services (2022).



SCOPE: Coach Survey *Mathematica*

Attribute	Value
Variable Name:	C1D07D
Variable Label:	C1:Chnge way: provide T feedback
Universe:	100
N:	99

Value	Label	Frequency	Cum. Freq.	Percent	Cum. Percent
1	Never changes/always the same for each	6	6	6.00	6.00
2	Sometimes needs to be changed	36	42	36.00	42.00
3	Often needs to be changed	9	51	9.00	51.00
4	Always needs to be changed	45	96	45.00	96.00
5	I do not do this activity with teachers/ providers	3	99	3.00	99.00
.s	Logical Skip	1	100	1.00	100.00

Attribute	Value
Variable Name:	C1D07E
Variable Label:	C1:Chnge way: model behavior for T
Universe:	100
N:	99

Value	Label	Frequency	Cum. Freq.	Percent	Cum. Percent
1	Never changes/always the same for each	8	8	8.00	8.00
2	Sometimes needs to be changed	38	46	38.00	46.00
3	Often needs to be changed	10	56	10.00	56.00
4	Always needs to be changed	37	93	37.00	93.00
5	I do not do this activity with teachers/ providers	6	99	6.00	99.00
.s	Logical Skip	1	100	1.00	100.00

Figure 7.20: Example codebook from the SCOPE Coach Survey

In addition to being an excellent resource for users to review your data without ever opening the file, this document may also help you catch errors in your data is out of range or unexpected values appear.

You can create separate codebooks per dataset or have them all contained in one document, clickable through a table of contents. Unlike a data dictionary, which I recommend creating manually, a codebook should be created through an automated process. Automating codebooks will not only save you tons of time, but it will also reduce errors that are made in manual entry. You can use many tools to create codebooks, including point and click statistical programs such as SPSS, or with a little programming knowledge you can more flexibly design codebooks using programs like R or SAS. For example, the R programming language has many packages that will create and export codebooks in a variety of formats from your existing dataset by just running a few functions (Lewis 2023).

Last, you may notice as you review codebooks, that many will start with several pages of text, usually containing information about the project. When it comes time to share their data, it's common for people to combine information from their research protocol or Readme files, into their codebooks, rather than sharing separate documents.

Template and Resources

Source	Resource
ICPSR	Guide to Codebooks [23]
National Center for Health Statistics	Example codebook [24]

7.5 Metadata

The last type of documentation to discuss is metadata, which is created in the "prepare for archiving" phase. When depositing your data in a repository, you will submit two types of documentation, human-readable documentation, which includes any of the documents we've previously discussed, and metadata. Metadata, data about your data, is documentation that is meant to be processed by machines and serves the purpose of making your files searchable (CESSDA n.d.b; Danish National Forum for Research Data Management n.d.). Metadata aids in the cataloging, citing, discovering, and retrieving of data and its creation is a critical step in creating FAIR data (GO FAIR n.d.; Logan, Hart, and Schatschneider 2021; UK Data Service n.d.a).

For the most part, no additional work is needed to create metadata when depositing your data in a repository. It will simply be created as part of the depositing process (CESSDA n.d.b; University of Iowa Libraries n.d.). As you deposit your data, the repository may have you fill out a form that contains descriptive (description of project and files), administrative (licensing and ownership as well as technical information), and structural (relationships between objects) metadata (Cofield n.d.; Danish National Forum for Research Data Management n.d.). The information from this form will become your metadata. Figure 7.21 is an example of an intake form, captured January 13, 2023, for the the Figshare repository (https://figshare.com/).

The most common metadata elements (Dahdul n.d.; Hayslett n.d.) are included in Figure 7.22.

Depending on the repository, at minimum, you will enter basic project-level metadata similar to above, but you may be required or have the option to enter

[23] https://www.icpsr.umich.edu/files/deposit/Guide-to-Codebooks_v1.pdf
[24] https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2020/adult-codebook.pdf

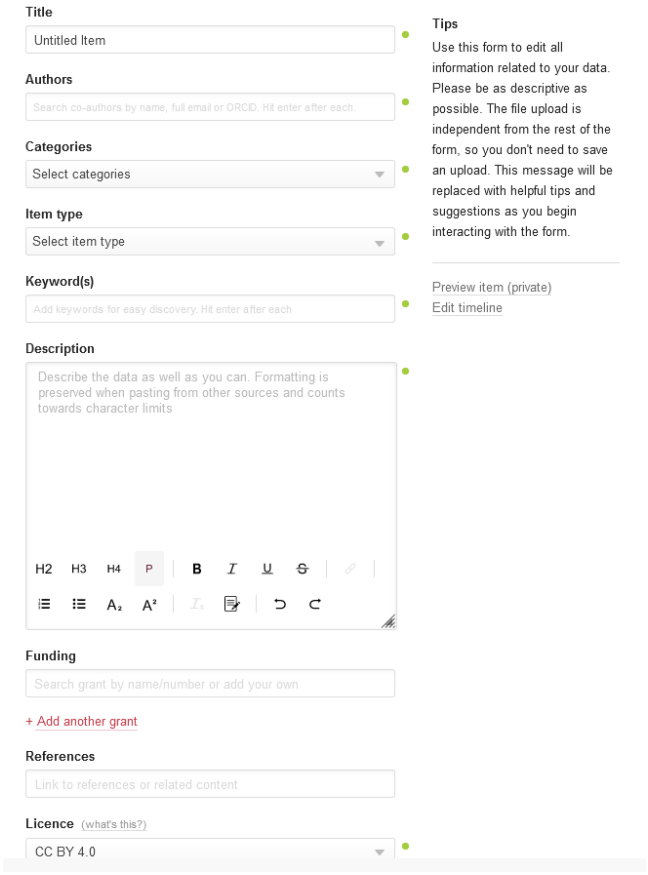


Figure 7.21: Example intake metadata form for Figshare repository

Title	Name of the project or collection of datasets
Creator	Names and institutions of the people who created the data
Date	Key dates associated with the data, such as dates covered by the data or date of creation
Description	Description of the resource
Keywords or subjects	Keywords or subjects describing the content of the data
Identifier	Unique identification code, such as a Digital Object Identifier (DOI), assigned to the resource, usually generated by the repository
Coverage (if applicable)	Geographic coverage
Language	Language of the resource
Publisher	Entity responsible for making the dataset available
Funding Agencies	Organization or agency who funded the research
Access restrictions	Where and how your data can be accessed by other researchers
Copyright	Copyright date and type
Format	What format your data is in

Figure 7.22: Common metadata elements

more comprehensive information, such as project-level information covered in your research protocol. You may also have the option to enter additional levels of metadata that will help make each level more searchable, such as file-level or variable-level metadata (Gilmore, Kennedy, and Adolph 2018; ICPSR n.d.b; LDbase n.d.b). All of the information needed for this metadata can be gathered from the documents we've discussed earlier in this chapter.

Once entered into the form, the repository converts entries into both human-readable and machine-readable, searchable formats such as XML (ICPSR n.d.b) or JSON-LD. We can see what this metadata looks like to humans once it is submitted. Figure 7.23 is an example of how ICPSR Open displays the metadata information on a project page (Page, Lenard, and Keele 2020). Notice we even have the option to download the XML formatted metadata files in one of two standards (see Section 7.5.1) if we want as well.

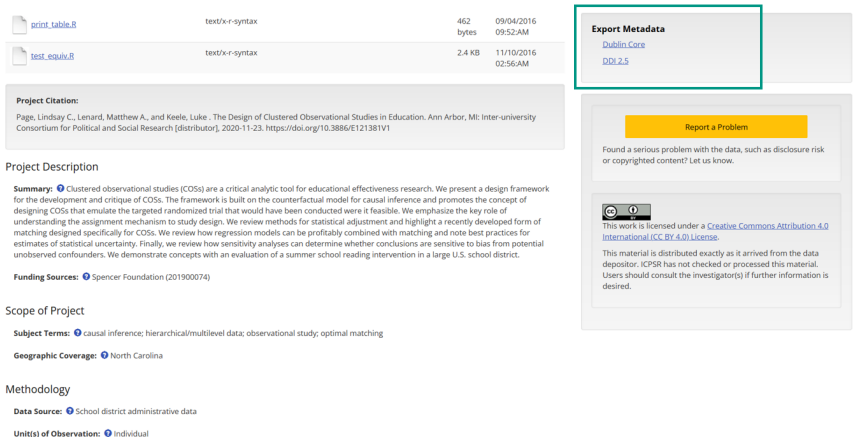


Figure 7.23: Example metadata displayed on an ICPSR Open project page

There are other ways metadata can be gathered as well. For instance, for variable-level metadata, rather than having users input metadata, repositories may create metadata from the deposited statistical data files that contain inherent metadata (such as variable types or labels) or from deposited documentation such as data dictionaries or codebooks (ICPSR n.d.b).

If your repository provides limited forms for metadata entry, you can also choose to increase the searchability of your files by creating your own machine-readable documents. There are several tools to help users create machine-readable codebooks and data dictionaries that will be findable through search engines such as Google Dataset Search (Arslan 2018; Buchanan et al. 2021; USGS, n.d.a).

7.5.1 Metadata standards

Metadata standards, typically field specific, establish a common way to describe your data which improves data interoperability as well as the ability of users to find, understand, and use data (Bolam n.d.). Metadata standards can be applied in several ways (Bolam n.d.; DDI Alliance n.d.a).

1. Formats: What machine-readable format should metadata be in?

2. Schema: What fields are recommended verses mandatory for project, dataset, and variable level metadata?
3. Controlled vocabularies: A controlled list of terms used to index and retrieve data.

Many fields have chosen metadata standards to adhere to. Some fields, like psychology (Kline 2018), are developing their own metadata standards, including formats, schemas, and vocabularies grounded in the FAIR principles and the Schema.org schema (Schema.org n.d.). Yet, the Institute of Education Sciences recognizes that there are currently no agreed upon metadata standards in the field of education (Institute of Education Sciences n.d.a).

Discipline	Metadata standard
General	Dublin Core (DC) Metadata Object Description Schema (MODS) Metadata Encoding and Transmission Standard (METS) DataCite Metadata Schema
Arts and Humanities	Categories for the Description of Works of Art (CDWA) Visual Resources Association (VRA Core) Text Encoding Initiative Guidelines (TEI)
Astronomy	Astronomy Visualization Metadata (AVM)
Biology	Darwin Core
Ecology	Ecological Metadata Language (EML)
Geographic	Content Standard for Digital Geospatial Metadata (CSDGM)
Social sciences	Data Documentation Initiative (DDI)

Figure 7.24: A sampling of field metadata standards

It can be helpful to see how standards differ as well as overlap. The DDI Alliance (n.d.b) put together this table in Figure 7.25 for instance, mapping the DDI Elements (and vocabularies) to the Dublin Core, two commonly used standards.

We can see what this metadata comparison actually looks like if we download the Dublin Core and the DDI 2.5 XML format metadata files from the ICPSR Open project we saw above (Page, Lenard, and Keele 2020). You can start to see the differences and similarities across standards.

If you plan to archive your data, first check with your repository to see if they follow any standards. For example, the OSF repository currently uses the Dat-

DC Element	DDI Element	Notes
Title	<titl> 2.1.1.1	Title of Data Collection
Creator	<AuthEnty> 2.1.2.1	Authoring Entity of Data Collection
Subject	<keyword> 2.2.1.1	Keyword(s)
	<topcClas> 2.2.1.2	Topic Classification
Description	<abstract> 2.2.2	Abstract
Publisher	<producer> 2.1.3.1	Producer of Data Collection
Contributor	<othId> 2.1.2.2	Other Identification/Acknowledgements - Data Collection
Date	<prodDate> 2.1.3.3	Production Date - Data Collection
Type	<dataKind> 2.2.3.10	Kind of Data
Format	<fileType> 3.1.5	Type of File
Identifier	<IDNo> 2.1.1.5	ID Number - Data Collection
	<holdings location="" callno="" URI=""> 2.1.8	Holdings Information - Data Collection
Source	<sources> 2.3.1.8	Sources - Used for Data Collection
Language		
Relation	<othrStdyMat> 2.5	Other Study Description Materials
Coverage	<timePrd> 2.2.3.1	Time Period Covered
	<collDate> 2.2.3.2	Date(s) of Data Collection
	<nation> 2.2.3.3	Country
	<geogCover> 2.2.3.4	Geographic Coverage
Rights	<copyright> 2.1.3.2	Copyright - Data Collection

Figure 7.25: A comparison of DDI Version 2 standards to Dublin Core standards

DC Standard

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<oaipmHtype xmlns="http://www.openarchives.org/OAI/2.0/"
xmlnsterms="http://purl.org/dc/terms/"
xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns="http://purl.org/dc/elements/1.1/">
    <responseDate>2023-01-18T21:47:09</responseDate>
    <request verb="GetRecord"
set="https://pcms.icpsr.umich.edu/pcms/api/1.0/oai/studies"/>
    <GetRecord>
        <record>
            <header>
                <identifier>121381</identifier>
                <datestamp>Wed Jan 18 21:47:10 EST
2023</datestamp>
            </header>
            <metadata>
                <dc:oaiDcType>
                    <dc:title>The Design of Clustered
Observational Studies in Education</dc:title>
                    <dc:creator>Lindsay C. Page</dc:creator>
                    <dc:creator>Matthew A. Lenard</dc:creator>
                    <dc:creator>Luke  Keele</dc:creator>
                    <dc:identifier>121381</dc:identifier>
                    <dc:identifier>
https://doi.org/10.3886/E121381V1</dc:identifier>
                    <dc:description>Clustered observational
```

DDI 2.5 Standard

```
<?xml version="1.0" encoding="UTF-8" ?>
<codeBook xsi:schemaLocation="ddi:codebook:2_5
http://www.ddialliance.org/Specification/DDI-
Codebook/2.5/XMLSchema/codebook.xsd" xmlns="ddi:codebook:2_5"
version="2.5" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance">
    <docDscr>
        <citation>
            <titlStmt>
                <titl>Metadata record for The Design of Clustered
Observational Studies in Education</titl>
                <IDNo agency="ICPSR">121381</IDNo>
            </titlStmt>
            <prodStmt>
                <producer abbr="ICPSR">Inter-university Consortium
for Political and Social Research</producer>
                <copyright>ICPSR metadata records are licensed
under a Creative Commons Attribution-Noncommercial 3.0 United
States License.</copyright>
            </prodStmt>
            <verStmt>
                <version date="2023-01-19">V1</version>
            </verStmt>
        </citation>
    </docDscr>
    <stdyDscr>
        <citation>
            <titlStmt>
                <titl>The Design of Clustered Observational
Studies in Education</titl>
                <IDNo agency="ICPSR">121381</IDNo>
                <IDNo agency="dara">
```

Figure 7.26: Metadata comparison from an AERA Open project

aCite schema (Gueguen n.d.), while ICPSR uses the DDI standard (ICPSR n.d.b). If the repository does use certain standards, work with them to ensure your metadata adheres to those standards. Some repositories may even provide curation support free or for a fee. But as I mentioned earlier, depending on your repository, adding metadata to your project may require no additional work on your part. The repository may simply have you enter information into a form and convert all information for you.

If no standards are provided by your repository and you plan to create your own metadata, you can choose any standard that works for you. Oftentimes researchers may choose to pick a more general standard such as DataCite or Dublin Core (University of Iowa Libraries n.d.), and in the field of education, most researchers are at least familiar with the DDI standard so that is another good option. Remember, if you do choose to adhere to a standard, this decision should be documented in your data management plan.

7.6 Wrapping it up

At this point your head might be spinning from the amount of documents we've covered. It's important to understand that while each document discussed provides a unique and meaningful purpose, you don't have to create every document listed. In data management we walk a fine line between creating sufficient documentation, and spending all of our working hours perfecting and documenting every detail of our project. Choose the documents that help you record and structure your processes in the best way for your project while also giving yourself grace to stop when the documents are "good enough". Each document you create that is well organized and well maintained will improve your data management workflow, decrease errors, and enhance your understanding of your data.

Chapter 8

Style guide

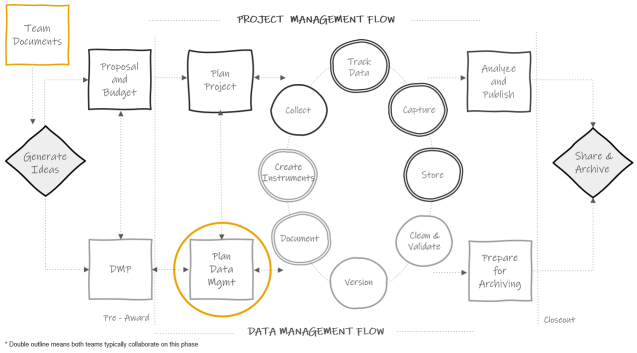


Figure 8.1: Style guide in the research project life cycle

A style guide provides general agreed upon rules for the formatting of information ("Style Guide" 2023). As mentioned in the previous Chapter 7, style guides can be created to standardize procedures such as variable naming, variable value coding, file naming, file versioning, file structure, and even coding practices.

Style guides create standardization within and across projects. The benefits of using them consistently include:

- Creating interoperability: This allows data to easily be combined or compared across forms or time.
- Improving interpretation: Consistent and clear structure, naming, and coding allows your files and variables to be findable and understandable to both humans and computers.
- Increasing reproducibility: If the organization of your file paths, file naming, or variable naming constantly change it undermines the reproducibility of any data management or analysis code you have written.

Style guides can be created for individual projects, but they can also be created at the team level, to be applied across all projects. Most importantly, they should be created before a project kicks off so you can implement them as soon as your project begins. If you do not have a team-wide style guide already created, you most likely will want to create a project-level style guide during your planning phase so that you can begin setting up your directory structures and file naming standards before you start creating and saving project-related files.

Style guides can be housed in one large document, with a table of contents used to reference each section, or they can be created as separate documents. Either way, style guides should be stored in a central location that is easily accessible to all team members (such as a team or project wiki), and all team members should be trained, and periodically retrained, on the style guide to ensure adherence to the rules. If all team members are not consistently implementing the style guide, then the benefits of the guide are lost.

For the remainder of this chapter, we will spend time reviewing some good practices for rules to add to your style guides for the following purposes:

1. Structuring directories
2. Naming files
3. Naming variables
4. Assigning variable values
5. Styling your syntax files

While some best practices will be provided below, ultimately the rules you choose to add to each style guide should be chosen based on which practices work best for your projects and your team. Whatever rules you settle on, write them in a style guide so that everyone is following the same rules within and across projects.

8.1 General good practices

Before we dive in to particular types of style guides, there are a few things to understand about how computers read names in order to understand the "why" behind some of these practices.

1. Avoid spaces.
    - While some applications (like Windows) recognize spaces, command line operations and some operating systems still do not support them so it is best to avoid them all together. Furthermore, they can often break a URL when shared
    - The underscore _ and hyphen - are good delimiters to use in place of spaces
        - It is worth noting that _ can be difficult to read when file paths are shared in links that are underlined to denote that the path

is clickable (for example when sharing a SharePoint link to a document)

2. With the exception of `_` and `-`, avoid special characters
   - Examples include but are not limited to `?`, `.`, `*`, `\`, `/`, `+`, `'`, `&`, `"`
   - Computers assign specific meaning to many of these special characters
3. There are several existing naming conventions that you can choose to add to your style guide. Different naming conventions may work better for different purposes. Using these conventions help you to be consistent with both delimeters and capitalization which not only makes your names more human-readable but also allows your computer to read and search names easier.
   - Pascal case (ScaleSum)
   - Snake case (scale_sum)
   - Camel case (scaleSum)
   - Kebab case (scale-sum)
   - Train case (Scale-Sum)
4. Character length matters. Computers are unable to read names that surpass a certain character length. This applies to file paths, file names, and variable names. Considerations for each type of limit are reviewed below.

8.2 Directory structure

When deciding how to structure your project directories (the organization of your operating systems folders and files), there are several things you want to consider.

When structuring your folders:

- First, consider organizing your directory into a hierarchical folder structure to clearly delineate segments of your projects and improve searchability
  - The alternative to using a folder structure is using metadata and tagging to organize and search for files (Cakici 2017; Fuchs and Kusniemi 2018; Krishna 2018)
- When creating your folder structure, strike a balance between a deep and shallow structure
  - Too shallow leads to too many files in one folder which is difficult to sort through
  - Too deep leads to too many clicks to get to one file, plus file paths can max out with too many characters. A file path includes the full length of both folders and file names
    - An example file path with 69 characters `W:\team\projecta\data\wave1\student\survey\projecta-`
  - Examples of file path limits:
    - SharePoint/OneDrive path limit is 400 characters (Microsoft n.d.)

    - Windows path limit is 260 characters (Ashcraft 2022)
- Create folders that are specific enough that you can limit access
  - For example you will want to limit user access to folders that hold Personally Identifiable Information (PII)
  - To protect any files that you don't want others to accidentally edit (for example your clean datasets), also consider making some files "read only"
- Decide if you want an "archive" folder to move old files into or if you want to leave previous versions in the same folder

When naming your folders:

- Consider setting a character limit on folder names (again to reduce problems with hitting path character limits)
- Make your folder names meaningful and easy to interpret
- Never use spaces in your folder names
  - Use `_` or `-` to separate words
- With the exception of `-` and `_`, don't use special characters in your folder names
- Be consistent with delimiters and capitalization. Follow an existing naming convention (as mentioned above).

Example directory structure style guide

