Part I

Introduction

CHAPTER 1. DATA PREPARATION: THE NEED FOR STRATEGY AND TRANSPARENCY

Importance of Data Preparation

> In 2016, a groundbreaking study published in a leading sociology journal reported a striking pattern: After a high-profile case of police brutality in Milwaukee (the Jude incident), 911 calls from Black neighborhoods sharply declined (Desmond, Papachristos, & Kirk, 2016). This finding appeared to expose another consequence of police brutality—a troubling breakdown in community trust toward law enforcement. This study garnered attention from local newspapers such as the *Milwaukee Journal Sentinel* and large national publishers like *The Atlantic*.
>
> Subsequent analysis revealed a potential flaw. One particular data point, an outlier from 47 weeks after the incident, appeared to skew the results (Zoorob, 2020). When this outlier was removed, the relationship between the incident and the drop in 911 calls weakened. Thus, questions were raised about the reliability of the original findings. Upon learning of this issue, the original authors revisited the data, this time including additional control variables and accounting for seasonal variations in call patterns. After running a series of sensitivity analyses, they confirmed that their core findings still held strong, despite the initial oversight (Desmond, Papachristos, & Kirk, 2020).

This opening example illustrates just one of many ways in which data preparation decisions—and errors—can significantly impact research outcomes (Aguinis, Gottfredson, & Joo, 2013; Allison, 2000; Osborne, 2013; Osborne & Waters, 2002). Throughout this book, I provide numerous instances of data preparation mistakes in published research. These examples are not presented to shame individual researchers. Rather, many are drawn from scholars renowned for their rigorous work—such as Matthew Desmond, a MacArthur “genius” fellow and exceptional researcher—to demonstrate that even the most skilled and meticulous experts can occasionally overlook critical aspects of data preparation.

Indeed, errors in data preparation should come as no surprise. Researchers often devote significant effort to mastering data analysis techniques, but they typically spend far less time learning how to properly prepare data for analysis (Firebaugh, 2008; Leahey, 2006; Leahey, Entwisle, & Einaudi, 2003). As a result, researchers adopt a range of data preparation practices (Leahey, 2006, 2008; Leahey et al., 2003; Sana & Weinreb, 2008), some of which are more or less error-prone.

Without a systematic, evidence-based approach to data preparation, mistakes are more likely to occur and compromise the accuracy of research inferences. To minimize the risk of error during data preparation, Part II of this book presents a framework—the CLEANR method.

Transparency

Even with systematic approaches in place, human error remains inevitable, and careful researchers who are doing their best will make mistakes. Unfortunately, even honest and seemingly minor mistakes can have serious repercussions. The two main consequences of errors in published research are: (a) an unclear research record and (b) diminished trust in scientific research (Anvari & Lakens, 2018; Cherlin, 1990; Hendriks, Kienhues, & Bromme, 2020; Koehler & Pennycook, 2019; Mede, Schäfer, Ziegler, & Weißkopf, 2021; Wingen, Berkessel, & Englich, 2020). As a result, the potentially positive impact of researchers' work may be significantly diminished. Fortunately, adopting transparency[1] at both institutional and individual levels may help mitigate these issues (Moody, Keister, & Ramos, 2022; Peng, 2011).

Clarifying the Research Record

Replication and reproducibility[2] are critical for clarifying the research record by distinguishing regular occurrences from anomalies. That is, since nearly anything can happen once, for researchers to make causal claims and confidently predict occurrences, a phenomenon must be repeatable. Therefore, the inability to replicate studies draws significant attention and concern.

Importantly, some issues with replication suggest that science is working as intended (Jamieson, 2018). For example, when slightly different

[1] By transparency, I am referring to openly sharing data, code, and other materials that enable researchers to better understand original authors' data preparation and analysis decisions (Freese & Peterson, 2017).

[2] Although there is some confusion and debate about the terminology, replication typically means that a study can be repeated and generate the same results. On the other hand, reproducibility generally means that the same data and analyses should yield the same results (National Academies of Sciences, Engineering, and Medicine, 2019).

versions of studies or new datasets produce different results, this can allow researchers to refine a theory—such as clarifying scope conditions and definitions (Sell, 2018). Even when the same data are reanalyzed and produce different findings, they can provide important insights about data preparation and analysis decisions.

