Making Research Reproducible

Jason S. McCarley & Raechel N. Soicher

School of Psychological Sciences

Oregon State University

Corresponding Author:

Dr Jason McCarley

Professor

School of Psychological Sciences

Oregon State University

Corvallis, OR 97331

541-737-1175

Email: Jason.Mccarley@oregonstate.edu

Copyright 2022 by **Jason McCarley and Raechel N. Soicher**. All rights reserved. You may reproduce multiple copies of this material for your own personal use, including use in your classes and/or sharing with individual colleagues as long as the author’s name and institution and the Society for the Teaching of Psychology (STP) heading or other identifying information appear on the copied document. No other permission is implied or granted to print, copy, reproduce, or distribute additional copies of this material. Anyone who wishes to produce copies for purposes other than those specified above must obtain the permission of the authors.

Contents

[Teaching Guide 2](#_Toc119328263)

[Relevance to the Teaching of Psychology 2](#_Toc119328264)

[Teaching Materials 3](#_Toc119328265)

[Student Learning Outcomes: 3](#_Toc119328266)

[Full Resource References 5](#_Toc119328267)

[Reproducibility: What it Is and Why It Matters 9](#_Toc119328268)

[What’s needed to make an analysis reproducible? 9](#_Toc119328269)

[What makes reproducibility important? 10](#_Toc119328270)

[To help us spot errors 10](#_Toc119328271)

[To boost robustness and replicability 10](#_Toc119328272)

[To discourage fraud 10](#_Toc119328273)

[To allow more discoveries 10](#_Toc119328274)

[What makes reproducibility difficult? 11](#_Toc119328275)

[Unavailability of data 11](#_Toc119328276)

[Unavailability of analytic code 11](#_Toc119328277)

[What does it mean when reproducibility fails? 11](#_Toc119328278)

[References 12](#_Toc119328279)

[Making Your Research Reproducible: JASP Users 15](#_Toc119328280)

[Availability 15](#_Toc119328281)

[Ethics of Data Sharing 15](#_Toc119328282)

[Organization and Documentation 15](#_Toc119328283)

[References 18](#_Toc119328284)

[Making Your Research Reproducible: R Users 20](#_Toc119328285)

[Availability 20](#_Toc119328286)

[Available Upon Request 20](#_Toc119328287)

[Publicly Available 20](#_Toc119328288)

[Ethics of Data Sharing 20](#_Toc119328289)

[Permission to Share 20](#_Toc119328290)

[De-Identifying Data 20](#_Toc119328291)

[Organization and Documentation 20](#_Toc119328292)

[References 23](#_Toc119328293)

[Exercise: Organizing and Documenting Data for Reproducibility with JASP 25](#_Toc119328294)

[1. Understand data structure 25](#_Toc119328295)

[Start from the Very Beginning 25](#_Toc119328296)

[Data Preparation (from Codebook Cookbook and Research Methods in Psychology) 25](#_Toc119328297)

[Variable Names 26](#_Toc119328298)

[Variable Codes 26](#_Toc119328299)

[2. Create a set of folders 26](#_Toc119328300)

[3. Download the original data and create data dictionary 27](#_Toc119328301)

[About the Study 27](#_Toc119328302)

[About the Data 27](#_Toc119328303)

[Create a data dictionary for the original data 27](#_Toc119328304)

[4. Create the analysis data by transferring the original data into a spreadsheet 28](#_Toc119328305)

[Familiarize Yourself with the Data 28](#_Toc119328306)

[Put the Data into a Spreadsheet 28](#_Toc119328307)

[Saving the Dataset 28](#_Toc119328308)

[5. Create a data dictionary for your analysis data 28](#_Toc119328309)

[6. Zip your folders & submit the assignment 28](#_Toc119328310)

[Exercise: Organizing and Documenting Data for Reproducibility with R 29](#_Toc119328311)

[1. Understand data structure 29](#_Toc119328312)

[Start from the Very Beginning 29](#_Toc119328313)

[Data Preparation (from Codebook Cookbook and Research Methods in Psychology) 29](#_Toc119328314)

[2. Create a set of folders 30](#_Toc119328315)

[3. Download the original data and create data dictionary 31](#_Toc119328316)

[About the Study 31](#_Toc119328317)

[About the Data 31](#_Toc119328318)

[Create a data dictionary for the original data 31](#_Toc119328319)

[4. Create the analysis data by transferring the original data into a spreadsheet 32](#_Toc119328320)

[Familiarize Yourself with the Data 32](#_Toc119328321)

[Put the Data into a Spreadsheet 32](#_Toc119328322)

[Saving the Dataset 32](#_Toc119328323)

[5. Create a data dictionary for your analysis data 32](#_Toc119328324)

[6. Zip your folders & submit the assignment 32](#_Toc119328325)

Making Research Reproducible

# Teaching Guide

## Relevance to the Teaching of Psychology

Statistics and research methods are a core component of the undergraduate psychology curriculum. Their purpose is not just to teach the particular knowledge and skills required of psychological scientists—research design, quantitative analysis, scientific communication—but through that teaching, to train students in the more general skills of scientific reasoning and critical thinking (American Psychological Association, 2013).