```
1. All project directories follow this hierarchical metadata structure
    - Level 1: Name of project
    - Level 2: Life cycle folders
    - Level 3: Data collection wave folders (if relevant)
    - Level 4: Participant folder (if relevent)
    - Level 5: Specific content folder
    - Level 6: Archive folders
2. All folders should be named according to these rules
    - Meaningful name but no longer than 20 characters
    - No spaces or special characters in folder names
    - Only use lower case letters
    - Use ‘-‘ to separate words
3. All previous versions of files must be placed into their respective "archive" folde
    - A changelog should be placed in each "archive" folder to document changes betwee
```

Example directory structure created using a style guide

```
                              levelName
1  project-new
2   ¦--intervention
3   ¦   °--cohort-1
4   ¦       °--coaching_materials
5   ¦           °--archive
6   ¦               °--changelog.txt
7   ¦--project-mgmt
```

```
8   ¦   °--cohort-1
9   ¦       °--scheduling-materials
10  ¦           °--archive
11  ¦               °--changelog.txt
12  ¦--documentation
13  ¦   ¦--sops
14  ¦   ¦   °--archive
15  ¦   ¦       °--changelog.txt
16  ¦   °--data-dictionaries
17  ¦       °--archive
18  ¦           °--changelog.txt
19  ¦--data
20  ¦   °--cohort-1
21  ¦       °--student
22  ¦           °--survey
23  ¦               °--archive
24  ¦                   °--changelog.txt
25  °--tracking
26      °--cohort-1
27          ¦--participant-database
28          ¦   °--archive
29          ¦       °--changelog.txt
30          °--parent-consents
```

8.3 File naming

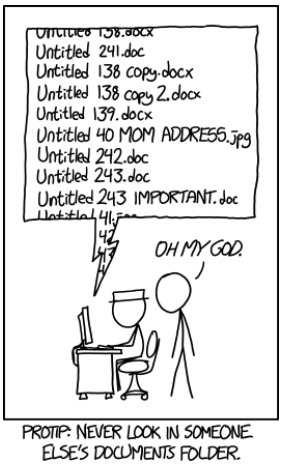


Figure 8.2: xkcd comic on naming files

As xkcd (n.d.) so aptly points out in Figure 8.2, many of us are pretty bad at naming files in a consistent and usable way. We are often in a rush to save

our files and maybe don't consider how unclear our file names will be for future users (including ourselves).

Our file names alone should be able to answer questions such as:

- What are these documents?
- When were these documents created?
- Which document is the most recent version?

A file naming style guide helps us to name files in a way that allows us to answer these questions. You can have one overarching file naming guide, or you may have file naming guides for different purposes that need different organizational strategies (for example one naming guide for project meeting notes, another naming guide for project data files). Let's walk through several conventions to consider when naming your files.

- Make names descriptive (a user should be able to understand the contents of the file without opening it)
- Never use spaces between words
  - Use `-` or `_` to separate words.
- With the exception of `_` and `-`, never use special characters
- Be consistent with delimiters and capitalization. Follow an existing naming convention.
- Consider limiting the number of allowable characters to prevent hitting your path limit (as mentioned above)
- Do not use `/` in dates and format them consistently. It is beneficial to format dates using the ISO 8601 standard in one of these two ways:
  - `YYYY-MM-DD` or `YYYYMMDD`
  - While the first way adds more characters to your variable names, it also may be clearer for users to interpret. Either of these date formats will be sortable.
- When versioning your files, pick a format and add it to your style guide
  - If you plan to version using a number, consider left padding with 0 before single digit numbers to keep the file name the same length as it grows (`v01`, `v02`).
  - As mentioned in our Chapter 7, it is possible to version programatically using tools like Git and GitHub. However, these tools are not always practical for education research. A more practical means of versioning may be to manually version files and track changes in a changelog.
- If your files need to be run in a sequential order, add the order number to the beginning of the file name, with leading zeros to ensure proper sorting (`01_`, `02_`)
- Choose abbreviations and/or consistent terms to use for common names/phrases and add them to your style guide (`student` = `stu`).
  - This helps reduce file name character lengths and also creates standardized, searchable metadata, which can allow you to more easily, programmatically retrieve files (for example, retrieve all files contain-

ing the phrase "stu_obs_raw").

- Keep redundant metadata in the file name
  - This reduces confusion if you ever move a file to a different folder or send a file to a collaborator. It also makes your files searchable.
  - For example, always put the data collection wave in a file name, even if the file is currently housed in a specific wave folder. Or always put the project name in the file name, even if the file is currently housed in a project folder.
- Choose an order for file name metadata (e.g., project -> time -> participant -> measure)

Example file naming style guide

```
1. Never use spaces between words
2. Never use special characters
3. Use _ to separate words
4. Only use lower case letters
5. Keep names under 35 characters
6. Use the following metadata file naming order:
  - Order of use (if relevant-and always add a 0 before single digits)
  - Project
  - Cohort/Wave (if relevant)
  - Participant
  - Measure
  - Further description
  - Date (always add)
  - Version (if necessary)
7. Format dates as YYYY-MM-DD
8. If there are multiple versions of a document on the same date, version using v# with a leading
9. Use the following abbreviations
  - student = stu
  - survey = svy
  - wave = w
  - project math efficacy = me
```

Example file names created using a style guide

```
me_stu_svy_sop_2022-08-01.docx
me_w1_stu_svy_data_raw_2022-11-03.csv
me_w1_stu_svy_cleaning_syntax_2023-01-22_v01.R
me_w1_stu_svy_cleaning_syntax_2023-01-22_v02.R
```

8.4 Variable naming

This style guide will be a necessary document to have before you start to create your data dictionaries. Below are several considerations to review before developing your variable naming style guide. These are broken into two types

of rules, those that are non-negotiable requirements that really should be included in your style guide (if you do not follow these rules you will run into serious problems in interpretation for both humans and machines), and then best practices suggestions that are recommended but not required.

Mandatory:

- Don't name a variable any keywords or functions used in any programming language (such as `if`, `for`, `repeat`) (R Core Team 2023; Stangroom 2019)
- Set a character limit
  - Most statistical programs have a limit on variable name characters
    - SPSS is 64
    - Stata is 32
    - SAS is 32
    - Mplus is 8
    - R is 10,000
  - With this said, do not limit yourself to 8 characters based on the fact that one future user may use a program like Mplus. Consider the balance between character limit and interpretation. It is very difficult to make good human-readable variable names under 8 characters. It is much easier to make them under 32. And the majority of your users will be using a program with a limit of 32 or more. If you have one potential Mplus user, they can always rename your variables for their specific analysis.
- Use the same variable name across time in a project
  - If an item is named `anx1` in the fall, name that same item `anx1` again in the spring
- Don't use spaces or special characters (except_), they are not allowed in most programs
  - Even the `-` is not allowed in programs such as R and SPSS as it can be mistaken for a minus sign
  - While `.` is allowed in R and SPSS it is not allowed in Stata so it's best to avoid using it
- Do not start a variable name with a number. This is not allowed in many statistical programs.
- All variable names should be unique
  - This absolutely applies to variables within the same dataset, but it should also apply to all variables across datasets within a project. The reason is, at some point you may merge data across forms and end up with identical variable names (which programs will not allow).
  - So, for example if you collect student gender from a survey and you also collect student gender from school records, differentiate between the two (`s_gender` and `d_gender`)
- If you substantively change an item (substantive wording OR response options change) after at least one round of data has been collected, version your variable names in order to reduce errors in interpretation.
  - For example revised `anx1` becomes `anx1_v2`

Suggested:

- Names should be meaningful
  - Instead of naming gender `q1`, name it `gender`
  - If a variable is a part of a scale, consider using an abbreviation of that scale plus the scale item number (`anx1`, `anx2`, `anx3`)
    - Not only does this allow you to easily associate an item with a scale, but it also allows you to programatically select and manipulate scale items (for example, sum all items that start with "anx")
- If you have used the question/scale before, consider keeping the variable name the same across projects. This can be very useful if you ever want to combine data across projects.
- Be consistent with delimiters and capitalization. Follow an existing naming convention. Most programming languages are case sensitive so consider this when choosing a convention that is feasible for your workflow.
  - Snake case (scale_sum) – preferred method for variable names
  - Kebab case (scale-sum) – don't use for variable names
  - Train case (Scale-Sum) – don't use for variable names
- Consider denoting reverse coding in the variable name to reduce confusion (`anx1_r`)
- Choose abbreviations and standard phrases to use across all variables. Using controlled vocabularies improves interpretation and also makes data exploration and manipulation easier (Riederer 2020).
  - mean = mean
  - scaled score = ss
  - percentile rank = pr
- Include an indication of the measure in the variable name (for example as a prefix) so you always know what instrument the item came from. This can also help with the unique variable name requirement above.
  - s = student self-report
  - t = teach report on students
  - `s_anx1`, `t_conf2`

Example variable naming style guide

```
1. Use snake case
2. Keep names under 32 characters
3. Use meaningful variable names
4. If part of a scale, use scale abbreviation plus item number from the scale (not order number)
5. Include an indication of the measure as a prefix in the variable name
  - student self-report survey = s_
  - teacher self-report survey = t_
  - district student level data = d_
6. Denote reverse coded variables using suffix '_r'
```

Example variable names created using a style guide

```
s_anx1
s_anx1_r
s_gender
d_gender
t_stress5
```

8.4.1 Time

Before moving on there is one last consideration for variable names. If your data is longitudinal, you may need to add rules for accounting for time in your variable names as well.

Depending on how you plan to combine your data over time, there are two different ways to account for time.

1. Concatenate time to your variable names. You do this if you plan to merge your data across time in wide format (see Chapter 3). The reason you need to concatenate time to your variable names here is because your variable names will repeat (`anx1` in wave 1, `anx1` in wave 2). And remember from our guidelines above, all variable names in a dataset **must** be unique. In order to create unique variable names and correctly interpret when items were given, we add time to our variable names. The only variables you will not assign time to are your linking variables (such as student unique identifier, teacher unique identifier, and so on). Those variables need to stay identical for linking purposes and will only appear once in your data after merging.
2. Create time variables and add them to your data. You do this if you plan to append your data over time in long format (see Chapter 3). Appending your data in long format requires no additional work in terms of variable naming. As discussed in Chapter 3, you actually want your variables to be identically formatted and named across time when appending. So here, in order to differentiate when items were asked, we add a new variable such as `time` or `wave` and add the appropriate value for each row.

During an active project, it is actually best to not combine data and to store all datasets as distinct files until you are either ready to internally use your data or you are ready to publicly share your data (during the preparing for archiving phase). At that time you can make a decision on the best way to combine your data (if you need to combine them at all), and programatically add time to variable names (if necessary) (Reynolds, Schatschneider, and Logan 2022). Waiting to combine data has benefits:

1. Having variables named consistently over time (with no time component added) allows you to easily reuse your data collection and data capture tools, as well as your cleaning code, each wave.
2. Storing files separately prevents you from potentially wasting time combining your data in a way that ends up not actually being useful or from

wasting time merging datasets that later need to be re-combined because you find an error in an individual dataset at some point.

8.4.1.1 Time in variable names

While combining your datasets across time should not happen early on in your project, it is helpful to consider early on how you *might* combine data in the future. If you do plan to potentially merge data in a wide format, it can be helpful to go ahead and plan your rule for adding time to variable names, and add that rule to your style guide. Just be abundantly clear in your guide that this time component should only be added when datasets are combined.

There is no right or wrong way to assign time in your variable names necessarily. Just make sure you continue to follow the rules from above (such as never starting a variable name with a number). Below are some options for adding time to a sample variable, `s_gender`.

- As a prefix or suffix with a generic abbreviation, such as `w1` for wave 1, added with a delimiter `_`
  - `w1_s_gender` or `s_gender_w1`
- As a prefix or suffix with a meaningful abbreviation, such as `f21` for fall 2021, added with a delimiter (`_`)
  - `f21_s_gender` or `s_gender_f21`
- One of the above options with no delimiter
  - `w1s_gender` or `s_genderw1`
- As a number embedded into your variable at a certain location, for instance, after an existing prefix such as `s` for `student survey`
  - `s1_gender`, `s2_gender`

While the first and second method do add additional characters to your variable name, there are also benefits to adding time in these ways. First, it can be easier to visually spot and interpret the time component when it is separated with a delimiter. Second, adding time as a standalone component also allows you to more easily, programmatically, manipulate the time component of your variable. This gives you more flexibility in working with your data, especially in selecting variables and restructuring your datasets.

8.5 Value Coding

In addition to naming variables in a standardized way, variables values also need to be added consistently. Value codes apply to any of your categorical variable. This may be numeric categorical values with associated labels (e.g., "no" = 1) or it may be character categorical values with associated labels (e.g., "no" = 'n').

First, if you are using a pre-existing measure, assign values and labels in the manner that the technical documentation tells you to assign codes. That will

be important for any further derivations you need to make later on based on those measures. Otherwise, you will be assigning your own values and labels. Some guidelines for assigning codes and labels (as well as examples for how to apply those guidelines) are below.

- Values must be unique
  - Do: Assign "yes" = 1 and "no" = 0
  - Don't: Assign "yes" = 1 **and** "no" = 1
- Values must be consistent within a variable
  - Do: For `gender` assign "male" = 'm'
  - Don't: For `gender` allow "male" = 'm' or 'M' or 'Male' or 'male'
- Values must be consistent across time
  - Do: Assign `anx1` values of "yes" = 1 and "no" = 0 in wave 1 **and** wave 2
  - Don't: Assign `anx1` values of "yes" = 1 and "no" = 0 in wave 1 and values of "yes" = 1 and "no" = 2 in wave 2
- Values should be consistent across the project
  - Do: Assign "yes" = 1 and "no" = 0 as the value for all yes/no items
  - Don't: Assign "yes" = 1 and "no" = 0 for some variables, and "yes" = 1 and "no" = 2 for others
    - Unless a pre-existing measure determines how some variables are coded
- Order Likert-type scale response options in a logical way
  - Do: Assign "Strongly Disagree" = 1; "Disagree" = 2; "Agree" = 3; "Strongly Agree" = 4
  - Don't: Assign "Strongly Disagree" = 1; "Disagree" = 3; "Agree" = 4; "Strongly Agree" = 2
    - Unless a pre-existing measure tells you to code variables in a different way

8.6 Missing Value Coding

There is little agreement about how missing data should be coded (White et al. 2013). There are essentially two options.

1. You can choose to leave all missing values as blank.
   - Benefits of this are that there is no chance of extreme values being mistaken as actual values
   - The concern with this method is that there is no way to discern if the value is truly missing, or was potentially erased by accident or skipped over during data entry (Broman and Woo 2018)
   - There is also the consideration that some statistical programs do not allow blank values (e.g., MPlus), and therefore missing values will need to be assigned at some point. Yet, as I mentioned earlier in this chapter, it is best to not make decisions based on one potential

use case. It is better to make decisions based on what is the most reasonable way to assign missing values for a general audience.

2. The other option is to define missing values and add them to your data. This may be one consistent value (i.e., the word NULL or the letters NA, or it may be extreme numeric values such as -999 or -98)
    - A benefit of this method are that using defined values allows you to specify distinct reasons for missing data (e.g., -97 = Not Applicable, -98 = Skipped) if that is important for your study.
    - Another benefit is that this removes the uncertainty that we had with blank cells. If a value is filled, we are now certain the the value was not deleted for skipped over during data entry.
    - The biggest problem that can occur with this method is that either your extreme values could be mistaken for actual values (if someone misses the documentation on missing values), or if you use a value that does not match your variable type, then you introduce new variable type issues (e.g., if NULL is used in a numeric variable, that variable will no longer be numeric)

Whichever method you choose, ultimately just make sure to adhere to these guidelines:

- If you decide to fill with defined missing values, use values that match your variable type (e.g., numeric missing values for numeric variables) (Tourangeau 2015; ICPSR n.d.a)
    - I will say that there is, however, some merit to using text to define missing values in numeric variables to prevent incorrect use of missing values. If you try to run a mean on your variable, you will be immediately notified that this is not possible because your variable will be stored as a character (or string) column. If you do not care about the different types of missingness, you could easily then choose to change all missing values to blank. However, if you do care about the types of missingness and want to keep that included in your variable, you will need to match variable type.
- If you use numeric values, use extreme values that do not actually occur in your data
- Use your values consistently within and across variables

Value	Description
-1	Not applicable, including legitimate skips
-7	Refused (a type of item nonresponse)
-8	Don't know (a type of item nonresponse)
-9	Not ascertained (a type of item nonresponse)
(blank)	System missing (unit nonresponse)

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K: 2011), fall 2010 and spring 2011.

Figure 8.3: Missing values assigned in the ECLS-K:2011 data file

8.7 Coding

If your team plans to clean data using code, it can be very helpful to create a coding style guide. This style guide can be tailored to a specific language that all staff will use (such as R or Stata), or it can be written more generically to apply to any coding language staff use to clean data. Below is a small sampling of good coding practices to consider adding to your guide. If you are looking for guides for a specific language, it can be very helpful to google existing style guides in that language.

- Consider building and implementing coding templates (Daskalova n.d.; Farewell 2018)
  - Templates can standardize the format of syntax files (such as using standard headers to break up code)
  - They also standardize the summary information provided at the beginning of your syntax (code author, project name, date created)
- Use comments throughout your code to clearly explain the purpose of each code chunk
  - The format of these comments will be dependent on your coding language
    - R uses `#`
    - SPSS and Stata use `*`
- Improve code readability by using (Wickham n.d.; San Martin, Rodriguez-Ramirez, and Suzuki 2023)
  - spaces
  - indentation
  - setting a line limit for your code (e.g., 80 characters)
- Use relative file paths for reproducibility
  - Setting absolute file paths in syntax reduces reproducibility because future users may have different file paths. It is important to set file paths relative the directory you are working in (Wickham and Grolemund 2017).
- If you create objects in your program (like you do in R or Python), consider adding object naming rules similar to variable naming rules
  - No spaces in object names
  - No special characters except `_` to separate words
  - No names that are existing program keywords (`if`, `for`, etc.)
- Reduce duplication, improve efficiency, and increase your ability to troubleshoot errors by using functions, loops, or macros for repetitive code chunks
- Record session information for future users
  - Record both version information as well as operating system information relevant to your code to increase the reproducibility of your code

Chapter 9

Data Tracking

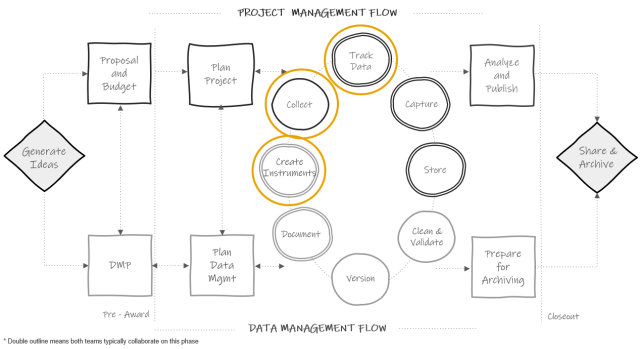


Figure 9.1: Tracking in the research project life cycle

During your project you will want to be able to answer both progress and summary questions about your recruitment and data collection activities.

1. How many participants consented to be in our study? How many have we lost during our study and why?
2. How much progress have we made in this cycle of data collection? How much data do we have left to collect?
3. How many forms did we collect each cycle and why are we missing data for some forms?

Questions like these will arise many times throughout your study for both your own project coordination purposes, as well as for external progress reporting and publication purposes. Yet, how will you answer these questions? Will you dig through papers, search through emails, and download in progress data, each time you need to answer a question about the status of your project activities? A better solution is to track all project activities in a participant tracking database.

A participant tracking database is an essential component of both project management and data management. This database contains all study participants, their relevant study information, as well as tracking information about their completion of project milestones. This database has two underlying purposes.

1. To serve as a roster of study participants and a "master key" (Pacific University Oregon 2014) that houses both identifying participant information as well as assigned unique study identifiers.
2. To aid in project coordination and reporting, tracking the movement of participants as well as completion of milestones throughout a study.

This database is considered your single source of truth concerning everything that happened throughout the duration of your project. Any time a participant consents to participate, drops from the study, changes their name, completes a data collection measure, is provided a payment, or moves locations, a project coordinator, or other designated team member, updates the information in this one location. Tracking administrative information in this one database, rather than across disparate spreadsheets, emails, and papers, ensures that you always have one definitive source to refer to when seeking answers about your sample and your project activities.

> **Note** I want to reiterate this single source of truth concept. Information is often coming in from multiple sources (e.g., data collectors in the field, emails to project coordinators from teachers, conversations with administrators). It is important to train your team that all relevant contact information (e.g., name change, new email, moved out of district) that is gleaned must be updated in the participant tracking database alone. If people track this information in other sources, such as their own personal spreadsheets, there is no longer a single source of truth, there are multiple sources of truth. This makes it very difficult to keep track of what is going on in a project. Whether a single person is designated to update information in this database, or multiple, make sure team members know either how to update information or who to contact to update information.

9.1 Benefits

A thorough and complete participant database that is updated regularly is beneficial for the following reasons:

1. Data de-identification
    - Assigning unique study identifiers that are only linked to a participant's true identity within this one database is necessary for maintaining participant confidentiality. This database is stored in a restricted secure location (see Chapter 12), separate from where de-identified study datasets are stored, and is typically destroyed at a period of time after a project's completion.

2. Project coordination and record keeping
    - This database can be used as a customer relation management (CRM) tool, storing all participant contact information, as well as tracking correspondence. It can also be used as a project coordination tool, storing scheduling information that is useful for planning activities such as data collection.
    - Integrating this database into your daily workflow allows your team to easily report the status of data collection activities (e.g., as of today we have completed 124 out of 150 assessments). Furthermore, checking and tracking incoming data daily, compared to after data collection is complete, reduces the likelihood of missing data.
    - Last, thorough tracking allows you to explain missing data in reports and publications (e.g., teacher 1234 went on maternity leave).
3. Sample rostering
    - At any time you can pull a study roster from this database that accurately reflects a participant's current status. The tracking information contained in this tool also aids in the creation of documentation including the flow of participants in your CONSORT diagram.
4. Data cleaning
    - As part of your data cleaning process, all raw dataset sample sizes should compared against what is reported as complete in your participant database to ensure that no participants are missing from your final datasets
    - Furthermore, this database can be used for de-identifying data. If data is collected with identifiers such as name, a roster from the tracking database can be used to merge in unique study identifiers so that name can be removed. A similar process can be used to merge in other assigned variables contained in the database such as treatment or cohort.

9.2 Building your database

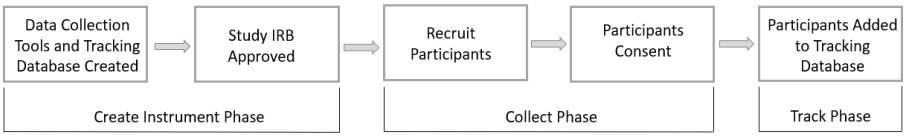


Figure 9.2: Timeline for constructing and using a tracking database

It is beneficial to build this database before you begin recruiting participants, typically during the same time that you are building your data collection tools. This way, as your team recruits and consents participants, you can record their name, assign them a study ID, and enter any other necessary identifying contact information into the participant database (see Figure 9.2). Depending on your

database system, you may even be able to scan and upload copies of your consent forms into the database.

While a project coordinator can build this database, it can be helpful to consult with a data manager, or someone with relational database expertise, when creating this system. This ensures that your system is set up efficiently and comprehensively.

This database may be a standalone structure, used only for tracking and anonymization purposes. Or it may be integrated as part of your larger study system, where all study data is collected and/or entered as well.

9.2.1 Relational databases

Before we discuss how to build this database, it is helpful to have a basic understanding of the benefits of relational databases, first introduced in Chapter 3. Using a relational database to track participant information, compared to disparate, non-connected spreadsheets, has many benefits including reducing data entry errors and improving efficiency. A relational database organizes information into tables, made up of records (rows) and fields (columns), and tables are related through keys (Bourgeois 2014b; Chen 2022). The general steps for building a relational database are below.

1. Decide what fields you want to collect and on whom you want to collect them.
2. Group those fields by entity (e.g.,students, teachers, schools) and purpose. Create tables for those groups.
3. Choose one or more fields to uniquely identify rows in those tables as primary keys. These keys should not change at any point. Typically these keys are your assigned unique study IDs.
4. Create relationships between tables through both primary and foreign keys

We can also further refine our database through normalization, structuring our database according to normal form rules (Bourgeois 2014a; Nguyen 2017; The Nobles 2020) to reduce redundancy and improve data integrity. Going in to more detail about normalization is outside of the scope of this book and building a database that follows all the normal form rules requires specific expertise, which most teams may not have. So with that said, it is perfectly acceptable to build a database that is not perfectly optimized but that works well for your team! The most important thing to consider when building a relational database is to not duplicate information across tables. Any one field should only need to be updated in one location, never more than one.

Let's compare a very simple example of building a tracking database using a relational model and a non-relational model.

9.2.1.1 Relational model

In Figure 9.3 we have three entities we need to track in our database: schools, teachers, and students. We built a very simple database with one table for each entity. Within each table we added fields that we need to collect on these participants. We have also set up our tables to include primary keys (which uniquely identify rows in each table) and foreign keys (which includes values that correspond to the primary key of another table). Our keys are all unique study identifiers that we have assigned to our study participants.

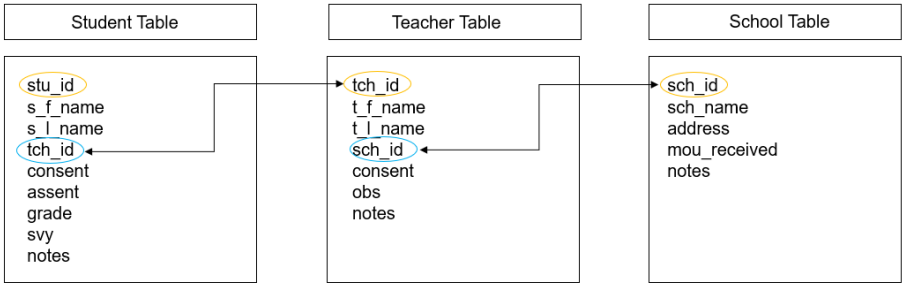


Figure 9.3: Participant database built using a relational model

We can see here that in each table we have no duplicated, repeating information. The student table only contains student level information, the teacher table only contains teacher level information, and the school table only contains school level information. This is a huge time saver. Imagine if a teacher's last name changes. Rather than updating that name in multiple places, we now only update it once, in the teacher table. If we want to see a table with both student and teacher information, we can simply query our database to create a new table. In some programs, this type of querying may be a simple point and click option, in other programs it may require someone to write some simple queries that can then be used at any time by any user.

Say for example, we needed to pull a roster of students for each teacher. We could easily create and run a query, such as this SQL query that joins the student and teacher tables above by `tch_id` and then pulls the relevant teacher and student information from both tables, seen in Table 17.1.

