Data Management in Large-Scale Education Research

Crystal Lewis

2022-09-16

Contents

Preamble

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

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

Introduction

In 2013, without knowing that the term research data management existed, I accepted a job as a Research Associate 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 PIs, other full-time staff, part-time data collectors, and graduate students. While I had some experience analyzing and working with education data, i.e. ECKLS-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 collection tools, organize and document data inputs, and produce usable data outputs; but I didn't learn to do these 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 PIs, 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, what these type of training methods really lack are standards, leading to inconsistencies across the field (J. Borghi & Van Gulick, 2022).

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). 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.

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 (J. A. Borghi & 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 (Ceviren & 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.

Without training, resources and formal support systems are the next best option for learning best practices. During my data management journey I have discovered an excellent support system of professionals in university systems, i.e. research data librarians, who can consult with research teams in their data management journey, and I have also come across some solid existing research data management books and manuals which I will link to in this book. However, while education researchers are starting to put out some excellent resources (Neild et al., 2022; Reynolds et al., 2022), I still find there is a dearth of tangible 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 data management ("Psych-DS Specification", n.d.), the field of education has yet to develop agreed upon rules for things such as data documentation or data formats. With this lack of standards, researchers are left to create practices that work for their team, leading to inconsistent data management practices across the field.

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 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 & 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 & Fyfe, 2013). In their 2021 study of 488 researchers who had published in a psychology journal between 2010 and 2018, Kovacs et al. (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 Open Science Foundation, the team has found that of the education researchers surveyed who are currently not publicly sharing their research data, about 10% mentioned "being nervous about mistakes" as a reason for not sharing (OSF, 2022). 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 (Eisenstein, 2022).

About this book

My hope is that this book can be a foundation to help researchers think through how to build a consistent and standardized data management workflow that works for their team and their projects. While the field as a whole may not have agreed upon rules for data management, there are still best practices that are proven to result in more reproducible, reliable, and secure data. While this book cannot remove barriers to implementing good data management practices, such as the complexity of your project or the novelty of the technology you are using (Alston & Rick, 2021) it hopefully provides you the knowledge and skills necessary to work in these complex environments.

This book should be viewed as a handbook to be referred to 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.

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.

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 or calculating analysis specific variables. This book is written from the perspective of a data manager, and that perspective is to implement practices that keep data in its most true, but usable form, for any future researcher to analyze in a way that works best for them.

Who this book is for

This book is for anyone involved in a research study involving original data collection. This book in particular focuses on quantitative data collection, while I do think that many of the practices covered can also apply to qualitative 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.

Final note

Planning and implementing new data management practices on top of planning the implementation of your entire research grant can feel overwhelming. How-

ever, the idea of this book is to find and implement the practices that work for you and your team, and that may be just a few of the suggestions mentioned or 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!

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 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.

Interviewees:

Others:

Research Data Management

What is research data management?

Research data management (RDM) involves the broad process of planning and implementing standardized practices across the research life cycle (J. Borghi & Van Gulick, 2022). The practices of managing data begin long before data is ever collected, during the planning phase, and continue well after a research project ends during the archiving phase.

While organizations like Data Documentation Initiative and Dublin Core have developed metadata standards for fields to adopt, it is common knowledge that there are no agreed-upon norms for managing data within and across disciplines within the field of education. The rules for how data should be collected, organized, stored, described, and shared is often left up to each individual team, as long as external requirements of the IRB and funders are met (Tenopir et al., 2016). With a growing interest in open science practices and expanding requirements for federally funded research to make data publicly available (of Science and Technology Policy, 2022), data repositories will most likely begin to play a stronger role in promoting standards for many data management practices around data formats and documentation (J. Borghi & Van Gulick, 2022).

Why care about research data management?

Without 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. There are both external pressures and personal reasons to care about developing research data management standards.

External Reasons

1. **Funder compliance**: Since 2013, even earlier for the National Science Foundation, most federal agencies that education researchers work with have required a data management plan as part of their funding application. 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, 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" (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 (of Science and Technology Policy, 2022). Along with this mandatory data sharing policy, comes the incentive to manage your data for the purposes of data sharing (J. Borghi & 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. Again, along with data sharing, comes the incentive to manage your data in a thoughtful, responsible, and organized way.

3. **Compliance with legal and ethical mandates**: If you are required to submit your research project to the Institutional Review Board, they will monitor how you manage your data. They care about the welfare, rights, and privacy of research participants and will have rules for how data is managed and stored securely.

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

Personal reasons

Even if you never plan to share your data outside of your research group, there are still many compelling reasons to manage your data.

1. **Contributes to reproducibility**: Reproducible research "is a by-product of careful attention to detail throughout the research process" (Alston & Rick, 2021, p.2 (Alston & Rick, 2021)). Even if you are not concerned with others being able to reproduce your work (which is

unlikely), you most likely want you and your team members to be able to reproduce each other's work, ensuring you can trust your results.