The troubling and ongoing crisis of replication, unfortunately, has revealed weaknesses in the research practices that psychologists have been taught and have passed onto their own students (Pashler & Wagenmakers, 2012). In keeping with the goals of the stats/methods curriculum, teachers of psychology have responded to the replication crisis with lessons acknowledging scientific failures, highlighting the soft spots in traditional research practices (Chopik et al., 2018), and even recommending the use of replications as projects for undergraduate methods classes (Frank & Saxe, 2012).

This discussion has been invaluable but has focused largely on conditions necessary for empirical replicability, that is, the ability of future researchers to repeat a study and obtain results similar to the original findings. Less attention has been paid to the ancillary problem of statistical reproducibility, the ability of future researchers to simply repeat analyses on an existing set of data and obtain the same results as originally reported (Stodden, 2015). A majority of published research fails to meet the conditions necessary for statistical reproducibility (Stodden et al., 2018), hindering efforts to spot analytic blunders and other constraints on validity.

## Teaching Materials

These materials are designed for teaching analytic reproducibility to undergraduate students in introductory statistics and research methods courses in psychology. They discuss the necessity of statistical reproducibility, with examples drawn from psychology and related social sciences.

## Student Learning Outcomes:

* document data though metadata/codebooks;
* document analyses through analytic recipes/code documentation;
* organize data and analyses using intelligible folder structures and readable filenames;
* share materials and data ethically, with protections for participant confidentiality.

These materials are suitable for integration into a wide range of courses or even a short within-course sequence on research transparency, from large lecture courses with lab sections led by graduate teaching assistants to online and hybrid courses. The materials have been developed for two software types: one tailored to courses using GUI-driven software (e.g., SPSS, JASP, Jamovi) and one tailored to courses using the statistical scripting language R.

This document includes the following resources:

* A short reading defining reproducibility and discussing the critical considerations around data sharing (e.g., ethics) and the importance of reproducibility.
* A primer for students on how to make their research reproducible (one primer for GUI-driven software, one for R).
* A class exercise for students to learn and practice the basics of data entry, coding, organization, and storage (one for each software type).

The other materials needed for the exercise are:

* The PowerPoint presentation for students discussing the importance of reproduction to transparency in psychology research and an introduction to methods for making their research reproducible.
* The PDF containing a set of simulated raw data for use in the exercise ([MusicAndRoomColorData.pdf](https://osf.io/xnq6j))

## Full Resource References

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, *73*(2), 146–156. <https://doi.org/10.1037/amp0000258>

American Psychological Association. (2013). *APA guidelines for the undergraduate psychology major: Version 2.0*. Washington, D.C.

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. <https://doi.org/10.1037/met0000365>

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. <https://doi.org/10.3758/s13428-011-0089-5>

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., Berthold, A., … Nguyen, H. H. V. (2021). *Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty*. MetaArXiv. <https://doi.org/10.31222/osf.io/cd5j9>

Chang, A. C., & Li, P. (2015). *Is economics research replicable? Sixty published papers from thirteen journals say “usually not”* (SSRN Scholarly Paper No. 2669564). <https://doi.org/10.2139/ssrn.2669564>

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology*, *45*(2), 158–163. <https://doi.org/10.1177/0098628318762900>

Eubank, N. (2016). Lessons from a decade of replications at the Quarterly Journal of Political Science. *PS: Political Science & Politics*, *49*(2), 273–276. <https://doi.org/10.1017/S1049096516000196>

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604. <https://doi.org/10.1177/1745691612460686>

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Henry, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (n.d.). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(180448). <http://dx.doi.org/10.1098/rsos.180448>

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 104332. <https://doi.org/10.1016/j.jml.2022.104332>

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. <https://doi.org/10.1371/journal.pone.0233107>

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. <https://doi.org/10.1177/2515245917747656>

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. A. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in The BMJ and PLOS Medicine. *BMJ*, *360*, k400. <https://doi.org/10.1136/bmj.k400>

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), annurev-psych-020821-114157. <https://doi.org/10.1146/annurev-psych-020821-114157>

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. <https://doi.org/10.3758/s13428-015-0664-2>

Pashler, H., & Wagenmakers, E. (2012). Editors’ introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. <https://doi.org/10.1177/1745691612465253>

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, *12*(3), 30–32. <https://doi.org/10.1111/j.1740-9713.2015.00827.x>

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. <https://doi.org/10.1073/pnas.1708290115>

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, *2*(1), 1–19. <https://doi.org/10.1146/annurev-statistics-010814-020127>

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. <https://doi.org/10.1038/s41597-021-00981-0>

*TIER Protocol | Project TIER | Teaching Integrity in Empirical Research*. (2022, July 12). <https://www.projecttier.org/tier-protocol/>

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*(1), 3. <https://doi.org/10.1525/collabra.13>

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS ONE*, *9*(12), e114876. <https://doi.org/10.1371/journal.pone.0114876>

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLOS ONE*, *6*(11), e26828. <https://doi.org/10.1371/journal.pone.0026828>

Wolins, L. (1962). Responsibility for raw data. *American Psychologist*, *17*(9), 657–658. <https://doi.org/10.1037/h0038819>

# Reproducibility: What it Is and Why It Matters

Scientists collect and analyze data to learn about the world. Ideally, our findings will lead to new treatments, services, and products that make peoples’ lives better. But before we can translate scientific discoveries into useful innovations, we need to be confident that those discoveries are trustworthy. In psychology, for example, we will want to make sure that the design of the study was sound and that participant sample was unbiased.