```
SELECT t_l_name, t_f_name, s_l_name, s_f_name,  grade
FROM Student INNER JOIN Teacher ON Student.tch_id = Teacher.tch_id
ORDER BY t_l_name, t_f_name, s_l_name, s_f_name
```

Depending on the design of your study and the structure of the database model, writing these queries can become more complicated. Again, this is where you want to strike a balance between creating a structure that reduces inefficiencies in data entry but also isn't too complicated to query based on the expertise of your team.

Table 9.1: Example roster created by querying our relational database tables

t_l_name	t_f_name	s_l_name	s_f_name	grade
Hoover	Elizabeth	Simpson	Lisa	2
Hoover	Elizabeth	Wiggum	Ralph	2
Krabappel	Edna	Prince	Martin	4
Krabappel	Edna	Simpson	Bart	4
Krabappel	Edna	Van Houten	Milhouse	4

9.2.1.2 Non-relational model

Now imagine that we built a non-relational database, such as three tabs in an Excel spreadsheet,to track our participant information (see Figure 9.4). Since we are unable to set up a system that links these tables together, we need to enter redundant information into each table (such as teacher or school name) in order to see that information within each table without having to flip back and forth across tables to find the information we need. Using this method we now have to enter repeating teacher and school names in the student table, and if any teacher names change, we will need to update it in both the teacher table and in the student table for every student associated with that teacher. This requires more entry time and creates the opportunity for more data entry errors.

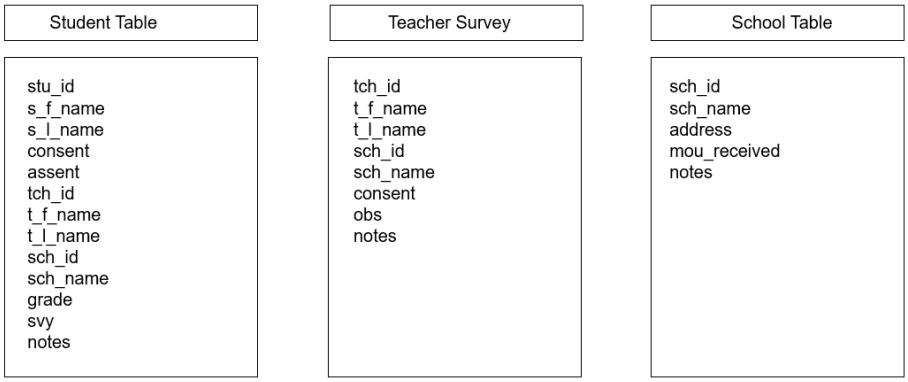


Figure 9.4: Participant database built in using a non-relational model

Note If your study includes a variety of related entities, tracked over waves of time, a relational database will be very helpful to build. If however, you are only tracking one entity (e.g., just students) for

one wave of data collection, then a database might be overkill and a simple spreadsheet will work just fine.

9.2.2 Structuring the database

Before you can begin to construct your database, you will need to collect the following pieces of information.

1. Who are your entities/units of analysis?
    - students, teachers, classrooms, districts, and so on
2. Are you collecting data longitudinally, across more than one wave?
3. Do you want to use a relational table structure?
    - If yes, how do you want to construct and relate your tables?
4. What fields do you want to include in each table?

Once you collect those pieces of information, you can begin to design your database structure. It can be helpful to visualize your database model during this process. In Figure 9.5 I am designing a database structure for a scenario where I will be collecting information from teachers in schools, over two waves of data collection.

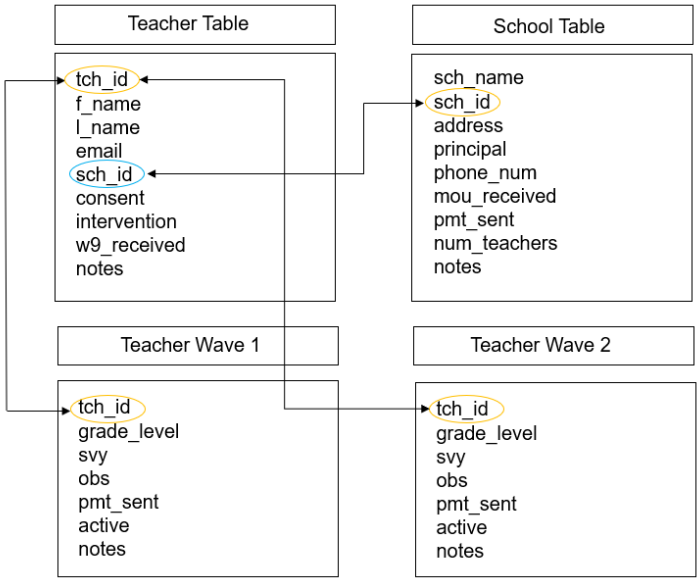


Figure 9.5: Example participant database model

I have designed this database model in this way:

1. I have four tables total
    - Two tables have information that should be constant based on my project assumptions (name, email, consent, one time payments sent

out, one time documents received)
  - If at any time this constant information changes (e.g., new last name, new principal), I would update that information in the appropriate table and make a note of when and why the change occurred in my "notes" field
- Two tables are for my longitudinal information
  - This is where I will track my data collection activities each wave, as well as any information that may change each wave, again based on the assumptions of my project. In this example, I assume that grade level may change, maybe because my data collection waves occur across school years and teachers may move around. I also assume that a participant can drop at any point in the study and I want to track their status each wave.

2. I have connected my tables through primary and foreign keys ("tch_id" and "sch_id")

The model above is absolutely not the only way you can design your tables. There may be more efficient or more appropriate ways to design this database, but again as long as you are not duplicating information, build what works for you. As an example of a potentially more efficient way to structure this database, I could combine all waves of data collection into one table and create a concatenated primary key that uses both "tch_id" and "wave" to uniquely identify rows since "tch_id" would now be duplicated for each wave of data collection (see Figure 9.6).

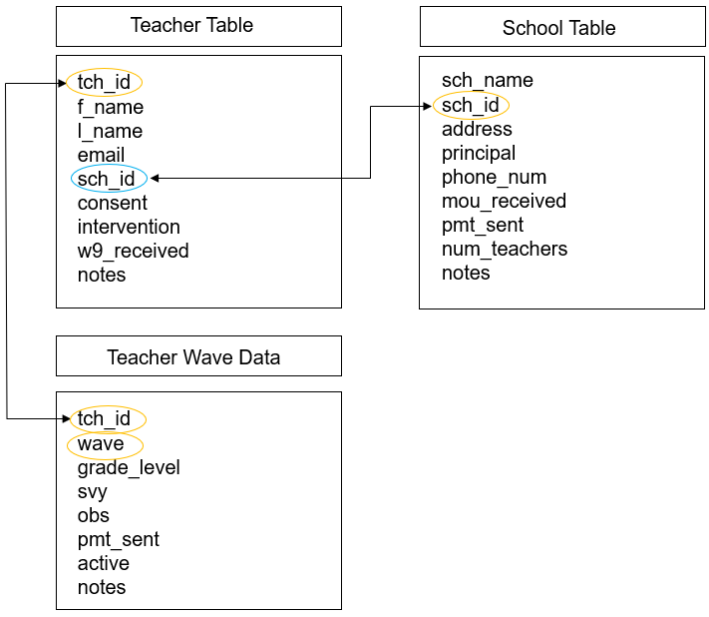


Figure 9.6: Example participant database model

While these examples are for a fairly simple scenario, you can hopefully see how you might extrapolate this model to more entities and more waves of data collection, as well as how you might modify it to better meet the needs of your specific project.

> **Note** If your study involves anonymous data collection, you will no longer be able to track data associated with any specific individual. However, it is still helpful to create some form of a tracking system. Creating a simplified database, with tables based on your sites for instance (school table, district table) allows you to still track your project management and data collection efforts (e.g., number of student surveys received per school per wave, MOU received).

9.2.3 Choosing fields

As you design your database model, you will also need to choose what fields to include in each table. The fields you choose to include will be dependent on your particular study design. While your participant tracking database may be the same database you enter all of your study data, for the purposes of this chapter we are only considering fields that are relevant for project coordination and participant de-identification. We are not concerned with fields that are collected as part of your data collection measures (i.e. survey items). You can consider your participant tracking database as an **internal** database that is only used for coordination, summary, and linking purposes. This is not a database where you would export data for external data sharing.

Below are ideas of field you may consider adding to your database. Depeding on the design and assumptions of your study, some of these may be collected once, others may be collected more than once, longitudinally.

Ideas of fields to collect:

- Study IDs (primary and foreign keys for a relational database)
- Names (participants and sites)
- Contact information
- Information relevant to project coordination (grade level, class periods, block schedules)
- Other necessary linking identifiers (double IDs, district/school IDs)
- Information helpful for data collection scheduling (blocks, class times)
- Consent/assent status
- Randomization (treatment/control)
- Grouping information (cohort)
- Summary information for rates (# of consents sent out, # of students in class)
- Administrative data status (W-9 received, MOUs received)
- Movement/drop out status
- Data collection status (unique fields for each instrument)
- Incentive status (gift cards sent out)

- Notes
  - Reasons for changes (for example changes in name, email)
  - Reasons for movement/drop out
  - Communication with participants
  - Reasons for missing data
  - Errors in data

9.2.3.1 Structuring fields

As you choose your fields you also need to make some decisions about how you will structure those fields.

1. Set data types for your fields (e.g., character, integer, date)
   - Restrict entry values to only allowable data types to reduce errors
2. Set allowable values and ranges
   - For example, a categorical status field may only allow "complete", "partially complete" or "incomplete"
3. Do not lump separate pieces of information together in a field
   - For example separate out first name and last name into two fields
4. Name your fields according to the variable naming rules we discussed in Chapter 8

9.2.4 Choosing a tool

There are many criteria to consider when choosing a tool to build your database in.

- Choose a tool that is customizable to your needs
  - Can you build a relational table structure?
  - Can you export files? Can we connect to the database via application programming interfaces (APIs)?
  - Can you query data?
- Choose a tool that is user-friendly
  - You don't want a tool with a steep learning curve for users.
- If you are running a project across multiple sites, consider the accessibility of the tool
  - For example, you may want a tool that is cloud-based so that all site coordinators can access it
  - You may also want to make sure multiple users can access it at the same time
- Choose a tool that is interoperable
  - For instance, some tools may have difficulties running on certain operating systems
- Consider cost and licensing
  - There are many free tools, but they may not provide all of the functionality you want

    - What products do you already have access to? Your institution has a license for?
- Consider security
    - Security in terms of participant confidentiality
        * Does the tool meet HIPAA/FERPA requirements?
        * Can we limit access to the entire database? To specific tables?
            · Since this database contains PII you will want to place restrictions on who can access and enter data
    - Protect data loss
        * Can we backup the system?
        * Can we protect against overwriting data?
        * Can we keep versions of the database in case a mistake is ever made and we need to go back to an older version?
- Data quality protection
    - Can we set up data quality constraints? For example, restrict input types/values

There are many tool options you can choose from. A sampling of those options are below. These tools represent a wide range from the criteria above. Take some time to review your options to see which one best meets your needs.

- Microsoft Access
- Microsoft Excel
- QuickBase
- Airtable
- REDCap
- Claris FileMaker
- Google Sheets and Google Forms
- Forms that feed into a relational database, maintained using a SQL (structured query language) database engine such as SQLite, MySQL, or PostgreSQL

9.3 Entering data

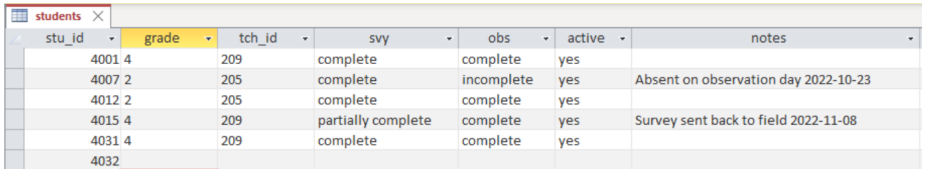
Your last consideration when building your database will be, how do you want your team to enter data into your database? There are many ways to enter data including using SQL statements, importing data, integrating your data collection platform and your tracking database, or even scanning forms using QR codes. While some of those options may work great for your project, here we are going to talk about the two simplest and most common options: manually entering data into a spreadsheet view, and manually entering data into a form.

9.3.1 Entering data in a spreadsheet view

Your first option is to manually enter data in a spreadsheet format for each participant in a row. This would be the most common (or only) option when

using tools such as Microsoft Excel or Google Sheets. However, you can also use this option when entering into other database tools such as Microsoft Access. There are both pros and cons to this method.

- Pros: This is the quickest and easiest method. It also allows you to view all the data holistically.
- Cons: This method can lead to errors if someone enters data on the wrong row/record.



students

stu_id	grade	tch_id	svy	obs	active	notes
4001	4	209	complete	complete	yes	
4007	2	205	complete	incomplete	yes	Absent on observation day 2022-10-23
4012	2	205	complete	complete	yes	
4015	4	209	partially complete	complete	yes	Survey sent back to field 2022-11-08
4031	4	209	complete	complete	yes	
4032						

Figure 9.7: Example spreadsheet view data entry

9.3.2 Entering data in a form

Your second option is to create a form that is linked to your tables. As you enter data in your forms, it automatically populates your tables with the information. This option is possible in many systems including Microsoft Access, RedCap, and even Google Forms which populates into Google Sheets.

- Pros: This method reduces data entry errors as you are only working on one participant form at a time
- Cons: Takes some time, and possibly expertise, to set up the data entry forms

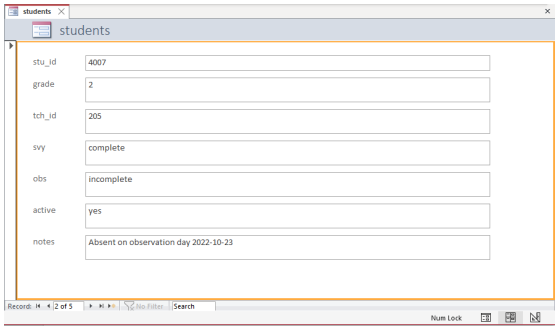


Figure 9.8: Example form view data entry

Note If your participant tracking database is separate from your data collection tools, all information will need to be entered by your team using one of these ways. However, if your participant tracking tool is also your data collection/data capture tool (such as those who collect data using RedCap), fields such as data collection status (e.g.,

survey completed) may not need to be manually entered. Rather they may be automated to populate as "complete" once a participant submits their responses in the data collection tool.

9.4 Creating unique identifiers

Participant unique identifiers are numeric or alphanumeric values and typically range from 2-10 digits. Assigning these identifiers is an important part of protecting the privacy of human participants. When publicly sharing your study data, all personally identifying information will be removed and these identifiers are what will allow you to uniquely identify and link participants in your data.

While there are several ways participant identifiers can be assigned (e.g., created by participants themselves, assigned by your data collection software), most commonly, the research team assigns these identifiers to participants. As participants are recruited and added to your participant database, you will assign them a unique participant ID. If confidentiality was promised to schools or districts, you will also want to assign identifiers to sites as well.

It can be very helpful to develop an ID schema during your planning phase, and document that schema in an SOP (see Chapter 7). In developing that schema, there are several best practices to consider.

1. Participants must keep this same identifier for the entire project.
    - This even applies in circumstances where a participant has the opportunity to be re-recruited into your study (as seen in Figure 9.9). The participant still keeps the same ID throughout the study. You can use other variables to identify the unique instances of that participant (e.g., cohort associated with that participant)
    - Having a static participant ID allows you to track the flow of each participant through your study and provides the added benefit of helping to measure dosage.

stu_id	cohort	grade	stress1	stress2
56987	1	4	2	3
54482	1	5	1	2
55574	1	3	4	1
56987	2	5	4	3

Figure 9.9: Example of keeping participant IDs for the entire study

2. Participant identifiers must be unique within and across entities

    - For example, no duplicating IDs within students or across teachers and schools
    - Not duplicating within entities is imperative to maintain uniqueness of records, while not duplicating across reduces confusion about who a form belongs to and reduces potential errors
    - If you are running multiple studies at the same time, using identical forms across studies, it can even be helpful to assign unique schemas across projects so that forms are not accidentally mixed up across projects.
3. The identifier should randomly assigned and be completely distinct from any personal information. This ensures confidentiality
    - The ID should not be associated with name, dob, income, grade level, and so forth. Some examples of what not to do include:
        - Do not sort names alphabetically and then assign IDs in sequential order
        - Do not sort names by DOB and then assign IDs in sequential order
        - Do not include initials as part of an identifier
4. Do not embed project information into the ID that has the potential to change
    - Some researchers prefer to embed some project level information into an ID to help with tracking of information. This is absolutely okay as long as the included project information is not expected to change (e.g., a project code). So for example, if the project code is "02", that may be the first 2 digits of all student identifiers. As long as the actual student IDs are still unique and randomly assigned, adding the project ID to the identifier works just fine because that value does not change.
    - However, embedding information such as wave or session into an identifier variable guarantees that your identifiers will not remain constant. This information should be added to your dataset in other ways (i.e., either as as its own variable or concatenated to variable names)
    - Embedding information such as teacher IDs, school IDs, treatment, or cohort also has the potential to cause problems. In longitudinal studies, depending on the study design, it is possible that students move to other study teachers, teachers move to other study schools, or teachers get re-recruited into other cohorts. Any of these issues would cause problems if this information was embedded into an ID because the ID would no longer reflect the accurate information and would require IDs to be changed, breaking best practice #1. Again, these additional identifiers can be tracked as separate variables (e.g., `stu_id`, `tch_id`, `sch_id`, `cohort`, `treatment`, `wave`) and added to forms and datasets as needed
5. Last, while less important during the data tracking phase, in your study datasets these identifiers should be stored as character variables. Even if

an ID variable is all numbers, it should be stored as character type. This helps prevent people from inappropriately working with these values (i.e., taking a mean of an ID variable).

stu_id	tch_id	sch_id
12000 - 13000	5000 - 6000	100 - 200

Figure 9.10: Example of a study id schema created using best practices

Note The only time you will not assign unique identifiers is when you collect anonymous data. In this situation you will not be able to assign identifiers since you will not know who participants are. However, it is still possible to assign identifiers to known entities such as school sites if anonymity is required.

Chapter 10

Data Collection

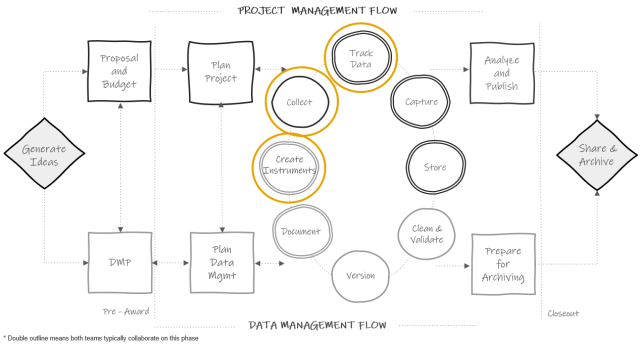


Figure 10.1: Data collection in the research project life cycle

When collecting original data as part of your study (i.e., you are administering your own survey or assessment as opposed to using existing data), data management best practices should be interwoven throughout your data collection process. The number one way to ensure the integrity of your data is to spend time planning your data collection efforts. Not only does planning minimize errors, it also keeps your data secure, valid, and relieves future data cleaning headaches.

If you have ever created a data collection instrument and expected it to export data that looks like the image on the left (Figure 10.2), but instead you export data that looks like the image on the right, then you know what I mean. Collecting quality data doesn't just happen because you create an instrument, it takes careful consideration, structure, and care on the part of the entire team.

sch_name	tch_years	stress1	stress2
Silver Oak Elementary	2	1	3
Silver Oak Elementary	10	4	1
Sun Valley Middle	3	2	2
Sun Valley Middle	1	5	5

Q1	Q2	Q3	Q6
Silver Oak elmenatry	two years	1	
silver oak	10	14	
sunvalley	2 years high school, 1 year middle	2	2
Sunvalley Middle	1 yr 2 months	15	5

Figure 10.2: A comparison of data collected without planning and data collected with planning

10.1 Quality assurance and control

In addition to planning data collection logistics (i.e. how will data be collected, who will collect it, and when), teams should spend time prior to data collection anticipating potential data integrity problems that may arise during data collection and putting procedures in place that will reduce those errors (DIME Analytics 2021a; Northern Illinois University n.d.). As shown in Figure 10.1, creating data collection instruments is typically a collaborative effort between the project management and data management team members. Even if the project management team builds the tools, the data management team is overseeing that the data collected from the tool aligns with expectations set in the data dictionary. In this chapter we will review two types of practices that both project management and data management team members can implement that will improve the integrity of your data.

1. Quality assurance practices that happen before data is collected
    - Best practices associated with designing and building your data collection instruments
2. Quality control practices implemented during data collection
    - Best practices associated with managing and reviewing data during collection

Before we dive into collecting data, it's important to first review the ethical and legal considerations of your data collection effort. When working with human subjects it is likely that the Institutional Review Board (IRB) will need to review and approve all of your data collection instruments as well as any agreement forms that will be collected as part of your study. Our next section will provide an overview of the IRB and its requirements as well as best practices for creating agreement forms for participants and partners.

