Data Management in Large-Scale Education Research

Crystal Lewis

2023-09-26

Contents

Chapter 1

Preamble

This is the in-progress version of *Data Management in Large-Scale Education Research*. When completed, this book will be published by CRC Press. 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, the team has found that of the education researchers surveyed who are currently not publicly sharing their research data, approximately 15% mentioned "being nervous about mistakes" as a reason for not sharing (Beaudry et al. 2022). 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 data, typically collected from human participants, although I do think that many of the practices covered can also apply to other types of data 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, Allyson Hanson, Rohan Alexander, and Peter Higgins. 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 2023). 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 2023). 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 2023), other fields have adopted standards specifically around metadata, such as the TEI standards used in digital humanities (Burnard 2014) or the ISO 19115 standard 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.; 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 5) 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**: Depending on your research design and the sensitivity level of the data you are collecting (see Chapter 4), there are a variety of policies as well as legal or contractual obligations you may

need to consider when managing data. If you are required to submit your project to an Institutional Review Board (see Section 11.2.5), the board will review and monitor your data management practices. Concerned with the welfare, rights, and privacy of research participants, your IRB will have rules for how data is collected, managed, and shared securely (Filip 2023). Your data may also be subject to laws, such as HIPAA or FERPA, which regulate the privacy and exchange of personal information (see Section 4.3). If working with research partners, you may also need to monitor and honor any conditions laid out in data sharing or other legal agreements. 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, 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. 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 our responsibility to have well-designed research studies with data collection, management, ownership, and sharing practices that consider the environmental, social, cultural, historical, and political context of the data we are working with (Alexander 2023). Furthermore, 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 compromised data is a huge disservice to research participants and erodes trust in the research process (Feeney, Kopper, and Sautmann 2022).

2.3.2 Personal reasons

There are also many compelling personal 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 docu-

mented 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 reuse 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, ensure 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; Princeton University 2023b). 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 2021). 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 assurance and 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 were published in Scientific Data (Wilkinson et al. 2016), 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. 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

The SEER principles, developed in 2018 by Institute of Education Sciences (IES), provide Standards for Excellence in Education Research (Institute of Education Sciences 2022). 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:

- Preregister 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 proponent 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 preregistration of study methods.

> **Note** When working with specific populations, there may be other principles to consider that complement FAIR principles and open science practices and provide further guidance for working with and protecting data collected from those specific communities. As an example, when conducting research with Indigenous populations, it is important to consider Indigenous data sovereignty which recognizes the rights of Indigenous peoples to own, control, access, and use data collected about their communities and lands, and to engage Indigenous communities when planning data management for your study (Carroll et al. 2020; National Institutes of Health 2022)

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 Glossary to gain a better understanding of how terms will be used throughout this book.

2.6 The Research Life Cycle

The remainder of this book will be organized into chapters that provide best practices for each phase of the research data life cycle. It is imperative to understand this life cycle in order to see the flow of data through a project, as well as to understand 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.

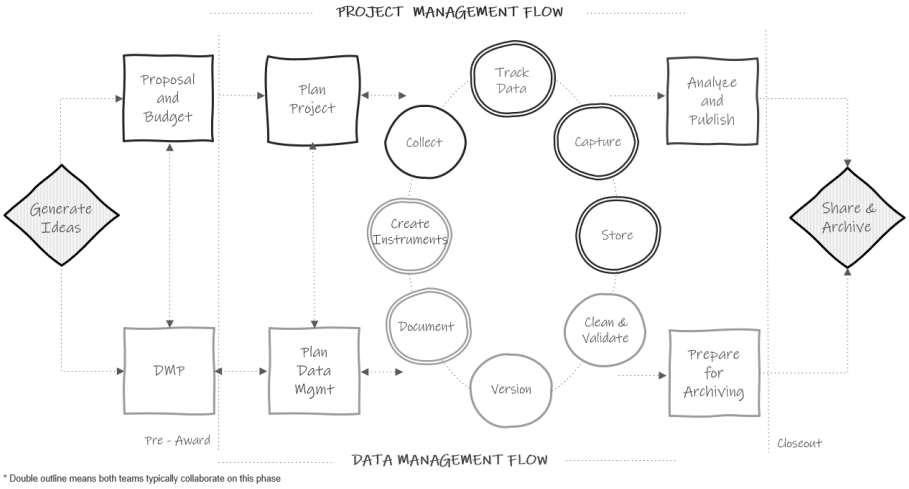


Figure 2.1: The research project life cycle

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 5) 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 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 long-term storage and 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

Before we jump into the data life cycle, 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 information.

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, data files), represented as text, numbers, or multimedia (USGS 2023). 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 the following types of variables:

- Variables you collect (from an instrument or from an external source)
- Variables you create (e.g., cohort, intervention, time, derivations)

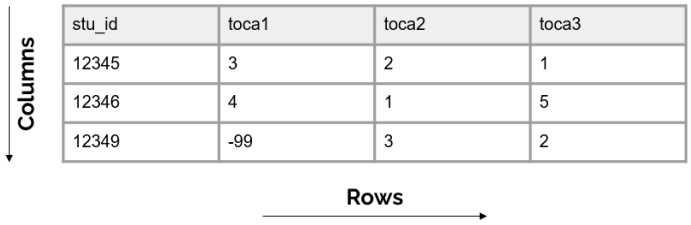


Figure 3.1: Basic format of a dataset

- Unless your data is collected anonymously, one of these variables must include values that uniquely identify subjects in your data (e.g., a student unique identifier).

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 in Chapter 9.
2. A measurement type
   - Examples include numeric, character, or date, which can also be more narrowly defined as needed (e.g., continuous, categorical)
3. Acceptable values
   - Examples include categorical 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 it may be the actual wording of an item (e.g., "Do you enjoy pizza?").

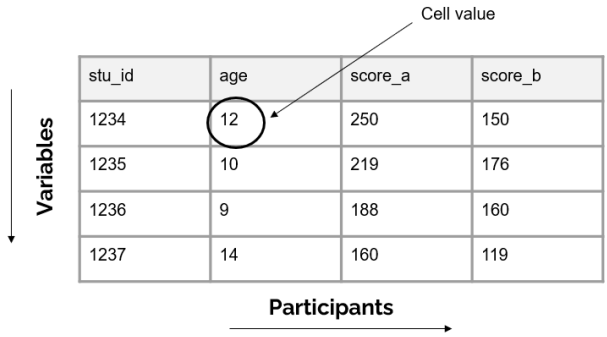
3.1.2 Rows

The rows in your dataset are aligned with subjects, or cases, in your data. Subjects in your dataset may be students, teachers, schools, locations, and so forth. The unique subject identifier variable mentioned above will denote which row belongs to which subject.

3.1.3 Cells

The cells are the observations associated with each case in your data. 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 (i.e., that participant and that question).



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

Figure 3.2: 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.

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

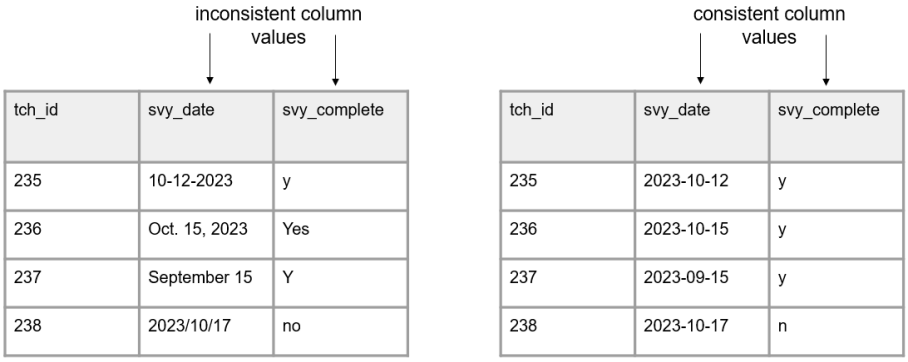
rectangle

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

2. Column values should be consistent. Both humans and machines have difficulty categorizing information that is not measured, coded, or formatted consistently.

- For text categorical values, use controlled vocabulary and keep consistent spelling, case, and spacing
- For date values, keep consistent format
- For numeric values, measure in consistent units and keep consistent decimal places

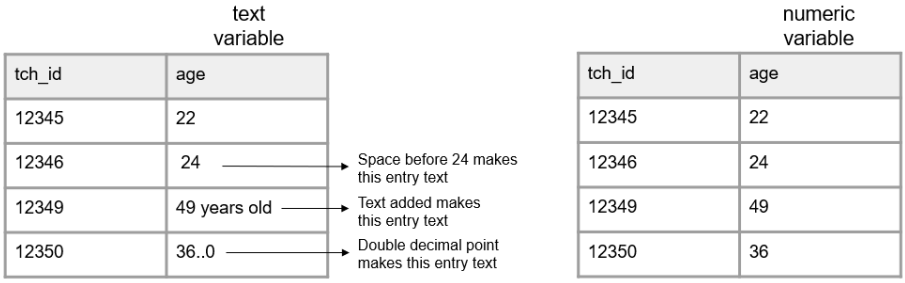


tch_id	svy_date	svy_complete
235	10-12-2023	y
236	Oct. 15, 2023	Yes
237	September 15	Y
238	2023/10/17	no

tch_id	svy_date	svy_complete
235	2023-10-12	y
236	2023-10-15	y
237	2023-09-15	y
238	2023-10-17	n

Figure 3.4: A comparison of inconsistent and uniform variable values

3. 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.



tch_id	age
12345	22
12346	24
12349	49 years old
12350	36..0

tch_id	age
12345	22
12346	24
12349	49
12350	36

Figure 3.5: A comparison of variables adhering and not adhering to a data type

4. 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)
   - 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 3.6: A comparison of two things being measured in one variable and two things being measured across two variables

5. All cell values should be explicit. This means all cells that are not missing values 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 Section 9.5.1 for ideas on coding missing values.
     - 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 an empty cell is implied to be 0, fill the cells with an actual 0

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

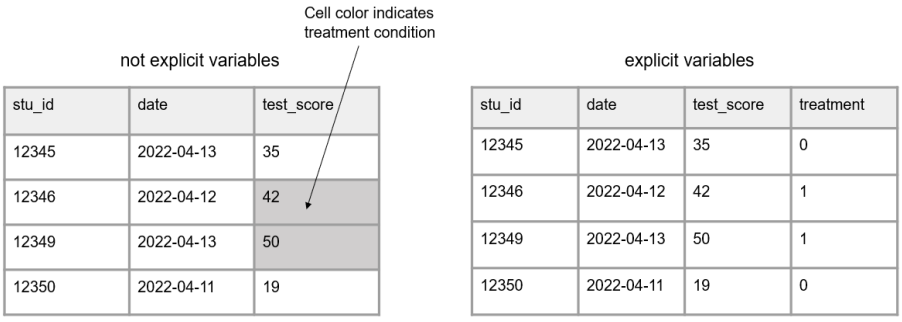
sch_id	year	grade	n_students
204	2020	3	100
204	2020	4	80
204	2020	5	90
205	2020	3	98
205	2020	4	88
205	2020	5	91

Figure 3.7: A comparison of variables with empty cells and variables with not empty cells

6. All variables should be explicit. No variables should be implied using color

coding.

- If you want to indicate information, add an indicator variable to do this rather than cell coloring



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

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 3.8: A comparison of variables with implicit values and variables with explicit values

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. 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, database design and data structure.

3.3.1 Database design

A database is "an organized collection of data stored as multiple datasets" (USGS 2023). Sometimes this database is actually housed in a database software 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 includes one or more variables that uniquely define rows in your data (i.e., a primary key). Tables may also contain variables associated with unique values in another table (i.e., foreign keys). 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 10).

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 (denoted by rectangles).

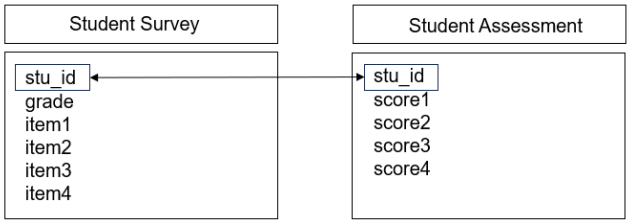


Figure 3.9: Linking data through primary keys

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). 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 being denoted by rectangles) and how each table can be linked together through a foreign key (denoted by ovals).

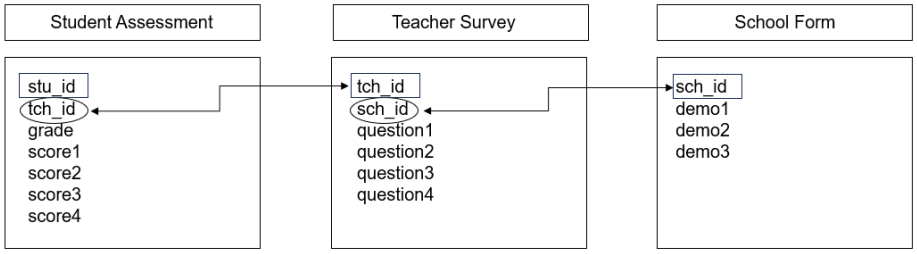


Figure 3.10: Linking data through foreign keys

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.

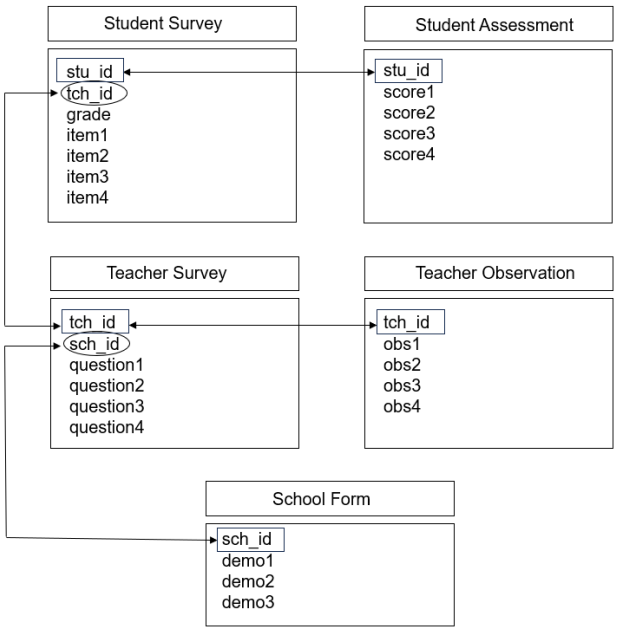


Figure 3.11: Linking data through primary and foreign keys

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 subject will be in one row. Subjects 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 same participant type, within and/or across time
  - This is commonly used to create comprehensive subject-level datasets (e.g., all student data combined into a student-level dataset)
- To link forms across different participant types
  - Such as a student survey and teacher survey

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. This is typically a one-to-one merge where each participant will only appear once in each dataset.

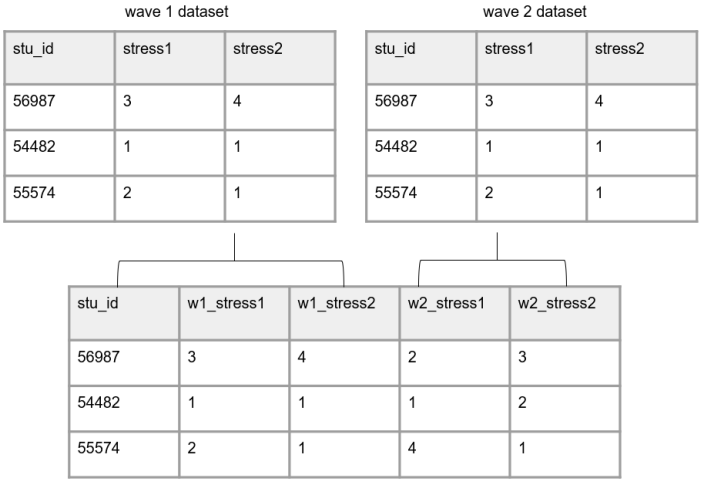


Figure 3.12: Example linking forms across time in wide format

A more complicated scenario is merging across participants (e.g., merging teacher data into student data). This is often a many to one join (e.g., multiple students are associated with the same teacher), meaning upon merging, teacher data will be repeated for all students in their classroom. Although describing all of the different types of joins will be outside the scope of this book, there are many resources available to help you decide which join type is appropriate for your needs [1].

> **Note** It is important to note here, that if your data do not have unique identifiers, as is in the case of anonymous data, you will be unable to merge data in a 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 often will, repeat in your dataset, and unique rows will now be identified through a combination of variables (e.g., `stu_id` and `wave` together will be your primary key).

Again, the most straightforward way to think about this is with repeated measure data, where repeating subject IDs in a row will represent different time points for a participant. Here instead of joining 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 "linked" by variable names. It is important here

[1] https://r4ds.hadley.nz/joins

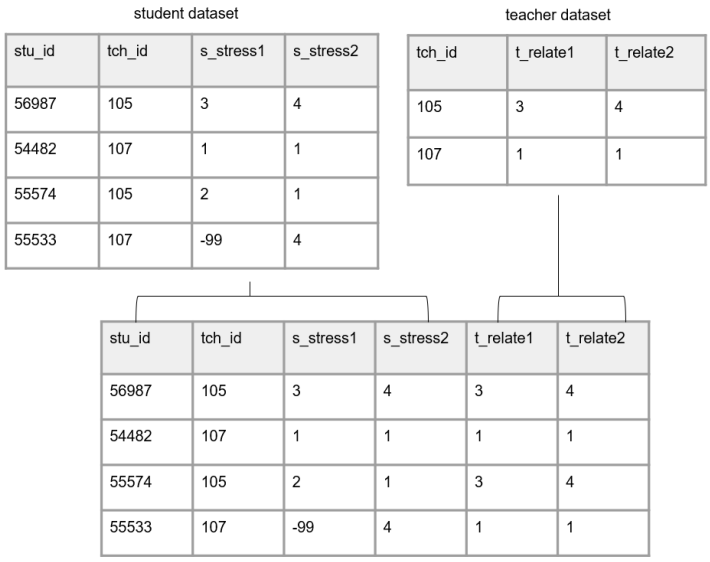


Figure 3.13: Example linking forms across participants in wide format

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.

3.3.2.3 Choosing wide vs long

There are different reasons for structuring your linked data one way or another. Storing linked data in long format is usually considered to be more efficient than storing in wide format, 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). It may be that you structure data in one format for one reason (e.g., storing or sharing), and then restructure data into another format a different reason (e.g., analysis). Luckily, this type of restructuring can be done fairly quickly in many statistical programs. We will further review decision making around data structure in Chapters 14 and 15.

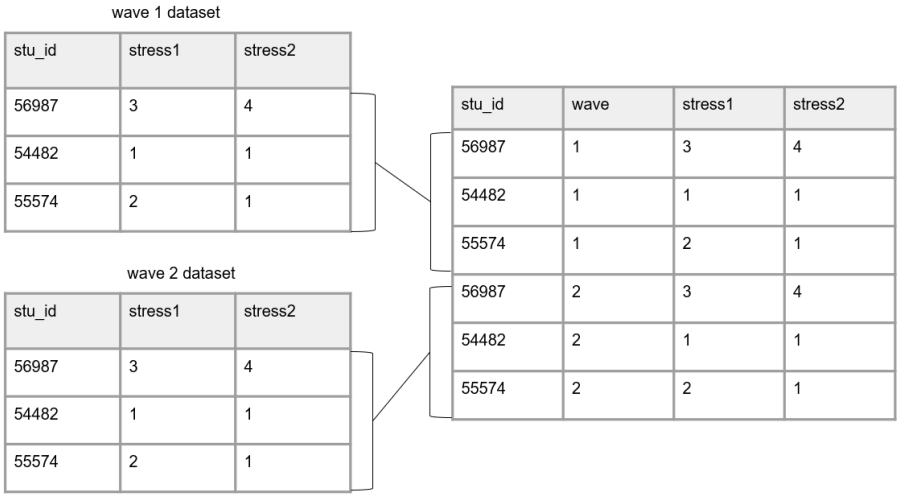


Figure 3.14: Data structured in long format

3.4 File formats

These rectangular datasets can be saved in a variety of file formats. Some common formats used in education research include interoperable formats such as CSV or TSV, or proprietary formats such as XLSX, SPSS, or Stata files.

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 an SPSS file, will greatly increase your file size. We will talk about the pros and cons to different file formats in Chapters 13 and 15.

Chapter 4

Human Subjects Data

In addition to understanding data structure, we also need a foundational understanding of the types of data we may collect. In the field of education research, we are often working with data that is collected from human subjects. Along with collecting data from people, comes the responsibility to secure that data. Data from humans may contain identifiable information increasing the risk that participants can be revealed in a dataset. Human subjects data sometimes also contains information on sensitive topics such as mental health, drug use, or criminal behavior, further increasing risks if participants are identified. Before beginning your project, it is important to assess the type of data you will be collecting and understand the protections that will need to be in place to secure your data. This chapter will briefly review the types of human subject data you may work with as well as any regulations, organizations, policies, or agreements that may impact how you need to secure your data.

4.1 Identifiability of a dataset

When working with human subjects there are two types of identifiers you may collect in your study, direct and indirect. Direct identifiers are unique to an individual and can be used to identify a participant. Indirect identifiers do not alone identify a particular individual, but if combined with other information or if category numbers are small, could be used to identify a participant.

A term often used when discussing identifiable information is personally identifiable information (PII). This term broadly refers to information that can be used to identify a participant. There is no agreed upon list for what fields should be included in a list of PII but generally it includes both the direct and indirect types of information shown in Figure 4.1.

When collecting data and creating datasets, you will be working with one or more of these four types of data files.

Direct Identifiers	Indirect Identifiers
• Name • Initials • Address • Phone number • Email address • Social security number • IP address • ID numbers (student ID, state ID)	• Age • Verbatim responses • Race • Ethnicity • Income • Education • Gender • Data collection date • Date of birth • ZIP code • Special education services

Figure 4.1: Examples of direct and indirect identifiers

1. Identifiable: Data includes personally identifiable information. It is common for your raw research study data to be identifiable.
2. Coded: In this type of data file, PII has been removed or distorted and names are replaced with a code (i.e., a unique participant identifier). 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 (see Chapter 10). Coded data is typically the type of file you create after cleaning your raw study data.
3. De-identified: In this type of file, 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). This is typically what you create when publicly sharing your research study data.
4. Anonymous: In an anonymous dataset, no identifying information is ever collected and so there should be little to no risk of identifying a specific participant.

4.2 Data classification

Data is often classified based on the level of sensitivity. These levels of sensitivity dictate how the data can be collected, stored, and shared, as well as what the response should be to any data breach. Depending on the institution, the names for these levels, the number of levels, what is included in these levels, and the rules applied to the levels all vary. While there is variation, here is a general summary of how information may be categorized.

1. Low sensitivity: This data is considered to have no or low-risk if disclosed. This includes de-identified and anonymous data.

2. Moderate sensitivity: This data is considered to have moderate risk if disclosed, meaning it could adversely affect people. This data either includes

identifiable information or information that could allow participants to be re-identified within the data itself or using an external source. This data is typically required to be kept confidential by law or other agreements. These data should be protected against unauthorized access.

3. High sensitivity: This data should be under the most stringent security and could cause great harm if disclosed. This data includes PII or information that could allow participants to be re-identified, as well as private or highly sensitive information (e.g., illegal behaviors, medical records) and are typically required to be kept confidential by law or other agreements. These data should be protected against unauthorized access.

It is important to review your institution's data classification levels, or data sensitivity levels, to determine how your specific institution classifies data. These rules may come from an information technology department, an institutional review board, or a combination of both. Note that different data collection efforts in the same project can be classified in different ways.

4.3 Human subjects data oversight

When collecting identifiable data, there are laws, policies, departments, and agreements that may impact how you collect and manage that data. Below we will review some of the most commonly encountered oversight in education research.

4.3.1 Regulations and laws

1. FERPA: The Family Educational Rights and Privacy Act (FERPA) is a federal law protecting the privacy of student education records. The law applies to elementary and secondary schools, as well as post-secondary institutions which receive federal funds from the Department of Education. FERPA provides a list of personally identifiable information often contained in education records [1].

2. HIPAA: The Health Insurance Portability and Accountability Act (HIPAA) provides federal protection for the privacy of protected health information (PHI) collected by covered entities serving patients. The HIPAA Privacy Rule provides a list of 18 identifiers that should be protected [2].

3. Common Rule: 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

[1] https://www.ecfr.gov/current/title-34/subtitle-A/part-99

[2] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

Rule", 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 2018).

4.3.2 Institutions and departments

1. IRB: An Institutional Review Board (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). In particular the IRB is concerned with three ethical principles established in the Belmont Report (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979); respect for persons (i.e., protecting the autonomy of participants), beneficence (i.e., minimizing harm and maximizing good), and justice (i.e., fair distribution of burdens and benefits)(Duru and Sautmann 2023; Gaddy and Scott 2020). When conducting human subjects research, it is important to review your local IRB's policies and procedures to determine if your study requires IRB approval.

2. IT department: Institutional information technology (IT) departments often vet data collection, transfer, and storage tools and are the authority on what tools are approved for research use. They may also be your source for determining classification levels for data security.

4.3.3 Agreements

1. Informed consent/assent: Consent involves informing a participant of what data will be collected for your research study and how it will be handled and used, as well as obtaining a participant's voluntary agreement to participate in your study. 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.

2. DUA: A data use agreement (DUA), also sometimes referred to as a data sharing agreement (DSA), is a contractual agreement that provides the terms and conditions for sharing data. DUAs are commonly written for data sharing when partnering with school districts or state agencies. As an example, a DUA may include the terms for sharing, working with, and storing education records data. However, DUAs can be used to provide guidance for outgoing data as well (i.e., a researcher is sharing their original data with an agency). DUAs can be standalone documents or may

be incorporated into other documents such as a memorandum of understanding (MOU).

3. NDA: Non-disclosure agreements (NDAs), which also may be synonymous with confidentiality agreements, restrict the use of proprietary or confidential information (University of Washington 2023) and are legally enforceable agreements.

4.3.4 Funders

1. Federal funders: Along with requiring data management plans, federal funders may have their own data protection procedures and may require applicants to submit additional documents agreeing to specific guidelines or requiring applicants to submit additional documents outlining their security plans for human subjects data.

4.4 Protecting human subjects data

Throughout the remaining chapters of this book we will review ways to keep identifiable human subjects data secure in each phase of the research life cycle. With that said, below is a quick review of some of the most important things to remember if you are collecting data that contain PII.

1. In most situations it will be important to get consent to collect identifiers. Consult with your local IRB to determine what is required. See Section 11.2.5
2. Collect as few identifiers as possible. Only collect what is necessary. See Section 11.2.1 for more information.
3. Follow rules laid out in applicable laws, policies, and agreements when collecting, storing, and sharing data. This includes, but is not limited to, using approved tools for data collection, capture, and storage, assigning appropriate data access levels, and transmitting data using approved methods. See Chapters 11, 12, 13 for more information.
4. Remove names in data and replace them with codes (i.e., unique study identifiers). See Chapter 10 for more information.
5. Fully de-identify data before data sharing. See Chapter 14 for more information.
6. Use data sharing agreements and controlled-access as needed when publicly sharing data. See Chapter 15 for more information.