An aspect of research quality that is sometimes overlooked is the need for *reproducible* results. To understand reproducibility, we can compare it to two related concepts, *replicability* and *robustness* (Nosek et al., 2022).

* The results of a study are replicable if researchers can repeat the study over and over again and (most of the time) get a similar outcome. Say, that I run an experiment and find an effect, then I run the experiment again and get the same effect, and then you run the experiment yourself and find the same effect again. After finding the same result in three consecutive runs of the experiment, by two different researchers, we can start to be confident that the effect is replicable.
* The results of a study are robust if researchers can analyze the data in a different way and reach the same meaningful conclusions. Data analysis often requires lots of small decisions that don’t necessarily have a single correct answer (Breznau et al., 2021). For instance, we might have choices to make about which statistical approach to use for our study, how to identify outliers (weird data points that look like they might be errors), or how to handle missing data (e.g., what to do if a participant fails to respond to a couple of questions in a larger survey). If we can show that different, reasonable choices don’t affect our conclusions, we know our findings are robust.
* The results of a study are reproducible if someone can take the data from the study, repeat the original analyses, and get the same outcome. Imagine, for example, that I run an experiment, analyze the data, and report my findings. We would call the analysis reproducible if you can take the same data, perform the same analyses, and get the same results.

In short (Nosek et al., 2022):

* Replicability = new data, same analyses, same results.
* Robustness = same data, new analyses, same results.
* Reproducibility = same data, same analyses, same results.

Here, our focus will be on the last of these, reproducibility: What it entails, why it’s important, and what makes it difficult.

## What’s needed to make an analysis reproducible?

At least two things are required for an analysis to fully reproducible. First, the data and analysis code—our research *artifacts* (Stodden et al., 2018)—have to be accessible. By analytic code, we mean any computer scripts or automated routines that are used to carry out an analysis. If we are trying to reproduce our own analyses, accessibility means that we know where the artifacts are stored (on our hard drive, in the cloud…) and we are able to retrieve them. If we are trying to reproduce someone else’s analysis, accessibility requires that they have shared their research artifacts with us.

Second, the research artifacts have to be organized and documented in a way that lets us make sense of them. If we start with a jumble of unexplained data files and scripts, we aren’t likely to have much luck reproducing an analysis. This is true whether the analysis we want to reproduce is somebody else’s or one of our own.

## What makes reproducibility important?

It might not be obvious why reproducibility matters. After an analysis has been conducted and published in a journal article, why would it be necessary for someone else to exactly repeat it? It turns out, reproducibility is important in several ways.

### To help us spot errors

Analyzing data and reporting the results of a study can be a difficult process, and no matter how careful we are, we're bound to make a mistake every now and then. It’s easy to mislabel a variable, mistype an equation, misread a statistical output, or make some other small but critical error. Audits have found that roughly half of the journal articles published in psychology contain at least one statistical error, and that 10–20% or more of published articles contains a statistical error big enough to change the article’s conclusions (Bakker & Wicherts, 2011; Nuijten et al., 2016; Veldkamp et al., 2014).

By making our analyses reproducible, we make it easier for ourselves and others to find these kinds of errors. Preserving our data and analytic code makes it possible to go back and check our own work. Allowing other researchers to see our data and analytic code lets them check that we have not made any mistakes, and helps them to feel confident in the findings we report.

### To boost robustness and replicability

Reproducibility makes it easier to check for replicability and robustness. By preserving, documenting, and sharing our data and analytic code, we let other researchers follow up on our original analyses, tweaking them to see if our conclusions are robust. We also enable other people to reuse our work when they try to replicate our findings in new experiments.

### To discourage fraud

Most researchers are honest, but some aren't. In the most extreme cases, scientists fake or alter their data in order to produce a "finding" that isn't really there. Reproducibility lets other researchers check our work for signs of fraud, should they wish to (e.g., <https://datacolada.org/98>), and helps them be confident that our data are real and our findings are truthful.

Making our data and analyses available to other researchers also models good scientific behavior and helps build the expectation of reproducibility. This will make it more difficult for future cheaters to hide their fraud.

### To allow more discoveries

Sometimes, after reading about our research, other people may think of new questions to ask with the same data. By documenting and sharing our data, we allow other researchers to go beyond our analyses, making new discoveries without having to spend time and money on new data collection. That makes science more efficient.