10.2 Institutional Review Board

The IRB is a formal organization designated to review and monitor human participant research and ensure that the welfare, rights, and privacy of research participants are maintained throughout the project (Oregon State University 2012). If you are conducting education research with human participants you will most likely have some interaction with and oversight from the IRB. Before reviewing potential requirements, lets review the history of this administrative body.

10.2.1 Background

In 1974 the IRB was established as part of the National Research Act in response to a long history of unethical research that had been conducted with human participants (Qiao 2018). In 1979, the Belmont Report (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979) outlined a set of ethical principles for doing research with human participants. Those ethical principles included the following (Duru and Sautmann 2023; Huisman n.d.):

1. Respect for persons
    - This included both protecting autonomy of participants by acquiring consent as well as providing a plan to protect participant privacy
        - In practice this means acquiring consent in a way that ensures participants can comprehend what is being asked of them, ensuring that they understand that their participation is voluntary, and ensuring that they understand the plan to protect their privacy
2. Beneficence
    - This involved maximizing good and minimizing harm in the study, for both participants and society at large
        - In practice this means taking time to assess risk and benefits of your study for both the intervention itself as well as the data collection efforts (e.g., how burdensome is the survey)
3. Justice
    - This included providing additional care and consideration when working with subjects who are vulnerable to coercion or undue influence (e.g., children, prisoners), as well as making sure practices are non-exploitative and that there is fair distribution of costs and benefits across all participants
        - In practice this involves fairness in the selection of participants

Heavily influenced by the Belmont Report, in 1991 the Federal Policy for the Protection of Human Subjects was published, establishing core procedures for human subject protections. The policy, 45 CFR part 46 (Office for Human Research Protections 2016), included four subparts. Subpart A, known as the "Common Rule" for the 15 federal departments and agencies which codified the

policy in separate regulations, provided a set of protections for human subjects research including informed consent, review by an IRB, and compliance monitoring (National Institute of Justice 2007; Office for Human Research 2009).

In 2018 the Common Rule was revised in order to better protect research participants and to reduce administrative burden (Office for Human Research Office for Human Research 2018; U.S. Department of Health and Human Services n.d.). While many revisions were made, some changes that are applicable to education researchers include the following (Fordham University n.d.):

- Revisions and additions to exempt categories, many of which are applicable to research conducted in educational settings
- Reduced burden of continuing review, particularly for exempt and expedited studies
- Clarifications on how informed consent should be organized, written, and provided

10.2.2 Requirements

While each institution's IRB submission process is different, typically if your study involves working with human subjects you are required to submit an application to the IRB. As part of your application you will be asked to state what review category your study falls under (Lafayette College n.d.; Northwestern University n.d.; University of California Berkeley 2022).

1. Exempt
    - These studies usually involve minimal risk and fit within categories predefined by your IRB (e.g., evaluating the use of accepted or revised standardized tests). These studies typically involve a shorter review process and a quicker review than non-exempt studies.
2. Expedited
    - These studies also involve minimal risk but do not meet criteria for exempt status (e.g., collection of voice, video, or image data from non-vulnerable populations).
3. Full Review
    - If a study does not fall into one of the two categories above (e.g., collection of information about illegal behavior), it requires full review, discussed by the full board at a convened meeting.

As part of your application, common documents you may be required to submit include the following (Cabrini University n.d.; Duru and Sautmann 2023):

1. Certificates from human subjects training (e.g., CITI training[1])
2. Research protocol (see Chapter 7)
    - When writing your protocol, make sure to review your IRB's rules around data handling and include this information in your plan. IRBs

[1] https://about.citiprogram.org/

typically have specific rules for things such as how paper and electronic data must be stored and backed up, how long data should be retained, how data can be transferred and shared, and how data should be anonymized (Filip, n.d.).

3. Study materials (e.g., recruitment materials)
4. Copies of your instruments (e.g., surveys, interview guides, observation forms)
    - Note that these will need to be created before you can submit to your IRB so make sure to consider timing and start building your instruments early enough to give you time to submit to your IRB before data collection
5. Copy of informed consent/assent forms
    - Same as above, give yourself plenty of time to submit before you start participant recruitment
6. If collecting data from sites (e.g., school districts) or sharing data between sites, supporting documentation from those partners may be required (MOUs, data use/sharing agreements, letters of support, confidentiality agreements)
7. If partnering with other institutions, IRB approval letters from partner institutions may also be required

The review process can take several weeks and it is common for the IRB to request revisions to materials. Make sure to review your timeline and give yourself plenty of time to work through this process before you need to begin recruitment and data collection.

10.2.3 Agreements

There are several types of agreements that may be required for your research study for both ethical and legal reasons. Here we will discuss the most common type of agreements, informed consent and assent, as well as other agreements used when working with external partners including data sharing agreements, memorandum of understanding documents, and confidentiality agreements.

10.2.3.1 Consents

Informed consent involves obtaining a participant's voluntary agreement to participate in your research study. As described in the Belmont Report (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979), informed consent should meet the following criteria (Huisman n.d.):

- Describe the study and what is expected of the participant
- Use accessible language to ensure comprehension. Avoid technical jargon and explain terms that may not be easily understood.
- Explain that participation is voluntary
- Review how participant privacy will be maintained

With the revised Common Rule, additional requirements for informed consent were added (Fordham University n.d.).

- The top of the consent must begin with a concise review of key information that allows participants to make informed decisions
- All information must be presented with sufficient detail to make decisions, not just bulleted lists of facts
- The form must disclose any plans to use participant data for other future research

Figure 10.3 shows common elements that are included in a participant consent form (Bellevue College, n.d.; The Turing Way Community 2022).

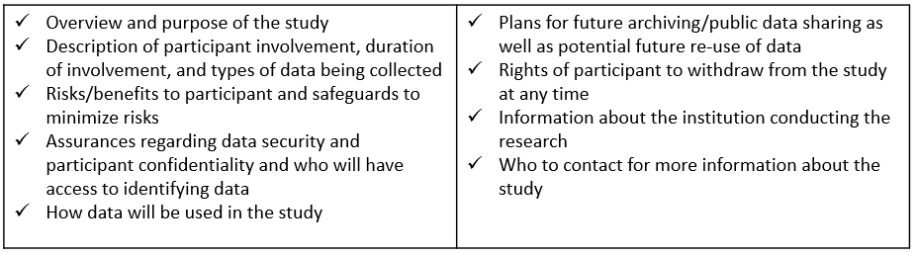


Figure 10.3: Common topics to include in an informed consent information sheet

Depending on the type of research study, a participant signature or a check box denoting consent may be required. If so, it can be helpful to put the above information on a cover/information sheet, and then have a separate page for signed consent. Before signing, participants should be required to acknowledge that they

- Have read and understood the information provided
- Have been given the opportunity to ask questions
- Understand that their participation is voluntary
- Understand that they may withdraw from the study at any time

Not all studies require active consent (University of Virginia n.d.). Some studies may allow passive consent which may be obtained by providing an information sheet to all participants with the following type of information:

If you consent to be in this study, no additional action is required; simply move forward with the study. If you choose to withdraw, you can notify a specified contact.

Your institution's IRB will let you know which type of consent is required for your study and what language is required.

10.2.3.1.1 Data sharing With an increase in federal data sharing requirements, it is very important to consider how you want to gain consent for public

data sharing. Meyer (2018) provides some general best practices to consider when adding language about public data sharing to a consent form.

- Don't promise to destroy your data (unless your funder/IRB explicitly requires it)
  - Do incorporate data-retention and sharing plans including letting participants know who will have access to their data
- Don't promise to not share data
  - Do get consent to retain and share data (consider adding the specific repository you plan to share your data in).
  - Consider offering tiered levels of consent for participants who may not want all of their data publicly shared but will allow some.
- Don't promise that research analyses of the collected data will be limited to certain topics
  - Do say that data may be used for future research and share general purposes (e.g., replication, new analyses)
- Do review the ways you plan to de-identify data but be thoughtful when considering risks of re-identification (ex: small sample size for sub-groups)

There are essentially three different ways you can go about obtaining consent for data sharing (Gilmore, Kennedy, and Adolph 2018).

1. Include a line about public data sharing in your consent to participate to research.
   - With this method, a participant who consents is agreeing to both participate in the research study and have their data shared publicly.
2. Have participants consent to data sharing at the same time you provide the research study consent, but provide a separate consent form for the purposes of public data sharing.
3. Have participants consent to data sharing on a separate consent form, at a later time, after research activities are completed.
   - Obtaining consent this way ensures the participant's are fully aware of the data collected from them and can make an informed decision about the future of that data.

A limitation of using method 1, as discussed by Gilmore, et al. (2018), is that if a participant is uncomfortable with their data being publicly shared, you will then also loose them as a study participant. So method 2 or 3 may be your best option. If you choose to go with method 2 or 3, it is very important that you not only track your participant study consent status in your tracking database (as discussed in Chapter 9), but that you also add a field to track the consent status for data sharing so that you only publicly share data for those that have given you permission to do so.

10.2.3.2 Assents

If your study involves participants under the age of 18, you may also be required to obtain a participant assent form, in addition to a parent/guardian consent

form. The guidelines for when assent is needed varies across IRBs, but typically if a child is age 7 or older (Duru and Sautmann 2023), both assent and parent consent is needed. While including similar information as provided in the consent, these are usually shorter forms that require much more simplistic language depending on the age of the child.

10.2.3.3 Collecting consent and assent

Last, many institutions have started collecting electronic consent rather than paper consents, especially with a rise in remote data collection efforts. There are benefits to this method including reducing the manual labor of collecting paper forms and removing the need to store paper forms or scan them into an electronic form. However, there are still a few things to consider before collecting electronic consent (Lee, Hughes, and Marsh 2020; Malow et al. 2021).

- Make sure your IRB approves this method
- Use institution and IRB approved tools to collect consent (e.g., Qualtrics, DocuSign)
- Find out what information is required by your IRB (e.g., signature, typed name, check box, date)
- Consider how those consents will be stored (e.g., download PDFs, download spreadsheet, store in collection tool)

If you are collecting paper consent or assent, there are still some additional things to consider.

- If consents are sent out as packets, say to schools, make sure to have a system in place to track who each form belongs to. When consents start coming back, it's possible that names are illegible, or there are duplicate names across sites. Tracking the origin of each form could look something like this:
  - Collecting class rosters ahead of time and pre-printing names and other identifiers (e.g., teacher, school) on consents before sending packets out (if this is allowed by both your IRB and the school)
  - Asking teachers to print student and teacher name on each form before the consents/assents are handed out
- If consents are collected by in-person data collectors, you will want a similar process
  - Either pre-print names on forms or have data collectors print names and other identifiers (e.g., teacher, school) on forms as they are collected

Templates and Resources

Source	Resource
Anja Sautmann	Annontated informed consent checklist [2]

[2] https://www.povertyactionlab.org/sites/default/files/research-resources/rr_irb_

Source	Resource
Holly Lane, Wilhemina van Dijk	Example parent consent [3]
Jeffrey Shero, et al.	Informed consent and waiver of consent cheat sheet [4]
Jeffrey Shero, Sara Hart	Informed consent template with a focus on data sharing [5]
University of Virginia	A collection of consent and assent templates [6]

10.2.3.4 Other agreements

As we discussed in Chapter 7, a data use agreement (DUA) is a contractual document that lays out expectations for how data will be shared between two or more parties. While the terms data use agreement and data sharing agreement (DSA) are often used interchangeably, I want to differentiate between the two documents. Data use agreements are typically legally binding agreements that provide terms and conditions for working with restricted use data. DUAs are commonly written for data sharing with school districts. In the case of education research, a DUA may include the terms for sharing, working with, and storing identifiable district-level data.

When working with de-identified, non-sensitive data though, a data sharing agreement is a good option. A DSA is a less formal agreement but is still beneficial if you want to provide terms for how data is used, such as limiting the types of projects that use the data (LDbase n.d.a). We will talk more about these types of agreements in the Chapter 14.

Another type of agreement, commonly signed when working with partners such a school districts, is a memorandum of understanding (MOU), which establishes the framework for collaboration (National Center for Education Statistics n.d.b; REL West, n.d.). This document is typically not legally binding, but establishes agreements around things such as responsibilities, communication, and expectations (Duru and Kopper 2021). An MOU can be a standalone document or can include a DSA or DUA as part of the document.

Last, confidentiality agreements and non-disclosure agreements (NDAs) are other types of agreement that may be needed. These documents restrict the use of proprietary or confidential information (University of Washington n.d.) and are legally enforceable agreements.

annotated-informed-consent-checklist_0.pdf

[3] https://www.ldbase.org/system/files/documents/2021-04/HS-ParentConsent.txt

[4] https://osf.io/3czbx

[5] https://figshare.com/articles/preprint/Informed_Consent_Template/13218773

[6] https://research.virginia.edu/irb-sbs/consent-templates

Templates and Resources

Source	Resource
Amy O'Hara	Sample text for data use agreements [7]
Florida State University	Example data use agreement [8]
REL West	Data use agreement checklist [9]
University of North Carolina	Data use agreement decision making flow chart [10]
Wilhelmina van Dijk, Sara Hart	Example data sharing agreement [11]

10.3 Quality Assurance

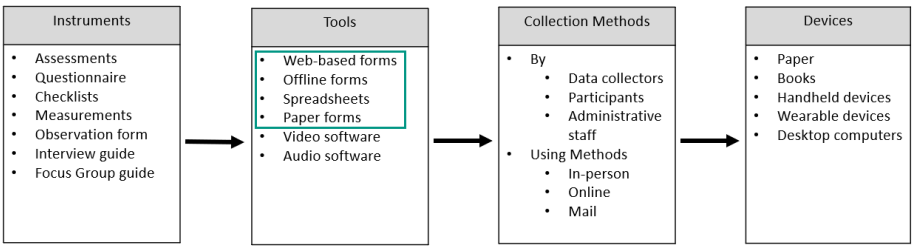


Figure 10.4: Common education research data collection methods

Now that we have a baseline understanding ethical and legal considerations, we can dive in to protecting data quality during data collection. Education researchers collect original data in many ways (see Figure 10.4). The focus of this chapter will be on data collected via forms (i.e., a document with spaces to respond to questions). Forms are widely used to collect data in education research (think surveys, assessments, observation forms, or a progress monitoring form on a website), yet if developed poorly, they can produce some of the most problematic data issues. On the flip side, if the practices discussed in this chapter are implemented, forms can also be the easiest tool to remedy issues with.

The focus on forms is not to discount the importance of data collected through other means such as video or audio recording, where issues such as participant privacy and data security and integrity should absolutely also be considered.

[7] https://admindatahandbook.mit.edu/book/v1.0-rc4/appendix/dua_appendix.pdf
[8] https://www.research.fsu.edu/media/1091/hipaadatause.doc
[9] https://ies.ed.gov/ncee/rel/regions/west/relwestFiles/pdf/CRP_Data_Sharing_Agreements_and_MOUs.pdf
[10] https://research.unc.edu/wp-content/uploads/sites/61/2013/04/CCM3_039360.pdf
[11] https://figshare.com/articles/preprint/Example_Data_Sharing_Agreement/14576049

However, even with those types of data collection efforts, often teams are ultimately still coding that data using some sort of form (e.g., observation form), further supporting the need to build forms that collect quality data.

When collecting information using forms you can certainly do your best to fix data errors after data collection during a cleaning process. However, one of the most effective ways to ensure quality data is to correct it at the source. This means designing items and creating data collection tools in a way that produces valid, reliable, and more secure data. When creating your original data collection instruments, there are four ways to collect higher quality data.

1. Using good questionnaire design principles
2. Implementing a series of pilot test
3. Choosing data collection tools that meet your needs
4. Building your instrument with the end in mind

We will discuss each of these phases below.

> **Note** If you are collecting data using a standardized assessment, along with a provided instrument (e.g., a computer-adaptive testing program), most of the information in this section will not be applicable. In those situations, it is best to adhere to all guidelines provided by the assessment company.

10.3.1 Questionnaire design

In Chapter 7 we discussed the importance of documenting all instrument items in your data dictionary before creating your data collection instruments. As you develop items to add to your data dictionary, it is vital to consider questionnaire design.

While some instruments (e.g., cognitive assessments) typically have standardized items, other instruments, such as surveys, are often not predefined, allowing researchers freedom in the design of the instrument which can lead to negative effects such as errors, bias, and potential harm (DIME Analytics 2021a; Northern Illinois University n.d.). While it is outside of the scope of this book to fully cover questionnaire design, below are a few tips to help you collect more valid, reliable, and ethical survey data.

1. Use existing standards if possible
    - Organizations such as the National Institutes of Health (n.d.b) and the National Center for Education Statistics (n.d.a) have developed repositories (Common Data Elements[12] and Common Education Data Standards[13]) of standardized question wording paired with a set of allowable response options for commonly used data elements. Using standards when collecting commonly used variables, such as

[12] https://www.nlm.nih.gov/oet/ed/cde/tutorial/03-100.html
[13] https://ceds.ed.gov/

demographics, provides the following benefits (ICPSR 2022; Kush et al. 2020):
  - Reduces bias
  - Allows for harmonization of data across your own research studies and also across the field