Chapter 5

Data Management Plan

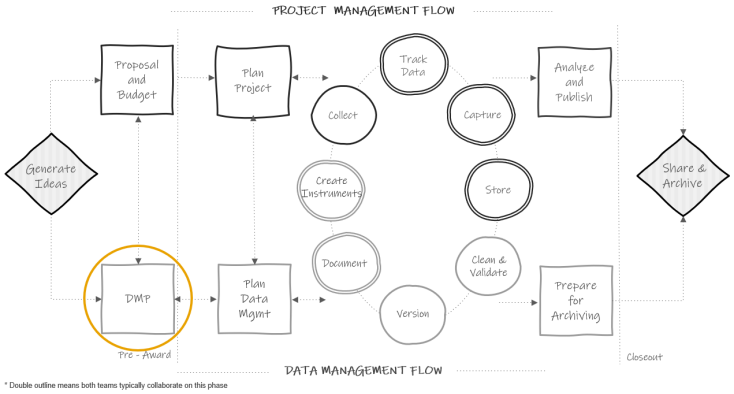


Figure 5.1: Data management plan in the research project life cycle

5.1 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) 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 (NIH) mandated that grant applicants, beginning January 2023, must submit a plan for both managing and sharing project data (National Institutes of Health n.d.). The National Science Foundation (NSF) also released version 2.0 of their public access plan in February of 2023, describing how the agency plans to ensure that all scientific data, funded by the NSF and associated with peer-reviewed publications, is publicly shared (National Science Foundation 2023).

> **Note** In the last year, agencies have begun revising the phrase "data management plan" to include the word "sharing" to better represent the shifting emphasis on sharing publicly funded data. As an example, NIH now uses the term Data Management and Sharing Plan (DMS Plan), while the Institute of Education Sciences (IES) has chosen to use the term Data Sharing and Management Plan (DSMP) [1]. For the sake of simplicity, the term DMP is used throughout this book to generally represent these plans, no matter the precise name, across all federal agencies.

5.1.1 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 2023). Even if a DMP is not required by a funder, it should always be the first step of your planning process. Although brief, this document serves as the foundation for all future planning and provides your team with a shared understanding of data management expectations.

5.2 What is it?

Typically, a data management plan is a supplemental 2-5 page document, submitted with your grant application, that contains high level decisions about

[1] https://ies.ed.gov/funding/pdf/2024_84305a.pdf

how you plan to collect, 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. Although this document is usually submitted to your funder, it should be considered a living document to be updated as plans change throughout a study.

5.2.1 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 generally 10 common categories covered in a data management plan (Center for Open Science 2023; Gonzales, Carson, and Holmes 2022; ICPSR 2020; Michener 2015) which we will review below.

1. Description of data to be shared (See Chapters 11, 12, 14, 15)
    - What is the source of data? (e.g., surveys, assessments, observations, extant data)
    - How will data be cleaned and curated data prior to data sharing?
    - What will the level of aggregation be? (e.g., item-level, summary data, metadata only)
        - Datasets from a project may need to be shared in different ways due to legal, ethical, or technical reasons.
    - Will both raw and clean data be shared?
    - What are the expected number of files? Expected number of rows/cases in each file?
2. Format of data to be shared (See Chapters 14 and 15)
    - 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., SPSS and CSV)
    - Are there any tools needed to manipulate or reproduce shared data? (e.g., software, code)
        - Provide details for those tools. (e.g., how they can be accessed, version number, required operating system)
3. Documentation to be shared (See Chapters 8 and 15)
    - What documentation will you share?
        - Consider project-level, dataset-level, and variable-level documentation.
    - What format will your documentation be in? (e.g., XML, CSV, PDF)
4. Standards (See Chapters 8 and 11)
    - Do you plan to use any standards for things such as metadata, data formatting, terminology, data collection (e.g., common data elements), or persistent identifiers (PIDs)?
5. Data preservation (See Chapter 15)

- Where will data be archived for public sharing?
  - Many agencies are now requiring applicants to name a specific data repository in this section.
- What are the desirable characteristics of the repository? [2] (e.g., unique persistent identifiers assigned to data, metadata collected, records provenance, licensing)
- When will you deposit your study data in the repository and for how long will data remain accessible?
- How will you enable discoverability and reuse of data?

6. Access, distribution, or reuse considerations (See Chapters 4 and 15)
   - Are there any legal, technical, or ethical factors affecting reuse, access, or distribution of your data?
   - Will any data be restricted?
   - Are access controls required (e.g., a data use agreement, data enclave)?
7. Protection of privacy and confidentiality (See Chapters 4, 14, and 15)
   - Do participants sign informed consent agreements? Does the consent communicate how participant data are expected to be used and shared?
   - How will you prevent disclosure of personally identifiable information when you share data?
8. Data security (See Chapter 13)
   - How will security and integrity of data be maintained during a project? (e.g., consider data storage, access, backup, and transfer)
9. Roles and responsibilities (See Chapter 7)
   - What are the staff roles in management and 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?
10. Pre-registration
    - 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 [3] [4]
- National Institutes of Health [5]
- National Institute of Justice [6]
- National Science Foundation [7]

[2] https://repository.si.edu/bitstream/handle/10088/113528/Desirable%20Characteristics%20of%20Data%20Repositories.pdf

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

[4] https://ies.ed.gov/funding/pdf/2024_84305a.pdf

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

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

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

5.3 Creating a data sources catalog

In preparation for writing your DMP, it can be helpful to create a data sources catalog that allows you to visually see what data sources you are collecting, what the sensitivity level of those sources are, and how they will be collected, managed, stored, and shared (Filip 2023). While this catalog will not cover all the questions required in a DMP, this table allows you to strategically plan the details for how you will manage and share data for each source. This type of data inventory cannot only help you write your DMP, but can also serve as an excellent planning or discussion tool throughout your entire project. Ultimately, each data source in your catalog, multiplied by the number of cohorts and/or waves it is collected, will give you an approximate estimate of your final number of distinct data files at the end of your study.

Some fields you can add to this catalog include:

- Instrument type
- Record level (who is this measure about)
- Source (who completes the instrument)
- Measures included in the instrument
- Collection and capture method
- Data collection waves
- Planned number and size of data files for each source (e.g., two student assessment files (T1, T2), with ~500 rows per file)
- PII included
- Sensitivity level based on your institution's policies
- Data storage and access plan
- Data ownership
- How confidentiality will be secured
- Data sharing method

Figure 5.2 is a simplified example of building this catalog for a hypothetical study. In this hypothetical example, if we only collected data for one year, we would end up with six datasets at the end of our study, three teacher-level files and three student-level files. In chapter 15, we will discuss whether to share these as unique datasets, or larger merged files combined by unit of analysis (e.g., student-level merged file, teacher-level merged file).

5.4 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

Instrument	Measures	Collection/Capture Method	Time Periods Collected	Direct or Indirect PII	Sensitivity Level	Data Storage and Access	Secure Confidentiality	Data Sharing Method
Teacher demographics survey	Demographics, professional development, teaching experience	Collected in Qualtrics, Exported as SPSS	T1	Name, Open-ended responses	Moderate	Stored on institution network drive, Access limited to need-to-know personnel	Consent, Coded data, PII removed, Open-ended responses categorized	Clean only, Item-level, SPSS and CSV format, OSF repository
Student assessment	Math achievement test	Collected on paper, Raw values entered into scoring program, Scores exported to CSV	T1, T2	Name, DOB	Moderate	Paper stored in locked filed cabinets, Electronic files stored on institution network drive, Access limited to need-to-know personnel	Consent, Coded data, PII removed	Clean only, Summary scores, SPSS and CSV format, OSF repository
Teacher rating of student survey	Math confidence, Math anxiety	Collected in Qualtrics, Exported as SPSS	T1, T2	Name	Moderate	Stored on institution network drive, Access limited to need-to-know personnel	Consent, Coded data, PII removed	Clean only, Item-level, OSF repository
Student school records	Demographics, attendance, discipline, state scores	CSV file received from districts	T2	Name, DOB, Student ID	Moderate	Stored on institution network drive, stored according to DUA, Access limited to need-to-know personnel	Consent, Coded data, PII removed, collapse small demographic categories	Clean only, Item-level, OSF repository, Data sharing agreement required

*T1 = Oct; T2 = May

Figure 5.2: Example data sources catalog

a wealth of knowledge about writing comprehensive data management plans. Also, if you plan to share your final data with a repository or institutional archive you will want to contact their team 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. 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.

As mentioned earlier, 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.

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 [8]
Figshare	DMP prompts specific to depositing data with Figshare [9]

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

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

Source	Resource
Hao Ye, et al.	NIH DMS Plan checklist [10]
Harvard Longwood Medical Area RDM Working Group	Annotated DMP template [11]
ICPSR	NIH DMS Plan template with specific recommendations for depositing data with ICPSR [12]
NIH	Sample DMS Plan for human survey data [13]
Sara Hart	A submitted DMP that is publicly available for example purposes [14]
UMN Libraries	Submitted DMP examples from University of Minnesota researchers [15]

5.5 Budgeting

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 2023) and time frame for incurring costs. Examples of potential allowable costs include (National Institutes of Health 2023a):

- 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

[10] https://osf.io/awypt/
[11] https://osf.io/ztjf2
[12] https://www.icpsr.umich.edu/files/ICPSR/nih/FINAL_ICPSR-NIH-DMS-Plan-Template_2023.docx
[13] https://www.nichd.nih.gov/sites/default/files/inline-files/Example_DMS_Plan-Human-Survey-NIH_Format_Page_V2.pdf
[14] https://figshare.com/articles/preprint/Example_of_a_Data_Management_Plan/13218743
[15] https://www.lib.umn.edu/services/data/dmp-examples

Source	Resource
UK Data Service	Data management costing tool and checklist [16]
University of Twente	Estimating RDM costs review list [17]
Utrecht University	Estimating the costs of data management review list [18]
DataOne	Considerations for providing budget information for a DMP [19]
J-PAL	Research proposal budget considerations [20]

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

Chapter 6

Planning Data Management

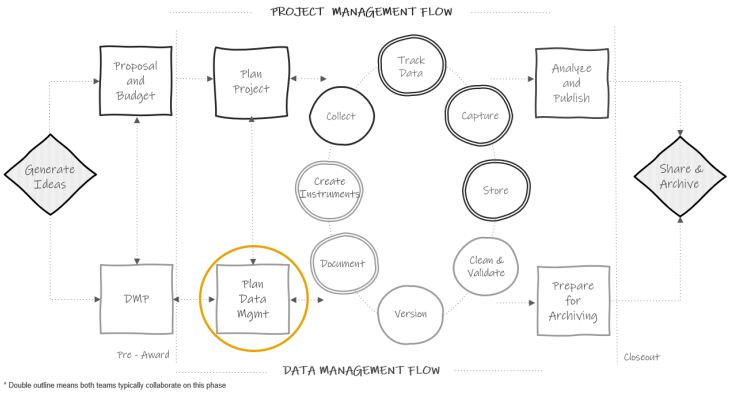


Figure 6.1: Planning in the research project life cycle

Planning data management is distinct from the 2-5 page data management plan (DMP) discussed in the Chapter 5. 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 (e.g., how to collect data, how to hire staff, what supplies are needed, how to recruit participants, how to communicate with sites). Team members such as investigators, project coordinators, and data managers, may be assisting in both planning processes.

6.1 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 2021). 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, 4).

6.2 Goals of planning

This planning phase should include a series of regular meetings with core decision makers. There are several goals to accomplish during these meetings.

1. Further flesh out project goals laid out in a grant proposal (e.g., confirm measures being collected in your study)
2. Finalize a timeline for goals (e.g., when will data be collected)
3. Lay out specific tasks needed to accomplish data management plans
4. Assign roles and responsibilities for specific tasks
5. Make decisions around how to manage tasks and communication

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 (see Chapter 8). All meeting notes should be stored in a central location where team members can reference them as needed (e.g., a planning folder with notes ordered by date, a centrally located 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.

6.3 Planning checklists

Along with your existing data management plan and other grant application materials, checklists are great tools to help inform your meeting agendas as you work through this planning process with your team. Below are 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]

> **Note** If this is your first time working through this book, these checklists are a great way to summarize content from each chapter. As you learn best practices for a phase, pull up the checklist specific to that chapter to begin thinking through which practices are feasible for your specific project.

6.3.1 Decision-making process

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 all 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 Microsoft Excel because that is the only tool they have access to, and double enter 20% because they don't have the

[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

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 6.2 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!

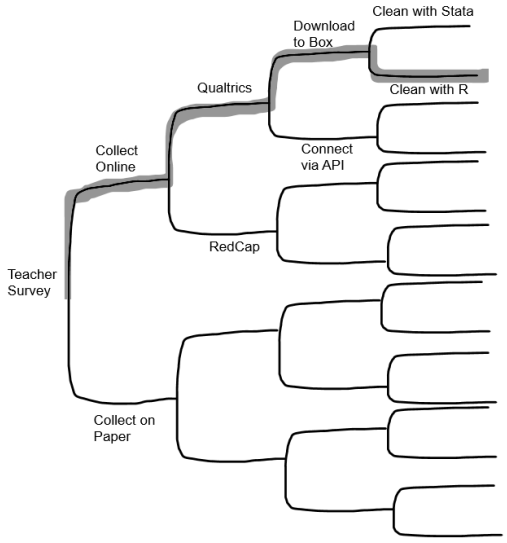


Figure 6.2: A simplified decision-making process

6.3.2 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 compliance requirements (e.g., IRB requirements, IT department requirements, consent agreements, data sharing agreements)?
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?

6.4 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 2023). 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 school records

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 6.3 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.

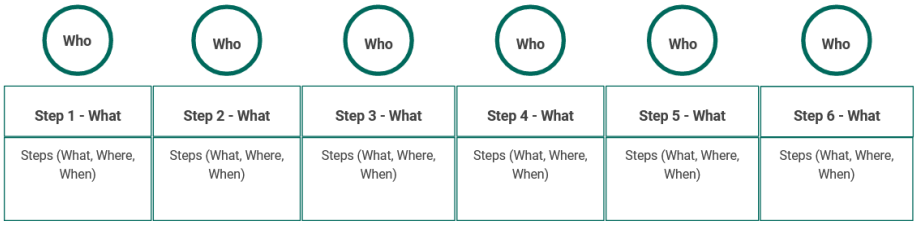


Figure 6.3: A simple workflow template

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

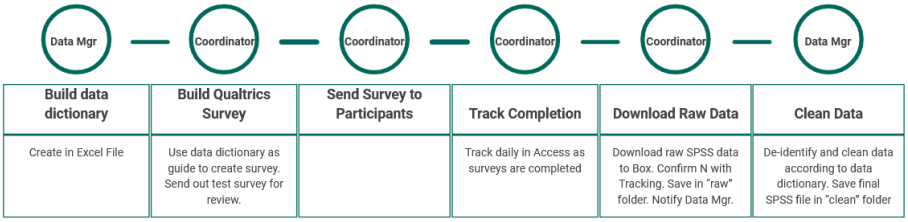


Figure 6.4: Example student survey workflow

But the format truly does not matter. Figure 6.5 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 Section 8.2.6) already created, you can even build time into your workflow. Figure 6.6 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 Section 8.2.7.

6.4.1 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

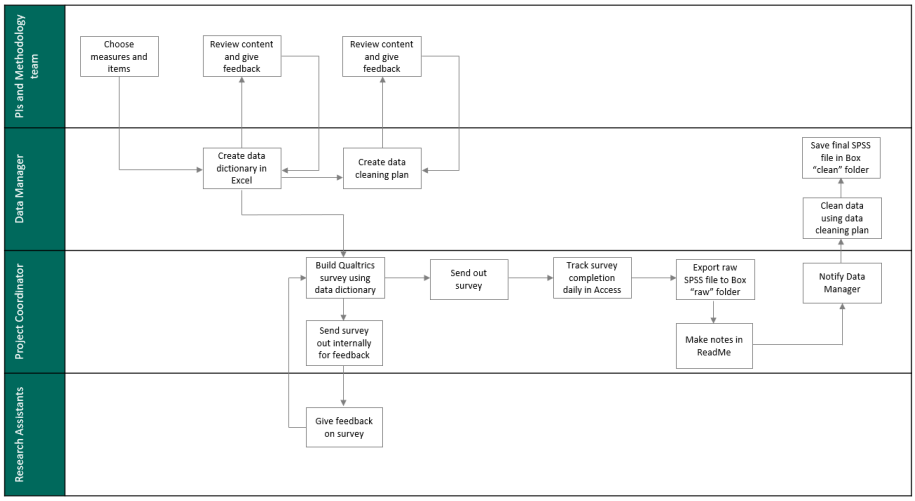


Figure 6.5: Example student survey workflow using a swimlane template

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 Box "clean" folder							■					■	Data Manager

Figure 6.6: Example student survey workflow using a Gantt chart

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.

6.4.2 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?

6.5 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 7

Project Roles and Responsibilities

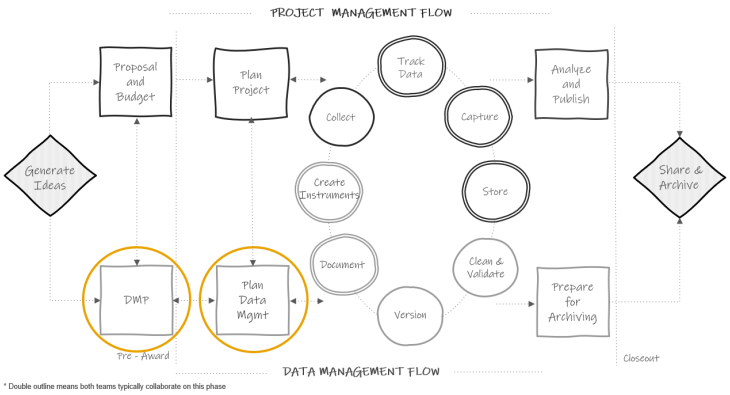


Figure 7.1: Planning in the research project life cycle

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 2023):

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

7.1 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 several 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.

7.1.1 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.

7.1.2 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.

7.1.3 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.

7.1.4 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.

7.1.5 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 (see Figure 7.2). Their tasks aren't shown on the original diagram but their work is happening alongside the data collection cycle.

7.2 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 2011).

1. Required skillset

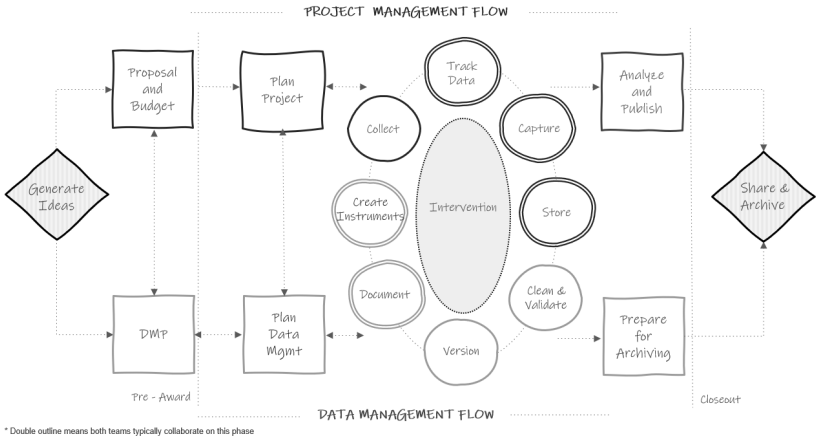


Figure 7.2: Life cycle diagram updated to show hidden processes

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.

7.3 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 7.3 and Figure 7.4 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 Section 8.2.7) 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 7.3: 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 7.4: 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)
- Name of team members who will take on responsibilities in case of a team member's absence

Chapter 8

Documentation

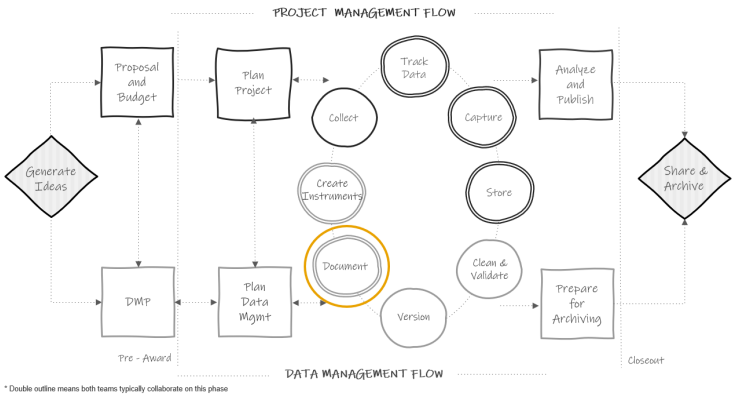


Figure 8.1: Data documentation in the research project life cycle

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, TXT, Google Doc, XLSX, 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 Section 13.2 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 and other governing bodies (see Chapters 4 & 5). 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). Ultimately though, all of your documents should be reviewed as a team in order to gather feedback and reach consensus. For these purposes, it can be helpful to create a data management working group (DMWG), consisting of PIs, key project staff, and other decision makers (e.g., methodologists), who can provide feedback as needed (Bochove, Alper, and Gu 2023).

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 8.1 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 5). With that said, the return for the investment is well worth the effort.

8.1 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 (see Figure 8.2).

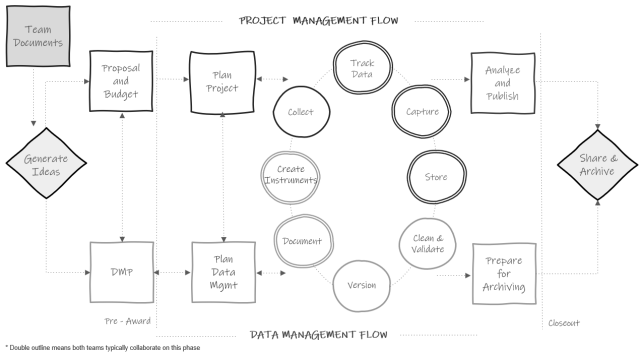


Figure 8.2: Team-level documentation in the research project life cycle

8.1.1 Lab manual

A lab manual, or team handbook, creates common knowledge across your team (Mehr 2020). It provides staff with consistent information about lab culture—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 2023; 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.

Templates 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]

[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

8.1.2 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 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.

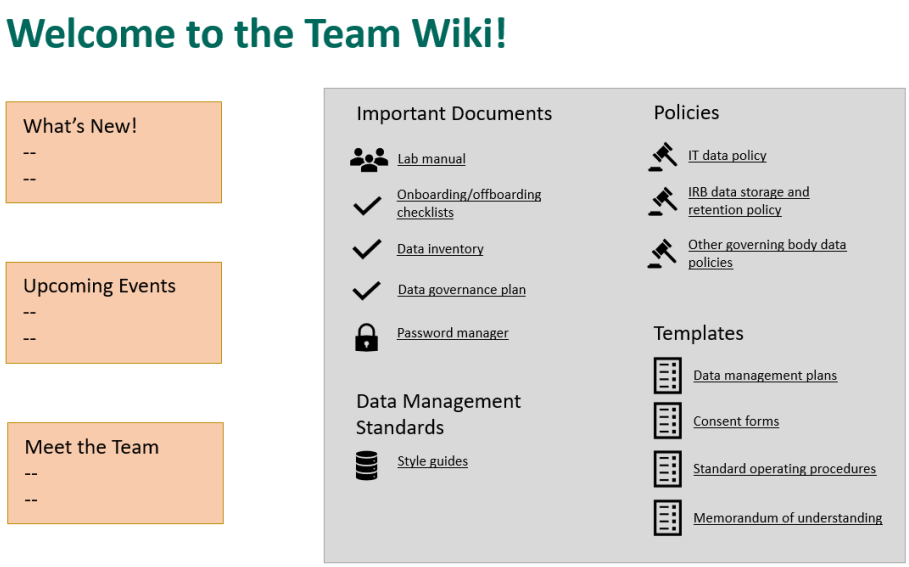


Figure 8.3: Example team wiki with links to frequently requested information

> **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]

[4] https://osf.io/mdh87/wiki/home/

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

8.1.3 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 8.4.

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]

8.1.4 Data inventory

A data inventory maps all datasets collected by the research team across all projects (Salfen 2018). As a team grows, the number of datasets typically expands as well. It can be very helpful to keep an inventory of what datasets are available for team members to use, as well as details about those datasets.

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

[5] https://www.notion.so/Company-home-240047f7526c4b0091681dc6c95b7e76
[6] https://eur-synclab.github.io/
[7] https://docs.google.com/document/d/1xyU5Q0uUD-PqRKRmMJKpD9lKaGQI6pjs
[8] https://docs.google.com/document/d/1W57cYuYyiqltQNXUITP-jVIf84jao4Ef

Onboarding	Offboarding
☐ Contacts ☐ For acquiring access to data storage spaces ☐ For data related questions ☐ Learning ☐ Where to go to learn more about existing data for current and past projects (e.g., data inventory) ☐ Relevant standard operating procedures to review ☐ Requirements and standards ☐ Any required training (e.g., CITI, IT training) ☐ Any required documents to review and complete (e.g., data security policy) ☐ Any standards to review (e.g., style guide) ☐ Tools ☐ What existing data tools are used ☐ How to access those tools ☐ Training needed for those tools	☐ Access ☐ Contacts for removing data access ☐ Tying up loose ends ☐ Making sure all standard operating procedures associated with your role are up to date ☐ Review all data files you have worked on to ensure they are ☐ Stored according to policy ☐ Documented adequately ☐ Named according to the style guide ☐ Accessible to someone on the team other than yourself

Figure 8.4: Sample data topics to add to onboarding and offboarding checklists

Project	Dataset Name	Location	Description	Linking variables
Math Stars: Description of project (2018-2021)	ms_stu_c1-c3_clean.sav	Link to SharePoint folder	Student survey, assessment, and district data	stu_id, tch_id
	tch_c1-c3_clean.sav	Link to SharePoint folder	Teacher survey and classroom observation data	tch_id
LEAN: Description of project (2016-2018)	ln_tch_c1-c2_clean.sav	Link to SharePoint folder	Teacher survey data	tch_id, sch_id
	ln_sch_c1-c2_clean.sav	Link to SharePoint folder	School enrollment and demographic data	sch_id

Figure 8.5: Example of a team data inventory

8.1.5 Data security policy

A data security policy is a set of formal guidelines for working with data within an organization. This policy, which can be named many other things (e.g., data governance plan, data security protocol or plan, data responsibility plan), should broadly cover how team members are allowed to work with data in a way that protects research participant privacy, ensures quality control, and adheres to legal, ethical, and technical guidelines. Documenting this information ensures a cohesive understanding among team members regarding the terms and conditions of project data use (CESSDA Training Team 2017). A data security policy can be added to a lab manual, or can be created as a separate document where team members can even sign (Filip 2023) or check a box acknowledging that they have read and understand the policy.

Ideas of content to include in a data security policy are included in Figure 8.6.