However, for science to work, researchers must be able to understand the reasons for differences between studies. Thus, researchers must be able to compare methodological, data preparation, and data analysis decisions. Without transparency, it is difficult to identify the reasons for variations between studies, and mistakes may go undetected, with no way to assess their impact on the scientific record. In contrast, transparency can help researchers identify why findings may differ across studies. In sum, transparency is essential for facilitating reproducibility, replication, and clarifying the research record (Christensen, Freese, & Miguel, 2019).

Increasing Trust in Research and Researchers

As noted above, research errors may decrease public trust in scientific research (Anvari & Lakens, 2018; Cherlin, 1990; Hendriks et al., 2020; Koehler & Pennycook, 2019; Mede et al., 2021; Wingen et al., 2020). Public trust in scientific research is crucial because research informs individuals' beliefs and behaviors. Without trust in science and in the scientific community, the public may be likelier to adopt beliefs or practices that are harmful to themselves or others (Druckman, 2022; Mann & Schleifer, 2020).

Fortunately, researchers can take steps to increase the public's trust in their work. Specifically, research suggests that the public is more likely to trust science and scientists that practice transparency (Rosman, Bosnjak, Silber, Koßmann, & Heycke, 2022; Song, Markowitz, & Taylor, 2022). Therefore, by adopting tools for transparency, researchers can promote reproducibility and public trust in research, ensuring that scientific knowledge continues to inform and benefit society.

Tools for Transparency

Depending on disciplinary norms, transparency in research design may be more or less common (Christensen et al., 2019; Freese, 2007; Freese & Peterson, 2017). For example, in many disciplines, researchers may not consistently report their data cleaning practices in published work. Additionally, publisher-imposed word limits further constrain the extent to which data management details are disclosed. Consequently, data management practices often escape scrutiny during peer review, leaving researchers solely responsible for identifying potential errors. As a result, data management errors are easy to make but notoriously difficult to detect, posing a persistent challenge to research integrity.

Transparency—in the form of preregistration, appendices (aka supplemental indices), and code/data sharing—can make it easier to detect errors. I discuss each of these tools below.

Preregistration and Registered Reports

Before collecting or preparing data for analysis, researchers may consider completing a preregistration.[3] Preregistration is a tool that researchers can use to plan and document their data preparation and analysis strategies. Specifically, in the process of completing a preregistration form, researchers are asked to answer questions about their hypotheses, desired samples, exclusion criteria, study design, and/or analysis plans. In short, preregistration involves stating hypotheses and/or research questions and prewriting the methods section of the paper, providing a rationale for methodological decisions. After publication (or during peer review, depending on the discipline), these answers are posted on a repository that is publicly accessible to other researchers (Kavanagh & Kapitány, 2019).

Preregistration offers numerous benefits, particularly for deductive research (Kavanagh & Kapitány, 2019; Lakens, 2019; Manago, 2023; Nosek et al., 2018; Wagenmakers & Dutilh, 2016). In terms of data preparation—a crucial step in all forms of research—preregistration is valuable as it encourages researchers to thoughtfully plan and reflect on their data preparation decisions. By carefully considering and documenting the rationale behind data cleaning and analysis choices, researchers can make more informed decisions that are grounded in theory and evidence-based practices (Manago, 2023). While the specific type of preregistration may vary according to the research context, its use as a planning tool may increase the likelihood of thorough and accurate data preparation across all research types.

In addition to its benefits for data preparation, there is some evidence that preregistration may help resolve the "file-drawer problem." The file-drawer

[3] Several misconceptions about preregistration may discourage researchers from adopting it. One of the most common concerns is the belief that preregistration imposes a rigid set of rules that must be followed without flexibility. This is not the case (DeHaven, 2017; Nosek, Ebersole, DeHaven, & Mellor, 2018). Researchers can and should still conduct additional exploratory or sensitivity analyses—whether planned or unplanned—even after preregistering their study. That is to say, adhering to preregistration should be in the spirit of ensuring careful consideration of decisions, not forgoing critical thinking and supplemental analyses.

problem refers to the nonpublication of papers with null findings.[4] For a variety of reasons, papers with null findings are less likely to be published. Since null findings can provide clarity about phenomena of interest, this publication bias can negatively impact scientific progress. There is some evidence that preregistration may reduce this bias, leading to a larger number of publications with null results, especially when preanalysis plans are included (Brodeur, Cook, Hartley, & Heyes, 2024; Kaplan & Irvin, 2015).

Perhaps an even better solution to the file-drawer problem that also facilitates transparency is the use of registered reports (Moody et al., 2022). Registered reports include the same information as preregistration, but the paper is reviewed *prior* to data preparation and analysis (and sometimes even prior to data collection). Then, a paper is contingently accepted on the basis of the research question and proposed methods—not the findings. If the paper receives contingent acceptance, then as long as the authors adhere to the proposed plans, the paper will be published.