    * This allows researchers to draw conclusions using larger samples or by comparing data over time
    * It also reduces the costs of integrating datasets
  - Improves interpretation of information

2. Make sure questions are clearly worded and answer choices are clear and comprehensive
   - Consider how the language might be interpreted. Is the question wording confusing? Can the response options be misinterpreted?
     - Rather than asking "What county are you from?" when looking for the participant's current location, be more specific and ask "What county do you currently reside in?"
     - Rather than asking "Which parent are you?" and providing the response options "m" and "f", where "m" and "f" could be interpreted as "male" or "female", clearly write out the response options and make sure they are comprehensive (mother, father, legal guardian, and so forth)
     - Is the question leading/biased?
       * Are the response options ordered in a leading way?
     - Is there no one way to answer this question?
       * Are categories mutually exclusive and exhaustive (ICPSR n.d.a)?
3. Consider data ethics in your questionnaire design (Kaplowitz and Johnson 2020; Kopper and Parry 2021; Mathematica n.d.; Narvaiz n.d.)
   - Consider the why of each item and tie your questions to outcomes
     - Don't cause undue burden on participants by collecting more data just to have more data.
     - If collecting demographic information, provide an explanation of why that information is necessary and how it will be used in your research
   - Review question wording
     - Does it have potential to do harm to participants?
     - If sensitive questions are included, make sure to discuss how you will protect respondent's information
   - Make questions inclusive of the population while also capturing the categories relevant for research
     - If a question is multiple choice, still include an "other" option with an open-text field
     - For demographic information, allow participants to select more than one option
   - Consider including one general free text field in your survey to allow participants to provide additional information that they feel was not

captured elsewhere

4. Limit the collection of personally identifiable information
    - Collecting identifiable information is a balancing act between protecting participant confidentiality and collecting the information necessary to implement a study. We often need to collect some identifying information either for the purposes of record linking or for purposes related to study outcomes (e.g., scoring an assessment based on participant's age).
    - As a general rule, you only want to collect personally identifiable information (PII) that is absolutely necessary for your project, and no more. As discussed in Chapter 2, PII can include both direct identifiers (e.g., name or email) as well as indirect identifiers (e.g., birthdate). Before sharing your data, all PII will need to be removed or altered to protect confidentiality.

Survey Design Resources

Source	Resource
Sarah Kopper, Katie Parry	Survey design [14]
Pew Research Center	Writing survey questions [15]
Stefanie Stantcheva	How to run surveys: A guide to creating your own identifying variation and revealing the invisible [16]
World Bank	Survey content-focused pilot checklist [17]

10.3.2 Pilot the instrument

Gathering feedback on your instruments is an integral part to the quality assurance process. There are three phases to piloting an instrument (DIME Analytics 2021b) (see Figure 10.5):

1. Gathering internal feedback on items
    - As discussed in Chapter 7, once all items for each instrument have been added to your data dictionary, have your team review the data dictionary and provide feedback
2. Piloting an instrument for content
    - Once the team has approved the items to be collected, the second phase of piloting can begin. Create a printable draft of your instrument that can be shared with people in your study population and gather feedback

[14] https://www.povertyactionlab.org/resource/survey-design
[15] https://www.pewresearch.org/our-methods/u-s-surveys/writing-survey-questions/
[16] https://www.nber.org/system/files/working_papers/w30527/w30527.pdf
[17] https://dimewiki.worldbank.org/Checklist:_Content-focused_Pilot

- If you are piloting your instrument with a small population (N < 10), and you are either gathering feedback from a checklist or collecting data using the instrument with no intent to disseminate outcomes as research data, then IRB approval will most likely not be required (Cornell University 2019; Stanford University n.d.). With that said, you should always consult with your institution's IRB because rules can vary.

3. Piloting the instrument for data related issues
    - Once the instrument is created in your chosen data collection tool, share the instrument with your team for review
    - Here we are most interested in whether or not the data we are collecting are accurate, comprehensive, and usable
    - We will discuss this phase in greater detail later in this chapter

Last, as you move through the piloting phases, remember to update any changes not only in your tool but also in your data dictionary and any other relevant documentation.

Pilot Phase	Phase 1 – Build data dictionary	Phase 2 - Content	Phase 3 - Data
Who tests	Team staff	People from your population	Team staff
What to provide to testers	A data dictionary for your instrument	General questions, response options for each question, and planned question order on paper	Fully built survey in chosen tool (e.g., web-based platform, paper form)
Example items to include in a feedback checklist	- Are all items included? - Are we in agreement about how items are named? - Are the items worded correctly? - Are response options correct? - Are response options coded correctly?	- Are the items clearly worded? Are they sensitive? - Are answer choices comprehensive? - Is the item order clear? - Is the time to complete survey acceptable?	- Were there any barriers to accessing the instrument? - Are all questions accounted for? - Are the items worded correctly? - Are all response options visible for each categorical question? - Is data validation working? Were you able to enter unallowable values, data types, or formats? - Is the skip logic working? - Are you allowed to skip items that you should not be able to?
Next steps	Make edits to data dictionary. Build paper	Make edits to items based on feedback. Update changes in data dictionary. Then create full instrument in chosen tool before moving on to Phase 2.	If data is collected electronically, download sample data. - Is the data organized as expected? Make edits to tool based on both feedback and findings from exported data if data is collected electronically.

Figure 10.5: Data collection instrument pilot phases

10.3.3 Choose quality data collection tools

Once content piloting is completed, teams should be ready to begin building their instruments in their data collection tools (see Figure 10.4). Research teams may be restricted in the tools they use to collect their data for a variety of reasons including limited resources, research design, the population being studied, or the chosen instrument (e.g., an existing assessment can only be collected using a provided tool). However, if you have the flexibility to choose how you collect

your data, pick a tool that meets the various needs of your project while also providing data quality and security controls. Things to consider when choosing a data collection tool are:

1. Pick the tool that meets the needs of your project
    - Is crowdsourcing required?
    - Is multi-site access required?
    - Who is entering the data (i.e., data collectors, participants)?
        - If participants are entering data, is the tool accessible for your population?
    - What are the technical requirements for the tool (i.e., will internet be available if you plan to use a web-based tool)?
    - Does the tool have customizable features that are necessary for your instrument (e.g., skip logic, automated email reminders, options to embed data, options to calculate scores in the tool)?
2. Compliance and security
    - If you collect identifiable data, is the tool HIPAA compliant? FERPA compliant? (see Chapter 2 for more information about these regulations)
    - Is the tool approved by your institution?
    - If collecting anonymous data, do you have the option to anonymize responses in the tool (e.g., remove IP Address and other identifying metadata collected by the tool)?
3. Training needed
    - Is any additional team training needed to allow your team to use and/or build instruments in the tool?
4. Associated costs
    - Is there a cost associated with the tool? Do you have the budget for the tool?
    - Will there be additional costs down the line (e.g., collecting data on paper means someone will need to hand enter the data later)?
5. Data quality features
    - Does the tool allow you to set up data validation?
    - Does the tool have version control?
    - Does the tool have features to deal with fraud/bots?

While there are a variety of tool options, in a nutshell when it comes to data collected via forms, we are collecting data in one of two ways—electronic or paper. In addition to choosing tools based on the above criteria, there are some general benefits associated with each method that should also be considered (Cohen, Manion, and Morrison 2007; Douglas, Ewell, and Brauer 2023; Gibson 2021; ICPSR n.d.a; Malow et al. 2021; Society of Critical Care Medicine 2018; Bochove, Alper, and Gu n.d.).

> **Note** If you choose to collect data in an electronic format, I highly recommend using a web-based tool that directly feeds into a shared database rather than through offline tools that store data on indi-

Electronic data collection benefits	Paper data collection benefits
• Ability to use data validation to collect accurate and uniform information • Scalable (easier to reuse, edit, and maintain) • Efficient (reduces both cost and effort associated with printing, collecting, and entering data) • Prevents inconsistencies with automated logic • Less missing data with request response options • Potential to reach broader populations (e.g., crowdsourcing) • Quicker turnaround of analysis-ready data and provides opportunities to build real-time reporting pipelines (e.g., using APIs) • Improves data integrity (ability to pre-load unique identifiers and verify identity)	• Intuitive to create (no training required) • Easy to do cognitive checks (eyeball for errors) • Easier to catch errors early on (for instance in the field)

Figure 10.6: Comparison of data collection tool benefits

vidual devices. Using a web-based tool, all data is stored remotely in the same database and can be easily downloaded or connected to at any time. No additional work is required. However, when collecting data on various tablets that are used in the field, if the forms are offline and cannot be later connected to a web-based form, then all data will be stored individually on each tablet. This not only may be less secure (e.g., a tablet becomes corrupted), it may also require additional data wrangling work including downloading data from each tablet to a secure storage location each day and then combining all files into a single dataset. If you use an electronic tool but your site does not have internet, consider using one of the many tools (e.g., Qualtrics, SurveyCTO) that allow you to collect data using their offline app and then upload that data back to the platform once you have an internet connection again.

Tool Comparison Resources

Source	Resource
Michael Gibson, Wim Louw	Survey platform comparison[18]
Washington State University Libraries	Software for sensitive data[19]
Benjamin Douglas, et al.	Data quality in online human-subjects research comparison of tools[20]

10.3.4 Build with the end in mind

Last, you want to build your tool with the end in mind. This means taking time to consider how the data you collect will be translated into a dataset (Beals and

[18] https://www.povertyactionlab.org/resource/survey-programming
[19] https://libguides.libraries.wsu.edu/rdmlibguide/ethics
[20] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279720

Schectman 2014; Lewis 2022b; UK Data Service n.d.b). Recall from Chapter 3, we ultimately need our data to be in a rectangular format, organized according to the basic data organization rules, in order to be analyzable.

The process for building your tools with the end in mind is fairly different for electronic tools compared to paper forms so we are going to talk about these two processes separately.

10.3.4.1 Electronic data collection

The first thing you will want to do before building your tool is bring out your data dictionary. This data dictionary will be your guide as you build your instrument. Some tools, such as REDCap, provide the option to upload your data dictionary which can then be used to automate the creation of data collection forms as opposed to building them from scratch (Patridge and Bardyn 2018).

However, if you are building your instrument manually, adhering to the following guidelines will ensure you collect data that is easier to interpret and more usable, and it will also reduce the amount of time you will need to spend on future data cleaning (Lewis 2022b).

1. Name all of your items the correct variable name from your data dictionary (UK Data Service n.d.b)
    - For example, instead of using the platform default name of "Q2", rename the item to "tch_years"
    - As we mentioned in Chapter 8, it's also best to not concatenate a time component to your variable names if your project is longitudinal. Doing so makes it difficult to reuse your instrument for other time periods, creating additional work for you or your team.
2. Code all values as they are in your data dictionary
    - For example, 1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree
    - Many times your tools assign a default value to your response options and these values may not align with what you've designated in your data dictionary
    - As you edit your survey, continue to check that your coded values did not change due to reordering, removal, or addition of new response options
3. Use data validation to reduce errors and missing data (UK Data Service n.d.b)
    - Content validation for open-text boxes
        - Restrict entry to the type assigned in your data dictionary (e.g., numeric)
        - Restrict entry to the format assigned in your data dictionary (e.g., YYYY-MM-DD)
        - Restrict ranges based on allowable ranges in your data dictionary (e.g., 1-50)

        * This could even include validating against previous responses (e.g., if SchoolA was selected in a previous question, grade level should be between 6-8, if SchoolB was selected, grade level should be between 7-8)
    - If duplicate responses are not allowed for a question (i.e., rank order), have options drop away after they are chosen
- Response validation
    - Set force response for variables that are absolutely necessary (e.g., unique identifier)
        * Do not add force response to nonessential questions. Doing so can be both harmful to participants and produce bad data. If you add forced response options to nonessential questions, ensure that those questions allow participants to still opt-out.
    - For nonessential questions, add a request response option which notifies the respondent if they skip a question and asks them if they still would like to move forward without responding

4. Choose an appropriate type and format to display the item
    - Become familiar with the various questions types available in your tool (e.g., rank order, multiple choice - select all, multiple choice - select one, text box, slider scale)
    - Become familiar with the various formats (e.g., radio button, drop-down, checkbox)
    - For example, if your item is a rank order question (ranking 3 items), creating this question as a multi-line, free-text entry form may lead to duplicate entries (such as entering a rank of 1 more than once). However, using something like a rank order question type with a drag and drop format ensures that participants are not allowed to duplicate rankings.
5. If there is a finite number of response options for an item, and the number isn't too large (less than $\sim$ 20) use controlled vocabularies (i.e., a pre-defined list of values) rather than an open-text field (OpenAIRE_eu 2018; UK Data Service n.d.b)
    - For example, list school name as a drop-down item rather than having participants enter a school name
        - This prevents variation in text entry (e.g., "Sunvalley Middle", "sunvalley", "Snvally Middle"), which ultimately creates unnecessary data cleaning work and may even lead to unusable values
6. If there is an infinite number of response options for an item or the number of options is large, use an open-text box
    - If you can create a searchable field in your tool, allowing your participants to easily sift through all of the options, you absolutely should. Otherwise, use a text-box as opposed to having participants scroll through a large list of options
    - Consider adding examples of possible response options to clarify what you are looking for
    - Using open-ended text boxes does not mean you can not regroup

this information into categories later during a cleaning process. It is just more time-consuming and requires interpretation and decision-making on the part of the data cleaner

7. Only ask for one piece of information per question
    - For example, rather than asking "Please list the number of students in your algebra class and geometry class", split those into two separate questions so those questions download as two separate items in your dataset
    - This also includes more simple examples such as splitting first name and last name into two separate fields
    - This also prevents confusion in case a participant or data collector swaps the order of information
8. To ensure the security and integrity of data, consider adding a line to the introduction of your web-based instrument, instructing participants to close their browser upon completion so that others may not access their responses
9. Last, if possible, export the instrument to a human-readable document to perform final checks
    - Are all questions accounted for?
    - Are all response options accounted for and coded as they should be?
    - Is skip logic shown as expected?

Once your tool is created, the last step is to pilot for data issues (see Figure 10.5). Collect sample responses from team members. Create a feedback checklist for them to complete as they review the instrument (Gibson and Louw 2020). Assign different reviewers to enter the survey using varying criteria (e.g., different schools, different grade levels). Let team members know that they should actively try to break things (Kopper and Parry 2020). Try to enter nonsensical values, try to skip items, try to enter duplicate entries. If there are problems with the tool, now is the time to find out.

After sample responses are collected from team members, export the sample data and review the data for the following:

1. Are there any unexpected or missing variables?
2. Are there any unexpected variable names?
3. Are there unexpected values for variables?
4. Are there missing values where you expect data?
5. Are there unexpected variable formats?
6. Is data exporting in an analyzable, rectangular format?

If any issues are found either through team feedback or while reviewing the exported sample data, take time to update the tool as well as your documentation as needed before starting data collection.

10.3.4.2 Paper data collection

There are many situations where collecting data electronically may not be feasible or the best option for your project. While it is definitely trickier to design a paper tool in a way that prevents bad data, there are still steps you can take to improve data quality.

1. Use your data dictionary as a guide as you create your paper form
    - Make sure all questions are included and all response options are accurately added to the form
2. Have clear instructions for how to complete the paper form (Kopper and Parry 2021)
    - Make sure to not only have overall instructions at the top of the form but also have explicit instructions for how each question should be completed
        - Where to write answers (e.g., not in the margin)
        - How answers should be recorded (e.g., YYYY-MM-DD, or 3 digit number)
        - How many answers should be recorded (e.g., circle only one answer, check all applicable boxes)
        - How to navigate item skip logic (e.g., include visual arrows)
3. Only ask for one piece of information per question to reduce confusion in interpretation

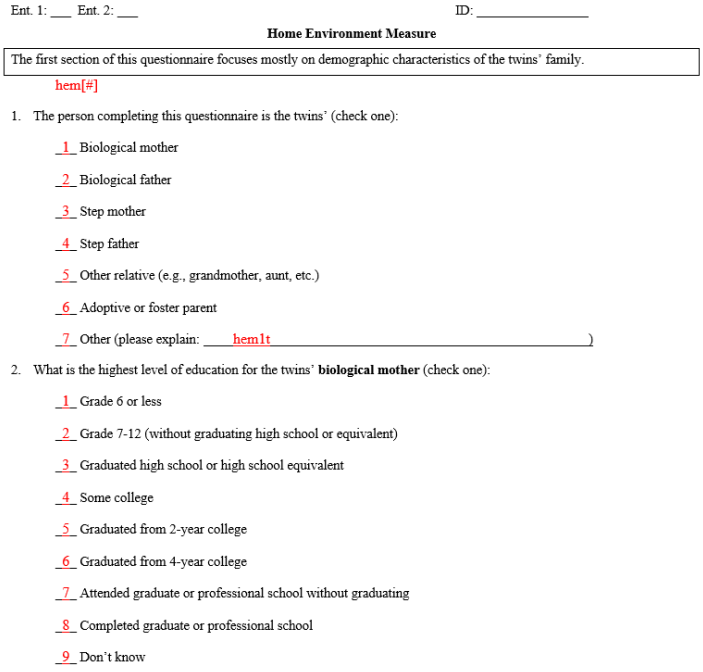
Once your tool is created, you will want to pilot the instrument with your team for data issues (see Figure 10.5). Using the feedback collected, edit your tool as needed before sending it out into the field.