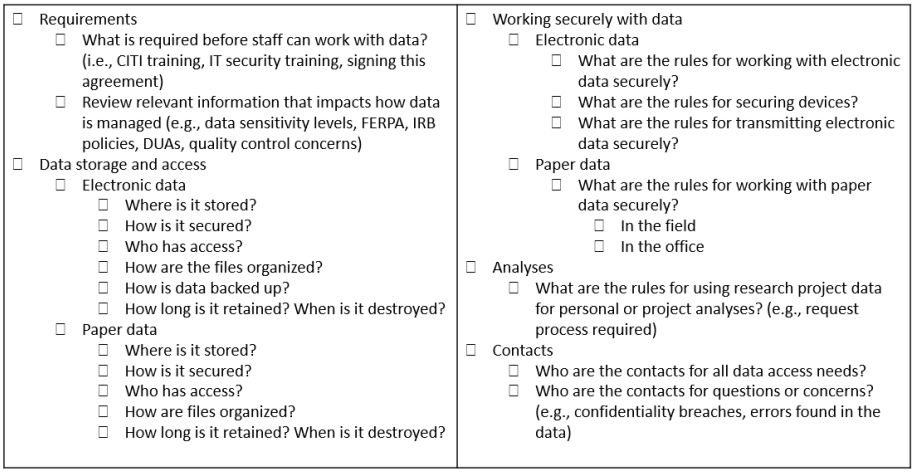


Figure 8.6: Example of content to include in a data security policy

Templates and Resources

Source	Resource
Crystal Lewis	Data security policy template [9]
SYNC Lab	Data security protocol [10]
University of Nebraska-Lincoln	Research data and security checklist [11]

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

[10] https://eur-synclab.github.io/data-management/data-security.html#keep-paper-data-logs-questionnaires-mri-checklists-locked-up

[11] https://uofnelincoln.sharepoint.com/:b:/s/ResearchComplianceServicesSharepoint/Ebr9awmFio1Mj3PTgb-GnSIBoS7xSua7uT-jePT2qtGTlw?e=RbRs50

8.1.6 Style guide

A style guide is a set of standards for the formatting of information. 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 9 for more information.

Templates and Resources

Source	Resource
Hadley Wickham	Example R coding style guide [12]
Strategic Data Project	Example style guide [13]

8.2 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.

8.2.1 Data management plan

As discussed in Chapter 5, 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.

8.2.2 Data sources catalog

Also, as reviewed in Section 5.3, a data sources catalog is an excellent project planning tool that should be developed early on during the DMP phase. This

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

spreadsheet helps you succinctly summarize the data sources you will collect for your project, as well as plan the details of how and when data will be collected and managed. This document serves as a referral source for the remaining planning phases of your project and should be a living document to be updated as needed.

8.2.3 Checklists and meeting notes

Checklists, as discussed in Section 6.3, are documents that are created, or copied from existing templates, 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.

8.2.4 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 7 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 [14]

8.2.5 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 an Institutional Review Board, you will most likely be required to submit this document as part of your application. A research protocol assists the board in determining if your methods provide adequate protection for human subjects.

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

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 8.7.

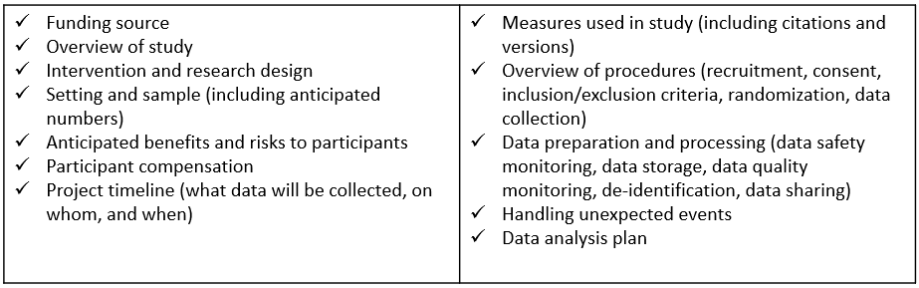


Figure 8.7: Common research protocol elements

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 (see Section 8.2.6).

Templates and Resources

Source	Resource
Crystal Lewis	A template to create a project-level summary document for data sharing (based on an IRB research protocol) [15]
Jeffrey Shero, Sara Hart	IRB protocol template with a focus on data sharing [16]
The Ohio State University	Protocol template [17]
University of Missouri	Protocol template [18]
University of Washington	Protocol checklist [19]

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

8.2.6 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 8.8 is an example of one way to visualize a data collection timeline.

	Year 1									Year 2								
	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Cohort 1 student survey																		
Cohort 1 teacher survey																		
Cohort 1 student assessment																		
Cohort 2 student survey																		
Cohort 2 teacher survey																		
Cohort 2 student assessment																		

Figure 8.8: 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. As seen in Figure 8.9, 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). They provide a means to understand how participants are randomized and assigned to treatment groups. As you can imagine though, this diagram cannot be started until participants are recruited and enrolled and must be updated as each wave of data is collected. Your participant tracking database, which we will discuss in Chapter 10, 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 include any technical documents associated with your instruments or measures

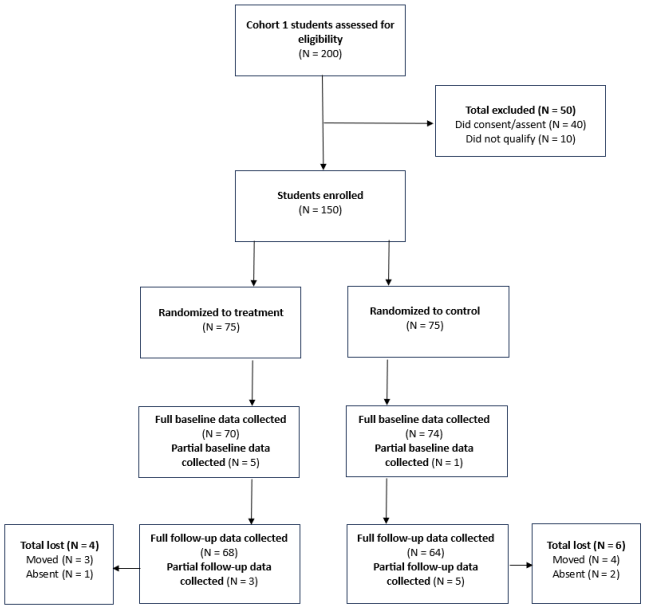


Figure 8.9: Example participant flow diagram

(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

Informed consent forms (see Chapter 11) 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.

8.2.7 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

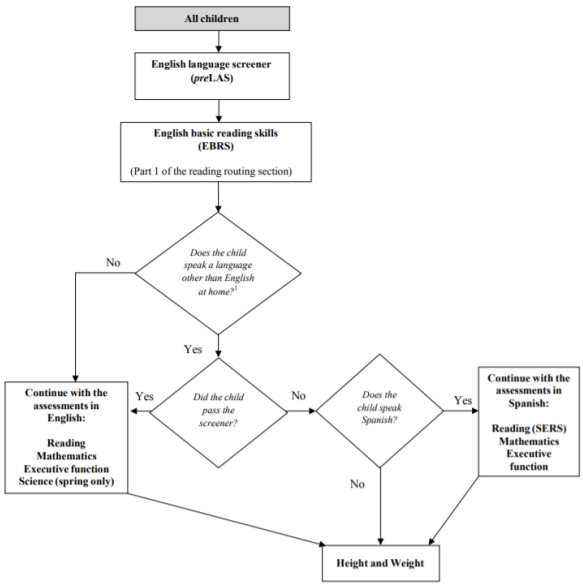


Figure 8.10: Flowchart of an ECLS-K:2011 kindergarten assessment

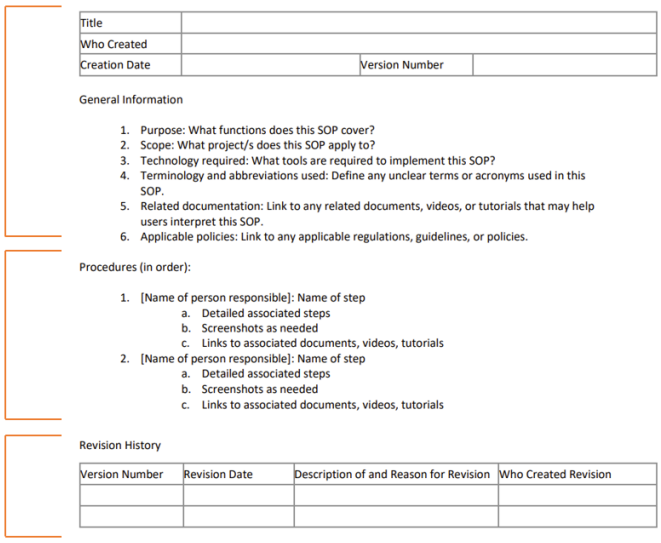
the procedures are actually implemented on a daily basis (NUCATS 2023). Standard operating procedures (SOPs) provide a set of detailed instructions for routine tasks and decision making processes. If you recall from Chapter 6, every step that we added to a data collection workflow is then added to an 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 8.11. 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.

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.

In developing your SOP template, like the one in Figure 8.12, you should begin with **general information** about the scope and purpose of the procedure, as

✓ Recruitment/screening 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 8.11: Examples of data management processes or decisions to develop an SOP for



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, videos, or tutorials 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. Screenshots as needed
   c. Links to associated documents, videos, tutorials
2. [Name of person responsible]: Name of step
   a. Detailed associated steps
   b. Screenshots as needed
   c. Links to associated documents, videos, tutorials

Revision History

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

Figure 8.12: Standard operating procedure minimal template

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 online tutorials or staff created how-to videos 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.

Templates and Resources

Source	Resource
Crystal Lewis	SOP template [20]

8.3 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 Training Team 2017; UK Data Service 2023). 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.

8.3.1 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

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

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 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 README 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, suffixes or acronyms used in datasets (e.g., w1_ = wave 1 - fall of study year), or instrument response rates. Figure 8.13 is an example README that can be used to describe all data sources shared in a project repository (Neild, Robinson, and Agufa 2022).

Templates and Resources

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 8.13: Institute of Education Sciences example README for conveying information on files in a directory

Source	Resource
Crystal Lewis	README template for sharing information about a set of files in a directory [21]
Crystal Lewis	README template for sharing project-level information [22]

8.3.2 Changelog

A changelog is a record of all of the major versions of your data and code (UK Data Service 2023; 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 identifiable human subjects 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 (e.g., 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:

[21] https://docs.google.com/document/d/1JWeKLDqtuk79beNJBv5xHueMwki0c7xD
[22] https://docs.google.com/document/d/1rbED1r0fGAk5CREslc8qQ5378EBV5759eSqdQbp4fHc

- 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 can also be helpful to record additional information such as who made the change and a link to any code used to transform the data (CESSDA Training Team 2017).

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 8.14: Example changelog for a clean student survey data file

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.

Templates and Resources

Source	Resource
Crystal Lewis	Changelog template [23]

8.3.3 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,

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

district student school records 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. Typically the person responsible for cleaning the data will write the data cleaning plans, but the documents can then be brought to a planning meeting allowing your DMWG 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 14.

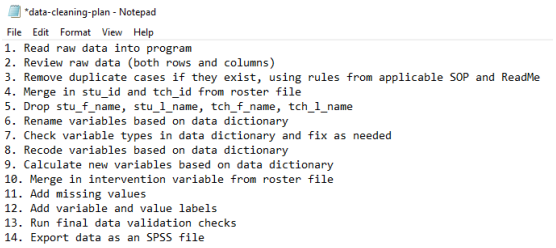


Figure 8.15: A simplistic data cleaning plan

8.4 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.

8.4.1 Data dictionary

A data dictionary is a rectangular formatted collection of names, definitions, and attributes about variables in a dataset (Bordelon 2023; Gonzales, Carson, and Holmes 2022; UC Merced Library 2023). This document is most useful if created during the documentation phase and used throughout a study for both planning and interpretation purposes (see Figure 8.16) (Lewis 2022a; Bochove, Alper, and Gu 2023).

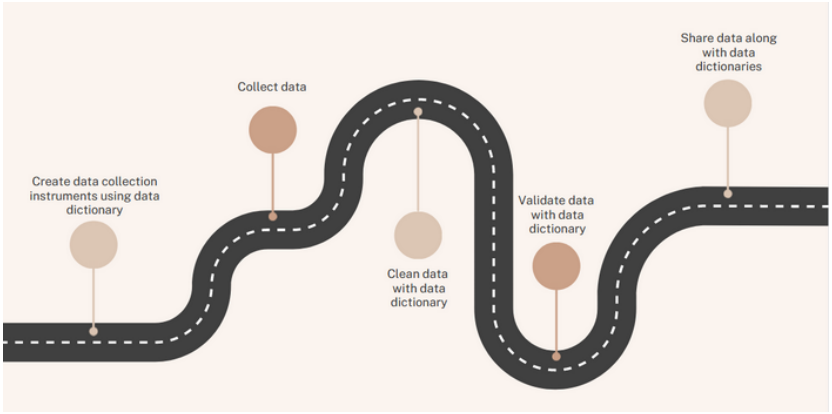


Figure 8.16: The many uses for a data dictionary

This document should be structured similar to a dataset, with variable names in the first row (Broman and Woo 2018). There are several necessary fields to include in this document, as well as several optional fields (see Figure 8.17) (Johns Hopkins Institute for Clinical and Translational Research 2020).

8.4.1.1 Creating a data dictionary for an original data source

When you are collecting data for an original source, there are a few things that are helpful to have when creating your data dictionaries:

1. Your data sources catalog
    - This document (see Section 8.2.2) will provide you an overview of all of the data sources you plan to collect for your project, including what measures make up each instrument
2. Your style guide already created
    - We will talk more about style guides in Chapter 9, but this document will provide team or project standards for naming variables and coding response values.

3. 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 doc-

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 (including labels associated with categorical codes)	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 8.17: Fields to include in a data dictionary

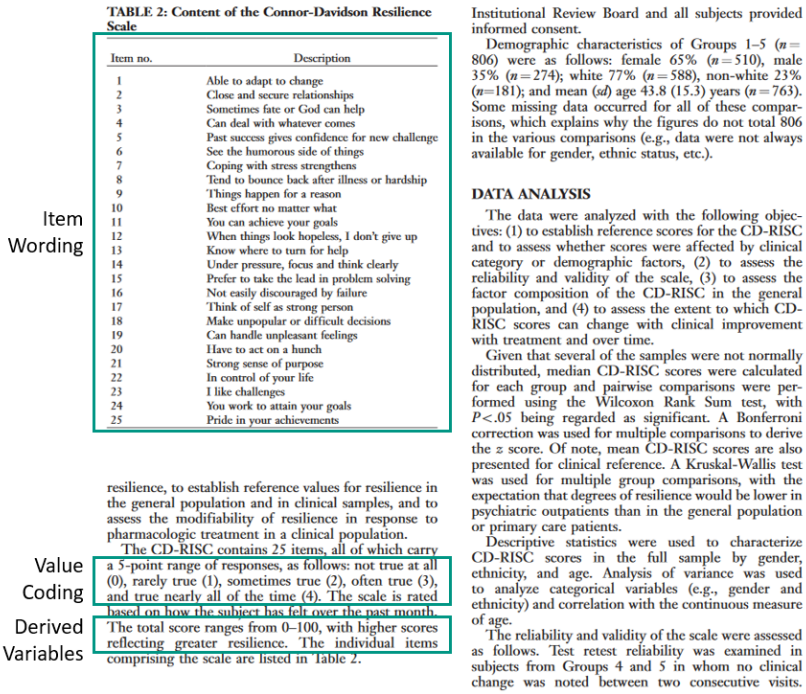
umentation on those measures such as technical documents or copies of instruments. You will want your documentation to provide information such as:
  - What items make up the measures/scales/assessment? What is the exact wording of the items?
  - How are items coded? What are allowable values?
  - Are there any calculations/scoring/reverse coding needed?
  - If items are entered into a scoring program and then exported, what variables are exported?
- See Figure 8.18 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).

4. Any data element standards that you plan to use
   - See Section 11.2.1 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 assessment data dictionary). If there are five data sources in your data sources catalog, and four of them are collected by your team (i.e., one is an extant dataset), you should end up with four data dictionaries for your original data. All measures/items for each instrument will be included in the data dictionary.

As you build your data dictionaries, consider the following:

- Item names
  - Are your variable names meeting the requirements laid out in your



Item Wording

Value Coding

Derived Variables

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

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 participant ids, rater ids)
    * Grouping variables (treatment, cohort)
    * 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 8.19 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.

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. Gather your DMWG and review the following types of questions:

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

- Does the team want to adjust any of the question/item wording?

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 8.19: Example student survey data dictionary

- 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.
- Does the dictionary include all the derived variables to create after collection? All the recoding of variables that is required?

8.4.1.2 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 capturing supplemental external data sources from organizations like school districts or state departments of education. If at all possible, start gathering information about your external data sources early on, during the documentation phase, and begin adding that information into your project data dictionary. Starting this process early will help you prepare for future data capture and cleaning processes.

- If the data source is public, you may be able to easily find codebooks or data dictionaries for the data source. If not, download a sample of the data to learn what variables exist in the source and how they are formatted.
- If the data source is non-public, request documentation ahead of time from your partner (see Section 12.3).

However, it is possible that you may not be able to access this information until you acquire the actual data during the data capture phase. If documentation is provided along with the data, begin reviewing the data to ensure that the documentation matches what you see in the data. Integrate that information into your project data dictionary.

If documentation is not provided it is important to review the data and begin collecting questions that will allow you to build your data dictionary.

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 (e.g., a 999 or 0 in an age variable)?
4. What data types are the items currently? What types should they be?

In most situations these questions will not be easily answered without documentation and may require further detective work.

- Contact the person who originally collected the data to learn more about the instrument or data collection process.
- Contact the person who cleaned the data (if cleaned) to see what transformations they completed on the raw data.

- If applicable, request access to the original instruments to review exact question wording, item response options, skip patterns, etc.

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.

8.4.1.3 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.

Templates and Resources

Source	Resource
Crystal Lewis	Data dictionary template [24]

8.4.2 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 (numeric, character, etc.)	Summary statistics
Value labels	Weighting

Figure 8.20: Codebook content that overlaps and is unique to a data dictionary

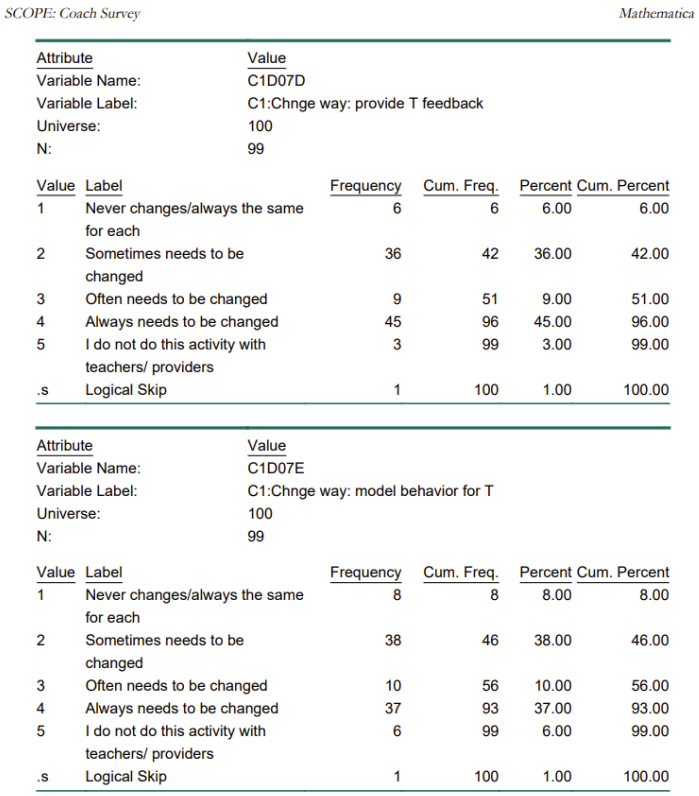
Figure 8.21 is an example codebook from the United States Department of Health and Human Services (2022).

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

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



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 8.21: Example codebook from the SCOPE Coach Survey

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.

Templates and Resources

Source	Resource
ICPSR	Guide to Codebooks [25]
National Center for Health Statistics	Example codebook [26]

8.5 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 Training Team 2017; 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 (Wilkinson et al. 2016; Logan, Hart, and Schatschneider 2021; UK Data Service 2023).

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 Training Team 2017; University of Iowa Libraries 2023). 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 2023; Danish National Forum for Research Data Management n.d.). The information from this form will become your metadata. Figure 8.22 is an example of an intake form for the the Figshare repository (https://figshare.com/).

The most common metadata elements (Dahdul 2023; Hayslett 2022) are included in Figure 8.23.

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 more comprehensive information, such as project-level information covered in your research protocol. You may also have the option to enter additional levels

[25] https://www.icpsr.umich.edu/files/deposit/Guide-to-Codebooks_v1.pdf

[26] https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2020/adult-codebook.pdf

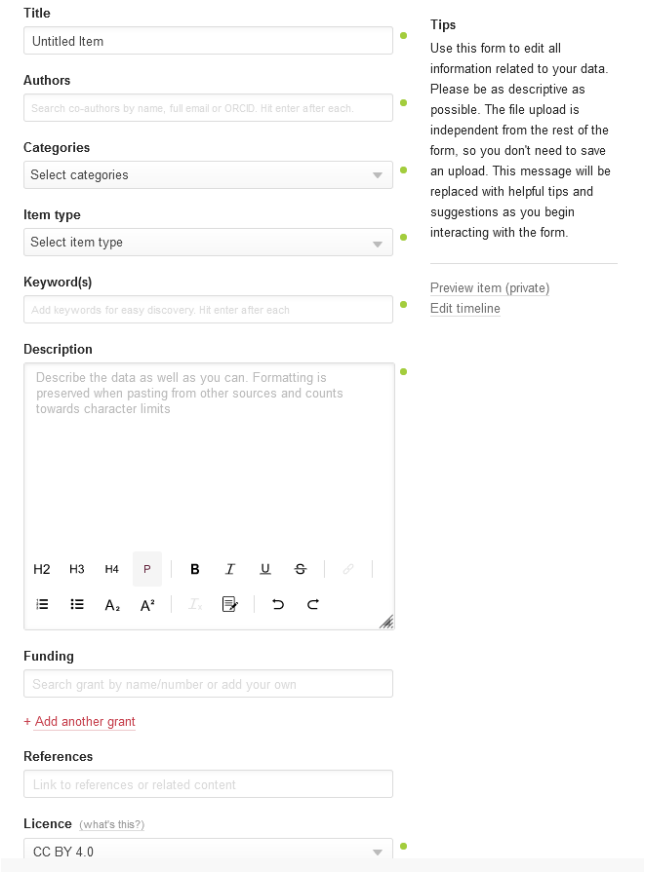


Figure 8.22: Example intake metadata form for Figshare repository, captured January 13, 2023

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 8.23: Common metadata elements

of metadata that will help make each level more searchable, such as file-level or variable-level metadata (Gilmore, Kennedy, and Adolph 2018; ICPSR 2023; LDbase n.d.). 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 2023) or JSON-LD. We can see what this metadata looks like to humans once it is submitted. Figure 8.24 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 8.5.1) if we want as well.

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 2023).

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 2019; Buchanan et al. 2021; USGS 2021).

8.5.1 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. Metadata standards can be applied in several

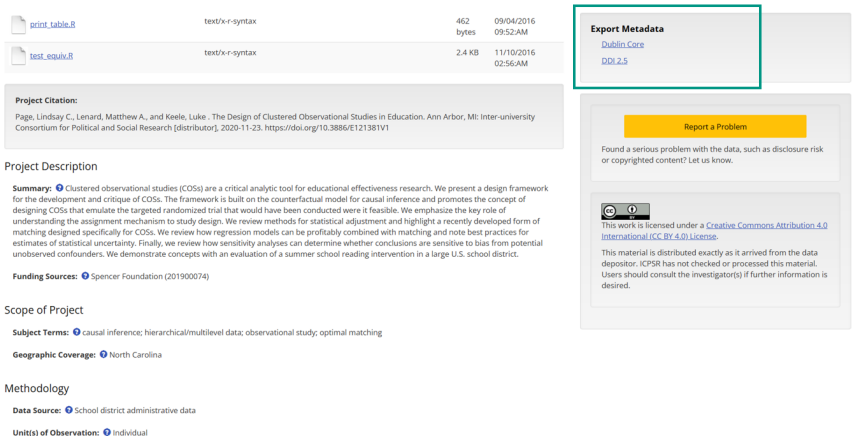


Figure 8.24: Example metadata displayed on an ICPSR Open project page

ways (Bolam 2022; DDI Alliance 2023a).

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 2023). 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.).

It can be helpful to see how standards differ as well as overlap. The DDI Alliance (2023b) put together this table in Figure 8.26 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, both the OSF (Gueguen 2023) and Figshare (Figshare 2023) repositories currently use the DataCite schema , while ICPSR uses the DDI standard (ICPSR 2023). If the repository does use certain standards, work with them to ensure your metadata adheres to those standards.

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 8.25: A sampling of field metadata standards

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 8.26: A comparison of DDI Version 2 standards to Dublin Core standards

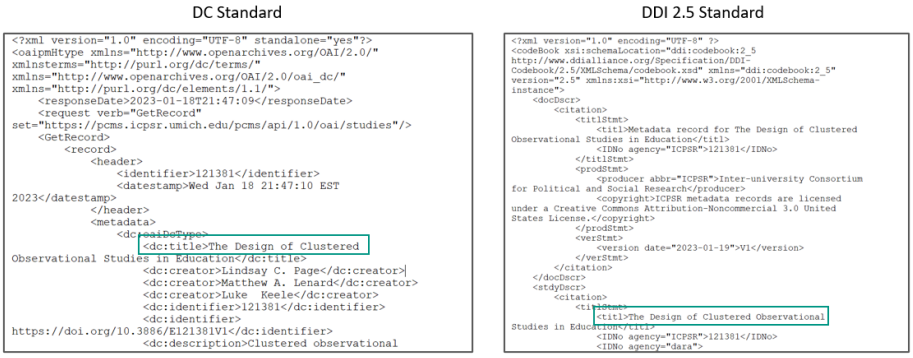


Figure 8.27: Metadata comparison from an AERA Open project

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 2023), 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.

8.6 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 9

Style guide

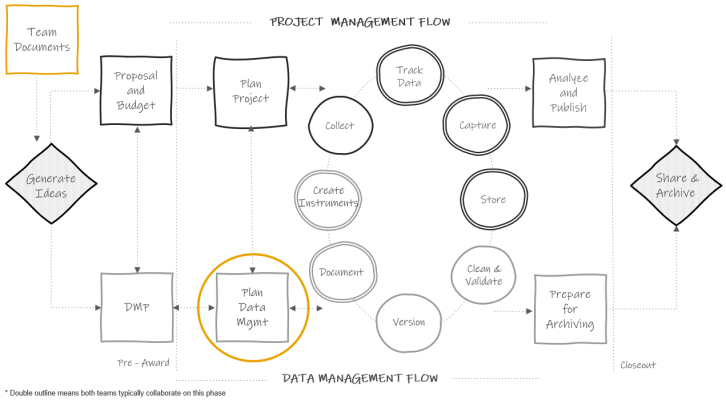


Figure 9.1: Style guide in the research project life cycle

A style guide provides general agreed upon rules for the formatting of information. As mentioned in Chapter 8, 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 nam-

ing, 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.

