Series Foreword

Why are you reading this book? Perhaps you have recently been assigned to write a research paper in an undergraduate course. Maybe you are considering graduate school in one of the behavioral, health, or social science disciplines, such as psychology, public health, nursing, or medicine, and know that having a strong research background gives you a major advantage in getting accepted. Maybe you simply want to know how to conduct research in these areas. Or perhaps you are interested in actually conducting your own study. Regardless of the reason, you are probably wondering—how do I start?

Conducting research can be analogous to cooking a meal for several people. Doing so involves planning (e.g., developing a menu), having adequate resources (e.g., having the correct pots, pans, carving knives, plates), knowing what the correct ingredients are (e.g., what spices are needed), properly cooking the meal (e.g., grilling vs. baking, knowing how long it takes to cook), adequately presenting the food (e.g., making the meal look appetizing), and so forth. Conducting research also involves planning, proper execution, having adequate resources, and presenting one's project in a meaningful manner. Both activities also involve creativity, persistence, caring, and ethical behavior. But just like cooking a meal for several people, conducting research should follow one of my favorite pieces of advice—"remember that the devil is in the details." If you want your dinner guests to find your meal tasty, you need to follow a recipe properly and measure the ingredients accurately (e.g., too much or little

of some of the ingredients can make the entrée taste awful). Similarly, conducting research without properly paying attention to details can lead to erroneous results.

Okay, but what about your question—"How do I start?" This American Psychological Association book series provides detailed but user-friendly guides for conducting research in the behavioral, health, and social sciences from start to finish. I cannot help but think of another food analogy here—that is, the series will focus on everything from "soup to nuts." These short, practical books will guide the student/researcher through each stage of the process of developing, conducting, writing, and presenting a research project. Each book will focus on a single aspect of research, for example, choosing a research topic, following ethical guidelines when conducting research with humans, using appropriate statistical tools to analyze your data, and deciding which measures to use in your project. Each volume in this series will help you attend to the details of a specific activity. All volumes will help you complete important tasks and will include illustrative examples. Although the theory and conceptualization behind each activity are important to know, these books will focus especially on the "how to" of conducting research, so that you, the research student, can successfully carry out a meaningful research project.

This particular volume, by Kathy Berenson, focuses on managing the data that you will eventually collect as part of your study, as well as analyzing the results in such a manner as to allow you to make reasonable and valid conclusions. Especially helpful is the inclusion of illustrative examples using SPSS statistical software. Emphasized throughout the volume is the need to be organized—more importantly, Dr. Berenson demonstrates how to do so. She also underscores the need to *document* everything you do from beginning to end. Conducting a study can be a daunting task—using this book can make it easier.

So, the answer to the question "How do I start?" is simple: Just turn the page and begin reading!

Best of luck!

—Arthur M. Nezu, PhD, DHL, ABPP
Series Editor

1

Introduction

If you needed a surgeon to operate on your brain, you would want to entrust your care to someone with training and expertise, perhaps even brilliance, in brain surgery. But brain surgery isn't the only thing you'd expect this person to do well and take seriously. Preoperative and postoperative procedures are important, even if they are not brain surgery per se. Indeed, if while lying on the operating table you were to hear your doctor say, "Never mind the handwashing: I'm a brain surgeon, not a cleaner," you would be smart to grab your paper hospital gown and make a run for the exit.

In the research world, data management and documentation can be seen as similar to essential pre- and postoperative tasks. They aren't data analysis per se; they are the crucial things that have to be done before and after data analysis. Students, professors, and other researchers all find themselves tempted to skip these procedures so they can focus their time

http://dx.doi.org/10.1037/0000068-001
Managing Your Research Data and Documentation, by K. R. Berenson

and attention on theories, predictions, and results. When you feel this way, remind yourself that you don't want to be like a brain surgeon who operates with dirty hands. Though data management and documentation may not be as intellectually stimulating or prestigious as other parts of the research process, even the greatest results will be worthless unless these basic steps are also done well.