## What makes reproducibility difficult?

Even if we agree that reproducibility is important, it’s not obvious why it might be difficult. How hard can it be to repeat an analysis that has already been done? Unfortunately, we often face obstacles when we try to reproduce an analysis.

### Unavailability of data

One big impediment to reproducing other peoples’ analyses is that the data aren’t available. In some cases, legal and ethical constraints prevent researchers from sharing their data (Meyer, 2018). We don’t expect researchers to share data, for example, if there is any possibility that individual participants might be identifiable.

But in the absence of legal and ethical barriers, researchers ought to share their data openly. Unfortunately, many don’t. In one investigation of data sharing practices in psychology (Vanpaemel et al., 2015), only 38% of authors contacted through email were willing to provide data upon request. A similar number (40%) ignored the request, and almost 20% flatly refused to share their data.

Many authors fail to provide data for published journal articles even when the journal supposedly requires it (Hardwicke et al., 2018; Laurinavichyute et al., 2022; Tedersoo et al., 2021). In February 2011, the journal *Science* implemented a policy making it mandatory for authors to share data and code for computational analyses. Despite the policy, fewer than half of the papers published in the following two years actually made their research artifacts available (Stodden et al., 2018).

Interestingly, papers for which authors are unwilling to share data are more likely to contain weak evidence or statistical errors than papers for which the authors share the data (Wicherts et al., 2011). This suggests that authors are more willing to share their data openly when they are confident in the strength of their research.

### Unavailability of analytic code

We might still find it difficult to reproduce an analysis even if we have access to the data. In theory, a scientific article is supposed to describe its statistical methods in enough detail for another person to repeat them. In reality, researchers often fail to describe their analyses in appropriate detail and sometimes even report them incorrectly. In one audit of the literature, for example, authors were found to have omitted details of their data exclusion procedures, or to have reported using a different statistical test than the one they actually used (Artner et al., 2021). An analysis is therefore more likely to be reproducible if the original researchers share their analytic code than if they don’t (Laurinavichyute et al., 2022).

## What does it mean when reproducibility fails?

Sometimes, for any of the reasons we’ve talked about—lack of data or analytic code, poor documentation of analyses—an attempt to reproduce earlier findings simply fails to get off the ground. An outcome like this is termed a *process reproducibility failure* (Nosek et al., 2022).

In other cases, we might be successful at re-creating the original analytic procedures but end up with results that are different from what the original researchers reported. This is termed an *outcome reproducibility failure* (Nosek et al., 2022). An outcome reproducibility failure implies that there was either an error in the original analyses and report, or an error in our reproduction. Follow up investigation is necessary.

## References

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. <https://doi.org/10.1037/met0000365>

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. <https://doi.org/10.3758/s13428-011-0089-5>

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., Berthold, A., … Nguyen, H. H. V. (2021). *Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty*. MetaArXiv. <https://doi.org/10.31222/osf.io/cd5j9>

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Henry, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(180448). <http://dx.doi.org/10.1098/rsos.180448>

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 104332. <https://doi.org/10.1016/j.jml.2022.104332>

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. <https://doi.org/10.1177/2515245917747656>

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), annurev-psych-020821-114157. <https://doi.org/10.1146/annurev-psych-020821-114157>

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. <https://doi.org/10.3758/s13428-015-0664-2>

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. <https://doi.org/10.1073/pnas.1708290115>

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. <https://doi.org/10.1038/s41597-021-00981-0>

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*(1), 3. <https://doi.org/10.1525/collabra.13>

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS ONE*, *9*(12), e114876. <https://doi.org/10.1371/journal.pone.0114876>

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE*, *6*(11), e26828. <https://doi.org/10.1371/journal.pone.0026828>

# Making Your Research Reproducible: JASP Users

At least two things are needed to make our work reproducible: our research artifacts must be available and must be organized and documented in a way that lets people make sense of them. We’ll tackle these two issues in turn.

## Availability

At a minimum, anyone who wants to reproduce our analyses needs access to our data. Preferably, we will share the details of our statistical analyses as well. How do we provide these materials?

#### Available Upon Request

Some researchers offer to make their artifacts “available upon request.” In other words, they attach a note when they publish their research, telling others to email and ask for the data and code if they want them. This approach is better than nothing, but a lot can go wrong. Some researchers promise to share their artifacts upon request but refuse to do so when they are actually asked (Hardwicke et al., 2018; Stodden et al., 2018; Laurinavichyute et al., 2022; Tedersoo et al., 2021; Vanpaemel et al., 2015). In other cases, the researcher who has the artifacts might change email addresses, retire, or otherwise become impossible to reach. In the most extreme case, if new researchers decide to reanalyze a set of data that is a few decades old (that happens from time-to-time), the original researcher might be long deceased.

#### Publicly Available

A better practice is to share the data and JASP files publicly, right from the start. A number of free online repositories will let us post our artifacts so that they can be permanently available for easy download by other scientists. Posting our artifacts publicly also makes it easy for *us* to find them ourselves, if we need them in the future. A popular repository for sharing research artifacts in psychology is the Open Science Framework, [osf.io](https://osf.io/).