9.1 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.

9.2 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 70 characters `W:\team\projecta\data\wave1\student\survey\projecta-`
    - Examples of file path limits:

    - SharePoint/OneDrive path limit is 400 characters (Microsoft 2023)
    - 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 all data "archive" folders to document changes b
```

Example directory structure created using a style guide

```
                                levelName
1  project-new
2   ¦--intervention
3   ¦   °--cohort-1
4   ¦       °--coaching_materials
5   ¦           °--archive
```

```
¦--project-mgmt
¦   °--cohort-1
¦       °--scheduling-materials
¦           °--archive
¦--documentation
¦   ¦--sops
¦   ¦   °--archive
¦   °--data-dictionaries
¦       °--archive
¦--data
¦   °--cohort-1
¦       °--student
¦           °--survey
¦               °--archive
¦                   °--changelog.txt
°--tracking
    °--cohort-1
        ¦--participant-database
        ¦   °--archive
        °--parent-consents
```

9.3 File naming

Figure 9.2: xkcd comic on naming files

As xkcd (n.d.) so aptly points out in Figure 9.2, 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)
- No PII should be used in a file name (e.g., participant name)
- 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 Section 8.3.2, it is possible to version programatically using tools like Git and GitHub. However, these tools are not always practical for education research, and the tools that are used in the field (e.g., SharePoint) often don't have a way of documenting differences between versions. A more practical means of versioning, at least major file changes, 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 containing 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, always put the project name in the file name, even if the file is currently housed in a project folder, or always put the word "raw" or "clean" in the file name, even if the file is housed in a "raw" or "clean" 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 40 characters
6. Use the following metadata file naming order:
  - Order of use (if relevant - add a 0 before single digits)
  - Project
  - Cohort/Wave (if relevant)
  - Participant
  - Instrument
  - 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_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
```

9.4 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 district student 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
    - While pascal case and camel case are also options, the use of underscores helps more clearly delineate relevant pieces of metadata in your variable names
  - Kebab case (scale-sum) – don't use for variable names
  - Train case (Scale-Sum) – don't use for variable names
- If a variable includes a "select all" option, start all associated variables with the same prefix (`cert_elem`, `cert_secondary`, `cert_leader`, `cert_other`). Again this allows you to easily see grouped items, as well as easily programatically manipulate those items as needed.
- 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 instrument in the variable name. This can also help with the unique variable name requirement above.
  - t = teacher self-report
  - s = student self-report
  - tr = teacher rating of student
  - d = district student records
  - `t_conf1`, `s_anx2`, `tr_relate4`, `d_gender`

Example variable naming style guide

```
1. Use snake case
2. Keep names under 32 characters
3. Use meaningful variable names
4. Use unique variable names within and across data sources
5. If part of a scale, use scale abbreviation plus item number from the scale
6. Include an indication of the instrument as a prefix in the variable name
  - student self-report = s_
  - teacher self-report = t_
  - parent report on students = p_
  - district student records = d_
7. 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_gender
t_stress5
p_relate
```

9.4.1 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. Recall from Chapter 3, there are two ways you can link data over time, in wide format or long format.

1. If combining data over time in long format, no changes need to be made to your variable names. Variable names should be identically named over time. To account for a time component, you will simply include a new variable (e.g., `time`, `year`, `wave`) and add the appropriate value for each row.

2. If combining data in wide format, you will need to concatenate time to all of your time varying variable names (i.e., not your subject unique identifier). This removes the problem of having non-unique variable names (e.g., `anx1` in wave 1 and `anx1` in wave 2) and allows you to interpret when each variable was collected. How you concatenate time to your variable names is up to you. Just make sure to continue adding time consistently to all variable names (i.e., same location, same format) and remember to follow variable naming best practices (e.g., never start a variable name with a number).

Before adding a time component, either as a new variable or as part of a variable name, it's important to decide what values you want to assign to time. This

will depend entirely on your study design and how you intend to use time in your analyses. When working with cohorts, it can be helpful to choose generic time values that allow you to combine samples collected in the same relative time periods (e.g., `wave 1` = fall of the study year, `wave 2` = spring of the study year, `wave 3` = fall of the follow up year). However, when not working with cohorts or if you have a dataset that does not fall within your pre-defined data collection periods, you can choose any values that work for you. Figure 9.3 provides just a few examples of how you might account for time in your data based on different scenarios.

Scenario	Time component added as a variable	Time component concatenated to variable names
A pre/post student survey, provided for one group of students Fall 2023 and Spring 2024	Variable name: time Values: pre, post	pre_varname, post_varname
A student survey collected over 3 time periods, for 2 different cohorts of students Cohort 1: Fall 2023, Spring 2024, Fall 2024 Cohort 2: Fall 2024, Spring 2025, Fall 2025	Variable name: wave Values: 1, 2, 3	w1_varname, w2_varname, w3_varname
School-level demographic data collected for two school years 2022-23 and 2023-24	Variable name: year Values: 23, 24	varname_23, varname_24

Figure 9.3: Examples of how time might be added to your data based on a variety of scenarios

With all of that said, during an active project, it is actually best to not add a time component to your data, and to store each dataset as a distinct file, with a clear file name that denotes the appropriate time period. There are a few benefits of this method:

1. Naming variables 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 (Reynolds, Schatschneider, and Logan 2022).
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.

So, with that said, add rules to your variable naming style guide around how to concatenate time to your variable names, but make an asterisk saying that this time component should only be added when you are ready to combine files. Once you are ready to combine files, either for an analysis or other reasons, you can fairly quickly add a time variable or programatically concatenate time to variable names using a statistical program (e.g., R) or even in a program like Microsoft Excel.

9.5 Value coding

Oftentimes in education research we codify categorical values. This coding of values helps in both data entry, data scoring, and data analysis. As an example, rather than referring to the lengthier values of "yes" or "no" in a variable, we may code those values into a code/label pair. The code can be numeric (e.g., "yes" = 1 | "no" = 0) or character (e.g., "yes" = 'y' | "no" = 'n'), depending on your needs. Ultimately, only the code appears in your data, while the code/label pair is represented in your data dictionary, allowing users to interpret the meaning of each code.

If you are planning to code any categorical variable values for your study, it can be helpful to include general guidelines in your style guide for assigning those codes. Some general good practices are outlined below.

First, if you are using a pre-existing measure, assign codes 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, if you are assigning your own codes:

- Codes must be unique
  - Do: Assign "yes" = 1 | "no" = 0
  - Don't: Assign "yes" = 1 | "no" = 1
- Codes 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'
- Codes must be consistent across time
  - Do: For `anx1` assign "yes" = 1 | "no" = 0 in wave 1 **and** wave 2
  - Don't: For `anx1` assign "yes" = 1 | "no" = 0 in wave 1 **but** "yes" = 1 | "no" = 2 in wave 2
- Codes should be consistent across the project
  - Do: Assign "yes" = 1 | "no" = 0 as the value for all yes/no items
  - Don't: Assign "yes" = 1 | "no" = 0 for some variables, and "yes" = 1 | "no" = 2 for others
    - The exception here is if a pre-existing measure determines how values are coded. In that case, there may be some inconsistency across items
- 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
    - The exception here is if a pre-existing measure tells you to code variables in a different way

9.5.1 Missing value coding

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

1. You can choose to leave all missing values as blank.
    - Benefits of this option is that there is no chance of assigned missing value codes (e.g., -999) 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 codes and add them to your data. This code can be numeric (e.g., "missing" = -999) or character (e.g., "missing" = 'NA') and it may be one consistent code applied to all missing data, or it may be multiple codes assigned for different types of missing data.
    - One benefit of this method is that this removes the uncertainty that we had with blank cells. If a value is filled, we are now certain the value was not deleted or skipped over during data entry.
    - Another benefit is that this allows you to specify distinct reasons for missing data (e.g., "Not Applicable" = -97, "Skipped" = -98) if that is important for your study.
    - The biggest problem that can occur with this method is that either your codes 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)

Ultimately, whichever method you choose, there are several guidelines you should follow.

1. If you decide to fill with defined missing codes, use values that match your variable type (e.g., numeric codes for numeric variables) (ICPSR 2020; White et al. 2013)
    - 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 column. If you do not care about the different types of missingness, you could easily then choose to change all missing codes 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.
2. If you use numeric values, use extreme values that do not actually occur in your data
3. Use your values consistently within and across variables

In your value coding style guide, you can add general rules to follow, or it may be an appropriate place to actually designate a missing value coding schema for your project (see Figure 9.4).

Code/Value	Label
-99	Unit nonresponse (entire instrument not completed)
-98	Item skipped
-97	Item not applicable
-96	Don't know

Figure 9.4: Example missing value code schema used for numeric variables

9.6 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 search online for existing style guides in that language.

- Consider building and implementing coding templates (Daskalova 2020; 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 `#` at the start of a comment
    - SPSS and Stata use `*` at the start of a comment
- Improve code readability by using (Wickham 2021; 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 following the DRY (don't repeat yourself) principle. Consider 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 10

Data Tracking

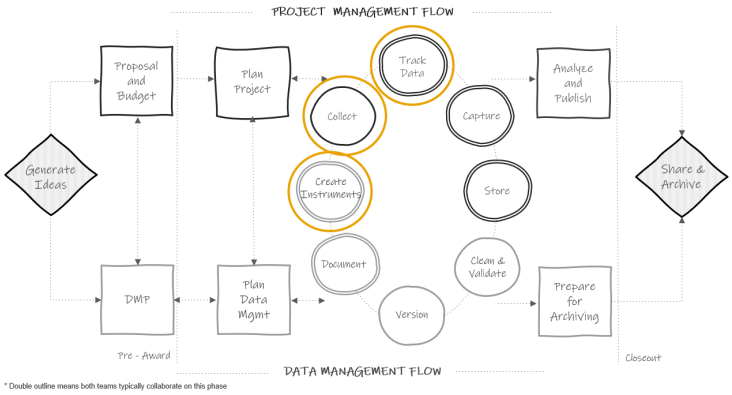


Figure 10.1: Tracking in the research project life cycle

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 (SSOT) 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 that is gleaned (e.g., name change, new email, moved out of district) 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.

10.1 Benefits

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

1. Protecting participant confidentiality
    - Assigning unique study identifiers (i.e., codes) 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 13), separate from where the identifiable and coded 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.

10.2 Building your database

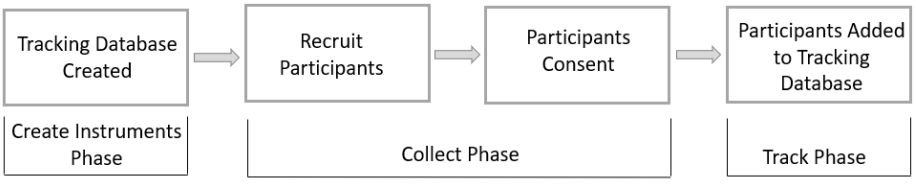


Figure 10.2: Example timeline for constructing and using a tracking database

While the tracking phase appears after collection in Figure 10.1, it is most ben-

eficial 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 participants, you can record information such as name, consent status, and any other necessary identifying contact information in the participant database and begin assigning participants study IDs (see Figure 10.2 for example of what this workflow may look like). 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.

10.2.1 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 Section 3.3.1. 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 2014; Chen 2022). There are three general steps for building a relational database.

1. Create tables made up of fields (i.e., variables)
2. 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.
3. 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 2014; 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 completely 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.

10.2.1.1 Relational model

In Figure 10.3 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 subjects. 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.