Although preregistration and registered reports may improve the quality of research, they do not guarantee it (Kavanagh & Kapitány, 2019; Lakens, 2019, p. 226). Even after careful planning, there are many ways for researchers to make errors. Fortunately, by making data preparation and analyses processes transparent, preregistration may facilitate the identification and resolution of errors.

Appendices

Data Management

During the typical journal review process, data analysis decisions are carefully scrutinized. However, data preparation decisions—which can substantially affect research findings (Aguinis et al., 2013; Hoekstra, Kiers, & Johnson, 2012; Kahn & Udry, 1986; Leahey et al., 2003; Munsch, 2018; Osborne, 2013; Rucker, McShane, & Preacher, 2015; Wicklin, 2017)—are often reported with less frequency and detail. This may be partly due to journal-imposed word limits.

A data management appendix can address the lack of detail by offering readers a deeper understanding of the rationale behind data preparation decisions and clarifying whether these decisions were predetermined (a priori) or influenced by the results (a posteriori). A data management appendix usually includes details about the merging/appending of data, examination of data, alterations made to data, creation of new variables, and so on. If a researcher follows preregistration protocols exactly, they can

[4] Null findings are findings that fail to reject the null hypothesis and therefore do not provide support for the alternative hypothesis.

simply copy and paste the preregistration information while presenting the results of planned examinations, such as the types and amount of missing data. Any deviations from a preregistration protocol should be reported, and the rationale for these deviations should also be provided.

Sensitivity Analysis Appendix

Sensitivity analysis appendices are supplemental materials used to report all analyses conducted while examining a research question. During the research process, meticulous and thorough researchers often perform supplemental analyses. Although there may not be room to delve into these details in the manuscript, reporting these supplemental analyses may reveal new and interesting insights or simply clarify the robustness of findings.

In the online appendix for this book, I provide data management and sensitivity analysis appendix templates.

Data and Code Sharing/Review

To make research decisions transparent, researchers should provide access to a study's data and code/script files. These files should be written and organized so that others can easily find and utilize them. In subsequent sections, I provide advice for ensuring files are easily understood by others. This includes suggestions about how to prepare script files, name files, and organize files within folders.

At times, data cannot be shared for privacy reasons. In these instances, it is still usually acceptable to share code/script files. The sharing of these files can still provide others with clear insights into data preparation and analysis decisions. In Chapters 2 and 3, I provide specific details on how to write, organize, and name these files in a way that makes them easy to share with others.

Ideally, journals will eventually adopt a practice of code review (Moody et al., 2022). Journal-level code review would look similar to the copyediting of a paper but look specifically for errors in code. These practices are effective at catching errors before publication (Eubank, 2016). If researchers adopt the best practices suggested below, the process of preparing code to share with others, including journal code reviewers/editors, will be made simpler.

Summary

In summary, data preparation decisions are often a source of errors in published research. To mitigate these issues, I have developed a comprehensive strategy for accurate and efficient data preparation, which I discuss in Part II of the book. Since errors can occur even when implementing a comprehensive strategy for data preparation, I first emphasize the critical importance of quickly and easily identifying errors when they arise, ideally before publication (Moody et al., 2022). Specifically, I stress the value of transparency in the research process. To foster transparency, I recommend utilizing tools such as preregistration, appendices, and data/code sharing.

Preregistration clarifies the original research question and intended data cleaning and analysis decisions. Appendices provide additional space to showcase specific analyses and decisions, allowing for further review and consideration by others. Finally, code/data sharing provides ultimate transparency, facilitating code review and scientific advancement. By adopting these strategies, researchers can ensure that data preparation is transparent.

Importantly, transparency requires a cooperative research climate that rewards intellectual humility. Unfortunately, some researchers use others' errors as opportunities to showcase their own intellectual prowess and shame their peers for supposed ineptitude. This adversarial environment may foster a reluctance to admit to errors, thereby reducing transparency, slowing the progress of scientific research, and further reducing public trust in research (Janz & Freese, 2021; Yarborough, 2014). To encourage researchers to practice open and transparent research, structural-level practices (e.g., registered reports) and positive cultural norms are crucial (Moody et al., 2022).

In the following two chapters, I cover best practices that support transparent data preparation, including the use of statistical software (not spreadsheet software), tips for script file composition, standardized file structures, and consistent naming conventions.