Last, all data that is collected on paper will need to be entered into an electronic format. While we will talk about data entry specifically in Chapter 11, this point in instrument creation is a great time to create an annotated instrument that can be used for future data entry (Neild, Robinson, and Agufa 2022). This includes taking a copy of your instrument and writing the associated codes alongside each item (i.e., variable name, value codes). This can be useful during the data entry process and serve as a linking key between your instrument and your data dictionary (see Figure 10.7) (Hart, Schatschneider, and Taylor 2018).

Another option for capturing paper data in an electronic format is using TeleForm rather than paper. These forms are designed to be scanned by machines. Using TeleForm removes the extra step of future data entry and errors associated with human entry.

10.3.4.3 Identifiers

When building data collection tools, no matter if they are paper or electronic, it is vitally important to make sure you are collecting unique identifiers (Kopper and Parry 2021). Whether you have participants enter a unique identifier into a form or you link study ID to each form in some other way, it's important to not



Ent. 1: ___ Ent. 2: ___ ID: _____

Home Environment Measure

The first section of this questionnaire focuses mostly on demographic characteristics of the twins' family.

hem[#]

1. The person completing this questionnaire is the twins' (check one):

_1_ Biological mother

_2_ Biological father

_3_ Step mother

_4_ Step father

_5_ Other relative (e.g., grandmother, aunt, etc.)

_6_ Adoptive or foster parent

_7_ Other (please explain: ____hem1t_____)

2. What is the highest level of education for the twins' **biological mother** (check one):

_1_ Grade 6 or less

_2_ Grade 7-12 (without graduating high school or equivalent)

_3_ Graduated high school or high school equivalent

_4_ Some college

_5_ Graduated from 2-year college

_6_ Graduated from 4-year college

_7_ Attended graduate or professional school without graduating

_8_ Completed graduate or professional school

_9_ Don't know

Figure 10.7: Annotated instrument from The Florida State Twin Registry project

accidentally collect anonymous data. Without unique identifiers in your data, you will be unable to link data across time and forms. If possible, you want to avoid collecting names as unique identifiers for the following reasons (McKenzie 2010):

- To protect confidentiality we want to use names as little as possible on forms
  - If they are used on forms, we want to remove them as soon as possible
- Names are not unique
  - If you do collect names, you'll want to ask for additional identifying information that when combined, make a participant unique (e.g., student name and email)
- Names change (e.g., someone gets married/divorced)
- There is too much room for error
  - If names are hand entered, there are endless issues with case sensitivity, spelling errors, special characters, spacing, and so forth

All of the above issues make it very difficult to link data. If you do decide to collect names, remember that you will need to remove names during data processing and replace them with your unique study identifiers.

Figure 10.8 shows what a data de-identification process looks like (O'Toole et al. 2018). Dataset 1 would be the incoming survey data with identifiers, Dataset 2 would be a roster exported from your participant database (see Chapter 9), and Dataset 3 is your clean, de-identified dataset, created by merging Dataset 1 with Dataset 2 on your unique identifier and dropping your identifying variables. I want to emphasize the importance of using a "merge" which we will discuss more in Chapter 13, as opposed to replacing names with IDs by hand entering identifiers. If at all possible, we want to completely avoid hand entry of study IDs. Hand entry is error-prone and can lead to many mistakes.

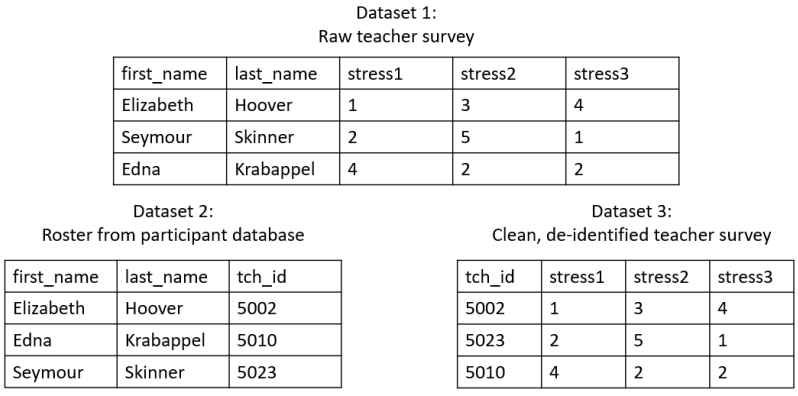


Figure 10.8: Process of creating a de-identified dataset

Rather than having to de-identify your data through this cleaning process, an-

other option is to collect a different type of unique identifier, or pre-link unique study identifiers and names in your instrument, removing many of the issues above (DIME Analytics 2021a; Gibson and Louw 2020). We will discuss these methods separately for electronic data and paper data below.

> **Note** If your study is designed to collect anonymous data, then you will not assign study identifiers and no participant identifying information should be collected in your instruments (e.g., name, email, date of birth). You will also want to make sure to not collect IP addresses as they can be used to identify an individual's computer. It's also important to recognize that if you collect anonymous data, you will not be able to link data across measures or across time. However, if your study randomizes participants by an entity (e.g., school or district), you will need to collect identifying information about that entity in order to cluster on that information (e.g., school name).

10.3.4.3.1 Electronic Data There are many ways you might consider collecting unique identifiers other than names. A few possible options are provided below. The method you choose will depend on your data collection design, your participant population, your tool capabilities, and your team expertise.

1. Create unique links for participants
    - Many tools will allow you to preload a contact list of participants (from your participant database) that includes both their names and study IDs. Using this list, the tool can create unique links for each participant. This is the most error-proof way to ensure study IDs are entered correctly.
    - When you export your data, the correct ID is already linked to each participant and you can choose to not export names in the data.
    - Make sure to build a data check into the system. When a participant opens their unique link, verify their identity by asking, "Are you {first name}?" or "Are your initials {initials}?". In order to protect other participant identities, do not share full names.
        - If they say yes, they move forward. If they say no, the system redirects them to someone to contact. This ensures that participants are not completing someone else's survey and IDs are connected to the correct participant.
2. Provide one link to all participants and separately, in an email, in person, or by mail, provide participants with their study ID to enter into the system.
    - This might be a preferred method if you are conducting surveys or assessments in a computer lab or on tablets at a school site
    - This can possibly introduce error if a participant enters study ID incorrectly.
        - Similar to the first option, after a participant enters their ID,

verify their identity by asking, "Are you {first name}?".
    - Note that participants are only becoming aware of their own study identifier, not the identifiers associated with other participants. However, if your team, or your IRB, is uncomfortable with participants knowing their study IDs you can also consider using a "double ID" which is yet another set of unchanging unique identifiers that you use for the sole purpose of data collection. Those identifiers will need to be tracked in your participant tracking database and will need to be replaced with study IDs in the clean data
3. If you have not previously assigned study identifiers (i.e., your consent and assent process is a part of your instrument), you can have participants enter their identifying information (e.g., name) and then have the tool assign a unique identifier to the participants
    - Using this method, you can potentially download two separate files
        - One with just the instrument data and assigned study ID, with name removed
        - One with just identifying information and assigned study ID (this information will be added to your participant tracking database)

10.3.4.3.2 Paper Data If you take paper forms into the field consider doing the following to connect your data to a participant (O'Toole et al. 2018; Reynolds, Schatschneider, and Logan 2022).

- Write the study ID, and any other relevant identifiers (e.g., school ID and teacher ID), on each page of your data collection form and then use either a removable label with participant name and other relevant information and place that over the ID or attach a cover sheet with this information. When you return to the office, you can remove the name label/cover sheet and be left with only the ID on the form.
    - It is this ID only that you will enter into your data entry form during the data capture process, no name.
    - Removing the label/cover sheet also ensures that your data entry team only sees the study ID when they enter data, increasing privacy by minimizing the number of people who see see participant name.
    - It is important to double and triple check study identifiers against your participant database to make sure the information is correct before removing the label or cover sheet
    - Make a plan for the labels/cover sheets (either shred them if they are no longer needed, or store them securely in a locked file cabinet and shred them at a later point)

10.4 Quality Control

In addition to implementing quality assurance measures during your planning phases, it is equally important to implement several quality control measures

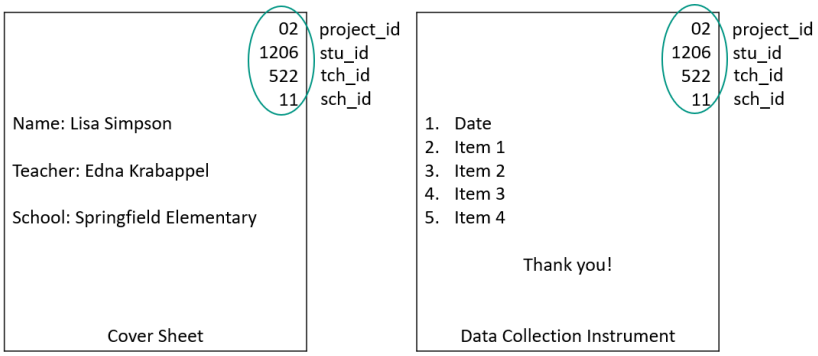


Figure 10.9: Example cover sheet for a paper data collection instrument

while data collection is underway. Those measures include:

1. Field data management
2. Ongoing data checks
3. Tracking data collection daily
4. Collecting data consistently

We will discuss each of these issues below.

10.4.1 Field data management

If your data collection efforts include field data collection (e.g., data collectors administering assessments in a school), there are several steps your team can implement that will keep your data more secure in the field, help a project coordinator keep better track of what happens in the field, and will lead to more accurate and usable data. Some best practices for field data collection include the following (DIME Analytics 2021a):

- Keep your data secure in the field
  - Make sure all paper forms are kept in a folder (or even a lock box) with you at all times and that they are promptly returned to the office (e.g., not left in a car, not left at someone's home)
  - Make sure all data collection devices (e.g., phone, tablets) are password protected and never left open and unattended. Keep all identifiable information encrypted on your in field devices (i.e., data is encoded so that only those with a password can decipher it). You may also consider remote wiping capabilities on portable devices in the case of loss or theft (O'Toole et al. 2018)
- Create tracking sheets to go out in the field
  - These sheets should include the names and/or identifiers of every participant who data collectors will be collecting data from
  - Next to each participant, include any other relevant information to

    track, such as
      * Was the data collected (i.e., a check box)
      * As well as a notes section to describe any potential issues with the data (e.g., "Student had to leave the classroom halfway through the assessment - only partially completed")
    - This tracking sheet allows the project coordinator to keep track of what is occurring in the field so that information can be accurately recorded in the participant tracking database and forms can be sent back out for completion as needed
- Check paper data in the field
    - Immediately upon completing a form, have data collectors do spot checks. If any problems are found, follow up with the participant for correction if possible.
      * Check for missing data
      * Check for duplicate answers given
      * Check for answers provided outside of the assigned area (e.g., answers written in the margins)
      * Check scoring (e.g., basals and ceilings)
- Assign a field supervisor
    - This person is assigned to
      * Do another round of data checks in the field once the data collector returns paper forms to the on-site central location (e.g., if data collectors have set up in the teacher's lounge)
      * Ensure that all data and equipment is accounted for and returned to the office
      * Is available for trouble shooting as needed
- Do another round of paper data spot checking as soon as the data is returned to the office (see Figure 10.10)
    - The project coordinator may do this round of checking as they are tracking information in the participant database
    - If any issues are found, note that in the tracking database and send the form back out to the field for correction
    - If your paper forms are mailed back to you from participants, rather than returned from field data collectors, it is still important to do those in-office spot checks. If at all possible, reach out to those participants for any corrections.
- When data collection wraps up for that period, collect feedback from data collectors to improve future data collection efforts
    - What went well? What didn't?

10.4.2 Ongoing data checks

If you are collecting data via a web-based form, you will want to perform frequent data quality checks, similar to the checks you performed during the content and data piloting phase. You will want to check for both programming errors (i.e., skip logic programmed incorrectly) as well as response quality errors (e.g. bots,

Figure 10.10: A series of spot checks that occur with paper data

survey comprehension) (DIME Analytics 2021a; Gibson 2021).

- Checks for bots/fraud
  - Are there surveys completed in a very short period of time?
  - Are there surveys in suspicious geolocations?
    - This information may not be available for anonymous data - consult with your IRB
  - Are there nonsensical responses for open-ended questions?
  - Are there nonsensical responses to attention or logic checking questions?
- Checks for comprehension
  - Are any questions being misinterpreted?
- Checks for missing data
  - Are items being skipped that should not be skipped?
  - Are surveys not being completed?
- Checks for ranges and formats
  - Are values in unexpected formats or falling outside of unexpected ranges?
- Checks for duplicate surveys
  - Are there duplicate surveys for participants?
- Is skip logic working as expected?
  - Are people being directed to the correct location based on their responses to items?

Some of these checks can be performed programatically (i.e., you can write a validation script in a program such as R, and run that script on a recurring schedule during data collection to check for things such as values out of range). Other checks may be a manual check of data (e.g., such as downloading your data on a recurring schedule and reviewing open-ended questions for nonsensical responses). If errors are found, consider revising your instrument to prevent future errors if this is possible without jeopardizing the consistency of your data.

10.4.3 Tracking data collection

Throughout data collection your team should be tracking the completion of forms (e.g., consents, paperwork, data collection forms). Your team may designate one person to track data (e.g., the project coordinator), or they may designate multiple. If you are working across multiple sites, with multiple teams,

you will most likely have one or more persons at each site tracking data as it comes in.

Some tracking best practices include:

1. Only track data that you physically have (paper or electronic)
    - Never track data as "complete" that someone just tells you they collected
        - You can always mark this information in a "notes" field and track when you have the physical data
2. Track daily during data collection
    - Do not wait until the end of data collection to track what data was collected
    - This helps ensure that you don't miss the opportunity to collect data that you *thought* you had but never actually collected
3. Only track complete data as "complete"
    - If a form is only partially completed and you plan to send it back out to the field for completion, mark this in the "notes" but do not mark it as "completed". If you have a "partially completed" option, you can mark this option.

10.4.4 Collecting data consistently

As mentioned in Chapter 8, it's important to collect data consistently for the entire project to ensure interoperability. Keep the following consistent across both time and forms (e.g., Spanish and English version of a form):

- Variable names
    - Use the same names for the same items (and remember it's best to not add a time component to your variable names at this time)
- Variable types
    - If gender is collected as a numeric variable, keep it as a numeric variable
- Value codes
    - Make sure response options are consistently coded using the same values
- Question type and format
    - If a slider question was used for "Percent of time on homework", continue to ask that question using a slider question

Failing to collect your data consistently has many consequences:

1. It can make it difficult or impossible to compare outcomes
2. It makes your work less reproducible
3. It reduces your ability to physically combine data (i.e., you cannot append dissimilar variables)
4. It can lead to errors in interpretation

Last, collecting data consistently also means measuring things in the same way

over time or across forms so that you don't bias your results. The slightest change in item wording or response options can result in dramatic changes to outcomes (ICPSR 2022).

10.5 Review

This chapter is a lot. There is so much to consider, both prior to data collection as well as during data collection, and the information can vary depending on how you collect your data and who you collect your data on. As I've said before, implement what you can. Implementing even some of these practices will lead to higher quality data compared to implementing none of these practices, and that's huge.

It's also important to consider your entire data collection workflow. Recall Figure 5.5, where we visualized a sample process for designing and collecting an online survey. Errors can happen at any point in this process so it is important to consider the entire workflow holistically and integrate both quality assurance and quality control procedures throughout (see Figure 10.11).

Once the workflow is developed, write the specifics of that plan into an SOP (see Chapter 7), including assigning roles and responsibilities for each task in the process. Last, train your team on how to implement the data collection SOP.

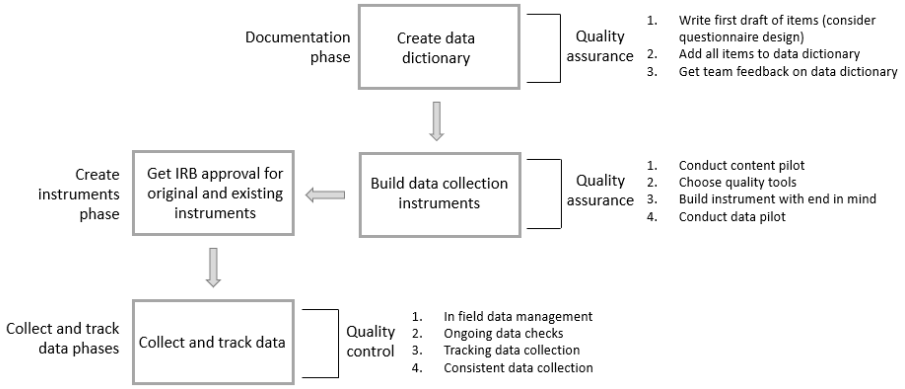


Figure 10.11: Integrating quality assurance and control into a data collection workflow

Instrument Workflow Resources

Source	Resource
DIME Wiki	Questionnaire design timeline[21]

[21] https://dimewiki.worldbank.org/Questionnaire_Design

Source	Resource
Sarah Kopper, Katie Parry	Five key steps in the process of survey design[22]

Note All of the web-based data collection efforts in this section assume you are making a private link that you are sharing with a targeted list (e.g., students in a classroom, teachers in a school). However, there may be times when you need to publicly recruit and collect data for your study and this opens your instrument up for a plethora of data quality issues. Bots, fraudulent data, and incoherent or synthetic responses are all issues that can plague your online data collection efforts, particularly with crowdsourcing platforms (Douglas, Ewell, and Brauer 2023; Veselovsky, Ribeiro, and West 2023; Webb and Tangney 2022). If possible though, avoid using public survey links. One possible workaround would be to first create a public link with a screener. Then after participants are verified through the screener, send a private, unique link. If a workaround is not possible and you need to use a public link, some suggestions that can help you both secure your instrument and detect fraud include the following (Simone 2019; Teitcher et al. 2015):
- Not posting the link on social media
- Using CAPTCHA verification
- Using tools that allow you to block suspicious geolocations
- Not automating payment upon survey completion
- Including open-ended questions
- Building attention/logic checks into the survey
- Asking some of the same questions twice (once early on and again at the end) Even with these additions to the survey, you will want to check your data thoroughly before analyzing it and before providing payments to participants.

[22] https://www.povertyactionlab.org/resource/survey-design

Chapter 11

Data Capture

11.1 Electronic data capture

11.2 Paper data capture

11.3 Extant data

Chapter 12

Data Storage and Security

Chapter 13

Data Cleaning

Chapter 14

Data Sharing

Chapter 15

Wrapping It Up

15.1 Connecting practices to outcomes

15.2 Putting in the work

Commonly used data management terms

Terms (Other Terms)	Definitions
Anonymous Data	Identifying information that was never collected. This data can not be linked across time or measures.
Append	Stacking datasets on top of each other (matching variables).