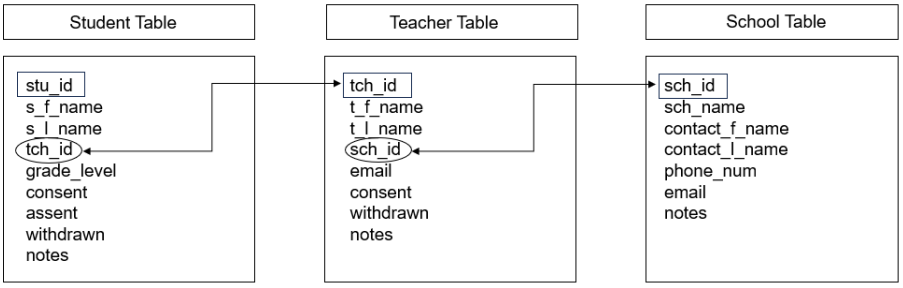


Figure 10.3: Participant database built using a relational model

We can see here that across each table we have no duplicated 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 10.1.

```
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 10.1: 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

10.2.1.2 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 10.4). 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. For example, 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 (Borer et al. 2009).

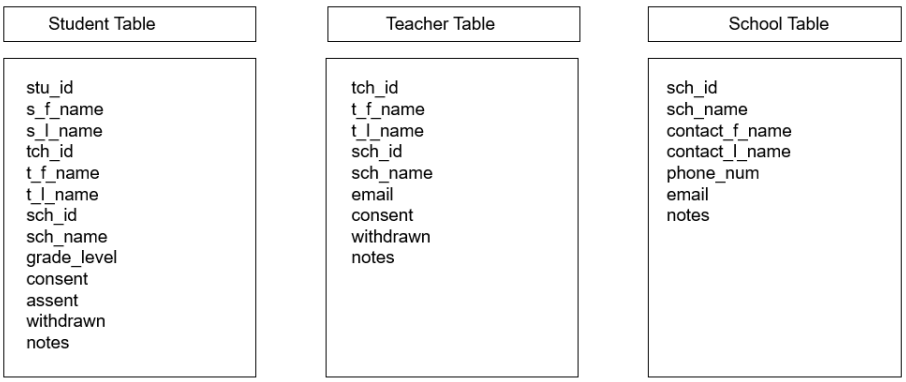


Figure 10.4: Participant database built in using a non-relational model

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.

10.2.2 Structuring the database

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

1. Do you want to use a relational table structure?
2. How many tables do you want to construct?
   - Consider entities (e.g., student, teacher, school)
   - Consider purpose (e.g., student enrollment table, student wave 1 data collection table, student wave 2 data collection table)
3. What fields do you want to include in each table?
4. If using a relational table structure, what fields will you use to relate tables?

Once you make decisions regarding these questions, you can begin to design your database structure. It can be helpful to visualize your database model during this process. In Figure 10.5 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.

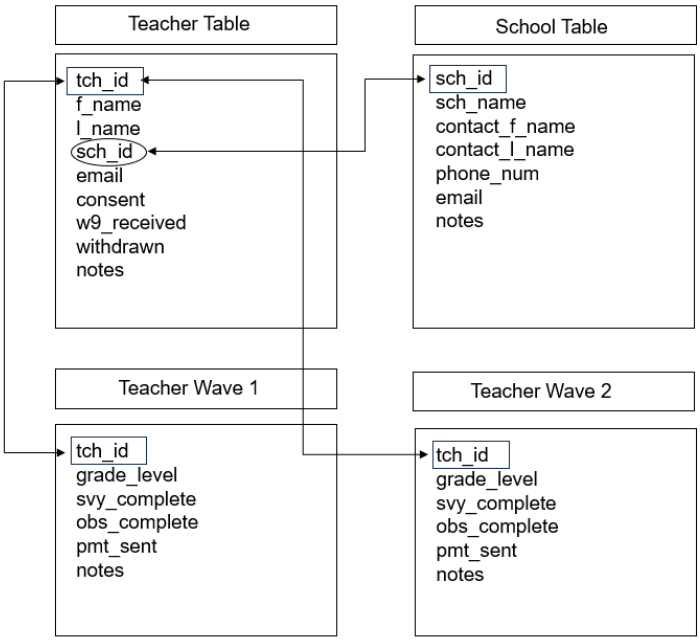


Figure 10.5: Example participant database model

I have designed this database model in this way:

1. I have four tables total
   - Two tables (the teacher table and school table) have information that should be fairly constant based on my project assumptions (name, email, consent, one time documents received)
     - If at any time this information changes (e.g., withdraw status, new last name, new contact person), 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. For example, I may put grade level in my longitudinal tables if I collect data across years and assume it's possible that teachers may switch grade levels.
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 10.6).

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, payment sent to school).

10.2.3 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

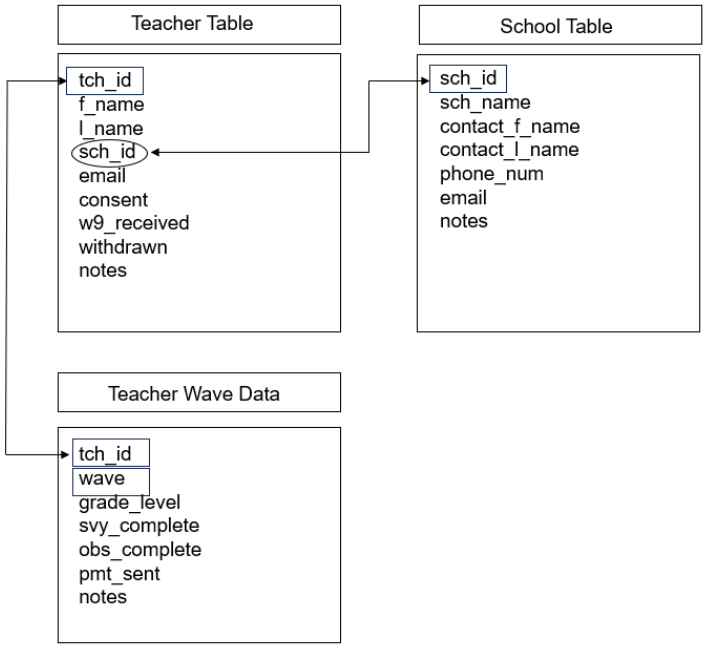


Figure 10.6: Example participant database model

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. Depending 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
- Inclusion/exclusion criteria status

- Enrollment status
- Randomization (treatment/control)
- Grouping information (cohort)
- Summary information not already accounted (# of consents sent out, # of students in class, # of teachers in school)
- Administrative data status (W-9 received, MOUs received)
- Movement/withdraw 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/withdraw
  - Communication with participants
  - Reasons for missing data
  - Errors in data

10.2.3.1 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 9

10.2.4 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 you 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 (i.e., your institution has a license for)?
- Consider security
    - Which tools are approved by your institution to protect the sensitivity level of this data (See Chapter 4)?
    - Can you limit access to the entire database? To specific tables?
        * If multiple people are entering data, you may want to restrict access/editing capabilities for some tables
    - Protect data loss
        * Can you backup the system?
        * Can you protect against overwriting data?
        * Can you keep versions of the database in case a mistake is ever made and you need to go back to an older version?
- Data quality protection
    - Can you set up data quality constraints (e.g., restrict input values/types)?

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

10.3 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 manually entering data, 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.

10.3.1 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 10.7: Example spreadsheet view data entry

10.3.2 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

  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 the ways mentioned in Section 10.3. 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.

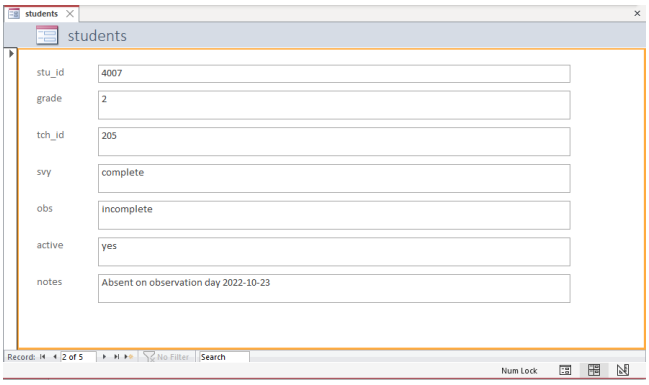


Figure 10.8: Example form view data entry

10.4 Creating unique identifiers

One of the most important parts of keeping this participant tracking database is assigning unique participant identifiers. As soon as participants are entered into your database, a unique study ID should be assigned. If confidentiality was promised to schools or districts, you will also want to assign identifiers to sites as well. 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 (i.e., codes), are what will allow you to uniquely identify and link participants in your data.

Participant unique identifiers are numeric or alphanumeric values and typically range from 2-10 digits. 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.

Before assigning identifiers, it can be very helpful to develop an ID schema during your planning phase, and document that schema in an SOP (see Section 8.2.7). 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 10.9). The participant still keeps the same ID throughout the study. In these cases, you will use a combination of variables to identify the unique instances of that participant (e.g., `stu_id` and `cohort`).
    - 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.

2. Participant identifiers must be unique within and across entities

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 10.9: Example of keeping participant IDs for the entire study

- 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

3. The identifier should be randomly assigned and be completely distinct from any personal information to protect confidentiality.
   - Do not sort by identifying information (e.g., names, date of birth) and then assign IDs in sequential order
   - Do not group by identifying information (e.g., grade level, teacher) and then assign IDs in sequential order
   - Do not include identifying information (e.g., 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 a project-level ID or acronym into a participant ID to help with tracking of information, especially when running multiple studies using identical forms across studies. This is absolutely okay because it is assumed this information never changes.
   - 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 participants 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 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 10.10: 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 11

Data Collection

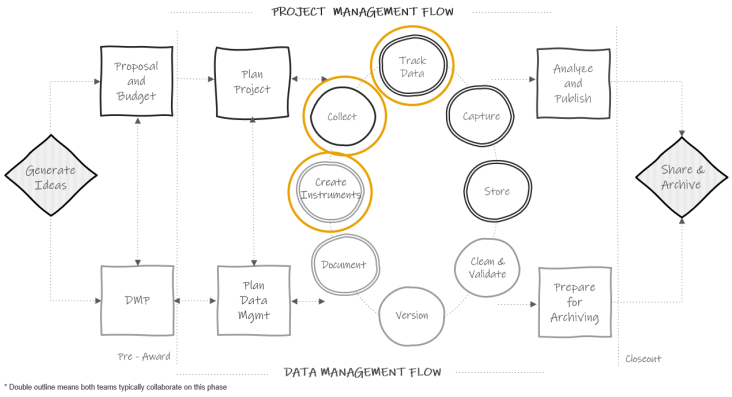


Figure 11.1: Data collection in the research project life cycle

When collecting original data as part of your study (i.e., you are administering your own survey or assessment as opposed to using an externally collected data source), 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 of Figure 11.2, 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 11.2: A comparison of data collected without planning and data collected with planning

11.1 Quality assurance and control

When planning your data collection efforts, first pull out your data sources catalog (see Section 8.2.2). This document will be a guide during your data collection planning period. Recall that every row in that document is an original instrument to be collected for your study. Some of your data sources may also include external datasets, which we will discuss in Chapter 12.

In addition to planning data collection logistics for your original data sources (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 2023). As shown in Figure 11.1, 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

11.2 Quality assurance

Education researchers collect original data in many ways (see Figure 11.3). The focus of this chapter will be on data collected via forms (i.e., a document with

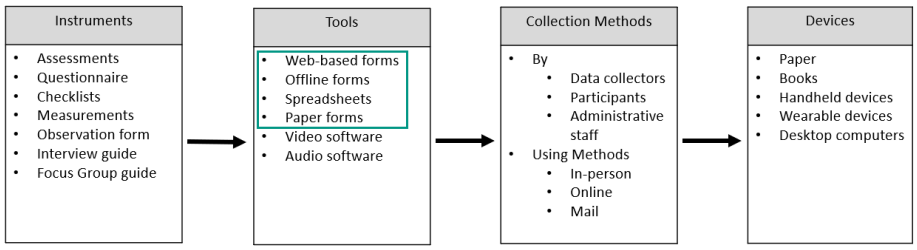


Figure 11.3: Common education research data collection methods

spaces to respond to questions). Forms are widely used to collect data in education research (i.e., think questionnaires, 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. 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 building data collection tools in a way that produces valid, reliable, and more secure data. When creating your original data collection instruments, there are five 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
5. Ensure compliance

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.

11.2.1 Questionnaire design

In Chapter 8 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 each data dictionary for each original data source, 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 2023). Question ordering, response option ordering, question wording, and more can all impact participant responses. While questionnaire design is actually outside of the scope of this book, I have a few tips to help you collect more valid, reliable, and ethical survey data. In addition to following these tips, make sure to consult a methodologist when designing your questionnaire.

1. Use existing standards if possible
    - Organizations such as the National Institutes of Health (2023b) and the National Center for Education Statistics (2023) have developed repositories (Common Data Elements[1] and Common Education Data Standards[2]) 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 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,

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

legal guardian, and so forth)
    - Rather than asking "Do your children not have siblings?" which can be confusing, remove the negative and ask "Do your children have siblings?" (Reynolds, Schatschneider, and Logan 2022)
    - Is the question leading/biased?
        * Are the response options ordered in a leading way?
    - Is there no one way to answer this question?
        * Are response categories mutually exclusive and exhaustive (ICPSR 2020)?
3. Consider data ethics in your questionnaire design (Gaddy and Scott 2020; Kaplowitz and Johnson 2020; Kopper and Parry 2021; Mathematica 2023; Narvaiz 2023)
    - 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? Do the benefits outweigh the risks?
        - 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 (PII)
    - 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 PII that is absolutely necessary for your project, and no more (Gaddy and Scott 2020). As discussed in Chapter 4, PII can include both direct identifiers (e.g., name or email) as well as indirect identifiers (e.g., date of birth). 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 [3]
Pew Research Center	Writing survey questions [4]
Stefanie Stantcheva	How to run surveys: A guide to creating your own identifying variation and revealing the invisible [5]
World Bank	Survey content-focused pilot checklist [6]

11.2.2 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 11.5):

1. Gathering internal feedback on items
    - As discussed in Section 8.4.1, 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. Consult with your IRB to determine if approval is required before piloting your instrument with your study population.
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 in Section 11.2.4.

Last, as you move through the piloting phases, remember to make updates not only in your tool but also in your data dictionary and any other relevant documentation (e.g., data cleaning plan).

11.2.3 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 11.3). Research

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

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 11.4: Data collection instrument pilot phases

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, sensitivity levels of data, 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., branching logic, automated email reminders, an alert system for ecological momentary assessments, options to embed data, options to calculate scores in the tool)?
2. Compliance and security
    - Consider the classification level of each data source (See Chapter 4)
        - Which tools are approved by your institution to protect the sensitivity level of your data?
    - 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, data collection tools can be categorized 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, especially when the research team has control over how the data collection tool is built (Cohen, Manion, and Morrison 2007; Douglas, Ewell, and Brauer 2023; Gibson 2021; ICPSR 2020; Malow et al. 2021; Society of Critical Care Medicine 2018; Bochove, Alper, and Gu 2023).

Electronic data collection benefits	Paper data collection benefits
• Able to build in 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) • Able to use automated logic to prevent inconsistencies in data • Response validation reduces the chances of missing data • Potential to reach broader populations (e.g., crowdsourcing) • Quicker turnaround of analysis-ready data and provides the opportunity to build real-time reporting pipelines (e.g., using APIs)	• Intuitive to create and use (no training required) • Easy to do cognitive checks (eyeball for errors) • Easier to catch errors early on (in the field)

Figure 11.5: Comparison of data collection tool benefits

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 individual 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 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[7]
Washington State University Libraries	Software for sensitive data[8]
Benjamin Douglas, et al.	Data quality in online human-subjects research comparison of tools[9]

11.2.4 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 Schectman 2014; Lewis 2022b; UK Data Service 2023). 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.

11.2.4.1 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).

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

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. Include all items from your data dictionary
    - This includes all substantive questions, as well as items that are necessary for linking purposes (e.g., participant identifiers, rater ids for inter-rater reliability)
    - This does not include any variables from your data dictionary that will be derived (e.g., sum scores) or any grouping variables that will be added in during the data cleaning phase (e.g., treatment, cohort)
2. Name all of your items the correct variable name from your data dictionary (UK Data Service 2023)
    - For example, instead of using the platform default name of `Q2`, rename the item to `tch_years`
    - As mentioned in Section 9.4, 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.
3. Code all values as they are in your data dictionary
    - For example, "strongly agree" = 1, "agree" = 2, "disagree" = 3, "strongly disagree" = 4
    - Many times 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
4. Use data validation to reduce errors and missing data (UK Data Service 2023)
    - 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*)
    - Response validation
        - Consider the use of forced-response and request-response options to reduce missing data
            * Forced-response options do not allow participants to move

forward without completing an item. Request-response options notify a respondent if they skip a question and ask if they still would like to move forward without responding
    - Be aware that adding a forced-response option to sensitive questions has the potential to be harmful and produce bad data. If adding a forced-response option to a sensitive question, consider allowing those participants to opt-out in another way (e.g., "Prefer not to answer").
5. 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, 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.
6. If there is a finite number of response options for an item, and the number isn't too large (less than ~ 20) use controlled vocabularies (i.e., a predefined list of values) rather than an open-text field (OpenAIRE_eu 2018; UK Data Service 2023)
    - 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
7. 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 cannot 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
8. 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 prevents confusion in case a participant or data collector swaps the order of information
9. To protect participant privacy and ensure the 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
10. 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 11.5). 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 using your chosen data capture process (see Chapter 12) 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.

Last, this is also the time to update your data dictionary. As you review your exported file, update your data dictionary to reflect any unexpected variables that are included (e.g., metadata), any unexpected formatting, as well as any newly discovered recoding or calculations that will be required during the data cleaning process. As an example, if upon downloading your sample data you learn that a "select all" question differently than you expected, now is the time to add this information, along with any necessary future transformations, to your data dictionary. This is also a great time to update your data cleaning plan with any new transformations that will be required.

11.2.4.2 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.

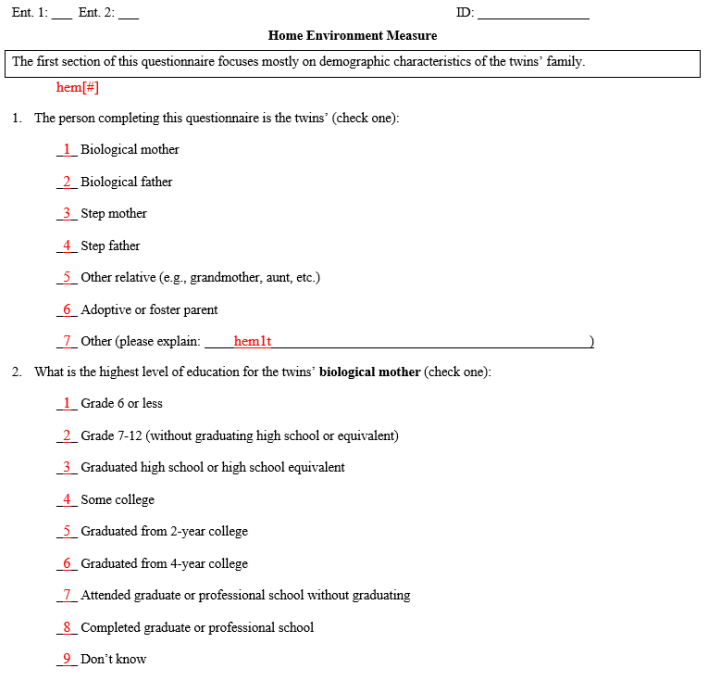
1. Use your data dictionary as a guide as you create your paper form
    - Make sure all questions are included and all response options are accurately added to the form
2. Have clear instructions for how to complete the paper form (Kopper and Parry 2021)
    - Make sure to not only have overall instructions 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*, 3 digit number)
        - How many answers should be recorded (e.g., circle only one answer, check all applicable boxes)
        - How to navigate branching 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 11.4). Using the feedback collected, edit your tool as needed before sending it out into the field.

Last, unless paper data it is collected using a machine-readable form, it will need to be manually entered into an electronic format during the data capture phase. While we will talk about data entry specifically in Section @ref(#capture-paper), this point in instrument creation is a great time to create an annotated instrument (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 annotated instrument can be useful during the data entry process and serve as a linking key between your instrument and your data dictionary (see Figure 11.6) (Hart, Schatschneider, and Taylor 2018).

11.2.4.3 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 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):



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 11.6: Annotated instrument from The Florida State Twin Registry project

- 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 (see Section 14.3.1 for more information about this process).

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

> **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 that if your tool collects identifying metadata such as IP Address or worker IDs in the case of crowdsourcing tools (e.g., MTurk), this information will not be included in your downloaded data. Remember 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 from that entity in order to cluster on that information (e.g., school name).

11.2.4.3.1 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 identifying information (e.g., names, emails) in the data.
    - If using this method, make sure to build a data check into the system. For example, 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 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 collecting data in a computer lab or on tablets at a school site, or if your tool does not have the option to create unique links
    - This can possibly introduce error if a participant enters their study ID incorrectly.
        - Similar to the first option, after a participant enters their ID, verify their identity
    - 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)

11.2.4.3.2 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 names.
- 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)

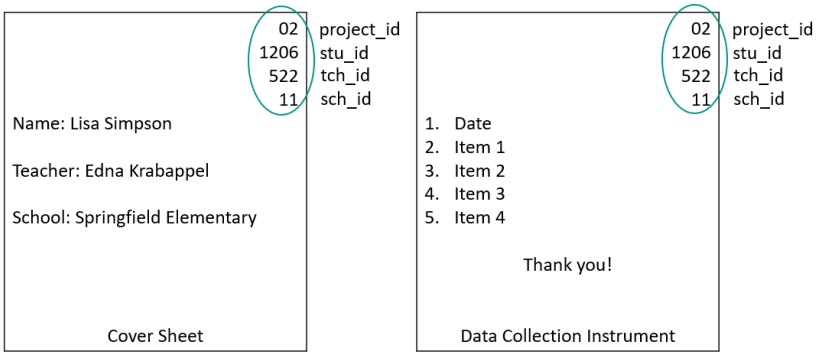


Figure 11.7: Example cover sheet for a paper data collection instrument

11.2.5 Ensure compliance

If you are collecting human subjects data and your study is considered research (see Chapter 17 for definitions of these terms), it is important to consult with your applicable institutional review board (IRB) about their specific requirements before moving forward with any data collection efforts. As discussed in Chapter 4, an IRB is a committee that assesses the ethics and safety of research studies involving human subjects. If an IRB application is required for your project, the review process can take several weeks and it is common for the IRB to request revisions to submission 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.

Informed consent agreements, and assents for participants under the age of 18, are commonly required by IRBs for research studies that collect human subjects data. As discussed in Chapter 4, these agreements ensure that participants fully

understand what is being asked of them and voluntarily agree to participate in your study. There are several categories of information that will be required for you to include in your consent form (e.g., description of study, types of data being collected, risks and benefits to participant, how participant privacy will be maintained) (The Turing Way Community 2022). Make sure to consult with your applicable IRB about what should be included. However, with an increase in federal data sharing requirements, it is very important at this time to also 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 (e.g., 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 that 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 participants are fully aware of the data collected from them and can make an informed decision about the future of that data.

Consult with your IRB to determine the preferred method for obtaining consent for public data sharing. If you use 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 10), 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. You will also want to consider who is included in your final analysis sample. If including all consented participants in your analysis, your publicly available dataset will not match your analysis sample if some people did not consent to data sharing. You may need to consider options such as using a controlled-access repository to share the full sample for purposes of replication. We will discuss different methods of data sharing in Chapter 15.

Templates and Resources

Source	Resource
Anja Sautmann	Annotated informed consent checklist [10]
Holly Lane, Wilhemina van Dijk	Example parent consent [11]
Jeffrey Shero, et al.	Informed consent and waiver of consent cheat sheet [12]
Jeffrey Shero, Sara Hart	Informed consent template with a focus on data sharing [13]
Melissa Kline Struhl	Lookit consent form template 5[14]
University of Virginia	A collection of consent and assent templates [15]

Note The security of consent and assent forms should be a top priority to your team. Not only does it contain identifiable particiant information, but without this form, you no longer have consent to collect a participant's data. Whether you collect consent on paper or electronically, make sure you have a clear plan that includes:
- Using institution and IRB approved tools to collect consent
- If collecting paper consent, being able to clearly read a participant's name and other relevant information collected (e.g., participant printed name or signature alone may not be sufficient due to duplicate names, nicknames used, or illegible handwriting). One option is to pre-print names and other relevant information on forms or have school staff write participant names on forms before handing them out.
- Capturing consent forms and storing them securely and consistently. If forms are collected in the field, make sure they are promptly returned to the office and stored securely.

[10] https://www.povertyactionlab.org/sites/default/files/research-resources/rr_irb_annotated-informed-consent-checklist_0.pdf
[11] https://www.ldbase.org/system/files/documents/2021-04/HS-ParentConsent.txt
[12] https://osf.io/3czbx
[13] https://figshare.com/articles/preprint/Informed_Consent_Template/13218773
[14] https://github.com/lookit/research-resources/blob/master/Legal/Lookit%20consent%20form%20template%205.md
[15] https://research.virginia.edu/irb-sbs/consent-templates

11.3 Quality control

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

11.3.1 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 electronic data collection devices (e.g., phones, tablets) are password protected and never left open and unattended. Keep all identifiable information encrypted on your 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 use 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)
    - Who collected the data (i.e., data collector initials or ID)
    - Date the data was collected
    - 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 physical 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 calculations and scoring (e.g., basals, ceilings, raw scores)
- Assign a field supervisor. This person is assigned to:
  - Do another round of data checks in the field once the data collector returns physical 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
  - Be available for trouble shooting as needed
- Do another round of physical data spot checking as soon as the data is returned to the office (see Figure 11.8)
  - 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 paper forms are mailed back to you from participants, rather than returned from field data collectors, it is still important to do in-office spot checks. If at all possible, reach out to those participants for any corrections.
- When a wave of data collection wraps up, collect feedback from data collectors to improve future data collection efforts
  - What went well? What didn't?

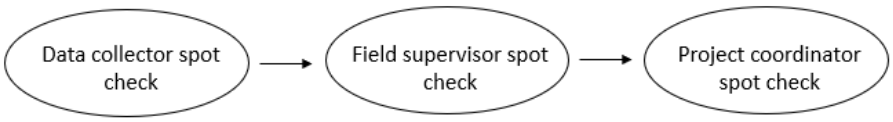


Figure 11.8: A series of spot checks that occur with paper data

Tracking sheet templates

Source	Resource
Crystal Lewis	Field tracking sheet template [16]

[16] https://docs.google.com/spreadsheets/d/1CeIXvTBtU9O3GNzfFaAMtb69Z2HyOLuPQWsxy7jOWdU/edit?usp=sharing

11.3.2 Ongoing data checks

If you collect 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, survey comprehension) (DIME Analytics 2021a; Gibson 2021).

- Checks for comprehension
  - Are any questions being misinterpreted?
- Checks for missing data
  - Are items being skipped that should not be skipped?
  - Are participants/data collectors not finishing forms?
- Checks for ranges and formats
  - Are values in unexpected formats or falling outside of unexpected ranges?
- Checks for duplicate forms
  - Are there duplicate entries 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 programmatically (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.

> **Note** All of the web-based data collection efforts in this chapter 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, 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 to the instrument. 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 (Arndt et al. 2022; Simone 2019;

Teitcher et al. 2015):
- Not posting the link on social media
- Using CAPTCHA verification, or a CAPTCHA alternative, to distinguish human from machine
- 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) Last, check your data thoroughly for bots or fraudulent responses before analyzing it and before providing payments to participants. The following types of things are worth looking into further:
- Forms being completed in a very short period of time
- Forms being collected from suspicious geolocations
- Duplicated or nonsensical responses to open-ended questions
- Nonsensical responses to attention or logic checking questions
- Inconsistent responses across repeated questions

11.3.3 Tracking data collection

Throughout data collection your team should be tracking the completion of forms (e.g., consents, paperwork, data collection forms) in your participant database (see Chapter 10). This includes paper forms, electronic forms stored on devices, as well as web-based data coming in. 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 people 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 but do not track it as "complete" until 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"
    - Review all data before marking it as complete, including consents, assents, and other administrative forms. 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.

11.3.4 Collecting data consistently

As mentioned in Chapter 9, 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, link for SchoolA and link for SchoolB):

- 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
  - For example, 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 (e.g., "no" = 0, "yes" = 1)
- 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; Pew Research Center 2023).

11.4 Review

Recall from Chapter 6, we discussed designing and visualizing a data collection workflow during your planning phase. As we've learned from this chapter, errors can happen at any point in the workflow so it is important to consider the entire data collection process holistically and integrate both quality assurance and quality control procedures throughout. Figure 11.9 helps us to see when these practices fit into the different phases our workflow.

Once your workflow is developed and quality assurance and control practices are integrated, consider how you will ensure that your team implements these practices with fidelity. Document the specifics of your plan in an SOP (see Chapter 8), including assigning roles and responsibilities for each task in the process. Last, train your team on how to implement the data collection SOP, and implement refresher trainings as needed.

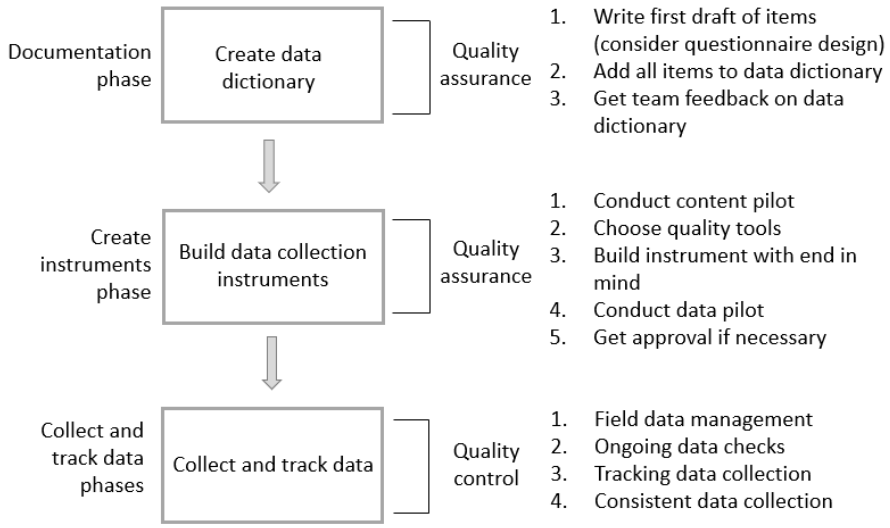


Figure 11.9: Integrating quality assurance and control into a data collection workflow

Instrument Workflow Resources

Source	Resource
DIME Wiki	Questionnaire design timeline[17]
Sarah Kopper, Katie Parry	Five key steps in the process of survey design[18]

[17]https://dimewiki.worldbank.org/Questionnaire_Design
[18]https://www.povertyactionlab.org/resource/survey-design

Chapter 12

Data Capture

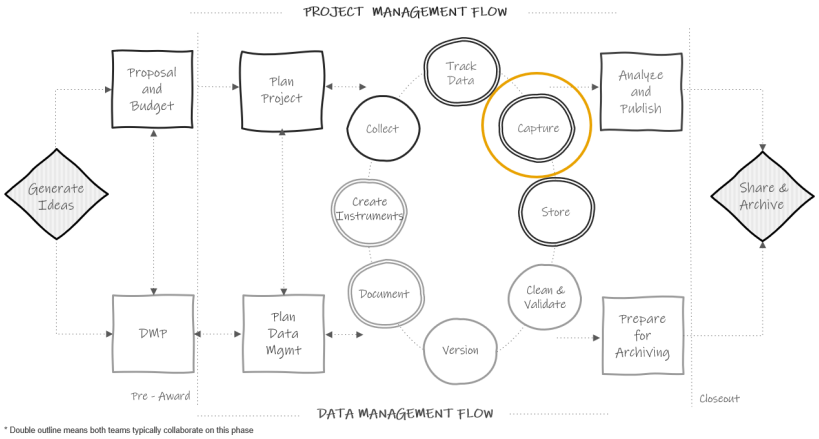


Figure 12.1: Data capture in the research project life cycle

After the data collection period is complete, the next phase in the cycle is to capture the data, meaning extracting, creating, or acquiring a file that we can save in our designated storage location. In quantitative research we typically want to capture data in an electronic, rectangular format (see Chapter 3). In this chapter we will review common ways to capture data based on three data collection methods (see Figure 12.2). Similar to data collection, it is possible for data errors to occur during this phase. In reviewing data capture methods, we will also cover how data quality can be managed during this phase.

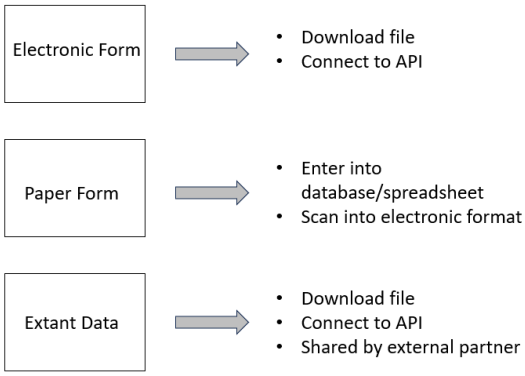


Figure 12.2: Common data capture methods

12.1 Electronic data capture

As discussed in Chapter 11, electronic data can be collected using a variety of software (either web-based or offline). Since electronic forms typically funnel data into a spreadsheet or database, it makes the process of data capture much easier compared to paper data. However, there is still much to consider.

1. How will the data be captured?
    - The most common way to capture data in an electronic format is to download it from your platform.
    - Another option you may have for web-based forms is to capture data via an API (application programming interface). If you regularly need to review your data before your final capture, using an API can be a great way to remove the burden of manually logging into a program and going through the point and click process of downloading a file. Instead you can write a script, in a program such as R, to extract the data. Once the script is created, you can run it as often as you want. However, this is only an option if your tool has an API available (e.g., Qualtrics).
2. What file type will the data be captured in?
    - Most electronic data collection tools provide an option to export to one or more file formats (e.g., SAV, CSV, TXT). It is important to choose a file type that is analyzable (i.e., rectangular formatted), as opposed to something like a PDF file. The rectangular file type you choose will mostly depend on your project plans. Things to consider might include:
        - Do you want the text values for responses or the numeric values? Your choice may limit your options (e.g., text values may be available in a CSV but not in an SPSS file).
        - Do you want embedded metadata, such as variable and value

labels, in your raw file? Again, your choice will narrow your options (e.g., an SPSS file allows you to export the numeric values while also being able to view the variable and value labels in the file).
  - Do you want a non-proprietary, interoperable format? If yes, you will not want to capture data in file types such as XLSX and SPSS that require proprietary software to view.
  - Will any file types create issues for your variables?
    * For instance, Microsoft Excel is well-known for applying unwanted formatting to values. As an example, if your assessment tool collects age in the format of "years-months", oftentimes Microsoft Excel will change this variable into a date, converting a value such as "10-2" (10 years and 2 months old) to "2-Oct". A more suitable file type in this situation may be a CSV or TXT file, which do not apply formatting.
  - Is there a file structure that you don't want to work with?
    * As an example, the structure of an SPSS file may look different compared to a XLSX file depending on the tool. In a tool like Qualtrics, an XLSX or CSV file may export with multiple header rows whereas an SPSS file does not.

3. What additional formatting options need to be considered?
   - In addition to choosing a file type, there may be other options that your tool allows you to consider. Examples of what these might look like are:
     - Do you want to export the text values or numeric values for categorical items?
     - How do you want to export "select all" questions?
       * Depending on your chosen file type, you may be allowed to choose how you want to format "select all" questions. Typically your options are to export them in one variable, where each option is separated by a comma, or you can split each option into its own column.
     - How do you want to recode seen but missing values?
       * This option is commonly provided because "select all" questions are often split out into multiple variables, where a `1` indicates the option was selected, and blank represents either the option was **not selected** or that the item was skipped entirely.
       * Typically tools provide the option to recode to *0* or *-99*. You can also choose not to recode and leave those responses as blank. If the types of missing data do not matter to your study, then leaving missing values as blank is typically the most straightforward option. However, adding an extreme value like *-99* can make it easier to know if those blank "select all" values are a "no" response (recoded to *-99*) or if those values were never seen and actually represent a missing value

(left as blank).

4. Where will the file be stored?
    - This decision should be based on guidelines laid out in your style guide (see Chapter 9) and your applicable data security documents and agreements (i.e., data management plan, data security plan, research protocol).
5. How will files be named?
    - While your tool may provide a name for your file, it may need to be renamed something more meaningful based on your style guide rules (e.g., `projecta_w1_stu_svy_raw_2023-09-01.csv`). Most importantly, name your files consistently across data sources and waves.
6. What documentation needs to accompany the data capture?
    - As discussed in Section 8.3, there are additional documents that can be helpful to include alongside the file.
        - A README can be very beneficial to include if there is anything in the file that a future person managing the data should be aware of.
        - A changelog can also be very beneficial. It is common to have to redownload a raw data file due to errors found or new participants added. A changelog can help the team both identify the most recent version of a raw data file, as well as understand the differences between files.
7. Who will capture the data?
    - It doesn't necessarily matter who takes on this responsibility. What matters most is that the person has the expertise to capture the data and that this responsibility is documented. If the person capturing the data is not the person who oversees data collection, it is important to still assign that person the responsibility of documenting any relevant information in the README.

8. What checks need to happen before this data is handed off?
    - It is important for the person responsible for data capture to do a basic review of the file before handing this data off for the next step.
        - Does the format of the file look as expected? Does it have data in it? Are all the variables there as expected?
        - Are all participants in the data? This is an excellent time to compare the number of unique participants in the file to the number of participants with completed data in your participant tracking database. If these numbers do not match, the person in charge of data capture should begin reconciling errors before handing off this data.
            * Was a participant accidentally dropped from the file? Is someone incorrectly marked as complete in the tracking database? Are there duplicate entries in the file?
                · If there are errors that can be corrected (e.g., someone incorrectly tracked a data point, a participant was left off

in the capture process), those corrections should be made now. If there are corrections that involve manipulating the raw data (e.g., reconciling duplicate IDs in the data), those corrections should not be made at this time. Instead, those should be added to a README file to be corrected in the data cleaning phase.

Note It is important to never make changes directly to the raw data files. This also includes not making changes directly to the data in your data collection tool. If you see errors in the raw data file that can't be fixed by simply re-downloading the data, make notes in a README for future correction as noted above. Those corrections can be made in the data cleaning process. The one exception to this rule is if you accidentally collect data on a non-consented participant. In this case, it may be best to delete data for this participant directly in your data collection tool so that no record is kept.

All of these decisions should be made and documented during the time you are developing data collection tools. Making these decisions early allows you to also implement them during the pilot testing and data checking processes. For instance, if you plan to capture your data by exporting a CSV file from your data collection platform with a variety of options selected, you will want to use this same method during your data piloting and data checking process. This allows you to know exactly what your data will look like once data collection is complete and make adjustments as needed.

As discussed in Chapter 6, your data capture process should be added to your workflow diagram and then detailed in an SOP. All of the decisions above should exist in the relevant SOP. This ensures that workflows are standardized and reproducible. As we've learned in this section, one deviation from the SOP has the potential to produce a very different data product (e.g., the format of a CSV file compared to an SPSS file can vary). Not only can this produce errors but it can also undermine the reproducibility of a data cleaning pipeline. Imagine a scenario where a data cleaning syntax is written to import a CSV file with an expected format, and that format changes. The pipeline is no longer reproducible. Last, documenting a timeline for when this data capture process should occur can also be beneficial for both the person responsible for data capture, as well as people responsible for subsequent phases such as data cleaning.

12.2 Paper data capture

The most common method for capturing paper forms is manual entry. While capturing electronic data is fairly quick and straightforward, planning for and implementing paper data entry is much more involved. Similar to electronic data collection, you will want to start planning data entry long before your

data is collected, and you will need to build your data entry tool before the data capture phase (e.g., when you are creating your data collection tools).

As you can imagine, manually entering data comes with the potential for many data quality issues. In developing a data entry process, it is important to implement quality assurance practices similar to those we discussed in Section 11.2.

1. Choose a quality data entry tool
2. Build your data entry form with the end in mind
3. Develop a data entry procedure

12.2.1 Choose a quality data entry tool

When choosing a data entry tool, if you are already using a relational database for your participant tracking, it may make the most sense to use this same database for data entry so that data can be stored in one location and tables can be linked (e.g., REDCap, FileMaker, Microsoft Access). However, if you need to choose a new tool for data entry, the criteria for choosing one will be similar to those reviewed in Section 11.2.3. Considerations for project needs, security, costs, and data quality should all still be reviewed.

In addition to reviewing those criteria, it can also be very beneficial to use a tool that allows you to create entry forms, similar to the form we saw in Figure 10.8, rather than entering directly into a spreadsheet. Building a data entry form that is laid out similar to the paper form can help reduce errors in data entry. Data that is entered into the form is then fed into a table that can be exported.

If however, you choose to use a spreadsheet program such as SPSS or Microsoft Excel for data entry, it is important to be aware of some of the limitations and possible issues with these tools including:

- Possible formatting issues
  - For example, Microsoft Excel formatting may cause errors in your data (e.g., dates get formatted as numeric, strings get formatted as dates, leading zeros get dropped from values)
- Potential to skip around
  - In a spreadsheet, the ability to click anywhere makes it very easy to enter data into the wrong cell or to skip cells completely (Eaker 2016). You may even write over existing data on accident. It's also possible to incorrectly sort data resulting in errors (Reynolds, Schatschneider, and Logan 2022).

12.2.2 Build with the end in mind

When you export or save a dataset from your data entry tool, it should meet all of our data structure rules (see Chapter 3), and all of the variables should

be formatted as we have described in our data dictionary, including correct name, variable type, and allowable values. In order to accomplish that goal, you need to build your data entry screens, whether in a spreadsheet or form layout, following rules similar to those discussed in Section 11.2.4.

1. Make sure that your items are laid out in the same order that they appear on the paper form so that people entering the data can easily follow the flow (Reynolds, Schatschneider, and Logan 2022).
2. Using the annotated instrument we discussed in Section 11.2.4.2, name all of the items on your data entry screen to match the final item names (e.g., instead of `Q2` use the final name `tch_years`).
3. For quicker data entry, with less errors, allow people to enter the numeric values associated with response options on the annotated instrument rather than the text values (e.g., enter *1* rather than "strongly disagree"). Or if you prefer to use text values, build those as drop-down values, removing variation in entry.
4. No matter which data entry tool you choose, make sure to include both content and response validation
    - Restrict data type, format, ranges, and values
    - Do not allow people to skip over items

Before releasing your data entry tool into the world, you will want to pilot it for issues, just like we did for electronic data collection tools (see Section 11.2.4.1). Collect sample responses from team members and collect feedback on what did or did not work well for them while entering data. Then download, using your chosen download format, or simply review the data if it is already in its final format (e.g., Microsoft Excel). Check that the data looks as you expect it to and make edits to the entry tool as needed.

12.2.3 Develop a data entry procedure

While building a reliable data entry tool is absolutely important in ensuring data quality, developing a clear and standard data entry process is even more important. Make sure to create a data entry process that includes the following things.

1. Where paper forms are stored, how they should be pulled, and how they should be returned to the storage location.
    - Consider organizing your forms in a way so that people entering data know what has been entered and what has not been entered
2. Where electronic entry databases or files are stored and how they will be named.
    - Similar to Section 12.1, you will want to name these files according to your style guide (e.g., `proja_w1_stu_svy_raw_entry1.xlsx`).
3. Specific data entry rules to follow
    - What values to enter for categorical variables (numeric values or text values)

- If any items allow free-text entry, provide specific data entry rules to prevent inconsistencies. While adding data validation will help remove some inconsistencies, further rules may be needed depending on the items. As an example:
  - Enter decimals with a leading zero (e.g. *0.4* not *.4*)
  - Enter "yes" values as "Y" (e.g., change any values of "y" or "yes" to "Y")
  - Only enter numeric values for measurements (e.g., *5* not "5cm")
- How to code missing data
- What to do if someone comes across a variety of common data errors. As an example:
  - Someone who has circled more than one response to an item
  - Someone who has written responses in the margin
  - Someone who has written a value out of range or an unallowable response

4. How to denote that a form has been entered
   - For example, staff can write their initials on a form after entry
5. Steps to be performed before handing off the entered data
   - Similar to the process in Section 12.1, it is imperative that whoever is overseeing the data entry process do a check of the data before handing it off for the next step of data cleaning. Most importantly, check to see that the correct number of participants exist in the file compared to your participant tracking database (e.g., no duplicate entries, no missing entries). If data entry or data tracking errors exist, fix mistakes as needed. If inherent data issues exist, make notes in a README to be corrected in the data cleaning phase.
6. Last, similar to electronic data capture (see Section 12.1), you will want to make decisions about final file types (e.g., this will be relevant if you use a database for data entry and need to export files), where final files will be stored, and how they will be named.
7. Who will oversee this process? (i.e., creating entry forms, answering questions, conducting final checks)

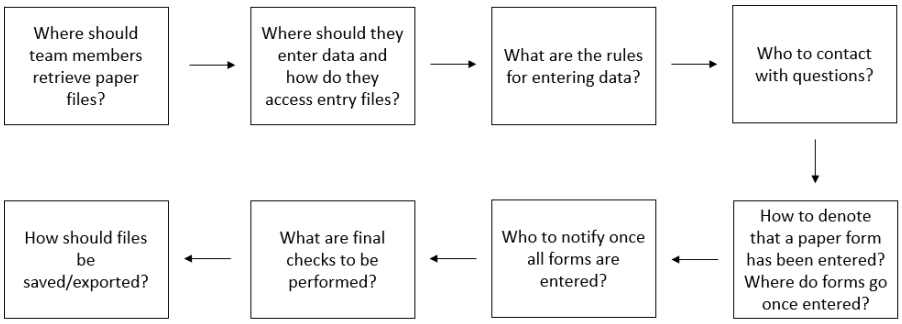


Figure 12.3: The flow of the decisions to make regarding the data entry process

Note As a reminder, the data capture phase is a time to do only that, capture the data that is already collected. This is not a time to score, calculate, or add additional fields. This is a time to enter the exact items that are found on the form. Creating additional variables or performing further data quality checks will occur during the data cleaning phase. The exception to this rule is if you collected an assessment that requires entry into a proprietary scoring program. Once data is entered, these tools often export a file that includes derived scores for the assessment and these are still considered raw captured data sources. As an aside, if only raw scores are entered into your scoring program, you may consider also entering raw item-level information using your own data capture process. Unless item-level variables are proprietary or include identifiable information, it can be beneficial to include both item-level and summary variables in your final project datasets.

12.2.3.1 Double entry

Last, it is important to integrate quality control into this data entry process. In their studies, Schmitt and Burchinal (2011) have found an error-rate between 5-10% when data are entered only once and that having a second person double-check data entry improves data quality. While there are several ways of double checking data including visual checking and read aloud methods, the double entry method has been shown to be the most reliable error-reducing technique (Barchard et al. 2020), ensuring that what is displayed on the paper form is what is entered into the database. A typical double entry process looks something like this:

1. A designated team member creates two identical entry forms. One person enters forms in the first entry screen, a different person enters forms in the second entry screen. Depending on your tool this might be two separate files, two separate tabs in a spreadsheet, or two separate tables or forms in a database.
    - It is important here that the second entry is completed by a different person so that systematic errors that are created by one person's interpretation of information are not repeated across files.
2. When both entries are complete, a system is used to check for inconsistencies across datasets.
    - This system varies across tools. Some tools have built in systematic ways to check for errors across entry screens. Other tools may require you to build your own system (e.g., write formulas to compare cells or draft syntax to compare spreadsheets). Ultimately, once those comparisons are done, you should have a report that tells you where errors exist across the two forms.
3. Using that report, a designated team member/members makes corrections (Yenni et al. 2019). This involves pulling out the original paper forms and

seeing what the correct value is for each error.

- There are varying ways you can make corrections at this point. You can make corrections just to one form, you can make corrections to both forms, or you can make corrections in a third, new form that contains all of the correct data. Different tools will handle this in different ways.
- However, if you are creating your own system, consider making corrections in both forms. In this way, you make a correction to which ever entry file has the error. Once all corrections are made, you can run your comparison system again, which will now let you know if all errors have been corrected. Once all errors are fixed, you can simply choose either file to be your "master" raw data file.
  - Figure 11.4 is an example using this process. Data has been entered in two spreadsheets. Then both files were imported into R where, in this particular example, a function from the `diffdf` package (Gower-Page and Martin 2020) was run to check for errors and a report was returned [1]. You can see that it identifies an error in our `stress1` variable. Entry file 1 (*BASE*) has a different value than entry file 2 (*COMPARE*). I would now need to go back to the original files to see what the actual reported answer was and fix the value in the corresponding file. If the value was incorrect in both files, I would correct it in both and then run my comparison system again to ensure no more errors exist before handing the file off.

Depending on the amount of data that is collected this can be a time consuming process. Double data entry is a matter of weighing costs and benefits. While double entering all of your data is the best way to reduce data errors, the cost of double entering all of your data might be too high, and you may decide to only double enter a portion of your data and gain a smaller benefit.

Whatever your decisions are throughout this process, document every decision in an SOP and assign team members for each step. This includes assigning someone to create double entry files, oversee data entry, create a double entry comparison system, conduct the comparison, make corrections, and do the final checks before handing the data off. Make sure to train your team on this system so that it is implemented consistently.

> **Note** If you are entering data into a proprietary scoring system that does not provide a double entry option, make sure to consider other ways you can reduce data entry errors (e.g., batch upload of double entered raw values).

[1] https://cghlewis.github.io/data-wrangling-functions/compare-data-frames/compare-df.html