### Ethics of Data Sharing

#### Permission to Share

Before sharing data with anyone, it’s important to make sure we have permission to share them. Ideally, we should have approval from your IRB or ethics board to share the data, and your informed consent form should tell participants where and why their data might be shared (Meyer, 2018). We might also need approval from research funders or partners, for example, if we are doing research that relates to matters of national security. And if we are using government, medical, or educational data, we need to follow federal and state privacy laws (Alter & Gonzalez, 2018).

#### De-Identifying Data

It’s also critical that we de-identify data before sharing them. In other words, we need to remove any information that might allow individual participants to be identified based on their data. This includes info like names, email addresses, IP addresses, and any other data that might let someone track down participants. If we can’t remove identifying information, then it is almost always inappropriate to share data (Meyer, 2018).

## Organization and Documentation

Even for a small-scale research project, computerized data analysis can involve multiple artifacts of different types. Working in JASP, these might include:

* *One or more files of raw data.* These are the data in their original electronic form. These files will be generated automatically If the data are collected electronically or might be generated manually or scanned in if the data are not collected electronically. For example, a computerized survey will automatically store the raw data in electronic files. A paper-and-pencil survey, on the other hand, will require a researcher to enter the data into a computer file by hand, or possibly to scan in the completed surveys. In any case, the raw data file(s) would contain the survey responses exactly as they were entered by the respondents. Note that once we have created these files, we never want to modify them. We need to retain them in the original form, permanently.
* *One or more files of data for analysis.* Sometimes, the raw data files might not be suitable to go directly into our statistical analysis software. They might contain information we don’t need, they might have errors or duplicate values, some of them might be incomplete, and they might be formatted in a way that our statistical software doesn’t like. To make analysis easier, we can “clean” the data by picking out only the information we need, removing the errors and duplicate values, and screening out the incomplete files. After that, if needed, we can reorganize the data into a format that our statistical software can use. Because we never want to modify our raw data files (remember?), we will save the cleaned and reformatted data in new files.
* *JASP Files.* There are three components to a JASP file.
  + *Spreadsheet.* Within the file is a spreadsheet that will hold your data. You can usually copy and paste the data columns from your analysis data files into this spreadsheet.
  + *Point-and-click controls.* To run our analyses in JASP, we will “point and click” on options in the software that tell it how to analyze the data. A menu of icons across the top of the JASP window lists different kinds of analyses for you to choose from. When you select an analysis, a panel will appear on the right side of the window, letting you select which variables to analyze and choose analysis settings. Each time you want to look at a new combination of variables or new analysis, you should start a new analysis, rather than simply moving around variables in the existing analysis. By doing this, you preserve all your decisions about how to analyze the data in the JASP file for someone else to see.
  + *Outputs.* When you run an analysis, the results appear in a panel on the right of the JASP window. JASP automatically updates the output every time you click a button. The output contains anything relevant to the statistical analysis, including figures, tables, or any other kind of statistical output we produce with our analysis. These outputs are saved with the JASP file, providing a record of your analyses.

We could just dump all these files into one big folder, but that might not be very helpful to someone who wants to reproduce our analyses. One problem is that it might not be obvious which files are which. Another problem is that the contents of the data files might not be easy to interpret. In most cases, the data will contain multiple columns of numbers or text. Where did those values come from? What do they mean? We need documentation to tell us.

So, to enable reproducibility, it’s best to organize all our artifacts, and to include documents that explain the contents of our data files. The way we choose to organize our files can vary from project to project, depending on our needs and personal preferences, but Figure 1 below shows a sample organizational system. This is based on a protocol recommended by the open-science organization Project TIER (“Tier Protocol”, 2022).



Figure 1 A folder structure for reproducibility. A top-level folder holds the entire project. Inside are subfolders for the original data, the analysis data, the JASP files, and written reports. ReadMe file and data dictionaries provide documentation.

Here, we have a top-level folder to hold our entire project. Inside that, we have subfolders for the original data files, the analysis data, the JASP files, and outputs exported from JASP. We also have three PDF files to help explain our project artifacts. (Note that it’s best not to use spaces in our folder and filenames. In Figure 1, we have eliminated the spaces from our folder and filenames and have capitalized the first letter of each word for readability. This is called *Pascal case*.)

* In the project folder, we have a *ReadMe* document that gives a brief, general description of our project, tells the reader what software we used for data analysis, and gives instructions for how to reproduce our analyses.
* An *OriginalDataDictionary* document that explains the files in our *OriginalData* folder. The exact contents of the *OriginalDataDictionary* will depend on our research project, but generally, the file should include:
  + an explanation of where the data came from (Did we collect them yourself? Did we download them?) and the date we got them;
  + a list of all the variables in data files, the definition of each variable, and an explanation of how the values are coded (i.e., what the values represent).
* An *AnalysisDataDictionary* document that explains the files in our *AnalysisData* folder. Usually—but not always--the analysis data are formatted in a table where each row represents an individual participant and each column represents a variable. At the top is a row of variable names. The *AnalysisDataDictionary* should explain what each row represents and what each variable name means. Just like the *OriginalDataDictionary*, it should also provide a definition of each variable, and an explanation of how the values are coded.