Archive	The transfer of data to a facility, such as a repository, that preserves and stores data long-term.
Attrition	The loss of study units from the sample, often seen in longitudinal studies
Clean data (processed data)	Data that has been manipulated or modified.
Cohort	A group of participants recruited into a study at the same time.

Terms (Other Terms)	Definitions
Confidential data (pseuodonymization, coded data, indirectly identifiable)	The status of this data is protected. Personally identifiable information (PII) in your data has been removed and names are replaced with a code and the only way to link the data back to an individual is through that code. The identifying code file (linking key) is stored separate from the research data.
Confidentiality	Confidentiality concerns data, ensuring participants agree to how their private and identifiable information will be managed and disseminated.
Control (business as usual)	The individual or group does not receive the intervention.
Cross-sectional	Data is collected on participants for a single time point.
Data (research data)	The recorded factual material commonly accepted in the scientific community as necessary to validate research findings. (OMB Circular A-110)
Data Type (measurement unit, variable format, variable class)	A classification that specifies what types of values are contained in a variable and what kinds of operations can be performed on that variable. Examples of types include numeric, character, logical, or datetime.
Database (relational database)	An organized collection of related data stored in tables that can be linked together by a common identifier.
Dataset (dataframe, spreadsheet)	A structured collection of data usually stored in tabular form. A research study usually produces one final dataset per entity/unit (ex: teacher dataset, student dataset).
De-identified data (anonymized data)	Identifying information has been removed or distorted and the data can no longer be re-associated with the underlying individual (the linking key no longer exists).
Derived data	Data created through transformations of existing data.

Terms (Other Terms)	Definitions
Direct identifiers (PII, PHI)	These identifiers can directly identify a participant and should always be removed from research study data. There should be no need to keep these identifiers for analysis (i.e. name, email, address).
Directory (file structure, file tree)	A cataloging structure for files and folders on your computer.
Experimental data	Data collected from a study where researchers randomly introduce an intervention and study the effects.
Extant data (secondary data)	Existing data sources created from external to the research team/study such as administrative data.
File formats	Education research data is typically collected in one of three file formats: text(.txt, .pdf, .docx), tabular (.xlsx, .csv, .sav) , multimedia (.mpeg, .wav).
Identifiable data	Data that includes personally identifiable information.
Indirect identifiers	Even though these identifiers are not necessarily uniquely tied to one individual (i.e., birthdate or place of birth), if combined, this information could indirectly identify a participant. Therefore this information should be managed before publicly sharing data.
Longitudinal	Data is collected on participants over a period of time.
Merge (join, link)	Combining datasets together in a side by side manner (matching on an identifier).
Missing data	Occurs when there is no data stored in a variable for a particular observation/respondent.
Observational data	Data collected from a study where researchers are observing the effect of an intervention without manipulating who is exposed to the intervention. This includes many formats that education researchers collect data with (ex: survey, observation, assessment).

Terms (Other Terms)	Definitions
Participant database (study roster, master list, master key, linking key, code key, key code, main list, identifiers dataset, crosswalk, record keeping, tracking, participant tracking)	This database, or spreadsheet, includes any identifiable information on your participants as well as their assigned study ID. It is your only own means of linking your confidential research study data to a participant's true identity. It is also used to track data collected across time and measures as well as participant attrition.
Path (file path)	A string of characters used to locate files in your directory system.
Personally identifiable information (PII, PHI)	This includes direct identifiers (e.g., name and email), as well as indirect identifiers that, if combined with other variables, could identify a participant (e.g., full birthdate and county of residence). Under FERPA, additional PII, such as a district or school ID, should also be removed. Protected health identifiers (PHI) is a similar protected category of information. There are 18 HIPAA protected health identifiers that should be removed from data in order to meet the Safe Harbor de-identification method (e.g., name, email, address).
Primary data (original data)	First hand data that is generated/collected by the research team as part of the research study.
Privacy	Privacy concerns people, ensuring they are given control to the access of themselves and their information.
Private data	Highly restricted data with limited access (i.e. passwords)
Qualitative data	Non-numeric data typically made up of text, images, video, or other artifacts.
Quantitative data	Numerical data that can be analyzed with statistical methods.

Terms (Other Terms)	Definitions
Randomized controlled trial (RCT)	A study design that randomly assigns participants to a control or treatment condition. In education research you often hear about two types of RCTs. The first being the Individual-Level Randomized Controlled Trial (I-RCT) in which individuals (such as students) are randomized directly to the treatment or control group. The second is a Cluster Randomized Controlled Trial (C-RCT), sometimes also called group-randomized, in which clusters of students (such as classrooms) are randomized.
Raw data	Unprocessed data collected directly from a source.
Replicable	Being able to produce the same results if the same procedures are used with different materials.
Reproducible	Being able to produce the same results using the same materials and procedures.
Research	The Common Rule definition of research is a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge.
Secondary data (extant data)	Existing data generated/collected by external organizations such as governments at an earlier point in time.
Sensitive data	Private information that could cause harm and should be protected from unwarranted disclosure
Simulation data	Data generated through imitations of a real-world process using computer models.
Standardization	Developing a set of agreed upon technical standards and applying them within and across all research projects.
Study	a single funded research project resulting in one more more datasets to be used to answer a research question.

Terms (Other Terms)	Definitions
Study ID (participant ID, location ID, site ID, unique identifier (UID), subject ID, participant code, record id)	This is a numeric or alphanumeric identifier that is unique to every participant, site or object in order to create confidential and de-identified data. These identifiers allow researchers to link data across time or measure.
Subject (case, participant, site, record)	A person or place participating in research and has one or more piece of data collected on them.
Syntax (code, program)	Programming statements written in a text editor. The statements are machine-readable instructions processed by your computer.
Treatment (experiment)	The individual or group receives the intervention.
Variable (column, field, question)	Any phenomenon you are collecting information on/trying to measure. These variables will make up columns in your datasets or databases.
Variable name (header)	A shortened symbolic name given the variable in your data to represent the information it contains.
Wave (time period, time point, event, session)	Intervals of data collection over time.

Chapter 16

Call to Action

Chapter 17

Glossary

Aczel, Balazs. n.d. "A Crowdsourced Effort to Develop a Lab Manual Template." *Google Docs*. Accessed January 12, 2023. https://docs.google.com/document/d/1LqGdtHg0dMbj9lsCnC1QOoWzIsnSNRTSek6i3Kls2Ik.

Alexander, Rohan. 2023. *Telling Stories with Data*. https://tellingstorieswithdata.com/.

Alston, Jesse M., and Jessica A. Rick. 2021. "A Beginner's Guide to Conducting Reproducible Research." *The Bulletin of the Ecological Society of America* 102 (2). https://doi.org/10.1002/bes2.1801.

Arslan, Ruben C. 2018. "How to Automatically Document Data with the Codebook Package to Facilitate Data Re-Use." Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/5qc6h.

Ashcraft, Alvin. 2022. "Naming Files, Paths, and Namespaces - Win32 Apps." https://learn.microsoft.com/en-us/windows/win32/fileio/naming-a-file.

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533 (7604): 452–54. https://doi.org/10.1038/533452a.

Beals, Laura, and Noah Schectman. 2014. "Data Formatting for Performance Management Systems." *AEA 365*. https://aea365.org/blog/laura-beals-and-noah-schectman-on-data-formatting-for-performance-management-systems/.

Bellevue College. n.d. "Elements of Informed Consent." https://www.bellevuecollege.edu/wp-content/uploads/sites/38/2016/03/Elements_of_Informed_Consent.doc.

BIDS-Contributors. 2022. "The Brain Imaging Data Structure (BIDS) Specification," October. https://doi.org/10.5281/ZENODO.3686061.

Bochove, Kees van, Pinar Alper, and Wei Gu. n.d. "Data Quality." Accessed June 12, 2023. https://rdmkit.elixir-europe.org/data_quality.

Bolam, Mike. n.d. "Guides: Metadata & Discovery @ Pitt: Metadata Standards." Accessed January 18, 2023. https://pitt.libguides.com/metadatadiscovery/metadata-standards.

Bordelon, Dominic. n.d.a. "Guides: Research Data Management @ Pitt:

Describing Data." Accessed March 14, 2023. https://pitt.libguides.com/managedata/describingdata.

———. n.d.b. "Guides: Research Data Management @ Pitt: Understanding Research Data Management." Accessed October 13, 2022. https://pitt.libguides.com/managedata/understanding.

Borghi, John, and Ana Van Gulick. 2021. "Data Management and Sharing: Practices and Perceptions of Psychology Researchers." *PLOS ONE* 16 (5): e0252047. https://doi.org/10.1371/journal.pone.0252047.

———. 2022. "Promoting Open Science Through Research Data Management." *Harvard Data Science Review*, July. https://doi.org/10.1162/99608f92.9497f68e.

Borycz, Joshua. 2021. "Implementing Data Management Workflows in Research Groups Through Integrated Library Consultancy." *Data Science Journal* 20 (1): 9. https://doi.org/10.5334/dsj-2021-009.

Bourgeois, David. 2014a. "Chapter 4: Data and Databases," February. https://pressbooks.pub/bus206/chapter/chapter-4-data-and-databases/.

———. 2014b. *Information Systems for Business and Beyond*. Published through the Open Textbook Challenge by the Saylor Academy. https://pressbooks.pub/bus206/.

Briney, Kristin. 2015. *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*. Research Skills Series. Exeter, UK: Pelagic Publishing.

Briney, Kristin, Heather Coates, and Abigail Goben. 2020. "Foundational Practices of Research Data Management." *Research Ideas and Outcomes* 6 (July): e56508. https://doi.org/10.3897/rio.6.e56508.

Broman, Karl W., and Kara H. Woo. 2018. "Data Organization in Spreadsheets." *The American Statistician* 72 (1): 2–10. https://doi.org/10.1080/00031305.2017.1375989.

Buchanan, Erin M., Sarah E. Crain, Ari L. Cunningham, Hannah R. Johnson, Hannah Stash, Marietta Papadatou-Pastou, Peder M. Isager, Rickard Carlsson, and Balazs Aczel. 2021. "Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set." *Advances in Methods and Practices in Psychological Science* 4 (1): 251524592092800. https://doi.org/10.1177/2515245920928007.

Burnard, Lou. 2014. *What Is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. OpenEdition Press. https://doi.org/10.4000/books.oep.426.

Butters, Oliver W, Rebecca C Wilson, and Paul R Burton. 2020. "Recognizing, Reporting and Reducing the Data Curation Debt of Cohort Studies." *International Journal of Epidemiology* 49 (4): 1067–74. https://doi.org/10.1093/ije/dyaa087.

Cabrini University. n.d. "Submissions for Research Protocol." Accessed April 27, 2023. https://www.cabrini.edu/about/departments/academic-affairs/institutional-review-board/submissions-for-research-protocol.

Cakici, Tatiana Baquero. 2017. "Folders v. Metadata in SharePoint Document Libraries." *Enterprise Knowledge*. https://enterprise-knowledge.com/

folders-v-metadata-sharepoint-document-libraries/.

Campos-Varela, Isabel, and Alberto Ruano-Raviña. 2019. "Misconduct as the Main Cause for Retraction. A Descriptive Study of Retracted Publications and Their Authors." *Gaceta Sanitaria* 33 (4): 356–60. https://doi.org/10.1016/j.gaceta.2018.01.009.

CDISC. n.d. "CDISC Standards in the Clinical Research Process." Accessed January 12, 2023. https://www.cdisc.org/standards.

Center for Open Science. 2022. "COS Engagement with the Education Community." https://docs.google.com/presentation/d/1LpyVOj8oJPr3SVkRM2GfCFnl2Qeo10YbbqcqwtwrVUM.

———. n.d. "Creating a Data Management Plan (DMP) Document - OSF Support." Accessed January 9, 2023. https://help.osf.io/article/144-creating-a-data-management-plan-dmp-document.

CESSDA. n.d.a. "Data Authenticity - Data Management Expert Guide." Accessed January 13, 2023. https://dmeg.cessda.eu/Data-Management-Expert-Guide/3.-Process/Data-authenticity.

———. n.d.b. "Documentation and Metadata." Accessed February 9, 2023. https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata.

Ceviren, A. Busra, and Jessica Logan. 2022. "Ceviren_logan_EHE_forum_2022.pdf." Presentation. https://doi.org/10.6084/m9.figshare.19514368.v1.

Chen, Lu. 2022. "Database Normalization Description - Office." https://learn.microsoft.com/en-us/office/troubleshoot/access/database-normalization-description.

Cofield, Melanie. n.d. "LibGuides: Metadata Basics: Key Concepts." Accessed February 10, 2023. https://guides.lib.utexas.edu/metadata-basics/key-concepts.

Cohen, Louis, Lawrence Manion, and Keith Morrison. 2007. *Research Methods in Education*. 0th ed. Routledge. https://doi.org/10.4324/9780203029053.

Connor, Kathryn M., and Jonathan R. T. Davidson. 2003. "Development of a New Resilience Scale: The Connor-Davidson Resilience Scale (CD-RISC)." *Depression and Anxiety* 18 (2): 76–82. https://doi.org/10.1002/da.10113.

"CONSORT 2010 Flow Diagram." n.d. Accessed April 14, 2023. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwi23o272_r2AhXjpFsKHb26BkAQFnoECDcQAQ&url=https%3A%2F%2Fwww.mdpi.com%2Fdata%2Fconsort-2010-flow-diagram.doc&usg=AOvVaw1vkBi-6QPcbnv9k40nfXfs.

Cornell University. 2019. "IRB FAQs." https://researchservices.cornell.edu/resources/irb-faqs.

Cowles, Wind. n.d. "Research Guides: Research Data Management at Princeton: Home." Accessed September 15, 2022. https://libguides.princeton.edu/c.php?g=102546&p=665862.

CSP Library Research. n.d. "CSP Library: Zotero Guide: Defining Your Research Workflow." Accessed November 1, 2022. https://library.csp.edu/Zotero/workflow.

Dahdul, Wasila. n.d. "Research Guides: Research Data Management: Describing Data." Accessed January 18, 2023. https://guides.lib.uci.edu/

datamanagement/describe.

Danish National Forum for Research Data Management. n.d. "Metadata - How to FAIR." Accessed January 18, 2023. https://howtofair.dk/how-to-fair/metadata/.

Daskalova, Gergana. n.d. "Coding Etiquette." *Coding Club*. Accessed February 17, 2023. https://ourcodingclub.github.io/tutorials/etiquette/.

DDI Alliance. n.d.a. "Controlled Vocabularies - Overview Table of Latest Versions | Data Documentation Initiative." Accessed January 19, 2023. https://ddialliance.org/controlled-vocabularies.

———. n.d.b. "Mapping to Dublin Core (DDI Version 2)." Accessed January 19, 2023. https://ddialliance.org/resources/ddi-profiles/dc.

Dijk, Wilhelmina van, Christopher Schatschneider, and Sara A. Hart. 2021. "Open Science in Education Sciences." *Journal of Learning Disabilities* 54 (2): 139–52. https://doi.org/10.1177/0022219420945267.

DIME Analytics. 2021a. "Data Quality Assurance Plan." https://dimewiki.worldbank.org/Data_Quality_Assurance_Plan.

———. 2021b. "Survey Pilot." https://dimewiki.worldbank.org/Survey_Pilot.

Doucette, Lise, and Bruce Fyfe. 2013. "Drowning in Research Data: Addressing Data Management Literacy of Graduate Students - PDF Free Download." https://docplayer.net/8853333-Drowning-in-research-data-addressing-data-management-literacy-of-graduate-students.html.

Douglas, Benjamin D., Patrick J. Ewell, and Markus Brauer. 2023. "Data Quality in Online Human-Subjects Research: Comparisons Between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." *PLOS ONE* 18 (3): e0279720. https://doi.org/10.1371/journal.pone.0279720.

Duru, Maya, and Sarah Kopper. 2021. "Grant Proposals." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/resource/grant-proposals.

———. n.d. "Gantt Chart Template." https://www.povertyactionlab.org/sites/default/files/research-resources/rr_grantprop_Template_Gantt_Chart.pdf.

Duru, Maya, and Anja Sautmann. 2023. "Institutional Review Board (IRB) Proposals." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/resource/institutional-review-board-irb-proposals.

Eaker, C. 2016. "What Could Possibly Go Wrong? The Impact of Poor Data Management." *In Federer, L. (Ed.). The Medical Library Association's Guide to Data Management for Librarians*. https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1023&context=utk_libpub.

Farewell, Timothy S. 2018. "My Easy R Script Header Template – Tim Farewell." https://timfarewell.co.uk/my-r-script-header-template/.

Feeney, Laura, Sarah Kopper, and Anja Sautmann. 2022. "Ethical Conduct of Randomized Evaluations." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/resource/ethical-conduct-randomized-evaluations.

Filip, Alena. n.d. "San Jose State University Institutional Review Board: Data

Management Handbook for Human Subjects Research." https://www.sjsu.edu/research/docs/irb-data-management-handbook.pdf.

Fordham University. n.d. "Revised Common Rule for Human Subjects Research." Accessed April 27, 2023. https://www.fordham.edu/academics/research/institutional-review-board/revised-common-rule/.

Foster, Erin D., and Ariel Deardorff. 2017. "Open Science Framework (OSF)." *Journal of the Medical Library Association : JMLA* 105 (2): 203–6. https://doi.org/10.5195/jmla.2017.88.

Fuchs, Siiri, and Mari Elisa Kuusniemi. 2018. "Making a Research Project Understandable - Guide for Data Documentation," December. https://doi.org/10.5281/zenodo.1914401.

Gentzkow, Matthew, and Jesse Shapiro. 2014. "Code and Data for the Social Sciences: A Practitioner's Guide." https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf.

Geraghty, Louise, and Laura Feeney. n.d. "Formalize Research Partnership and Establish Roles and Expectations." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. Accessed January 12, 2023. https://www.povertyactionlab.org/resource/formalize-research-partnership-and-establish-roles-and-expectations.

Gibson, Michael. 2021. "Data Quality Checks." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/resource/data-quality-checks.

Gibson, Michael, and Wim Louw. 2020. "Survey Programming." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/resource/survey-programming.

Gilmore, Rick O., Joy Lorenzo Kennedy, and Karen E. Adolph. 2018. "Practical Solutions for Sharing Data and Materials From Psychological Research." *Advances in Methods and Practices in Psychological Science* 1 (1): 121–30. https://doi.org/10.1177/2515245917746500.

GO FAIR. n.d. "FAIR Principles." Accessed October 21, 2022. https://www.go-fair.org/fair-principles/.

Gonzales, Sara, Matthew B. Carson, and Kristi Holmes. 2022. "Ten Simple Rules for Maximizing the Recommendations of the NIH Data Management and Sharing Plan." *PLOS Computational Biology* 18 (8): e1010397. https://doi.org/10.1371/journal.pcbi.1010397.

Grace-Martin, Karen. 2013. "The Wide and Long Data Format for Repeated Measures Data." *The Analysis Factor*. https://www.theanalysisfactor.com/wide-and-long-data/.

Gueguen, Gretchen. n.d. "New OSF Metadata to Support Data Sharing Policy Compliance." Accessed March 1, 2023. https://www.cos.io/blog/new-osf-metadata-to-support-data-sharing-policy-compliance.

Hansen, Karsten Kryger. 2017. "DataFlowToolkit.dk." https://doi.org/10.5278/16k4-4n24.

Hart, Sara, Chris Schatschneider, and Jeanette Taylor. 2018. "Florida Twin Project on Reading, Behavior, and Environment." http://ldbase.org/projects/c3ed1fba-b1fb-4fd0-89ff-42013957cccf.

Hayslett, Michele. n.d. “LibGuides: Metadata for Data Management: A Tutorial: Basic Elements.” Accessed March 10, 2023. https://guides.lib.unc.edu/metadata/basic-elements.

Holdren, John. 2013. “OSTP Memo: "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.