THE IMPORTANCE OF DOCUMENTING EVERYTHING YOU DO WITH YOUR DATA

Replicability is a central value in science. It is part of the overall idea that a meaningful result did not just happen randomly but can be counted on to happen repeatedly. If a study you conducted is replicable (or reproducible), someone else who conducts your study over again from scratch, copying exactly what you did, will get the same results (Open Science Collaboration, 2015). Psychology is one of several scientific fields in which low rates of replicability have received a lot of criticism from researchers and the popular media. Indeed, many well-known and cherished ideas in psychology are now being called into question because the famous studies that supported them have not been able to be reproduced. In a recent major attempt to replicate 100 published psychology studies, only 36% obtained the same results (Open Science Collaboration, 2015). One reason many attempts to replicate psychology research studies fail is that it is often difficult to match closely enough what the original researchers did. For example, it might matter whether participants complete the study on a computer rather than on paper or have a different cultural background from the participants in the original study (Gilbert, King, Pettigrew, & Wilson, 2016). Trying to replicate previous studies and make future studies more readily replicable is an important goal in psychology research today. It is also an incredibly daunting one because it is so complex and so broad.

When faced with a daunting goal, it's often helpful to break the goal down into specific parts that are more manageable. Before worrying about being able to replicate entire studies, researchers should perhaps first solve the problem of being able to replicate what one another does with their data. For example, if your friends conducted a study and gave you all their

original data to analyze yourself, would you be able to find the same results your friends did? Don't count on it. The procedures that are broadly called *data management* (preparing your data to be analyzed by checking and fixing problems and computing variables) and *data analysis* (obtaining descriptive and inferential statistics) involve making many decisions, and these decisions can have a large influence on the results you ultimately find (Simmons, Nelson, & Simonsohn, 2011). Unless you knew exactly everything your friends decided to do with their data from the beginning, you might not be able to retrace their steps or even know where to start. Although documentation of data management and analysis is not all there is to the replicability of research, this small part is a crucial one.

You have probably been taught that it is necessary to have the correct documentation to back up everything you say in psychology research papers—specifically, citations and references in the format required by the American Psychological Association (APA). When you're doing original research—conducting your own analyses and maybe even collecting your own data—the documentation relating to your data management and analyses is just as important. In fact, it is an ethical responsibility of psychologists to share necessary data and documentation with other researchers who wish to verify their results, as described in Ethical Standards 6.01, Documentation of Professional and Scientific Work and Maintenance of Records, and 8.14, Sharing Research Data for Verification, of the American Psychological Association's (APA's) *Ethical Principles of Psychologists and Code of Conduct* (APA, 2017; see http://www.apa.org/ethics/code/index.aspx) and in the sixth edition of the *Publication Manual of the American Psychological Association*, section 1.08 (APA, 2010).

Enabling outsiders to examine our work helps inform and improve future studies and therefore has possible repercussions for the lives of many people. Let's say you do a study and publish the results. When other researchers and students read your paper and develop projects that build on it, understanding exactly what you did may turn out to be important for their careers. Moreover, the results of many scientific studies can have important effects on the public at large. The population being investigated by your research could be hurt by a false positive or false negative result and would benefit from the ability of other researchers to clear up the scientific record.

Requiring you to properly document how you handled your data is not just about your project grade; it isn't even really about you or your reputation. Developing these skills promotes the progress of science and the greater good.

A renewed emphasis on teaching research documentation practices is necessary at this point because recent technological advances have made research documentation both more convenient and more complicated. Decades ago, everything researchers did was on paper, and documenting research largely meant storing a lot of file boxes and folders. Clunky old computers also required a built-in system of documentation; the only way they could analyze your data was if you typed up a set of commands, which would then be kept for future use and as a record of what was done. But what are the comparable procedures for documenting research today? Being able to access data anywhere means that files can easily end up scattered and lost. People are also often tempted to do things to their data with a point and click or by directly typing into their data file—conveniences that are quite dangerous to the integrity of scientific research. Today, new procedures for documenting data management and analysis are necessary to achieve the level of record keeping that at less convenient times we had by default.