Once you have created this folder structure, we can zip it up to share to a public repository. Or if we don’t plan to share the artifacts publicly, we can just store the project folder and its contents someplace where we can find them when needed.

## References

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, *73*(2), 146–156. <https://doi.org/10.1037/amp0000258>

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Henry, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(180448). <http://dx.doi.org/10.1098/rsos.180448>

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 104332. <https://doi.org/10.1016/j.jml.2022.104332>

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. <https://doi.org/10.1177/2515245917747656>

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. <https://doi.org/10.1073/pnas.1708290115>

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. <https://doi.org/10.1038/s41597-021-00981-0>

*TIER Protocol | Project TIER | Teaching Integrity in Empirical Research*. (2022, July 12). <https://www.projecttier.org/tier-protocol/>

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*(1), 3. <https://doi.org/10.1525/collabra.13>

# Making Your Research Reproducible: R Users

At least two things are needed to make our work reproducible: our research artifacts must be available and must be organized and documented in a way that lets people make sense of them. We’ll tackle these two issues in turn.

## Availability

At a minimum, anyone who wants to reproduce our analyses needs access to our data. Preferably, we will share our analytic code as well. How do we actually provide these materials?

### Available Upon Request

Some researchers offer to make their artifacts “available upon request.” In other words, they attach a note when they publish their research, telling others to email and ask for the data and code if they want them. This approach is better than nothing, but a lot can go wrong. Some researchers promise to share their artifacts upon request but refuse to do so when they are actually asked (Hardwicke et al., 2018; Stodden et al., 2018; Laurinavichyute et al., 2022; Tedersoo et al., 2021; Vanpaemel et al., 2015). In other cases, the researcher who has the artifacts might change email addresses, retire, or otherwise become impossible to reach. In the most extreme case, if new researchers decide to reanalyze a set of data that is a few decades old (that happens from time-to-time), the original researcher might be long deceased.

### Publicly Available

A better practice is to share the data and scripts publicly, right from the start. A number of free online repositories will let us post our artifacts so that they can be permanently available for easy download by other scientists. Posting our artifacts publicly also makes it easy for *us* to find them ourselves, if we need them in the future. A popular repository for sharing research artifacts in psychology is the Open Science Framework, [osf.io](https://osf.io/).

## Ethics of Data Sharing

### Permission to Share

Before sharing data with anyone, it’s important to make sure we have permission to share them. Ideally, we should have approval from our IRB or ethics board to share the data, and our informed consent form should tell participants where and why their data might be shared (Meyer, 2018). We might also need approval from research funders or partners, for example, if we are doing research that relates to matters of national security. And if we are using government, medical, or educational data, we need to follow federal and state privacy laws (Alter & Gonzalez, 2018).

### De-Identifying Data

It’s also critical that we de-identify data before sharing them. In other words, we need to remove any information that might allow individual participants to be identified based on their data. This includes info like names, email addresses, IP addresses, and any other data that might let someone track down participants. If we can’t remove identifying information, then it is almost always inappropriate to share data (Meyer, 2018).

## Organization and Documentation

Even for a small-scale research project, computerized data analysis can involve multiple artifacts of different types. Working in R, these might include:

* *One or more files of raw data.* These are the data in their original electronic form. These files will be generated automatically If the data are collected electronically, or might be generated manually or scanned in if the data are not collected electronically. For example, a computerized survey will automatically store the raw data in electronic files. A paper-and-pencil survey, on the other hand, will require a researcher to enter the data into a computer file by hand, or possibly to scan in the completed surveys. In any case, the raw data file(s) would contain the survey responses exactly as they were entered by the respondents. Note that once we have created these files, we never want to modify them. We need to retain them in the original form, permanently.
* *One or more files of data for analysis.* Sometimes, the raw data files might not be suitable to go directly into our statistical analysis software. They might contain information we don’t need, they might have errors or duplicate values, some of them might be incomplete, and they might be formatted in a way that our statistical software doesn’t like. To make analysis easier, we can “clean” the data by picking out only the information we need, removing the errors and duplicate values, and screening out the incomplete files. After that, if needed, we can reorganize the data into a format that our statistical software can use. Because we never want to modify our raw data files (remember?), we will save the cleaned and reformatted data in new files.
* *Scripts.* To run our analyses in R, we will create scripts.
* *Outputs.* These are the files that our analysis scripts generate. They might be figures, tables, or any other kind of statistical output we produce with our analysis.

We could just dump all these files into one big folder, but that might not be very helpful to someone who wants to reproduce our analyses. One problem is that it might not be obvious which files are which. Another problem is that the contents of the data files might not be easy to interpret. In most cases, the data will contain multiple columns of numbers or text. Where did those values come from? What do they mean? We need documentation to tell us.