2. **Improve continuity**: Implementing reproducible practices ensures project continuity through staff turnover. Having developed thorough protocols allows new staff to pick up right where the project left off, and implement the project with fidelity (J. A. Borghi & Gulick, 2021; Cowles, n.d.).
3. **Increases efficiency**: Documenting and automating tasks reduces duplication of efforts for repeating tasks, especially in longitudinal studies.
4. **Reduces burden and saves time, energy, and resources**: Taking the time to implement quality data management through the entire research study reduces data curation debt caused by suboptimal data management practices (Butters et al., 2020). Having poorly managed or documented data may make your data unusable, either permanently or until errors are corrected. Decreasing or removing this debt reduces the time, energy, and resources spent at the end of your study scrambling to fix errors made in poorly designed data collection instruments, gathering duplicate data that was lost, or documenting efforts long after most information has been forgotten.
5. **Improve efficiency**: Being able to find and understand your data when you need it is a huge benefit. It allows for the use and re-use of your data, and hastens efforts like the publication process. Not having to search around for numbers of consented participants or asking which version of the data you should use allows you to spend more time analyzing and writing and less time playing detective.
6. **Increases reliability**: Errors come in many forms, from both humans and technology. Implementing quality control procedures allow 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.

Existing Frameworks

Terminology

The Research Life Cycle

Data Management Plan

History and purpose

What is it?

Why are DMPs important?

What to include?

Getting help

Budgeting

Planning Data Management

Why spend time on planning?

Planning checklists

How to move from a planning checklist to a workflow

Project Roles and Responsibilities

Why it's important to assign roles

Typical roles in a research project

Documentation

What is documentation?

Why is documentation important?

Team Level

Project Level

Dataset Level

Variable Level

Data Collection

Why consider data management in data collection?

Consents

Electronic data collection instruments

Paper data collection instruments

Interviews/cocus groups

Data Capture

Electronic data capture

Paper data capture

Extant data

Data Storage and Security

Types of data you'll be storing

General security rules

Participant tracking database

Electronic data

Detachable media

Audio/visual data

Paper data

Sharing data

Data Cleaning

Foundational knowledge

Data structure

Data cleaning plan

Data validation

Why use code?

Data Sharing

Why share your data?

Considering FAIR principles

Best practices

Retractions and revisions

Wrapping It Up

Connecting practices to outcomes

Putting in the work

Call to Action

Last thoughts

Training for future researchers

Investing in data management and data managers

Appendices

Bibliography

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

Borghi, J., & Van Gulick, A. (2022). Promoting Open Science Through Research Data Management. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.9497f68e

Borghi, J. A., & Gulick, A. E. V. (2021). Data management and sharing: Practices and perceptions of psychology researchers. *PLOS ONE*, *16*(5), e0252047. https://doi.org/10.1371/journal.pone.0252047

Briney, K. (2015). *Data management for researchers: Organize, maintain and share your data for research success* [OCLC: ocn921133380]. Pelagic Publishing.

Butters, O. W., Wilson, R. C., & Burton, P. R. (2020). Recognizing, reporting and reducing the data curation debt of cohort studies. *International Journal of Epidemiology*, *49*(4), 1067–1074. https://doi.org/10.1093/ije/dyaa087

Campos-Varela, I., & Ruano-Raviña, A. (2019). Misconduct as the main cause for retraction. A descriptive study of retracted publications and their authors. *Gaceta Sanitaria*, *33*(4), 356–360. https://doi.org/10.1016/j.gaceta.2018.01.009

Ceviren, A. B., & Logan, J. (2022). Ceviren_logan_ehe_forum_2022.pdf. https://doi.org/10.6084/m9.figshare.19514368.v1

Cowles, W. (n.d.). Research Guides: Research Data Management at Princeton: Home. Retrieved September 15, 2022, from https://libguides.princeton.edu/c.php?g=102546&p=665862

Doucette, L., & Fyfe, B. (2013). Drowning in Research Data: Addressing Data Management Literacy of Graduate Students - PDF Free Download. Retrieved September 15, 2022, from https://docplayer.net/8853333-Drowning-in-research-data-addressing-data-management-literacy-of-graduate-students.html

Eisenstein, M. (2022). In pursuit of data immortality. *Nature*, *604*(7904), 207–208. https://doi.org/10.1038/d41586-022-00929-3

House, T. W. (2013). Executive Order – Making Open and Machine Readable the New Default for Government Information. Retrieved September 15, 2022, from https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

Hubbard, A. (2017). *Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step*. Retrieved September 15, 2022, from https://eric.ed.gov/?id=ED583982

Kovacs, M., Hoekstra, R., & Aczel, B. (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

Neild, R., Robinson, D., & Agufa, J. (2022). Sharing Study Data: A Guide for Education Researchers. (NCEE 2022-004). *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.* https://ies.ed.gov/ncee/pubs/2022004/pdf/2022004.pdf

of Science and Technology Policy, O. (2022). OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay. Retrieved September 15, 2022, from https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/

OSF. (2022). COS Engagement with the Education Community. Retrieved September 15, 2022, from https://docs.google.com/presentation/d/1LpyVOj8oJPr3SVkRM2GfCFnl2Qeo10YbbqcqwtwrVUM

Psych-DS Specification. (n.d.). Retrieved September 16, 2022, from https://docs.google.com/document/d/1u8o5jnWk0Iqp_J06PTu5NjBfVsdoPbBhstht6W0fFp0/edit?usp=embed_facebook

Reynolds, T., Schatschneider, C., & Logan, J. (2022). The Basics of Data Management. https://doi.org/10.6084/m9.figshare.13215350.v2

Tenopir, C., Allard, S., Sinha, P., Pollock, D., Newman, J., Dalton, E., Frame, M., & Baird, L. (2016). Data Management Education from the Perspective of Science Educators. *International Journal of Digital Curation*, *11*(1), 232–251. https://doi.org/10.2218/ijdc.v11i1.389