The culture of academic psychology has historically focused solely on the final products of research, so that careful behind-the-scenes documentation often went unrewarded and untaught. The chances are that many of your professors and mentors were left to figure out documentation procedures on their own. But now that culture is changing because of renewed attention to the replicability of science and rapid advances in technology. Gone are the days when papers covered with illegible scribbles were considered sufficient documentation for a research project. Doing research today requires knowing how to keep records of our work that others—supervisors, collaborators, reviewers, editors, and other researchers—can readily understand and use. As research documentation moves from musty storage facilities to downloadable files, it is time to take this process out of the dark and actually teach it. This is something we all do and that we all have to do. It should not have to be treated like a dirty little secret.

To ensure that research documentation does get done effectively and systematically, students and beginning researchers have to be taught to do it and how, step by step. Along these lines, Richard Ball and Norm Medeiros at Haverford College developed Project TIER (Teaching Integrity in Empirical Research) to help prepare faculty to better teach data management and documentation procedures to undergraduate students conducting original research projects in the social sciences. The suggestions offered in this book were informed and inspired by the TIER documentation protocol (which can be downloaded from their website at http://www.projecttier.org) and are compatible with it—but adapted to suit better the specific needs of psychological science specifically.[1]

GOAL AND PRINCIPLES BEHIND THIS BOOK

This book provides guidelines and step-by-step instructions for managing data and documentation in psychological research. It is designed to meet the needs of undergraduate or graduate level psychology students learning to conduct research, whether for a capstone or thesis, independent study, laboratory course, or research assistant position in a psychology lab. Though this book gives examples in the Statistical Package for the Social Sciences (SPSS®),[2] the procedures it describes can be carried out with any programmable statistical package, such as SAS, R, or Stata.[3] The procedures the book recommends are designed to help people conduct research in ways that are compatible with sound and

[1]Compared with the TIER protocol, this book places less emphasis on the creation of metadata and importable files because students learning to do psychology research typically collect their own data in SPSS or in formats that SPSS can easily read. (Students who are not collecting their own data or who require data importation commands will find these topics covered in appendices.) However, this book places greater emphasis on documentation of variable computations (e.g., making a scale from several individual items) because this is a common aspect of working with data in psychology. Missing data and cases requiring exclusion are also addressed because they are such common issues when using data collected from living research subjects. Finally, this book makes substantial modifications to the TIER documentation protocol to encourage consideration of the participant confidentiality requirements important in many types of psychological research.

[2]SPSS is a registered trademark of International Business Machines Corporation.

[3]Excel may be used for data entry, but it is not recommended for data management and analysis because it cannot be programmed with the executable command files that are essential for carrying out these procedures in a fully replicable and documented way. Excel is similarly considered incompatible with the Project TIER protocol (Owens, 2014).

ethical research practices for improving scientific replicability. Because these procedures help researchers stay more organized and keep better records, practicing them can also help people feel less overwhelmed and gain a greater sense of mastery when working with data, in turn allowing people to think more deeply and creatively about their data and get more out of the experience.[4]

The goal of the procedures described in this book is to have a researcher who reads your paper be readily able to recreate your final results, starting from the beginning with the same data file(s) with which you originally started. To make this goal possible, this book is designed on the following principles, which correspond to the four central components of the *project folder* that you will create to store what is needed for your project and that also correspond to the four upcoming chapters. Each principle is the focus of a subsequent chapter in this book.

- Keep well-organized copies of all project files, such as institutional review board (IRB) approvals, materials, logs, and participant information (Chapter 2).
- Keep read-only copies of your data file(s) in the original, untouched state they were in before you started modifying them in any way (Chapter 3).
- Create reader-friendly command files to document everything you do with your data (instead of saving endless pages of output and instead of typing up a separate set of descriptive documents; Chapter 4).
- Create a concise folder of replication documentation to correspond to the final version of your paper, suitable for use by outside researchers (Chapter 5).

The remainder of this book explains one folder at a time each of these components of your project folder and what makes them useful and important. Before we proceed, though, let's start with an overview of your project folder and its structure.