```
Differences found between the objects!

A summary is given below.

Not all Values Compared Equal
All rows are shown in table below

  ============================
   Variable  No of Differences
  ----------------------------
   stress1           1
  ----------------------------


All rows are shown in table below

  ================================
   VARIABLE  tch_id  BASE  COMPARE
  --------------------------------
   stress1    1235    2       4
  --------------------------------
```

Figure 12.4: A report displaying differences between two entry files

12.2.4 Scanning forms

Although less common now, it is possible that you may collect paper data using forms which can be scanned and converted automatically into a machine-readable dataset. Depending on whether your team is personally doing the scanning or whether an external company captures the data, this has the potential to save you time and energy compared to a manual data entry process. These may also have the potential to be less error-prone than manual entry, yet this process is still not error-free and caution should be taken when capturing this data (Jørgensen and Karlsmose 1998). It is still important to do data checks to ensure that the correct values were recorded in the electronic file.

12.3 Extant data

It is common in education research to also capture external supplemental data sources to either link to your original data sources or to describe information about your sample. The process for capturing this externally collected data will vary widely depending on the source. Furthermore, the quality and usability of the data can also vary widely. In this section we are going to review some practices that will help you acquire better, more interpretable data. We will divide this discussion between two types of data sources, non-public and public.

12.3.1 Non-public data sources

Non-public, or restricted-use, data sources are files that cannot be directly accessed from a public website (e.g., school records data, statewide longitudinal data systems). These data are typically individual-level and may contain sensitive, usually identifiable, information or a combination of variables that could enable identification. Acquiring these sources usually involves a data request process (see Figure 12.5). This process may or may not be part of larger request for research process (e.g., if also collecting original data in school districts). In addition to an application or proposal, this request process may also include the submission of one or more of the agreements discussed in Section 4 (e.g., informed consent, DUA, confidentiality agreement).

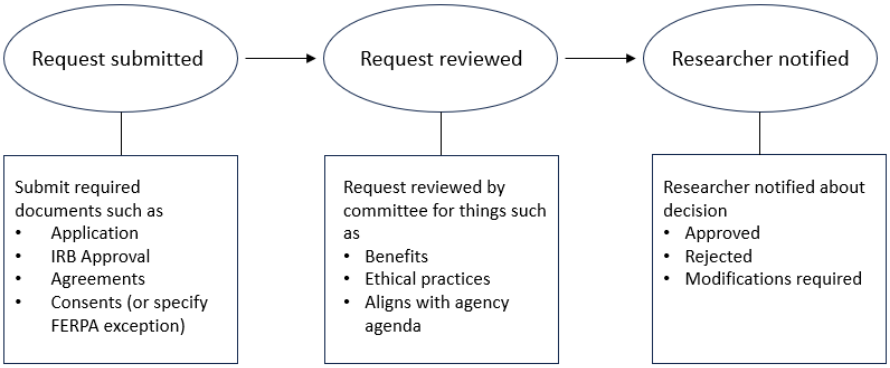


Figure 12.5: Example non-public confidential data request process

If not already included in the provider's data request process, it is important to share the following information:

1. A list of variables you are requesting
    - If you plan to link data, make sure to request a unique identifier that both you've collected and that exists in the external data (e.g., state student unique identifier), or a combination of identifiers (e.g., name and DOB), which allows you to link the external data to your existing original data.

    - If you are planning to combine data from multiple sources (e.g., multiple school districts), this can require hours of harmonization to make data comparable due to variations in how data is collected across agencies. If there is some flexibility in the request process, it can be helpful to provide details to your data provider about how you would like the variables to be formatted, helping to standardize inputs and removing any room for interpretation (Feeney et al. 2021).
        - Variable type (e.g., numeric, text)
        - Variable formats (e.g., `DOB` as YYYY-MM-DD)

  - Value coding (e.g., specify how to code FRPL categories)
  - How to handle missing data (e.g., leave cell blank)
  - How to aggregate summary data (e.g., number of days absent for the full year **or** by term)
  - For calculated variables (e.g., age at assessment) consider requesting the raw inputs to calculate your own values (e.g., request date of assessment and DOB)

Figure 12.6 is an example of how you might provide this information to a data provider.

Variable Requested	Variable Type	Description	Categorical Codes
State unique identifier	Numeric	Student state unique identifier	NA
Free or reduced-price lunch	Numeric	Free or reduced price lunch status	0 = none; 1 = reduced; 2 = free
Grade level	Numeric	Grade level for 23-24 school year	0 = pre-k; 1 = 1st; 2 = 2nd; 3 = 3rd; 4 = 4th; 5 = 5th
Date of birth	Date (YYYY-MM-DD)	Student date of birth	NA
State testing date - math	Date (YYYY-MM-DD)	Math MAP testing date for 23-24 school year	NA
State testing score - math	Numeric	Math MAP score for 23-24 school year	NA
Number of in-school suspensions	Numeric	Total number of in-school suspensions for the 23-24 school year	NA

Figure 12.6: Example variable request for an external data provider

2. Clarify the periods you are requesting data for
   - This may be the current year alone, or you may also need the previous year as well for comparison
3. Ask how and when data will be shared
   - Ask how many data files will be provided and what each file will contain (e.g., enrollment file, assessment file, attendance file)
   - Provide a preferred file format for the data (e.g., CSV file)
   - Request a timeline for when data will be shared

   - Ask how data will be shared (e.g., email, drop in a secure folder). If the data contain identifiable information, make sure to use a secure file transfer method (see Chapter 13). Once received, make sure to follow any data sharing agreements around how data should be stored.

4. Identify points of contact
   - Not only do you need contact information for acquiring the data, you also need to know who to contact for any questions or concerns that come up after the data is received.

5. Request documentation to accompany your file
    - Receiving data dictionaries or codebooks along with your data will be vital in allowing you to correctly interpret variables. This is especially important when observing variations in how variables are measured across sites or even within sites across time (e.g., a test score is measured differently in a subsequent year)
    - If documentation does not exist, provide the data provider with a form to complete that allows them to enter relevant, variable information.
        - What each variable represents
        - What each value represents if the variable is categorical
        - How each variable is captured or calculated (e.g., hand entered)
        - The universe for each variable (e.g., grades *3-5*)
        - Any data quality concerns about any of the variables
    - If you receive new exports each year, make sure to request documentation each year. It is possible that the way variables are collected or recorded change over time.

Figure 12.7 is an example of a document you can ask your data provider to complete.

Variable Requested	Variable Name	Description	Categorical Codes	Data Quality Concerns
State unique identifier				
Free or reduced-price lunch				
Grade level				
Date of birth				
State testing date - math				
State testing score - math				
Number of in-school suspensions				

Figure 12.7: Sample documentation form for an external data provider to complete

Note When working with external datasets, it is possible to encounter inconsistencies across data sources (e.g., a student is shown in a different school across two files), as well as duplicate records within a data source (e.g., a student has two state reading assessment scores) (Levesque, Fitzgerald, and Pfeiffer 2015). These anomalies can happen due to human error or due to circumstances such as student mobility. While you may be able to work with your data provider to solve some data issues, for others it may be important for you to develop and document your own data management rules

that you consistently apply to your external data sources during the data cleaning phase (e.g., if duplicate assessment records exist, the earliest assessment date is used).

12.3.2 Public data sources

Publicly available data sources are typically aggregated (i.e., state, district, or school level) or de-identified individual level datasets that are available through various agencies such as state departments of education or federal agencies. These datasets are often extracted by downloading a file, although some organizations may have more sophisticated API capabilities. The quality of these datasets may vary. A few tips for working with publicly available datasets are:

1. Extract the data early on in your project.
    - Even if it is not the most up to date data that you need, it's important to get a sense early on for what the data looks like (e.g., what variables are included, what file types data is stored in, how the files are structured). This helps you prepare for future data wrangling needs.
2. Find the associated documentation and read it thoroughly. Types of documentation to look for are:
    - Data dictionaries or codebooks
        - These documents will help you interpret and use variables correctly
    - Changelogs
        - Public data sources are constantly updating (e.g., new data is acquired, errors are found). It's important to understand what version of the data you working with.
    - Data quality documentation
        - This documentation helps make you aware of any known issues in the data
3. Do not hesitate to reach out for help
    - Typically the site will include contact information for questions. Never hesitate to reach out to that contact if there is something you do not understand in the data.
4. If extracting data across states (e.g., Missouri Department of Elementary and Secondary Education and Oklahoma State Department of Education), be aware that the information may not be easily comparable. While you may find that some states use similar standards, it is common for states to collect and store data in different ways (e.g., different state assessments, different ways of reporting enrollment). Depending on your data needs, it may be better to use a data source that aggregates information across states. Examples of such data sources include the Department of Education's Common Core of Data [2] or EDFacts [3]. Or if you are needing to

[2] https://nces.ed.gov/ccd/
[3] https://www2.ed.gov/about/inits/ed/edfacts/index.html

use multiple data sources, other tools, such as the Urban Institute's Education Data Portal [4], have even harmonized variables and documentation across several federal government datasets, allowing researchers to access multiple data sources in a single site.

[4] https://educationdata.urban.org/documentation/

Chapter 13

Data Storage and Security

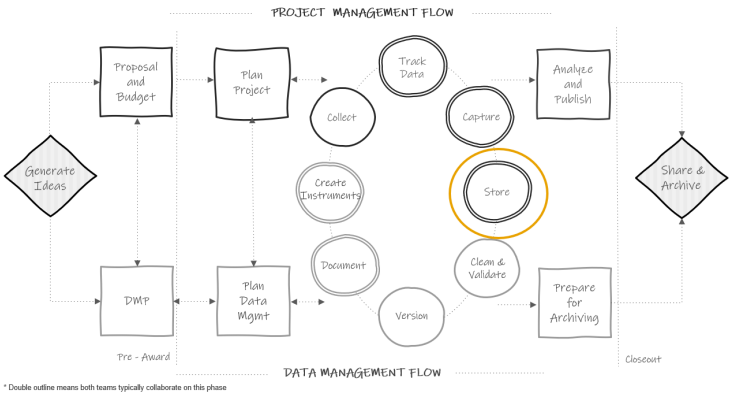


Figure 13.1: Data storage in the research project life cycle

As you begin to capture data, it is important to have a well-planned structure for securely storing and working with that data during an active study. Not only do you need a plan for storing data files, but you also need a plan for storing other project files (e.g., meeting notes, documentation, participant tracking databases). Your team should implement this structure early on so that files are stored consistently and securely for the entire project, not just once the data collection life cycle begins. There are several goals to keep in mind when setting up your file storage and security system for an active project.

1. File safety: Ensuring that your files are not lost, corrupted, or edited unexpectedly
2. Protecting confidentiality: Making sure that sensitive information is not seen or accessed by unauthorized individuals
3. Accessibility and usability of files: Making sure that your team can easily

find files and that they are able to understand what the files contain

13.1 Planning short-term data storage

When planning a storage and security process, for data files in particular, it is important to gather all relevant information before making a plan. A typical process for developing a plan may begin like this:

1. Review what data needs to be stored and how often
    - Use documents such as your data sources catalog (see Section 5.3) and your data collection timeline (see Section 8.2.6) to better understand your data storage needs.
2. Take an inventory of what data storage solutions are available to you
    - In terms of electronic data, institutions have different licenses or partnerships with varying software companies and they may approve and not approve different tools (e.g. Dropbox, SharePoint, Box, Google Drive).
3. Consider compliance
    - Make note of any data storage laws, policies, or agreements that your data is subject too (e.g., IRB policies, data sharing agreements, funder policies).
4. Review classification levels
    - Review each data source's classification level (see 4.2) to ensure that you are making decisions that are appropriate for the sensitivity level your data.

This process should help narrow down your data storage solutions for each data source. However, from there a series of decisions need to be made depending on the type of data you are working with, paper (e.g., a paper consent form) or electronic (e.g., a CSV file of questionnaire data, a Microsoft Access participant tracking database). The remainder of this section will review a series of decisions to make for each type of data, as well as provide some best practices along the way.

13.1.1 Electronic data

Once you have reviewed all relevant information from Section 13.1, several more decisions will need to be made when deciding on and setting up your structures for storing and securely working with electronic data.

1. Reviewing additional criteria
    - After narrowing down storage solutions based on available tools (e.g., cloud storage, institution network drive, personal device) that meet your compliance needs, electronic data storage locations can be further narrowed based on other criteria.
        - Versioning availability: While manually versioning is beneficial for major changes, it is very helpful to store your files in a location

that has automated versioning as a fail-safe in case of accidents such as unintended overwriting of files.
  - Size of the storage space: You will need to make sure your storage contains enough space for your files (recall Section 3.4. Consider how many files you will be storing, as well as the size of your files (e.g., number of rows and columns in each dataset).
  - Comfort level of your team: It is helpful to choose a storage space that your team is comfortable working in or you have the ability to train them in how to use it
  - Accessibility: Consider the accessibility of your storage location for users (e.g., how staff access the location off-site), as well as the compatibility with different operating system
  - Collaboration: Consider how the storage method handles multi-user editing of files
  - File sharing: It can be very beneficial to use a storage platform that allows sharing files through links, rather than sharing the actual file. That way if updates are made to the file, those changes are shown in the link, rather than having to send an updated version of the file.
  - Costs: Consider if there are any costs associated with any of your potential storage solutions
2. Choose a final storage location
   - While you may be allowed to store files in different locations depending on their sensitivity level, a more effective solution is to create a collaborative research environment (UK Data Service 2023). To do this, designate the highest level of security needed (e.g., an institution network drive), and keep all, or as many as possible, project-related files stored in that same location, assigning access to files and folders as needed. Keeping all files located in a central, consistent location often provides the benefit of data security (e.g., automated backups, not having different versions of documents on different computers) as well as accessibility (i.e., team members can find documents).
3. Set up your folder structure according to your style guide
   - Following your style guide, create a folder structure before team members begin storing files so that they are stored consistently.
     - If not already designated in your style guide, note that a best practice, and possibly a mandate from your institution, is to store your participant tracking database separately from your research study data (i.e., a separate folder with restricted access). Not only does the participant tracking database contain PII, but it is the one linking key between your study codes and your participant names and should not be stored alongside your datasets.
     - Similarly, any informed consent forms collected electronically should also not be stored alongside research study data. They should be stored in a separate, secure location.

4. Set up additional security systems
    - Data backups
        - It is important to regularly backup up your data. Consider using something similar to the 3-2-1 rule, keeping three copies of your data, on two different types of storage media, in more than one location (Briney 2015; UK Data Service 2023). Talk with your institution IT department for help with setting up this system.
    - User access
        - Assign user access to folders and files based on sensitivity levels, quality control needs, and applicable policies, agreements, or plans.
5. Designate rules for working securely with data
    - Complete required trainings (e.g., CITI trainings, IT training, internal training)
    - Consistently name folders and files
        - As you begin to save files in your project folder, you will also want to have team members consistently name and version files according to your style guide.
    - Do not keep copies of files
        - Outside of making data backups, do not keep copies of files in different folders. This opens the door for edits being made to one copy and not the other. If this happens, different team members may be working with different versions of files. If you want to have a copy of a file in more than one location (e.g., an SOP in the `documentation folder` and the `project coordination folder`), some storage systems allow you to link to documents from other location (i.e., the project coordination folder contains a link to the SOP in the documentation folder).
    - Secure your devices (O'Toole et al. 2018; Princeton University 2023a)
        - Choose safe passwords to protect devices
        - Do not leave devices open and unattended when working in the field
        - Have protection on your devices (e.g., up to date antivirus software, firewall, encryption)
        - When working remotely, use password protected wifi and use secure connections (e.g., VPN, 2FA) when working with data files
        - Any files stored on detachable media (e.g., external hard drive, CDs, flash drives) should typically be stored behind two locked doors when not in use
    - Securely transmit data files
        - When transmitting data, either internally or externally, it is important that you use secure methods, especially when data contain PII. As a general rule, no moderate or highly sensitive data should be transmitted via email. Use a secure, institution approved, file transfer method that includes encryption.

13.1.2 Paper data

Working with paper data involves reviewing another set of decisions while planning for data storage and security.

1. Choose a final storage location
    - After reviewing available locations as well as applicable laws, policies, and agreements, you will want to consider additional criteria such as accessibility of the storage site, your physical storage size needs, storage costs, and the security of the location. Most commonly required for any files containing PII, is to store them behind two locked doors (e.g., a locked file cabinet in a locked storage room).
2. Consistently structure your file cabinets and folders
    - While you may not have a style guide created for organizing physical folders and files, it is still important to consistently structure and name them for clarity. As an example, organize drawers by data source (e.g., student survey), and further organize folders by time period.
3. Securely work with files
    - As discussed in Section 11.3.1 and Section 12.2.3, as team members work with files, it is important that staff understand the rules and process for returning files back to the designated storage location when not in use (i.e., no files left on desks).

13.2 Planning long-term storage

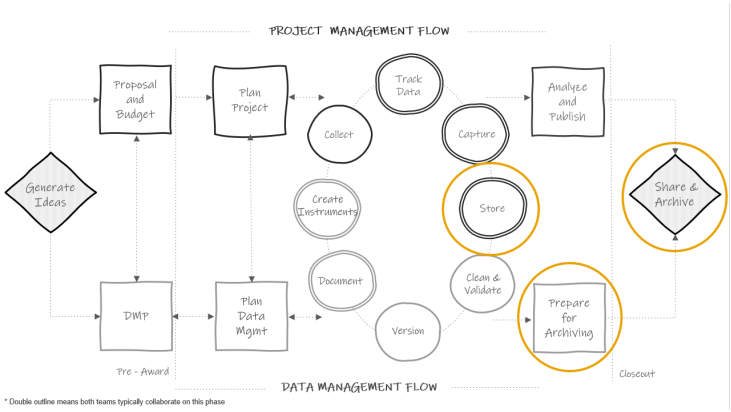


Figure 13.2: Long-term data storage in the research project life cycle

In addition to planning for short-term storage, during an active project, you will need to plan for long-term storage and use of files after a project is complete (see Figure 13.2). While your project may be ending, there are still many reasons to retain your data long-term including internal reuse, opportunities to make

corrections (e.g., going back to paper files if an error is found in the data), and data retention requirements from funders and institutions. In this section we will discuss how to care for internal files, while public data sharing and archiving will be discussed in Chapter 15.

1. First, review your requirements for data retention and destruction
    - There may be varying requirements for both data retention and destruction depending on your oversight (e.g., institution requirements, funder requirements, data sharing agreement requirements). It is common for oversight to require that you retain your data anywhere from 3-5 years and there may be specific destruction requirements for data that contain PII. Make sure to review documentation from relevant policies to determine what is required.
2. Make a plan for retention and destruction
    - If you are required to retain your data for a specified number of years, consider how you will continue to store data, and documentation, in a way that meets your original goals (e.g., data safety, protecting confidentiality, accessibility and usability of files).
        - Paper
            * When it comes to paper data, many institutions have records management departments that can assist you with long-term storage of paper files. These departments typically store your physical files for a designated set of time, as well as assist in destruction of files once that period has ended. Note that you may not want to use this solution until you are certain you will no longer need to easily access your paper files (e.g., for fixing errors, entering any additional information). If destroying paper files on your own (e.g., sometimes it is required to destroy documents containing PII, such as contact lists or classroom rosters, as soon as they are no longer needed), make sure to choose a quality destruction method such as paper shredding.
    - Electronic
        - For electronic data long-term storage, you will want to consider two things, file formats and storage location (Borer et al. 2009; Briney 2015).
            * File formats
                · First, choose file types that are widely used (i.e., don't require proprietary software) for both accessibility as well as preventing your file formats from becoming obsolete. This means that you can still keep copies of your files in a format such as SPSS if you prefer, but it is good practice to have a second copy of your data in a non-proprietary format such as CSV. Your documentation file formats should also be considered. Formats such as PDF or TXT are often recommended for long-term storage of text documents

while CSV is a good format for tabular data dictionaries.
    * Storage location
        · Similar to choosing file formats, choose a storage location that is accessible and not at risk of becoming corrupt or obsolete (e.g., think obsolescence of floppy disks). If your short-term storage solution meets these requirements (e.g., your institution network drive), you may not need to do anything different in preparing for long-term storage, but it will be important to continue implementing good practices to keep your data safe (e.g., continuing data backups). When it comes time to destroy data, make sure to permanently delete files, including all backups of files. When deleting PII, this often involves more than just moving files to the trash can on your computer. Work with your institution IT department during this process.
3. Consider how you will share data internally
    - At the end of a project, or possibly earlier in the project, it is important to consider where you will store final datasets in your storage location (i.e., a specific folder), how you will notify team members of their availability, and how you will allow team members, and other research partners, to access data. An example of this process may look like this:
        – Storing all finalized datasets (i.e., cleaned and de-identified) in a "master dataset" folder (see Chapter 14 for more information)
        – Adding descriptions of the finalized datasets to a data inventory (see Section 8.1.4)
        – Creating a data request process for team members, or external partners, to request access to finalized study datasets for various reporting and analysis purposes. Most likely you will not want researchers going in to "master data" folders and grabbing datasets without consulting with a core team member first. Therefore, it is important to develop a system for providing data to researchers on an as-needed basis. Some recommendations for setting up this system include:
            * Design a system for requesting access (e.g., designate a person to email, develop a survey form that is submitted to a designated person)
                · In that system, the researcher should describe what data they are requesting (i.e., what variables, from what time periods), as well as the purpose of their analysis. It may be helpful to build a data request process that involves providing data dictionaries and other documentation to researchers to review before requesting data.
            * Decide who needs to review the request to ensure all re-

quested information is available (e.g., a data manager), and who needs to give final approval for the data request submission (e.g., a PI)
    * Design a system for gathering data for requestors (e.g., will you provide researchers will full datasets or will you narrow datasets based on specific requests)
        - If narrowing datasets for researchers, where will new datasets be stored? (e.g., a "data request" folder) (see Figure 13.3)
    * Consider how you will share datasets with researchers (e.g., a secure link to a cloud folder, using secure file transfer)
    * Consider how you will track data requests
        - It is important to keep track of data requests in case of situations such as errors found in the data. In those cases, you can reach back out to researchers to inform them that errors were found and new versions of the data are available.