So, to enable reproducibility, it’s best to organize all our artifacts, and to include documents that explain the contents of our data files. The way we choose to organize our files can vary from project to project, depending on our needs and personal preferences, but Figure 1 below shows a sample organizational system. This is based on a protocol recommended by the open-science organization Project TIER (“Tier Protocol”, 2022).

Schematic illustration of the recommended folder structure.

Figure 1 A folder structure for reproducibility. A top-level folder holds the entire project. Inside are subfolders for the original data, the analysis data, the scripts, and the outputs. ReadMe file and data dictionaries provide documentation.

Here, we have a top-level folder to hold our entire project. Inside that, we have subfolders for the original data files, the analysis data, the scripts, and the analysis outputs. We also have three PDF files to help explain our project artifacts. (Note that it’s best not to use spaces in our folder and filenames. In Figure 1, we have eliminated the spaces from our folder and filenames and have capitalized the first letter of each word for readability. This is called *Pascal case*.)

* In the project folder, we have a *ReadMe* document that gives a brief, general description of our project, tells the reader what software we used for data analysis, and gives instructions for how to reproduce our analyses.
* An O*riginalDataDictionary* document that explains the files in our *OriginalData* folder. The exact contents of the O*riginalDataDictionary* will depend on our research project, but generally, the file should include:
  + an explanation of where the data came from (Did we collect them yourself? Did we download them?) and the date we got them;
  + a list of all the variables in data files, the definition of each variable, and an explanation of how the values are coded (i.e., what the values represent).
* An *AnalysisDataDictionary* document that explains the files in our *AnalysisData* folder. Usually—but not always--the analysis data are formatted in a table where each row represents an individual participant, and each column represents a variable. At the top is a row of variable names. The *AnalysisDataDictionary* should explain what each row represents and what each variable name means. Just like the O*riginalDataDictionary*, it should also provide a definition of each variable, and an explanation of how the values are coded.

Once you have created this folder structure, we can zip it up to share to a public repository. Or if we don’t plan to share the artifacts publicly, we can just store the project folder and its contents someplace where we can find them when needed.

## References

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, *73*(2), 146–156. <https://doi.org/10.1037/amp0000258>

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Henry, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(180448). <http://dx.doi.org/10.1098/rsos.180448>

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 104332. <https://doi.org/10.1016/j.jml.2022.104332>

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. <https://doi.org/10.1177/2515245917747656>

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. <https://doi.org/10.1073/pnas.1708290115>

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. <https://doi.org/10.1038/s41597-021-00981-0>

*TIER Protocol | Project TIER | Teaching Integrity in Empirical Research*. (2022, July 12). <https://www.projecttier.org/tier-protocol/>

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*(1), 3. <https://doi.org/10.1525/collabra.13>

# Exercise: Organizing and Documenting Data for Reproducibility with JASP

One of our big goals in this class is *reproducibility*. When we say an analysis is reproducible, we mean that another researcher can find our data, repeat the analyses exactly, and produce the same results.

The way to make our analyses reproducible is to organize and document our files carefully. The purpose of this assignment is to help you practice doing this. In this exercise, you will:

1. Learn the basics about data formatting.
2. Create a set of folders to hold your data and analyses.
3. Download a file of raw original data and create a data dictionary for it.
4. Create the analysis data.
5. Create the analysis data dictionary.
6. Zip up all your files/folders to submit them.

## 1. Understand data structure

### Start from the Very Beginning

For analyses to be reproducible, the data need to be organized and coded in an easy-to-understand way. Although most modern research is conducted electronically, there may still be instances in which data are collected on paper and then need to be put into a file that can be analyzed. Here, you are going to practice coding survey responses into an electronic spreadsheet. Going through this process one step at a time will help you understand how data should be coded and organized for easy and reproducible analysis.

### Data Preparation (from [Codebook Cookbook](http://www.medicine.mcgill.ca/epidemiology/joseph/pbelisle/CodebookCookbook.html) and [Research Methods in Psychology](https://opentextbc.ca/researchmethods/))

There is much freedom in the way data can be entered. A few rules, however, should be followed, to make both the data entry and subsequent data analysis as smooth as possible. The most common format is for **each row to represent a participant** and for **each column to represent a variable** (with the variable name at the top of each column). A sample data file is shown in the table below. The first column contains participant identification numbers. This is followed by columns containing demographic information (age), independent variables (mood, four ratings of self-esteem, and the total of the four self-esteem items), and finally two dependent variables (intentions and attitudes).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **age** | **mood** | **se1** | **se2** | **se3** | **se4** | **total** | **int** | **att** |
| 1 | 20 | 1 | 2 | 3 | 2 | 3 | 10 | 6 | 5 |
| 2 | 22 | 1 | 1 | 0 | 2 | 1 | 4 | 4 | 4 |
| 3 | 19 | 0 | 2 | 2 | 2 | 2 | 8 | 2 | 3 |
| 4 | 24 | 0 | 3 | 3 | 2 | 3 | 11 | 5 | 6 |

### Variable Names

A unique, unambiguous name should be given to each variable. Variables names MUST consist of a single string that beings with a letter, and they should only include letters, numbers, and underscores (\_). Spaces and symbols are not allowed in variables names in most statistical programs, even if data entry programs like Excel do allow them. In this exercise, we will use a convention called Pascal case to name variables. In Pascal case, all spaces are eliminated, and the first letter of each word in the variable name is capitalized for readability.