[4]Being able to manage data is also a skill of practical value to develop because it is associated with more job prospects and higher salary even for students who do not pursue a career in science (Blumenstyk, 2016).

YOUR PROJECT FOLDER

When you are starting your project, one of your first steps will be to create a series of computer folders (on a specific computer or in cloud storage) to contain and organize all your work while you are completing your project. If your paper is published, you will keep the project folder that is associated with it for at least 5 more years, as required by the APA (APA Ethical Standard 6.01, Documentation of Professional and Scientific Work and Maintenance of Records), in case researchers require additional information.

You will begin by creating a folder for your project and give it an informative, clearly identifying name (rather than a generic or vague one such as "Project folder" or "Psych project" or even "Developmental psych final project"). If you're handing this folder in for a class, keep in mind that professors are often frustrated when they receive many identical-looking assignments that cannot be easily matched to their creators (e.g., dozens of electronic documents that are all called "final paper"). A better name for your folder might include the last names of the individual(s) doing the project and/or an abbreviated project title, along with the course number and year (e.g., "Bond Psych 399 Spring 2017" or "Perfectionism and Rumination Psych 399 Fall 2016"). Research labs similarly use abbreviations, nicknames, or numbering systems to identify particular projects and distinguish them from one another (e.g., "Attributions Study 2"). What is important is that the project folder's name uniquely refers to the specific project in a way that is recognizable and sensible to everyone involved.

Next, create the subfolders needed within your project folder, as shown in Figure 1.1. The order in which you're creating these folders or files is similar to the order in which you'll fill them and use them. You'll first create files for private use by you and other members of your research team in carrying out and writing up your study. Afterward, you will prepare replication documentation for sharing and/or storing in a public archive. Notice that the subfolders in Figure 1.1 are numbered (e.g., 1, 1a, 1b, etc.) so that they will stay in this same fixed order on your computer.

The following is an overview of the various subfolders and the steps you will take to fill them as you complete your project from beginning to