```
|--data_requests
       |data-request_log.xlsx
       |--simpson_marge_2023-08-09
              |--pa_stu_svy_data-dictionary.xlsx
              |--pa_stu_svy_w1_clean_2023-05-02-2023.csv
       |--simpson_homer_2023-07-15
              |--archive
                     |--changelog.xlsx
                     |--pa_tch_svy_w1_clean_2023-05-15.csv
              |--pa_tch_svy_data-dictionary.xlsx
              |--pa_tch_svy_w1_clean_2023-06-10.csv
```

Figure 13.3: Example set up for a data request folder

Last, if maintaining your electronic data long-term sounds like too much effort for your team, there are other options. Many universities have institutional repositories that often include services such as data curation and preservation. Additionally, there are several external repositories that offer curation and preservation services where you may be able to deposit your data for long-term storage. It's possible that depositing your data in one of these two options may also align with publicly sharing your data, which we will review in Chapter 15.

13.3 Documenting and disseminating your plan

Once you make a plan for short and long-term data storage, that plan should be added to all required documentation (e.g., DMP, IRB research protocol, informed consent forms). Once your plan has been approved, it is important to not deviate from that plan unless your revisions have also been approved. This

is especially important in the case of informed consents. Once participants have agreed to the consent terms, those terms should be honored.

Make sure to assign responsibilities to team members for both short and long-term storage tasks such as creating directory structures, adding and removing storage access, overseeing data backups, monitoring training compliance, and facilitating internal data requests. Without oversight of these processes, it is easy for errors to occur.

Last, all information needs to be disseminated to team members in the form of documentation and training to ensure fidelity to your data storage and security plan. As discussed in Chapters 7 and 8, while team members can review data management plans and research protocols, this information may be more clearly disseminated, with full details outlined, in documents such as team data security policies and team or project roles and responsibilities documents. In addition, make sure to embed this information into any team or project related staff training.

Chapter 14

Data Cleaning

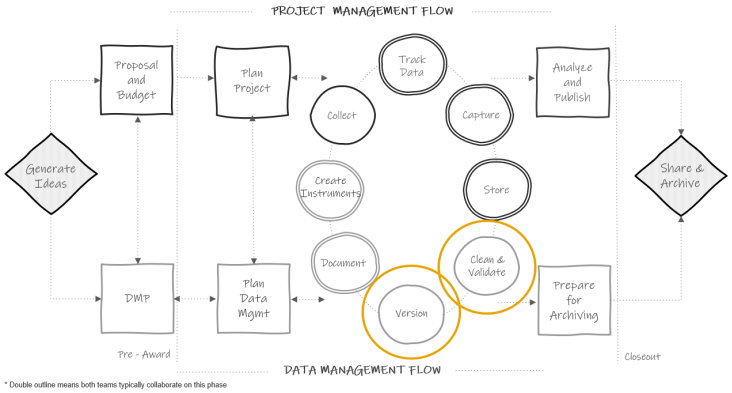


Figure 14.1: Data cleaning in the research project life cycle

Even with the most well-designed data collection and capture efforts, data still require at least some additional processing before it is in a format that you will confidently want to share for analysis. What is done in that data processing, or data cleaning, phase will depend on your project and your data. However, in this chapter we will review some standard data cleaning steps that should be considered for every education research project.

What is most important to emphasize here is that data cleaning needs to happen every wave of data collection. Once a wave of data has been collected and captured and the raw data has been stored, your data cleaning process should begin. In a best case scenario, the data cleaning is wrapped up before your next wave of data collection. Cleaning data each wave, as opposed to waiting until the end of your project, has two large benefits.

1. Allows you to catch errors early on and fix them

    - While cleaning your data you may find that all data is missing unexpectedly for one of your variables, or that values are incorrectly coded, or that you forgot to restrict the input type. If you are cleaning data each wave, you are able to then correct any errors in your instrument in order to collect better data next round.
2. Data is ready when you need it
    - Proposal, report, and publication deadlines come up fast. As various needs arise, rather than having to first take time to clean your data, or waiting for someone on your team to clean it, data will always be cleaned and available for use because it is cleaned on a regularly occurring schedule.

14.1 Data cleaning for data sharing

Data cleaning is the process of organizing and transforming raw data into a dataset that can be easily accessed and analyzed. Data cleaning can essentially result in two different types of datasets; a dataset cleaned for general data sharing purposes and a dataset cleaned for a specific analysis. The former means that the dataset is still in its truest, raw form, but has been minimally altered to allow the data to be correctly interpreted. A dataset cleaned for general data sharing means that it includes the entire study sample (no one is removed), all missing data is still labelled as missing (no imputation is done), and no analysis-specific variables have been calculated. Any further cleaning is taken care of in another phase of cleaning during analyses.

Ultimately, you can think of data in three distinct phases (see Figure 14.2).

1. Raw data
    - This is the untouched raw file that comes directly from your data collection source. If your data is collected electronically, this is the file you extract from your tool. If your data is collected on paper, this is the data that has been entered into a machine-readable format.
    - In education research this data is typically not shared outside of the research team as it usually contains identifiable information and often needs further wrangling to be decipherable by an end user.
2. The general clean study data
    - This is the dataset that you will publicly share and is the one we will be discussing in this chapter.
3. Your analytic data
    - This dataset is created from the general clean dataset (either by your team or by other researchers), but is further altered for a specific analysis. This dataset will typically also be publicly shared in a repository at the time of publication to allow for replication of the associated analysis. Since this dataset is analysis specific, we will not discuss this type of data cleaning in this book.

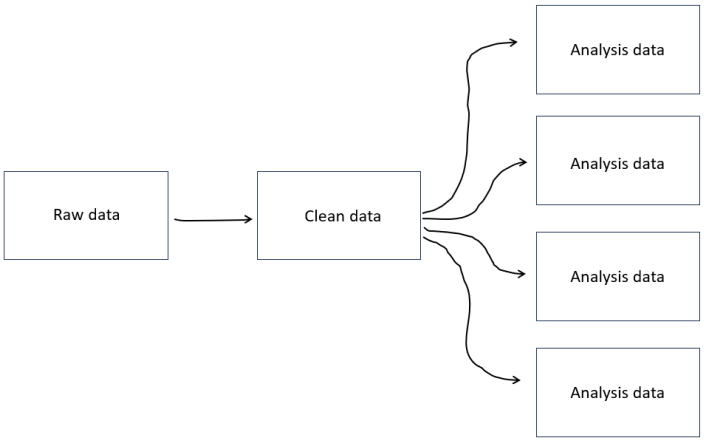


Figure 14.2: The three phases of data

14.2 Data quality criteria

Before cleaning our data, we need to have a shared understanding for what we expect our data to look like once it is cleaned. Adhering to common standards for data quality allows our data to be consistently cleaned and organized within and across projects. There are several data quality criteria that are commonly agreed upon (DeCoster 2023; Elgabry 2019; Schmidt et al. 2021; Bochove, Alper, and Gu 2023). Upon cleaning your data for general data sharing, your data should meet the following criteria.

1. Complete
    - The number of rows in your dataset should match the number of completed forms tracked in your participant tracking database. This means that all forms that you collected have been captured (either entered or retrieved). It also means that you have removed all extraneous data that doesn't belong (e.g., duplicates, participants who aren't in the final sample).
    - The number of columns in your data match the number of variables you have in your data dictionary (i.e., no variables were accidentally dropped). Similarly, there should be no unexpected missing data for variables (i.e., if the data was collected, it should exist in your dataset).
2. Valid
    - Variables conform to the constraints that you have laid out in your data dictionary (e.g., variable types, allowable variable values and ranges, item-level missing values align with variable universe rules and defined skip patterns)
3. Accurate

- Oftentimes there is no way to know whether a value is true or not.
  - However, it is possible to use your implicit knowledge of a participant or a data source (i.e., ghost knowledge) (Boykis 2021) to determine if values are inaccurate (e.g., a value exists for a school where you know data was not collected that wave).
  - It is also possible to check for alignment of variable values within and across sources to determine accuracy
    - For example, in a student-level dataset, if grade level = 2, their teacher ID should be associated with a 2nd grade teacher. Or, a date of birth collected from a student survey should match date of birth collected from a school district.

4. Consistent
   - Variable values are consistently measured, formatted, or coded within a column (e.g., all values of survey date are formatted as YYYY-MM-DD).
   - Across waves and cohorts of data collection, all repeated variables are consistently measured, formatted, or coded as well (e.g., free/reduced priced lunch is consistently coded using the same code/label pair across all cohorts).
5. De-identified
   - If confidentiality is promised to participants, data needs to be de-identified. At the early phases of data cleaning, this simply means that all direct identifiers (see Chapter 4) are removed from the data and replaced with study codes (i.e., participant unique identifier). Before publicly sharing data, additional work may be required to remove indirect identifiers as well and we will discuss this more in Chapter 15.
6. Interpretable
   - Variables are named to match your data dictionary and those variable names should be both human and machine-readable (see Section 9.4). Variable and value labels are added as embedded metadata as needed to aid in interpretation.
7. Analyzable
   - The dataset is in a rectangular (rows and columns), machine-readable format and adheres to basic data structure rules (see Section 3.2).

14.3 Data cleaning checklist

Recall from Section 8.3.3, that it is helpful to write up your data cleaning plan, for each raw dataset, before you begin cleaning your data. Writing this plan early on allows you to get feedback on your planned alterations, and it also provides structure to your cleaning process, preventing you from meandering and potentially forgetting important steps. This plan does not need to be overly detailed, but it should include actionable steps to walk through when cleaning your data (see Figure 8.15).

In many ways, writing this data cleaning plan will be a very personalized process. The steps needed to wrangle your raw data into a quality dataset will vary greatly depending on what is happening in your specific raw data file. However, in order to produce datasets that consistently meet the data quality standards discussed in Section 14.2, it can be helpful to follow a standardized checklist of data cleaning steps (see Figure 14.3). These steps, although very general here, once elaborated on in your data cleaning plan, for your specific data source, can help you produce a dataset that meets our data quality standards. Following this checklist helps to ensure that data is cleaned in a consistent and standardized manner within and across projects.

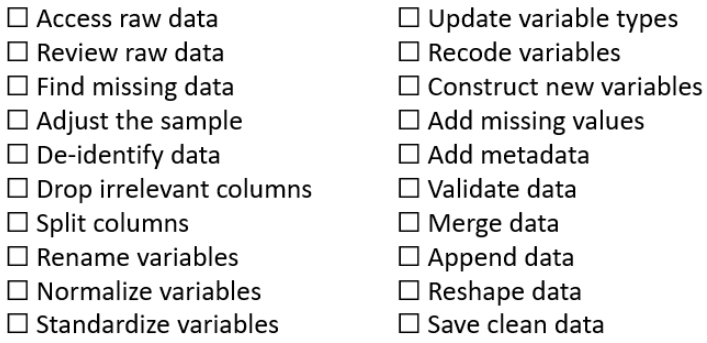


Figure 14.3: Data cleaning checklist

As you write your data cleaning plan, you can add the checklist steps that are relevant to your data and remove the steps that are not relevant. The order of the steps are fluid and can be moved around as needed. There are two exceptions to this. First, accessing your raw data will always be number one of course, and the most important rule here is to never work directly in the raw data file (Borer et al. 2009; Broman and Woo 2018). Either make a copy of the file or connect to your raw file in other ways where you are not directly editing the file. Your raw data file is your single source of truth (SSOT) for that data source. If you make errors in your data cleaning process, you should always be able to go back to your SSOT to start over again if you need to. Second, reviewing your raw data should always be step number two. Waiting to review your data until after you've started cleaning means that you may waste hours of time cleaning data only to learn later that participants are missing, your data is not organized as expected, or even that you are working with the wrong file.

14.3.1 Checklist steps

Let's review what each step specifically involves so that as you write your data cleaning plan, you are able to determine which steps are relevant to cleaning your specific data source.

1. Access your raw data
    - If you use code to clean your data, you will read your raw data file into a statistical program (e.g., R, Stata) and export a clean data file, ensuring the raw data file is never touched. If you manually clean your data, you should make a copy of the raw data file and rename it to your clean data file, ensuring you are not writing over your SSOT.
    - Part of accessing your raw data may also involve putting it into an analyzable format (e.g., if your second row of data is variable labels, you will want to drop that second row in this process so that you are only left with variable names in the first row and values associated with each variable in all remaining cells)
2. Review your raw data
    - Check the rows in your data
        - Do the number of cases in your data match the number of tracked forms in your participant tracking database?
    - Check the columns in your data
        - Do the number of variables in your data dictionary match the number of variables in your dataset? Remember we are only looking for variables that are captured directly from our source (i.e., not derived variables)?
        - Are the variable types and values as expected?

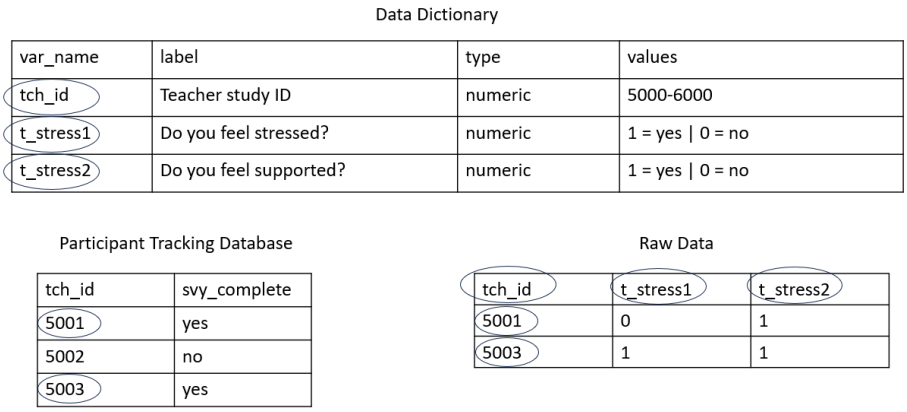


Figure 14.4: Reviewing rows and columns in a raw data file

3. Find missing data
    - Find missing cases
        - If cases are marked as complete in your tracking database but their data is missing, investigate the error. Was a form incorrectly tracked in your tracking database? Was a form not entered during the data capture phase?
            * If there is an error in your tracking database, fix the error at this time

        * Otherwise, search for missing forms, add them to your raw data, and start again at step number 1 of your data cleaning process.
    - Find missing variables
      - If you are missing any variables, investigate the error. Was a variable incorrectly added to your data dictionary? Or was a variable somehow dropped in the data capture process or in our data import or file copying process?
        * Fix the error in the appropriate location and then start again at step number 1
4. Adjust the sample
    - Remove duplicate cases
      - First, make sure your duplicates are true duplicates (not incorrectly assigned names or IDs). Any incorrect identifiers should be corrected at this time.
        * If you have true duplicates (participants who completed a form more than once or their data was entered more than once), duplicates will need to be removed
            · Follow the decisions written in your documentation (e.g., research protocol, SOP) to ensure you are removing duplicates consistently. An example rule could be to always keep the first complete record of a form.
    - Remove any participants who are not part of your final sample (i.e., did not meet inclusion criteria)

    Note In the special case where you purposefully collect duplicate observations on a participant (i.e., for reliability purposes), you will only want to keep one row per participant in your final study dataset. Again, a decision rule will need to be added to documentation so duplicates are dealt with consistently (e.g., always keep the primary observer's record).

5. De-identify data
    - If confidentiality was promised to participants, you will need to make sure your data is de-identified. If your data does not already contain your assigned study IDs, replace all direct identifiers (e.g., names, emails) in your data with study IDs using a roster from your participant tracking database. At this point we are focusing on removing direct identifiers only, but in Chapter 15, we will also discuss dealing with indirect identifiers before publicly sharing your data.
    - Figure 14.5 shows what a data de-identification process looks like (O'Toole et al. 2018). Dataset 1 would be the incoming raw data with identifiers, Dataset 2 would be a roster exported from your participant database, and Dataset 3 is your de-identified dataset, created by joining Dataset 1 with Dataset 2 on your unique identifier/s (e.g., `first_name` and `last_name`) and dropping your identifying variables. I want to emphasize the importance of using a join in your program

of choice, 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.

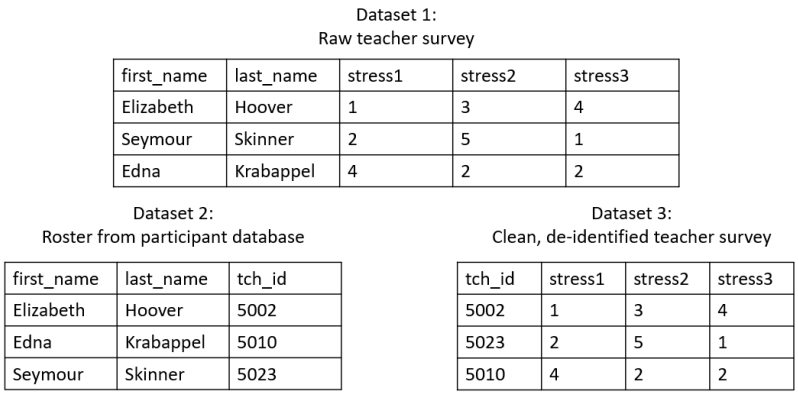


Figure 14.5: Process of creating a de-identified dataset

6. Drop any irrelevant columns not included in your data dictionary
   - Here you can think of examples such as the metadata collected by a survey platform. These columns may be completely irrelevant to your study and cause clutter in your final dataset.
7. Split columns as needed
   - As discussed in Section 3.2, a variable should only collect one piece of information. Here you will split one variable into multiple variables so that only one thing is measured per variable.

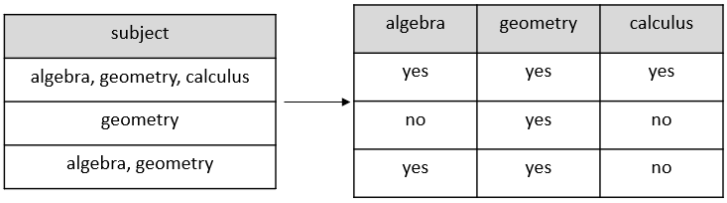


Figure 14.6: Splitting one column into multiple columns

8. Rename variables
   - Rename variables to correspond with the names provided in your data dictionary.
9. Normalize variables
   - Compare the variable types in your raw data to the variable types you expected in your data dictionary. Do they align? If no, why?
     - As an example, it may be that you need to remove unexpected characters such as `$` or `%` that are preventing your variables from

being a numeric type. Or it could be accidentally inserted white space or letters in your variable.

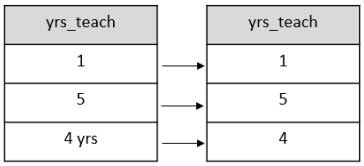


Figure 14.7: Normalizing a variable

10. Standardize variables
    - Are columns consistently measured, coded, and formatted according to your data dictionary? If no, they need to be standardized.
      - This may involve rescaling variables (e.g., age measured in months in wave 1 and age measured in years in wave 2 would need to be rescaled)
      - This may mean updating a variable format (e.g., converting to a consistent date format)
      - Or it may mean collapsing categories of free text categorical variables (e.g., 'm' | 'M' | 'male' = 'male')

    Note In the case of Figure 14.7, this kind of standardization needs to happen before you can perform steps such as joining on names for de-identification purposes. Linking keys need to be standardized across files before linking can occur.

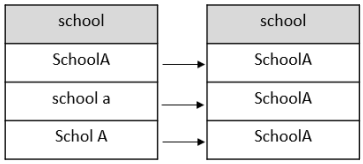


Figure 14.8: Standardizing a variable

11. Update variable types
    - After normalizing and standardizing variables, you can now convert any variable types that do not match the types you've listed in your data dictionary (e.g., convert a string to numeric)

    Note It's important to normalize before updating your variable types. Updating your variable types before normalizing could result in lost data (i.e., converting a character column to numeric, when the column still contains cells with character values, will often recode those cells to missing).

12. Recode variables

- If your categorical value codes (see Chapter 9.5) do not match your data dictionary, now is the time to recode those (e.g., you expected "no" = 1, but the data exported as "no" = 14)
- As discussed in Chapter 3.2, this also includes recoding implicit values, explicitly (e.g., if a missing value is implied to be 0, recode them to 0)
- You can also recode any variables as planned in your data dictionary (e.g., a reverse coded item)

t_stress4	t_stress4_r
1	5
2	4
5	1

Figure 14.9: Reverse coding a variable

13. Construct additional variables
    - This is not the time to construct analysis-specific variables. This is the time to create or calculate variables that should always be a part of the core study dataset. These variables should be chosen by your data management working group early on and added to your data dictionary.
    - Examples of variables to consider creating or calculating:
      - cohort
      - time component (e.g., `wave`, `time`, `year`)
      - treatment
      - measure composite or summary scores
      - completion variables or data quality flags
      - variables created for composite/summary scoring purposes (e.g., age)
      - variables that you want added to the core sharing dataset (e.g., categorizing an open-ended text response variable based on a documented pre-defined coding schema)

Note Some of these variables may exist in other sources (e.g., treatment may exist in your participant tracking database). If so, these variables won't need to be created or calculated, they can simply be merged into your clean dataset using a technique similar to the one described in data de-identification step. You can export a file from your participant tracking database that contains unique identifier/s as well as the variables you need, and join on similar unique identifiers across files (e.g., unique teacher ID), bringing in the necessary variables from an outside source.

14. Add missing values
    - Assign missing value codes based on your designated schema (as documented in your data dictionary and style guide).
15. Add metadata (UK Data Service 2023)
    - While interoperable file types (e.g., CSV) are highly recommended for storing data, it can be extremely helpful to create another copy of your clean data in a format, such as SPSS, that allows for embedded metadata. These file types allow you to embed variable and value code labels that can be very handy for a data user. This can be especially helpful if you plan to export your variables with numeric values (1 | 0), rather than text values ("yes" | "no"). In this case, rather than having to flip back and forth between a file and a data dictionary to interpret codes, users can review information about the variables within the file itself. While future data users may not have a license for the proprietary file type, these file formats can often be opened in free/open source software (e.g., GNU PSPP) or can usually be easily imported into a variety of other statistical programs which can interpret the metadata (e.g., importing SPSS files into R or Stata).
16. Data validation
    - Errors in the data can happen for many reasons, some of which come from the data collection and capture process, others come from the data cleaning process (e.g., coding errors, calculation errors, merging errors). At minimum you should validate, or check, your data for errors at the end of your data cleaning process. Ideally though, you should be checking every one of your transformations along the way as well.
    - Data validation should begin with the manual method of opening your clean data and eyeballing it. Believe it or not, this can actually be a very useful error-catching technique. However, it should not be your only error-catching technique. You should also create tables, calculate summary and reliability statistics, and create univariate and bivariate plots to search for errors. Codebooks are great documents for summarizing and reviewing a lot of this information (Arslan 2019).
    - You can organize your data validation process by our data quality criteria. The following is a sampling of checks you should complete during your validation process (CESSDA Training Team 2017; ICPSR 2020; Strand 2021; Reynolds, Schatschneider, and Logan 2022; UK Data Service 2023):
        - Complete
            * Check for missing cases/duplicate cases
                · It can also be helpful to check Ns by cluster variables for completeness (e.g., number of students per teacher, number of teachers per school) (DeCoster 2023)
            * Check for missing columns/too many columns

  - Valid and Consistent
    - Check for unallowed categories or values out of range
      - Checking by groups can also help illuminate issues (e.g., compare age and grade level) (Riederer 2021)
    - Check for invalid, non-unique, or missing study IDs
    - Check for incorrect variable types
    - Check for incorrect formatting
    - Check missing values (i.e., do they align with variable universe rules and skip patterns)
  - Accurate
    - Cross check for agreement across variables (e.g., a student in 2nd grade should be associated with a 2nd grade teacher)
    - Checks for other project-specific unique situations
  - De-identified
    - Are all direct identifiers removed?
  - Interpretable
    - Are all variables correctly named?
    - Is metadata applied to all variables?
- If during your validation process you find errors, you first want to determine where the errors originated (i.e., data entry, data export, data cleaning), and correct them in the appropriate location. If errors occurred in the data entry or data export/saving process, this may involve creating a new raw data file and starting the cleaning process again at step 1.
  - If, however, you find true values that are inaccurate, uninterpretable, or outside of a valid range (i.e., they represent what the participant actually reported), you will need to make a personal decision on how to deal with those. Some examples of how you might deal with true errors include:
    - Leave the data as is, make a note of the errors in documentation, and allow future researchers to deal with those values during the analysis process.
    - Assign a value code (e.g., "inaccurate value" = -90) to recode those values to
    - Create data quality indicator variables to denote which cells have untrustworthy values (e.g., `age` contains the true values and `age_q` contains 0 = "no concerns" | 1 = "quality concerns").
    - If you find inconsistencies across different sources, you could choose one form as your source of truth and recode values based on that form
    - If there are true errors where the correct answer can be easily inferred (e.g., a 3-item rank order question is completed as 1, 2, 4), sometimes logical or deductive editing can be used in those cases and the value is replaced with the logical correction (IPUMS USA 2023; Seastrom 2002).

  - No matter what your decision is, make sure it is documented in the appropriate places for future users (e.g., data dictionary, data cleaning plan, research protocol)

At this point, your dataset should be clean. However, there may be additional transformations to be performed depending on how you plan to store and/or share your datasets.

17. Merge data
    - Merging is joining forms horizontally, by one (e.g., `stu_id`) or more (e.g., `first_name` and `last_name`) unique identifiers (see Figure 14.10)
      - This is commonly used to link longitudinal data within participants in wide format. In this case it will be necessary to append a time component to your time varying variable names (e.g., "w1_", "w2_")
      - However this type of merging can also be used to link forms within time or link forms across participant types (e.g., merging student data with teacher data on `tch_id`)

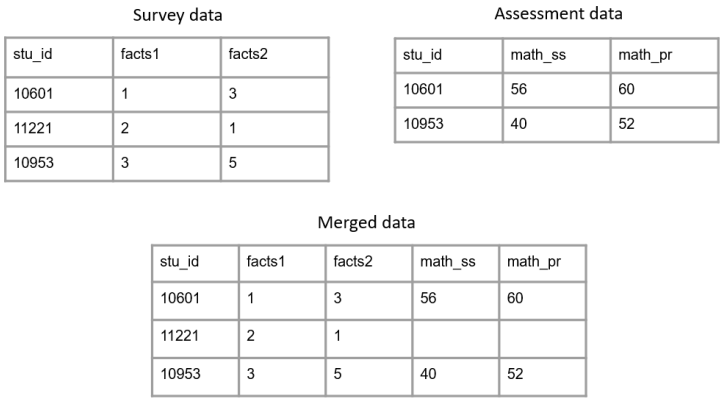


Figure 14.10: An example of merging data across forms from the same participant, in the same wave

18. Append data
    - Appending is stacking forms on top of each other and columns are matched by variable names. In this case, variable names and variable types should be identical across forms in order for the matching to work (see Figure 14.11).
      - Appending may be used to combine longitudinal data within participants in long format. Here it will be necessary to include a new variable that indicates the time period associated with each row.
      - However, appending is also often used for combining forms col-

lected from different links or captured using multiple databases (e.g., data collected across sites or cohorts)

- Once your merging or appending is complete, it will be very important to do additional validation checks
  - Do you have the correct number of rows and columns after merging or appending?

Survey data from site 1

stu_id	sch_id	facts1	facts2
114389	102		4
105678	102	2	5

Survey data from site 2

stu_id	sch_id	facts1	facts2
10601	100	1	3
11221	100	2	1
10953	100	3	5

Appended data

stu_id	sch_id	facts1	facts2
114389	102		4
105678	102	2	5
10601	100	1	3
11221	100	2	1
10953	100	3	5

Figure 14.11: An example of appending data collected on the same form across sites

Note Depending on how your data is collected or captured, as well as how you want to structure your data, you may use a combination of both merging and appending to create your desired dataset.

19. Reshape data
    - Recall Section 3.3.2 where we reviewed various reasons for structuring your data in wide or long format.