Hollmann, Susanne, Marcus Frohme, Christoph Endrullat, Andreas Kremer, Domenica D’Elia, Babette Regierer, and Alina Nechyporenko. 2020. “Ten Simple Rules on How to Write a Standard Operating Procedure.” *PLoS Computational Biology* 16 (9): e1008095. https://doi.org/10.1371/journal.pcbi.1008095.

Houtkoop, Bobby Lee, Chris Chambers, Malcolm Macleod, Dorothy V. M. Bishop, Thomas E. Nichols, and Eric-Jan Wagenmakers. 2018. “Data Sharing in Psychology: A Survey on Barriers and Preconditions.” *Advances in Methods and Practices in Psychological Science* 1 (1): 70–85. https://doi.org/10.1177/2515245917751886.

Hubbard, Aleata. 2017. “Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step.” In. https://eric.ed.gov/?id=ED583982.

Huisman, Dena. n.d. “3 Principles of Ethical Research That UX Can Take From Academia.” *Indeed Design*. Accessed April 27, 2023. https://indeed.design/article/three-basic-principles-of-ethical-research-that-ux-can-take-from-academia.

ICPSR. 2011. “Guide to Codebooks 1st Edition.” Ann Arbor, MI. https://www.icpsr.umich.edu/files/deposit/Guide-to-Codebooks_v1.pdf.

———. 2022. “An Introduction to Common Data Elements.” https://www.youtube.com/watch?v=GsnoiPzxC4g.

———. n.d.a. “Guide to Social Science Data Preparation and Archiving: 6th Ed.” Accessed January 9, 2023. https://www.icpsr.umich.edu/web/pages/deposit/guide/.

———. n.d.b. “ICPSR, Data Management, Metadata.” Accessed January 18, 2023. https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/metadata.html.

Institute of Education Sciences. n.d.a. “Frequently Asked Questions About Providing Public Access To Data.” Accessed October 21, 2022. https://ies.ed.gov/funding/datasharing_faq.asp.

———. n.d.b. “IES Funding Opportunities.” Accessed January 9, 2023. https://ies.ed.gov/funding/23rfas.asp.

———. n.d.c. “Standards for Excellence in Education Research.” Accessed October 21, 2022. https://ies.ed.gov/seer/index.asp.

Johns Hopkins Institute for Clinical and Translational Research. n.d. “Data Dictionary/Codebook.” Accessed February 10, 2023. https://ictrweb.johnshopkins.edu/ictr/dmig/Best_Practice/a8376318-ebd6-421f-be63-acf8c88376a1_6342a1c3-1a5d-4287-a46e-374824e3780e.html?v=65849&ip=hpdkvlttuiyioooqhw.

Kaplowitz, Rella, and Jasmine Johnson. 2020. “5 Best Practices for Equitable

and Inclusive Data Collection." *Schusterman Family Philanthropies*. https://www.schusterman.org/article/5-best-practices-for-equitable-and-inclusive-data-collection.

Kline, Melissa. 2018. "A Technical Specification for Psychological Datasets." *Google Docs*. https://docs.google.com/document/d/1u8o5jnWk0Iqp_J06PTu5NjBfVsdoPbBhstht6W0fFp0/edit?usp=embed_facebook.

Koos, Jessica. n.d. "Research & Subject Guides: Research Data Guide: Data Collection and Creation." Accessed April 14, 2023. https://guides.library.stonybrook.edu/research-data/collection.

Kopper, Sarah, and Katie Parry. 2020. "Questionnaire Piloting." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/resource/questionnaire-piloting.

———. 2021. "Survey Design." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/resource/survey-design.

Kovacs, Marton, Rink Hoekstra, and Balazs Aczel. 2021. "The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management." *Advances in Methods and Practices in Psychological Science* 4 (4): 251524592110459. https://doi.org/10.1177/25152459211045930.

Krishna, Vamsi. 2018. "How to Tag Files in Windows for Easy Retrieval." *Make Tech Easier*. https://www.maketecheasier.com/tag-files-in-windows/.

Kush, R. D., D. Warzel, M. A. Kush, A. Sherman, E. A. Navarro, R. Fitzmartin, F. Pétavy, et al. 2020. "FAIR Data Sharing: The Roles of Common Data Elements and Harmonization." *Journal of Biomedical Informatics* 107 (July): 103421. https://doi.org/10.1016/j.jbi.2020.103421.

Lafayette College. n.d. "The Three Types of IRB Review · ." Accessed April 27, 2023. https://irb.lafayette.edu/the-three-types-of-irb-review/.

LDbase. n.d.a. "Data Use Vs Data Sharing | LDbase." Accessed May 1, 2023. https://ldbase.org/resources/best-practices/data-use-vs-data-sharing.

LDbase. n.d.b. "Information to Gather Before Uploading Your Data | LDbase." Accessed January 18, 2023. https://www.ldbase.org/resources/user-guide/information-to-gather.

Lee, Amanda, Sarah Hughes, and Shawn Marsh. 2020. "Considerations for Collecting Electronic Signatures." *The Abdul Latif Jameel Poverty Action Lab (J-PAL)*. https://www.povertyactionlab.org/blog/6-30-20/considerations-collecting-electronic-signatures.

Lewis, Crystal. 2022a. "Using a Data Dictionary as Your Roadmap to Quality Data." *Crystal Lewis*. https://cghlewis.com/blog/data_dictionary/.

———. 2022b. "How to Export Analysis-Ready Survey Data." *Crystal Lewis*. https://cghlewis.com/blog/survey_data/.

———. 2023. "Codebook Package Comparison." https://github.com/Cghlewis/codebook-pkg-comparison.

Logan, Jessica, and Sara Hart. 2023. "Within & Between S4e2." *Within & Between*. http://www.withinandbetweenpod.com/.

Logan, Jessica, Sara Hart, and Christopher Schatschneider. 2021. "Data Sharing in Education Science." *AERA Open* 7 (January): 233285842110064. https://doi.org/10.1177/23328584211006475.

Malow, Beth A., Anjalee Galion, Frances Lu, Nan Kennedy, Colleen E. Lawrence, Alison Tassone, Lindsay O'Neal, et al. 2021. "A REDCap-Based Model for Online Interventional Research: Parent Sleep Education in Autism." *Journal of Clinical and Translational Science* 5 (1): e138. https://doi.org/10.1017/cts.2021.798.

Markowetz, Florian. 2015. "Five Selfish Reasons to Work Reproducibly." *Genome Biology* 16 (1): 274. https://doi.org/10.1186/s13059-015-0850-7.

Mathematica. n.d. "Tips for Conducting Equitable and Culturally Responsive Research." *Mathematica*. Accessed June 12, 2023. https://www.mathematica.org/features/tips-for-conducting-equitable-and-culturally-responsive-evaluation.

McKay Bowen, Claire, and Joshua Snoke. 2023. "Do No Harm Guide: Applying Equity Awareness in Data Privacy Methods." https://www.urban.org/research/publication/do-no-harm-guide-applying-equity-awareness-data-privacy-methods.

McKenzie, Patrick. 2010. "Falsehoods Programmers Believe About Names | Kalzumeus Software." https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/.

Mehr, Samuel. n.d. "How to... Write a Lab Handbook." *RSB*. Accessed January 12, 2023. https://www.rsb.org.uk//biologist-features/how-to-write-a-lab-handbook.

Meyer, Michelle N. 2018. "Practical Tips for Ethical Data Sharing." *Advances in Methods and Practices in Psychological Science* 1 (1): 131–44. https://doi.org/10.1177/2515245917747656.

Michener, William K. 2015. "Ten Simple Rules for Creating a Good Data Management Plan." *PLOS Computational Biology* 11 (10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525.

Microsoft. n.d. "Restrictions and Limitations in OneDrive and SharePoint - Microsoft Support." Accessed February 17, 2023. https://support.microsoft.com/en-us/office/restrictions-and-limitations-in-onedrive-and-sharepoint-64883a5d-228e-48f5-b3d2-eb39e07630fa.

Midgley, Carol. 2000. "Manual for the Patterns of Adaptive Learning Scales." http://websites.umich.edu/~pals/PALS%202000_V13Word97.pdf.

Nahmias, Allison S., Samantha Crabbe, Steven C. Marcus, and David S. Mandell. 2022. "The Effects of Community Preschool Characteristics on Developmental Outcomes for Students With Autism Spectrum Disorder." *Focus on Autism and Other Developmental Disabilities*, November, 108835762211334. https://doi.org/10.1177/10883576221133495.

Narvaiz, Sarah. n.d. "Data Ethics Statement." *Sarah Narvaiz*. Accessed June 12, 2023. https://www.sarahnarvaiz.com/ethics/.

National Center for Education Statistics. n.d.a. "Common Education Data Standards (CEDS)." Accessed April 13, 2023. https://ceds.ed.gov/Default.aspx.

———. n.d.b. "Memoranda of Understand and Other Data Use Agreements." Accessed April 25, 2023. https://nces.ed.gov/forum/dataethicscourse/additional-materials/memoranda-of-understanding.pdf.

National Endowment for the Humanities. 2018. “Data Managment Plans for NEH Office of Digital Humanities Proposals and Awards.” https://www.neh.gov/sites/default/files/2018-06/data_management_plans_2018.pdf.

National Institute of Justice. 2007. “The "Common Rule".” https://nij.ojp.gov/funding/common-rule.

National Institutes of Health. n.d.a. “Budgeting for Data Management & Sharing | Data Sharing.” Accessed January 11, 2023. https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/budgeting-for-data-management-sharing#after.

———. n.d.b. “Common Data Elements: Standardizing Data Collection.” Accessed April 13, 2023. https://www.nlm.nih.gov/oet/ed/cde/tutorial/03-100.html.

———. n.d.c. “Data Management & Sharing Policy Overview | Data Sharing.” Accessed March 13, 2023. https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview.

National Science Foundation. 2023. “NSF Public Access Plan 2.0.” https://nsf-gov-resources.nsf.gov/2023-06/NSF23104.pdf?VersionId=cSTD31SSPUEkM_Vm25HSlgZBDeiPvzdQ.

Neild, R. C., D. Robinson, and J. Agufa. 2022. “Sharing Study Data: A Guide for Education Researchers. (NCEE 2022-004).” *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.* https://ies.ed.gov/ncee/pubs/2022004/pdf/2022004.pdf.

Nelson, Alondra. 2022. “OSTP Memo: "Ensuring Free, Immediate, and Equitable Access to Federally Funded Research.” https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

Nguyen, Kim. 2017. “Relational Database Schema Design Overview.” *Medium.* https://medium.com/@kimtnguyen/relational-database-schema-design-overview-70e447ff66f9.

Nichols Hess, Amanda, and Joanna Thielen. 2017. “Advancing Research Data Management in the Social Sciences: Implementing Instruction for Education Graduate Students into a Doctoral Curriculum.” https://our.oakland.edu/handle/10323/6893.

Northern Illinois University. n.d. “Data Collection.” Accessed April 27, 2023. https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html.

Northwestern University. n.d. “Exempt Review.” Accessed April 27, 2023. https://irb.northwestern.edu/submitting-to-the-irb/types-of-reviews/exempt-review.html.

NUCATS. n.d. “Standard Operating Procedures (SOPs).” Accessed January 13, 2023. https://www.nucats.northwestern.edu/docs/cecd/overview-of-sops.pdf.

O’Toole, Elisabeth, Laura Feeney, Kenya Heard, and Rohit Naimpally. 2018. “Data Security Procedures for Researchers.” J-PAL North America. https://www.povertyactionlab.org/sites/default/files/Data_Security_

Procedures_December.pdf.

Office for Human Research. 2009. "Federal Policy for the Protection of Human Subjects ('Common Rule')." Text. *HHS.gov*. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.

Office for Human Research Office for Human Research. 2018. "Revised Common Rule Q&As." Text. *HHS.gov*. https://www.hhs.gov/ohrp/education-and-outreach/revised-common-rule/revised-common-rule-q-and-a/index.html.

Office for Human Research Protections. 2016. "45 CFR 46." Text. *HHS.gov*. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html.

OpenAIRE_eu. 2018. "Basics of Research Data Management." https://www.youtube.com/watch?v=3sDhQRIYUmA.

Oregon State University. 2012. "What Is the Institutional Review Board (IRB)?" *Research Office*. https://research.oregonstate.edu/irb/what-institutional-review-board-irb.

Pacific University Oregon. 2014. "Data Security and Storage." *Pacific University*. https://www.pacificu.edu/academics/research/scholarship-and-sponsored-projects/research-compliance-integrity/institutional-review-board/irb-policies-recommended-practices/data-security-storage.

Page, Lindsay, Matthew Lenard, and Luke Keele. 2020. "The Design of Clustered Observational Studies in Education." https://doi.org/10.3886/E121381V1.

Patridge, Emily F., and Tania P. Bardyn. 2018. "Research Electronic Data Capture (REDCap)." *Journal of the Medical Library Association : JMLA* 106 (1): 142–44. https://doi.org/10.5195/jmla.2018.319.

Qiao, Haiping. 2018. "A Brief Introduction to Institutional Review Boards in the United States." *Pediatric Investigation* 2 (1): 46–51. https://doi.org/10.1002/ped4.12023.

R Core Team. 2023. "R: The R Base Package." *R Foundation for Statistical Computing*. https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html.

REL West. n.d. "Data Sharing and Memorandums of Understanding: Considerations for Development." https://ies.ed.gov/ncee/rel/regions/west/relwestFiles/pdf/CRP_Data_Sharing_Agreements_and_MOUs.pdf.

Renbarger, Rachel, Jill L. Adelson, Joshua Rosenberg, Sondra M Stegenga, Olivia Lowrey, Pamela Rose Buckley, and Qiyang Zhang. 2022. "Champions of Transparency in Education: What Journal Reviewers Can Do to Encourage Open Science Practices." Preprint. EdArXiv. https://doi.org/10.35542/osf.io/xqfwb.

Reynolds, Tara, Christopher Schatschneider, and Jessica Logan. 2022. "The Basics of Data Management." figshare. https://doi.org/10.6084/m9.figshare.13215350.v2.

Riederer, Emily. 2020. "Column Names as Contracts." *Emily Riederer*. https://emilyriederer.netlify.app/post/column-name-contracts/.

Salfen, Jeremy. 2018. "Building a Data Practice from Scratch." *Locally Optimistic*. https://locallyoptimistic.com/post/building-a-data-practice/.

Samuel J. Wood Library. n.d. "Research Data Management, Retention, and Sharing | Weill Cornell Medicine Samuel J. Wood Library." Accessed January 18, 2023. https://library.weill.cornell.edu/research-support/research-data-management-retention-and-sharing.

San Martin, Luis Eduardo, Rony Rodriguez-Ramirez, and Mizuhiro Suzuki. 2023. "Stata Linter Produces Stata Code That Sparks Joy." https://blogs.worldbank.org/impactevaluations/stata-linter-produces-stata-code-sparks-joy.

Schema.org. n.d. "Schema.org." Accessed January 19, 2023. https://www.schema.org/.

Schulz, Kenneth F., Douglas G. Altman, David Moher, and CONSORT Group. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ (Clinical Research Ed.)* 340 (March): c332. https://doi.org/10.1136/bmj.c332.

Simone, Melissa. 2019. "How to Battle the Bots Wrecking Your Online Study." *Behavioral Scientist*. https://behavioralscientist.org/how-to-battle-the-bots-wrecking-your-online-study/.

Society of Critical Care Medicine. 2018. "Building an Efficient Database for Your Research." https://www.youtube.com/watch?v=9ELr2P2pQZg.

Stanford University. n.d. "Use of Human Subjects in Student Projects, Pilot Studies, Oral Histories and QA/QI Projects | DoResearch." Accessed June 27, 2023. https://doresearch.stanford.edu/policies/research-policy-handbook/human-subjects-and-stem-cells-research/use-human-subjects-student.

Stangroom, Jeremy. 2019. "Rules for Naming Variables in SPSS - Quick SPSS Tutorial." *EZ SPSS Tutorials*. https://ezspss.com/rules-for-naming-variables-in-spss/.

Strand, Julia. n.d. "Error Tight: Exercises for Lab Groups to Prevent Research Mistakes." Accessed October 31, 2022. https://psyarxiv.com/rsn5y/.

"Style Guide." 2023. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Style_guide&oldid=1131556370.

Teitcher, Jennifer E. F., Walter O. Bockting, José A. Bauermeister, Chris J. Hoefer, Michael H. Miner, and Robert L. Klitzman. 2015. "Detecting, Preventing, and Responding to 'Fraudsters' in Internet Research: Ethics and Tradeoffs." *The Journal of Law, Medicine & Ethics : A Journal of the American Society of Law, Medicine & Ethics* 43 (1): 116–33. https://doi.org/10.1111/jlme.12200.

Tenopir, Carol, Suzie Allard, Priyanki Sinha, Danielle Pollock, Jess Newman, Elizabeth Dalton, Mike Frame, and Lynn Baird. 2016. "Data Management Education from the Perspective of Science Educators." *International Journal of Digital Curation* 11 (1): 232–51. https://doi.org/10.2218/ijdc.v11i1.389.

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. "The Belmont Report." https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf.

The Nobles. 2020. "Normalization of Database, the Easy Way." *The Startup*. https://medium.com/swlh/normalization-of-database-the-easy-

way-98f96a7a6863.

The Turing Way Community. 2022. "The Turing Way: A Handbook for Reproducible, Ethical and Collaborative Research." Zenodo. https://doi.org/10.5281/ZENODO.3233853.

The White House. 2013. "Executive Order – Making Open and Machine Readable the New Default for Government Information." *Whitehouse.gov*. https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.

Tourangeau, Karen. 2015. "Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011)." Institute of Education Sciences. https://nces.ed.gov/pubs2015/2015074.pdf.

UC Merced Library. n.d. "What Is a Data Dictionary? | UC Merced Library." Accessed January 17, 2023. https://library.ucmerced.edu/data-dictionaries.

UK Data Service. 2022. "Data Management Costing Tool and Checklist." https://ukdataservice.ac.uk//app/uploads/costingtool.pdf.

———. n.d.a. "Metadata." *UK Data Service*. Accessed January 18, 2023. https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/metadata/.

———. n.d.b. "Quality." Accessed January 13, 2023. https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/quality/.

———. n.d.c. "Roles and Responsibilities." *UK Data Service*. Accessed January 10, 2023. https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/roles-and-responsibilities/.

———. n.d.d. "Versioning." *UK Data Service*. Accessed January 17, 2023. https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/versioning/.

United States Department Of Health And Human Services. 2022. "Study of Coaching Practices in Early Care and Education Settings (SCOPE), United States, 2019: Version 1 SCOPE Data User Guide." ICPSR - Interuniversity Consortium for Political; Social Research. https://doi.org/10.3886/ICPSR38290.V1.

University of California Berkeley. 2022. "Exempt Research." https://cphs.berkeley.edu/exempt.pdf.

University of Iowa Libraries. n.d. "Metadata." Accessed February 10, 2023. https://www.lib.uiowa.edu/data/share/metadata/.

University of Virginia. n.d. "When Consent Is Not Required | Research." Accessed May 3, 2023. https://research.virginia.edu/irb-sbs/when-consent-not-required.

University of Washington. n.d. "Sharing Information and Data." *UW Research*. Accessed April 27, 2023. https://www.washington.edu/research/myresearch-lifecycle/setup/collaborations/sharing-information-and-data/.

U.S. Department of Health and Human Services. n.d. "What's New in IRB Review Under the Revised Common Rule." Accessed April 27, 2023. https://www.youtube.com/watch?v=zDsUUs9j3sQ.

USGS. n.d.a. "Tools for Creating Metadata Records." *USGS*. https://www.usgs.gov/data-management/metadata-creation#tools.

———. n.d.b. "What Are the Differences Between Data, a Dataset, and a Database? | U.S. Geological Survey." Accessed October 17, 2022. https://www.usgs.gov/faqs/what-are-differences-between-data-dataset-and-database.

Valentine, Theresa. n.d. "Best Practice: Define Roles and Assign Responsibilities for Data Management." *DataOne*. Accessed January 10, 2023. https://dataoneorg.github.io/Education/bestpractices/define-roles-and.

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West. 2023. "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks." https://doi.org/10.48550/ARXIV.2306.07899.

Webb, Margaret A., and June P. Tangney. 2022. "Too Good to Be True: Bots and Bad Data From Mechanical Turk." *Perspectives on Psychological Science*, November, 174569162211200. https://doi.org/10.1177/17456916221120027.

White, Ethan, Elita Baldridge, Zachary Brym, Kenneth Locey, Daniel McGlinn, and Sarah Supp. 2013. "Nine Simple Ways to Make It Easier to (Re)use Your Data." *Ideas in Ecology and Evolution* 6 (2). https://doi.org/10.4033/iee.2013.6b.6.f.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (September): 1–23. https://doi.org/10.18637/jss.v059.i10.

———. n.d. *Welcome | The Tidyverse Style Guide*. Accessed February 17, 2023. https://style.tidyverse.org/index.html.

Wickham, Hadley, and Garrett Grolemund. 2017. *Welcome | R for Data Science*. https://r4ds.had.co.nz/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. "Good Enough Practices in Scientific Computing." *PLOS Computational Biology* 13 (6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.

xkcd. n.d. "Documents." Accessed February 16, 2023. https://xkcd.com/1459/.

Zhou, Xuan, Zhihong Xu, and Ashlynn Kogut. 2023. "Research Data Management Needs Assessment for Social Sciences Graduate Students: A Mixed Methods Study." *PLOS ONE* 18 (2): e0282152. https://doi.org/10.1371/journal.pone.0282152.

Table 17.1

nes are replaced with a code and the only way to link the data back to an individual is through that code. The identifying code file (linking
naged and disseminated.
gs. (OMB Circular A-110)
rformed on that variable. Examples of types include numeric, character, logical, or datetime.
r entity/unit (e.g., teacher dataset, student dataset).
g individual (the linking key no longer exists).
hould be no need to keep these identifiers for analysis (i.e. name, email, address).
inistrative data).
ormation in education records (e.g., name, address, DOB). The law applies to all public elementary and secondary schools, as well as post-se
.sav) , multimedia (.mpeg, .wav).
nformation.
ombined, this information could indirectly identify a participant. Therefore this information should be managed before publicly sharing data
isting scale, an existing academic assessment).