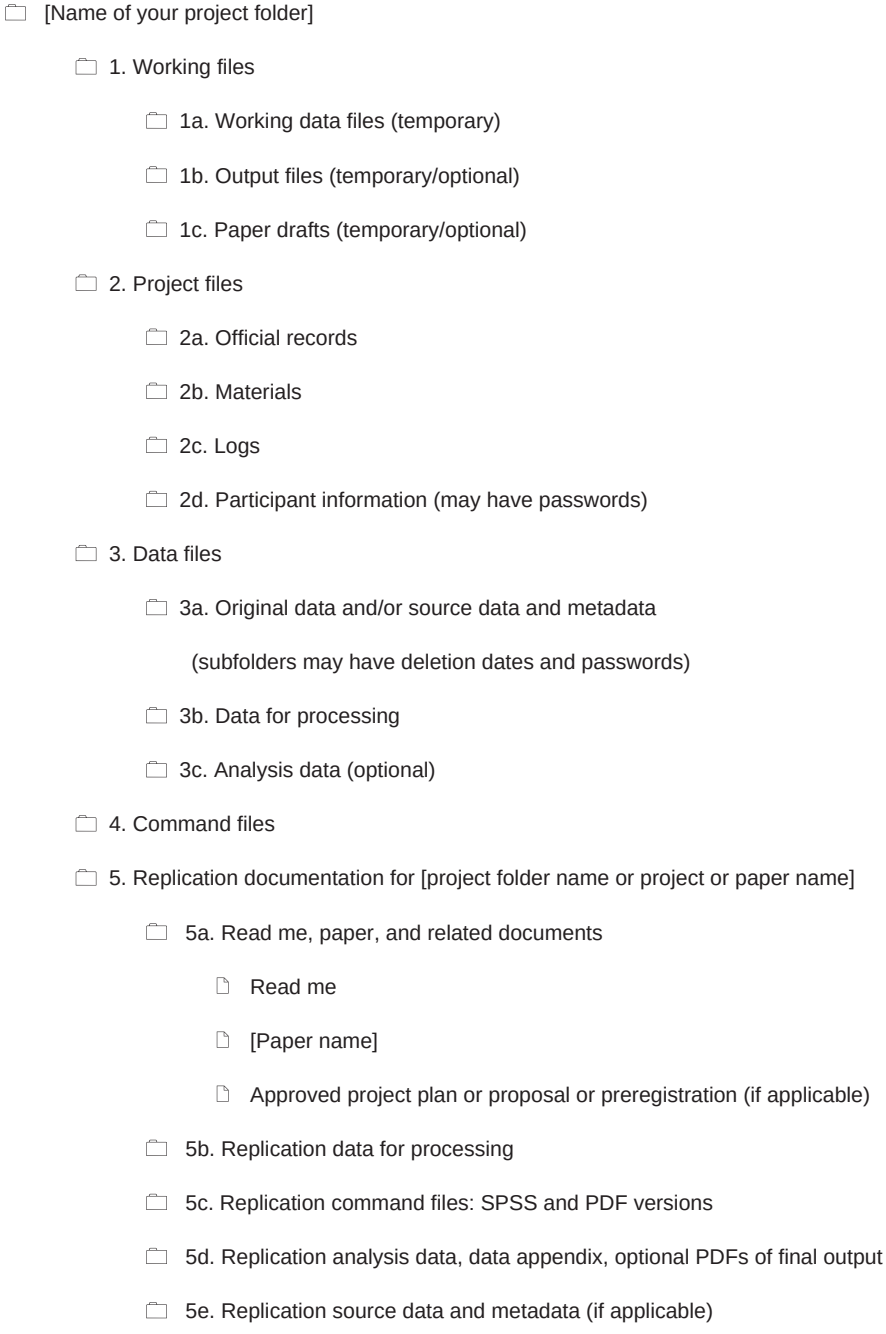

Figure 1.1

Project folder.

end. Many of these subfolders will stay empty for a while, but that's OK. You'll be learning how to fill up the subfolders within your project folder throughout the rest of this book, and you might want to refer back to this overview as you read further ahead.

Working Files

Your *Working Files* folder is a place to temporarily store the files you are actively working with while your project is in progress. The folder contains three major subfolders. The *Working Data* folder will be kept empty most of the time and serves only one purpose: to remind you never to modify directly any of the data files that you keep archived in your Data Files folder (described later). That is, when you are doing data management or analyses, you will always start by putting a copy of the data file that you want to work with into your Working Data folder and run the applicable command file(s) on that temporary copy. When you're finished working with it, you can choose to save that working data file as an optional Analysis Data file (described later) or to delete it. In addition, you may want to optionally store *Output Files* and *Paper Drafts* in progress in your Working Files folder. These files can be considered "working files" in that they are meant to be temporary and do not have to be saved, but it can be helpful to keep them readily accessible for looking up information while you're working on your paper. (See Chapter 2 for a suggestion on how to keep multiple drafts of the same document organized by labeling them with the date they were created.)

Project files, as described in Chapter 2, include various important documents that are not data or command files but that have to be preserved as records of your project. You will be able to start filling several of the Project Files subfolders with forms even before you collect any data. Your *Official Records* subfolder will contain your IRB submission and whatever other proposals or plans you submitted for approval prior to starting your study. Your *Materials* subfolder will store copies of things such as questionnaire packets, experiment instructions, and stimuli. Your *Logs* subfolder will contain a data issues log to keep track of potential problems with particular observations or cases. You will add to this log while collecting data

and perhaps while managing or analyzing it. Depending on the needs of your project, you may benefit from keeping additional logs—for instance a data collection log, data entry log, or data analysis log. Finally, you will use the *Participant Information* subfolder to store any potentially identifying records (such as lists of names, contact information, or ID numbers) in password-protected files.