      - In wide format, all data collected on a unique subject will be in one row. Here, unique identifiers should not repeat.
      - In long format, participant identifiers can repeat, and unique rows are identified through a combination of variables (e.g., `stu_id` and `wave` together).
    - If at some point after merging or appending your data, you find you need to reshape data into a new format, this restructuring process will need to be added to your data cleaning process.

Note Having your time component concatenated to the beginning or end of a variable name (as it is in Figure 14.12), rather than embedded into your variable name, makes this back and forth restructuring process much easier to do in statistical programs.

Wide Format

tch_id	w1_t_stress1	w2_t_stress1
5001	1	2
5003	2	4

Long Format

tch_id	wave	t_stress1
5001	1	1
5003	1	2
5001	2	2
5003	2	4

Figure 14.12: A comparison of long and wide format

20. Save your clean data
    - The final step of your cleaning process will be to export or save your clean data. You can save your files in one or more file types depending on your needs. It can be helpful to save your data in more than one format to meet various analysis, long-term storage, or data sharing needs (i.e., an interoperable format like CSV, and a format that contains embedded metadata such as SPSS).

14.4 Data cleaning workflow

Data cleaning is not a standalone process. It should be part of a larger, well-planned workflow that is designed to produce standardized, reproducible, and reliable datasets. Ignoring this planning and jumping into data cleaning in a haphazard way only leads to more work after the cleaning process, having to organize our messy work so that others can understand what we did.

14.4.1 Preliminary steps

The first part in creating a data cleaning workflow is making sure that your folder structure is set up according to your style guide, and that your folders and files are consistently named according to your style guide. It is important that the metadata in your names is always provided in the same order (e.g., `project_time_participant_instrument_type.ext`). Breaking away from a standardized naming convention begins to erode the reproducibility of your work.

Next, you will want to gather all of the necessary documentation that will be used throughout your cleaning process.

1. Data dictionary
    - In this document, variables should be named and coded according to your style guide and all variables and transformations approved by the data management working group.
2. Data cleaning plan

    - This should include a series of steps based off of our standardized data cleaning checklist, and all transformations have been reviewed by the data management working group.
3. ReadMe file
    - This includes any ReadMe files, stored alongside raw data files, that contain notes that may be relevant to your data cleaning process (e.g., a project coordinator notes that "ID 1234 should actually be ID 1235"). You will want to integrate this information into your data cleaning plan as needed.
4. Participant tracking database
    - Make sure that this database is up to date so that you can compare form completion status numbers to the Ns in your dataset.

Once you gather your documentation, you are ready to begin the data cleaning process.

14.4.2 Data cleaning practices

While you can clean data through a point and click method in a program like SPSS or Excel, cleaning data manually is typically not reproducible, leads to errors, and is time consuming. The number one practice that you can implement to improve the reproducibility, reliability, and efficiency of your work is to clean data using code (Borer et al. 2009). The code can be written in any program your team chooses (e.g., R, SAS, Stata). While writing code, or syntax, may seem time consuming up front, it has numerous benefits.

- It helps you to be more thoughtful in your data cleaning process
- It allows others to review your work and catch potential errors
- It can actually save you an enormous amount of time in the future if you plan to clean data for the same form multiple times (in say a longitudinal study)
- It allows others to reproduce your work. By simply re-running your code file, they should be able to get the same resulting dataset that you created.

However, writing code alone will not provide all of the desired benefits. There are several additional practices that must be implemented during this process.

1. Do all transformations in code
    - Cleaning data using code only improves reproducibility if you do all transformations, no matter how small, in the code. No transformations should be done to your data outside of code, even if you think it is something insignificant. Once you work outside of your code, your chain of processing is lost and your work is no longer reproducible. Code files should contain every transformation you make from the raw data to your clean data.

2. Follow a coding style guide
    - As we discussed in Chapter 9, creating a code style guide for your project ensures that all team members are setting up their files in a consistent manner. This reduces the variation across code files and allows your code to be more usable by others. Developing code templates for team members to use also helps to create further standardization.
3. Use relative file paths
    - In a point and click environment (e.g., Excel), we typically open or read in a file by going to `file -> open` and navigating to the file's location. However, when writing code, we import a file by writing out our file path in our syntax. A file path is the location where a file lives. When writing out those paths, it is a good practice to write paths relative the directory you are working in, as opposed to writing a full, absolute file path. Writing absolute file paths in our syntax reduces the reproducibility of our code because future users often have different file paths than us.
    - **Example absolute file path:** "/Users/username/projecta/data/raw/proja_stu_svy_raw.csv
    - **Example relative file path:** "raw/proja_stu_svy_raw.csv"
4. Review your data upon import
    - As we discussed in Section 14.3.1, it is imperative that you review your data before beginning to clean it to ensure you have a thorough understanding of what is happening in your file. This review process can become even more relevant if you are reusing a syntax file to clean data collected multiple times (e.g., in a longitudinal study). You may expect your syntax to run flawlessly each time period, yet if anything changes in the data collection or entry process (e.g., a variable name changed, a new item is added, a new variable category is added), your data cleaning syntax will no longer work as intended. It's best to find this out before you start the cleaning process so you can adjust your data cleaning plan and your code as needed.
5. Use comments
    - Code comments help you to organize and communicate your thought process. While your syntax may seem intuitive to you, it is not necessarily clear to others. As you clean your data according to your data cleaning plan, comment every step in your syntax, explaining what that specific line of code is doing.
6. Don't do anything random
    - Everything in your syntax must be replicable. Yet, there are a few scenarios where, without much thought, you could be producing different results each time you run your code.
        - If you randomly generate any numbers in your data (e.g., study IDs), use an algorithmic pseudorandom number generator (PRNG) (Klein et al. 2018). This can be easily done in most statistical programs by setting a seed. Every time the PRNG is

run with the same seed, it will produce the same results (i.e., the same set of random numbers). Without this, you will get a new random set of numbers each time your syntax is run.
  - Another example is when you are removing duplicate cases. Be purposeful about how you remove those duplicates. Do not assume your raw data will always come in the same order. Set parameters in your syntax before dropping cases (e.g., order by date then drop second occurrence of a case). Otherwise, if at some point, someone unexpectedly shuffles your raw data around and you re-run your syntax, you may end up dropping different duplicate cases.

7. Write functions for repeatable tasks
   - As best as you can, it is important to follow the DRY (don't repeat yourself) principle and never write the same code twice. Not only does it make your script more readable, but it reduces the errors that might be created through things like copy and paste.
   - Similarly, find ways to automate some of your tasks. For instance, rather than renaming all of your variables by hand, use your data dictionary to automate tasks like this. This not only increases efficiency but also reduces mistakes you might make when typing out variable names [1].
8. Check each transformation
   - As mentioned in Section 14.3.1, check your work along the way, don't wait until the end of your script. For each transformation in your data:
     - Review your variables/cases before and after the transformations
     - Review all errors and warning codes
       * Some warnings may be innocuous (just messages)
       * Some errors are telling you that your code did not run, you need to fix something
       * Other warnings are telling you that your code did run but it did not run as you expected it to. If you don't pay attention to these warnings, you may end up with unexpected results.
9. Validate your data before exporting
   - As we also discussed in Section 14.3.1, this is when you will want to run through your final list of sanity checks, based on our data quality criteria, to make sure no mistakes exist in the data before you export it.
10. Record your session info
    - Information about software/package versions and operating system used should be recorded in a text or log file so that future users can review the requirements needed for running your code. If users run into errors running your code, this information may help them troubleshoot.

[1] https://cghlewis.com/blog/dict_clean/

11. Do code review
    - If you have more than one person on your team who understands code, code review is a great practice to integrate into your workflow. This is the process of having someone, other than yourself, review your code for things such as the readability, usability, and efficiency. Through code review it's possible to create more interpretable code as well as catch errors you were not aware of. Code review checklists can be implemented to standardize this process [2].

While I highly recommend using code to clean your data for all the reasons previously mention, I acknowledge that code writing does require technical expertise that not everyone may have. If you need to clean your data manually, the most important thing in this case then, is to document every transformation. Here you have two options.

- If cleaning data using the point and click menu in a program such as SPSS, when performing a transformation you can use a "paste" type button to copy all associated commands into a syntax file that can easily be reused (Kathawalla, Silverstein, and Syed 2021).
- If using a program such as Excel for data cleaning, it will be important to add detailed notes into your data cleaning plan that would allow anyone to replicate your exact data cleaning process by hand (The Carpentries 2023).

14.4.3 Data versioning practices

The last part of the workflow to consider is where you will store your data and how you will version it. As we've discussed previously, as you export or save your clean datasets, make sure to name them appropriately to differentiate between raw and clean datasets. As discussed in Chapter 13, you may keep these clean datasets in their respective individual folders (e.g., wave 1 - student survey folder, wave 2 - teacher survey folder), or you may choose to move all finalized datasets to a "master folder" in order to keep all clean datasets in one accessible location. What is most important here is to not copy files across folders; keeping one single master dataset per data source for authenticity purposes (CESSDA Training Team 2017; UK Data Service 2023). Also, make sure to limit access as needed based on requirements covered in Chapter 13.

However, once your final datasets are saved, it is common that at some point you will find an error in your data and/or your code. Yet, once you've shared your data and code with others, it will be imperative that you do not save over existing versions of those files. You will need to version both your code and your data, following the guidelines laid out in your style guide. Versioning your final files, and keeping track of those different versions in a changelog (see Section 8.3.2), allows you to track data lineage, helping users understand where the data originated as well as all transformations made to the data. While you can

[2] https://github.com/tgerke/r-code-review-checklist

version any files that you choose, I am specifically referring to final files here, not in-progress, working files that have not yet been shared with others.

Last, along with assigning someone to oversee data cleaning, it will be important to assign someone to oversee this versioning process. Versioning files and updating documentation takes time and consideration, and that responsibility will need to be explicitly laid out in order to ensure it isn't forgotten.

Chapter 15

Data Sharing

Chapter 16

Wrapping It Up

Diagram of what is accomplished in each phase

16.1 Connecting practices to outcomes

16.2 Putting in the work

Chapter 17

Glossary

Term	Other Terms
Anonymous data	NA
Aggregated data	NA
Append	NA
Archive	NA
Attrition	NA
Clean data	processed data
Cohort	NA
Coded data	pseudonymized data, indirectly identifiable, confidential data
Confidential data	NA
Confidentiality	NA
Control	business as usual (BAU)
Cross-sectional	NA
Data	research data
Data type	measurement unit, variable format, variable class
Database	relational database
Dataset	data set, dataframe, spreadsheet, rectangular data, tabular data
De-identified data	anonymized data
Derived data	NA
Direct identifiers	NA
Directory	file structure, file tree
Experimental data	NA
Extant data	secondary data, administrative data
FERPA	NA
File formats	NA
Human subject	NA
HIPAA	NA
Identifiable data	NA
Indirect identifiers	NA
Instrument	NA
Limited data set	NA
Longitudinal	NA
Measure	NA
Merge	join, link
Missing data	NA
Observational data	NA
Participant database	study roster, master list, master key, linking key, tracking databa
Path	file path

Aczel, Balazs. 2023. "A Crowdsourced Effort to Develop a Lab Manual Template." *Google Docs*. 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): e01801. https://doi.org/10.1002/bes2.1801.

Arndt, Aaron D., John B. Ford, Barry J. Babin, and Vinh Luong. 2022. "Collecting Samples from Online Services: How to Use Screeners to Improve Data Quality." *International Journal of Research in Marketing* 39 (1): 117–33. https://doi.org/10.1016/j.ijresmar.2021.05.001.

Arslan, Ruben C. 2019. "How to Automatically Document Data With the Codebook Package to Facilitate Data Reuse." *Advances in Methods and Practices in Psychological Science* 2 (2): 169–87. https://doi.org/10.1177/2515245919838783.

Ashcraft, Alvin. 2022. "Naming Files, Paths, and Namespaces." 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.

Barchard, Kimberly A., Andrew J. Freeman, Elizabeth Ochoa, and Amber K. Stephens. 2020. "Comparing the Accuracy and Speed of Four Data-Checking Methods." *Behavior Research Methods* 52 (1): 97–115. https://doi.org/10.3758/s13428-019-01207-3.

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/.

Beaudry, Jennifer, Donna Chen, Bryan Cook, Timothy Errington, Laura Fortunato, Lisa Given, Krystal Hahn, et al. 2022. "The Open Scholarship Survey (OSS)," November. https://doi.org/10.17605/OSF.IO/NSBR3.

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. 2023. "Data Quality." https://rdmkit.elixir-europe.org/data_quality.

Bolam, Mike. 2022. "Guides: Metadata & Discovery @ Pitt: Metadata Standards." https://pitt.libguides.com/metadatadiscovery/metadata-standards.

Bordelon, Dominic. 2023. "Guides: Research Data Management @ Pitt: Describing Data." https://pitt.libguides.com/managedata/describingdata.

Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. 2009. "Some Simple Guidelines for Effective Data Management." *Bulletin of the Ecological Society of America* 90 (2): 205–14. https://doi.org/10.1890/0012-9623-90.2.205.

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

Boykis, Vicki. 2021. "The Ghosts in the Data." *Vicki Boykis*. https://veekaybee.github.io/2021/03/26/the-ghosts-in-the-data/.

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). 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.

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.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. "The CARE Principles for Indigenous Data Governance" 19 (1): 43. https://doi.org/10.5334/dsj-2020-043.

CDISC. 2023. "CDISC Standards in the Clinical Research Process." https://www.cdisc.org/standards.

Center for Open Science. 2023. "Creating a Data Management Plan (DMP)

Document - OSF Support." https://help.osf.io/article/144-creating-a-data-management-plan-dmp-document.

CESSDA Training Team. 2017. "CESSDA Data Management Expert Guide." Bergen, Norway: CESSDA ERIC. https://dmeg.cessda.eu/.

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. 2023. "LibGuides: Metadata Basics: Key Concepts." 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.

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. 2023. "CSP Library: Zotero Guide: Defining Your Research Workflow." https://library.csp.edu/Zotero/workflow.

Dahdul, Wasila. 2023. "Research Guides: Research Data Management: Describing Data." 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. 2020. "Coding Etiquette." *Coding Club*. https://ourcodingclub.github.io/tutorials/etiquette/.

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

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

DeCoster, Jamie. 2023. "Systematic Data Validation." *Prezi.com*. https://prezi.com/view/oOXBw0bPmlReD3T7LZJm/.

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 Qual-

ity 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. 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 *In Federer, L. (Ed.). The Medical Library Association's Guide to Data Management for Librarians.* Lanham, Maryland: Rowman; Littlefield Publishing Group. https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1023&context=utk_libpub.

Elgabry, Omar. 2019. "The Ultimate Guide to Data Cleaning." *Medium.* https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

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

Feeney, Laura, Jason Bauman, Julia Chabrier, Geeti Mehra, and Michelle Woodford. 2021. "Using Administrative Data for Randomized Evaluations." *The Abdul Latif Jameel Poverty Action Lab (J-PAL).* https://www.povertyactionlab.org/resource/using-administrative-data-randomized-evaluations.

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.

Figshare. 2023. "Figshare Metadata Schema Overview." https://help.figshare.com/article/figshare-metadata-schema-overview.

Filip, Alena. 2023. "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.

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.

Gaddy, Marcus, and Kassie Scott. 2020. "Principles for Advancing Equitable Data Practice." Urban Institute. https://www.urban.org/sites/default/files/publication/102346/principles-for-advancing-equitable-data-practice_0.pdf.

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.

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.

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.

Gower-Page, Craig, and Kieran Martin. 2020. "Diffdf: Dataframe Difference Tool." https://cran.r-project.org/web/packages/diffdf/index.html.

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. 2023. "New OSF Metadata to Support Data Sharing Policy Compliance." 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. 2022. "LibGuides: Metadata for Data Management: A Tutorial: Basic Elements." 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.

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

ICPSR. 2020. "Guide to Social Science Data Preparation and Archiving: 6th Ed." https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

ICPSR. 2022. "An Introduction to Common Data Elements." https://www.youtube.com/watch?v=GsnoiPzxC4g.

———. 2023. "ICPSR, Data Management, Metadata." https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/metadata.html.

Institute of Education Sciences. 2022. "Standards for Excellence in Education Research." https://ies.ed.gov/seer/index.asp.

———. n.d. "Frequently Asked Questions About Providing Public Access To Data." Accessed October 21, 2022. https://ies.ed.gov/funding/datasharing_faq.asp.

IPUMS USA. 2023. "Introduction to Data Editing and Allocation." https://usa.ipums.org/usa/flags.shtml.

Johns Hopkins Institute for Clinical and Translational Research. 2020. "Data Dictionary/Codebook." https://ictrweb.johnshopkins.edu/ictr/dmig/Best_Practice/a8376318-ebd6-421f-be63-acf8c88376a1_6342a1c3-1a5d-4287-a46e-374824e3780e.html?v=65849&ip=hpdkvlttuiyioooqhw.

Jørgensen, Carsten Krogh, and Bo Karlsmose. 1998. "Validation of Automated Forms Processing." *Computers in Biology and Medicine* 28 (6): 659–67. https://doi.org/10.1016/S0010-4825(98)00038-9.

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.

Kathawalla, Ummul-Kiram, Priya Silverstein, and Moin Syed. 2021. "Easing Into Open Science: A Guide for Graduate Students and Their Advisors." *Collabra: Psychology* 7 (1): 18684. https://doi.org/10.1525/collabra.18684.

Klein, Olivier, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C. Frank. 2018. "A Practical Guide for Transparency in Psychological Science." Edited by Michéle Nuijten and Simine Vazire. *Collabra: Psychology* 4 (1): 20. https://doi.org/10.1525/collabra.158.

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. 2023. "Research & Subject Guides: Research Data Guide: Data Collection and Creation." 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.

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

Levesque, Karen, Robert Fitzgerald, and Jay Pfeiffer. 2015. "A Guide to Using State Longitudinal Data for Applied Research (NCEE 2015–4013)." Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation; Regional Assistance, Analytic Technical Assistance; Development. https://ies.ed.gov/ncee/rel/regions/central/pdf/CE5.3.2-A-Guide-to-Using-State-Longitudinal-Data-for-Applied-Research.pdf.

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

McKenzie, Patrick. 2010. "Falsehoods Programmers Believe About Names | Kalzumeus Software." https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/.

Mehr, Samuel. 2020. "How to… Write a Lab Handbook." *RSB*. 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. 2023. "Restrictions and Limitations in OneDrive and SharePoint - Microsoft Support." 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. 2023. "Data Ethics Statement." *Sarah Narvaiz*. https://www.sarahnarvaiz.com/ethics/.

National Center for Education Statistics. 2023. "Common Education Data Standards (CEDS)." https://ceds.ed.gov/Default.aspx.

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. 2022. "Supplemental Information to the NIH Policy for Data Management and Sharing: Responsible Management and Sharing of American Indian/Alaska Native Participant Data." NOT-OD-22-214. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-214.html.

———. 2023a. "Budgeting for Data Management & Sharing | Data Sharing." https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/budgeting-for-data-management-sharing#after.

———. 2023b. "Common Data Elements: Standardizing Data Collection." https://www.nlm.nih.gov/oet/ed/cde/tutorial/03-100.html.

———. n.d. "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 Educa-*

tion, 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. 2023. "Data Collection." https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html.

NUCATS. 2023. "Standard Operating Procedures (SOPs)." 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.

Pew Research Center. 2023. "Writing Survey Questions." *Pew Research*

Center. https://www.pewresearch.org/our-methods/u-s-surveys/writing-survey-questions/.

Princeton University. 2023a. "Best Practices for Data Analysis of Confidential Data." *Research Integrity and Assurance*. https://ria.princeton.edu/human-research-protection/data/best-practices-for-data-a.

———. 2023b. "Research Lifecycle Guide | Princeton Research Data Service." https://researchdata.princeton.edu/research-lifecycle-guide/research-lifecycle-guide.

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.

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." 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/.

———. 2021. "Make Grouping a First-Class Citizen in Data Quality Checks." *Emily Riederer*. https://emilyriederer.netlify.app/post/grouping-data-quality/.

Salfen, Jeremy. 2018. "Building a Data Practice from Scratch." *Locally Optimistic*. https://locallyoptimistic.com/post/building-a-data-practice/.

Samuel J. Wood Library. 2023. "Research Data Management, Retention, and Sharing." 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. 2023. "Schema.org." https://www.schema.org/.

Schmidt, Carsten Oliver, Stephan Struckmann, Cornelia Enzenbach, Achim Reineke, Jürgen Stausberg, Stefan Damerow, Marianne Huebner, Börge Schmidt, Willi Sauerbrei, and Adrian Richter. 2021. "Facilitating Harmonized Data Quality Assessments. A Data Quality Framework for Observational Health Research Data Collections with Software Implementations in R." *BMC Medical Research Methodology* 21 (1): 63. https://doi.org/10.1186/s12874-021-01252-7.

Schmitt, Charles P., and Margaret Burchinal. 2011. "Data Management Practices for Collaborative Research." *Frontiers in Psychiatry* 2 (July): 47. https://doi.org/10.3389/fpsyt.2011.00047.

Schulz, Kenneth F., Douglas G. Altman, David Moher, and CONSORT Group. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Par-

allel Group Randomised Trials." *BMJ (Clinical Research Ed.)* 340 (March): c332. https://doi.org/10.1136/bmj.c332.

Seastrom, Marilyn M. 2002. "NCES Statistical Standards." https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003601.

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.

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. 2021. "Error Tight: Exercises for Lab Groups to Prevent Research Mistakes." https://psyarxiv.com/rsn5y/.

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 Carpentries. 2023. "Data Organization in Spreadsheets for Ecologists." https://datacarpentry.org/spreadsheet-ecology-lesson/01-format-data.html.

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. 2023. "What Is a Data Dictionary? | UC Merced Library." https://library.ucmerced.edu/data-dictionaries.

UK Data Service. 2022. "Data Management Costing Tool and Checklist." https:

//ukdataservice.ac.uk//app/uploads/costingtool.pdf.

———. 2023. "Research Data Management." *UK Data Service*. https://ukdataservice.ac.uk/learning-hub/research-data-management/.

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 Iowa Libraries. 2023. "Metadata." https://www.lib.uiowa.edu/data/share/metadata/.

University of Washington. 2023. "Sharing Information and Data." *UW Research*. https://www.washington.edu/research/myresearch-lifecycle/setup/collaborations/sharing-information-and-data/.

U.S. Department of Health and Human Services. 2018. "What's New in IRB Review Under the Revised Common Rule." https://www.youtube.com/watch?v=zDsUUs9j3sQ.

USGS. 2021. "Tools for Creating Metadata Records." *USGS*. https://www.usgs.gov/data-management/metadata-creation#tools.

USGS. 2023. "What Are the Differences Between Data, a Dataset, and a Database? | U.S. Geological Survey." https://www.usgs.gov/faqs/what-are-differences-between-data-dataset-and-database.

Valentine, Theresa. 2011. "Best Practice: Define Roles and Assign Responsibilities for Data Management." *DataOne*. 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.

———. 2021. *Welcome | The Tidyverse Style Guide*. 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/.

Yenni, Glenda M., Erica M. Christensen, Ellen K. Bledsoe, Sarah R. Supp, Renata M. Diaz, Ethan P. White, and S. K. Morgan Ernest. 2019. "Developing a Modern Data Workflow for Regularly Updated Data." *PLOS Biology* 17 (1): e3000125. https://doi.org/10.1371/journal.pbio.3000125.

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.