Data Files

You will keep all your data organized in your *Data Files* folder (see Chapter 3). When your data have been collected and/or entered, you will store your untouched data files in your *Original Data* folder. If you obtain your data from another source (rather than collect it yourself), you will store untouched copies of the source data along with metadata, as described in Appendix A. Your *Data for Processing* folder will contain your data files in SPSS format. If your original data files were in SPSS format to begin with, these will simply be duplicate copies. After all data management tasks have been completed, you can choose to save a copy of your data set(s) in ready-to-analyze form in an optional *Analysis Data* folder. In addition to a detailed discussion of each of these subfolders, Chapter 3 provides guidelines useful in creating data files, such as information about entering data and choosing names for your variables. The accompanying appendices address special topics related to data files, including metadata for data sets obtained from other sources (Appendix A), "tall" data files (with repeated observations of each variable; Appendix B), and instructions for ensuring accurate data entry (Appendix C).

Command Files

As described in Chapter 4, to manage and analyze your data, you will create command files for everything you do to manage and analyze your data. After an orientation to *command files* for people who may never have used (or even heard of) them before, Chapter 4 provides step-by-step instructions for many common data management tasks, including renaming variables, assigning variable labels and value labels, checking frequency tables

for data errors, using missing data codes, handling missing data, excluding cases, computing variables (including reversing items and creating groups, *z*-scores, and scales), analyzing scale reliability, and working with subsets of your sample. Additional instructions for specialized data management tasks are covered in the appendices, including labeling and renaming many variables efficiently (Appendix D), importing data files (Appendix E), merging data files (Appendix F), and estimating missing values (Appendix G).

Replication Documentation

After you finish your analyses and write your paper, you'll put your paper and all the materials necessary for replication of your results into a publicly shareable *Replication Documentation* folder (see Chapter 5). The *Read Me* document will list or describe the files in your Replication Documentation folder and provide instructions for recreating the results reported in your paper. You'll check that the steps listed actually work by reproducing your own results yourself. You'll also include the final version of your paper and any documentation of the plans you made for your study prior to conducting it, such as a formal preregistration or an approved proposal. The Data for Processing files that you'll put in your shareable Replication Documentation folder will often be identical copies of the data files you've saved for private use. However, to be consistent with participant confidentiality and copyright laws, you may have to omit some information from the data files you share. You will create *replication command file(s)* by copying and editing the relevant portions of the command files you used to manage and analyze your data. Replication commands for data analysis will appear in the order in which they appear in your paper and have comments indicating the page on which the relevant results appear. Your replication command files will be saved as PDFs (as well as the standard SPSS command file format) to make it easier for readers to access them. You will save shareable Analysis Data file(s) containing the data used for the analyses reported in your paper (with the data already cleaned and variables computed). You will also create a *data appendix* document to provide variable information and descriptive statistics for every variable in the Analysis Data file. You can optionally create a PDF version of the final output of the analyses in

your paper to store in this folder. Finally, if you obtained your data from another source and have permission to share it, you'll include a copy of your *Source Data* folder (and its Metadata subfolder) in your Replication Documentation folder. Submit your Replication Documentation folder along with your paper if requested by the professor, institution, or publication to which you are submitting it.

CONCLUDING COMMENTS

Some medical professionals might like to play loud, energizing music when they are getting ready to perform surgery, whereas others might prefer meditative silence, but these differences in style do not matter as long as all the essential cleaning and preparation gets done. Likewise, individual researchers may have their own tips and tricks for managing data and documentation or use different terms for their files and folders than the ones suggested here, but these differences in style don't matter either. What matters is that all researchers manage their data and documentation in ways that meet the increasingly strict standards that funding agencies, journals, and the field of scientific psychology at large are adopting to promote greater transparency and replicability of psychological research. This book will teach you a way to do this, step by step. And as you've just read, the first step is to create a well-organized set of empty subfolders for storing all the things you will be keeping in your project folder.

Before moving on, take a moment to consider that you're now the owner of the project folder you've just created. (Congratulations!) You will fill your folder with what it needs, update it, and keep it organized. You will ensure that your folder is regularly backed up, as a precaution against computer loss and other unthinkables. When you meet with your collaborators and/or professor to work on your project, you will have the materials you need from this folder ready (i.e., not assume that others will provide them or make others wait while you try to find them). You will put in the time it takes to become familiar with what is inside your folder and take responsibility for asking questions about things that are unclear to you. Owning this folder is a central part of making your research project truly yours.