It is good practice to enter variables names on top of each column. Variable names should be long enough to be meaningful, but short enough to be easy to type. A more detailed description of what the variable names mean is documented separately, in a document called the “data dictionary” (other names that are sometimes used for this kind of document are “metadata” and “codebook”). We’ll talk more about data dictionaries below.

### Variable Codes

Each nominal (or sometimes called categorical) variable has a set of exhaustive, mutually exclusive values. Although these variables aren’t actually numeric, you can usually code the values of a nominal variable with numbers. In the table above, for example, we have coded the variable mood using 0 for negative mood and 1 for positive mood. Nominal variables can also be entered as text labels, if you prefer. For instance, we could have coded mood here using “P” and “N” for positive and negative moods, respectively.

In some cases, you might also have a coding scheme for an ordinal variable. For example, imagine a scale with three options: *Disagree*, *Neither agree/disagree*, *Agree*. In this case, it would be natural to code the three options, respectively, with the values 1–3.

## 2. Create a set of folders

The assignment will require you to create a folder that contains four sub-folders. The top folder should be given a name that represents the assignment (e.g., *ReproducibilityAssignment*). The four subfolders should be named as follows:

1. *OriginalData*
2. *AnalysisData*
3. *JaspFiles*
4. *Outputs*



Throughout the rest of this assignment, you will be creating files to go in the subfolders. (The *JaspFiles* subfolder will remain empty. JASP is the statistical software we will be using for other assignments and activities in this class.)

## 3. Download the original data and create data dictionary

### About the Study

Researchers were interested in whether different kinds of music, listened to in differently colored rooms, would influence participants’ relaxation levels. Forty participants were randomly assigned to listen to either classical or heavy metal music for 15 minutes while seated in a room with either blue or red walls. After the 15 minutes, participants were asked to rate their level of relaxation on the following 1-7 scale:

1. Not at all relaxed
2. More anxious than relaxed
3. Somewhat anxious
4. Neutral
5. Somewhat relaxed
6. More relaxed than anxious
7. Very relaxed

The researchers predicted that listening to classical music would be more relaxing than listening to heavy metal music, and that this effect would be even bigger for participants in a blue room.

### About the Data

For every participant, the researchers recorded a participant number, their age, their year in college, the color of the room they were in, the genre of music they listened to, and their mood score. You’ll need to download and inspect these data. You can download the data from a repository we have created on the Open Science Framework (osf.io) website. The page holding the data is [here ](https://osf.io/4qkc7/?view_only=eec8506bf0084236afc488b4380f8761). The data are stored in a PDF. After you download the PDF, move it to the *OriginalData* folder that you created in Step 2.

### Create a data dictionary for the original data

The next step is to create a data dictionary file to accompany the original data file. The data dictionary file provides information like variable definitions and coding, sampling methods, and anything else a user would need to know to work with and interpret the data appropriately. In an electronic document (Word, Google Docs, Pages, etc.), provide the following information:

* The source from which you obtained the data. In this case, that would be URL from which the data were downloaded.
* The date on which you downloaded the original data file.
* A brief (1-2 sentence) summary of the research that produced the data.
* The following information for each variable:
  + name;
  + a definition/description, including units and level of measurement;
  + an explanation of the coding scheme, if it is not obvious.

Save/export your data dictionary file as a PDF with the name *OriginalDataDictionary*. Save this PDF file to the *OriginalData* subfolder you created in Step 2. There will now be two items in this folder: the original data (the PDF file) and the data dictionary.

## Create the analysis data by transferring the original data into a spreadsheet

Most software programs require that the data be organized inside a table or spreadsheet. The most compatible file type across different programs is a Comma Separated Values file (or .csv). These can be created and saved in software like Excel or in an online app like Google Sheets.

### Familiarize Yourself with the Data

Open the PDF of the *OriginalData*. Each participant’s data is separated from the next by a line. Ask yourself – how many variables are there for each participant? This is the number of columns you will have in your dataset. Which variables are nominal and will need to be coded? What codes will you use?

### Put the Data into a Spreadsheet

1. Open a new spreadsheet in your preferred spreadsheet software (e.g., Excel, Google Sheets, Numbers).
2. In your spreadsheet file, create a column for each variable with a short, easy-to-understand variable name in the top row. Remember the best practices for naming and coding variables described above.
3. Under the top row, every row of data will represent one participant. Carefully type in the data for each variable, one participant at a time. In other words, for Participant 1, you would enter 1 in the first column, 22 in the second column, 4 in the third column, your code for “blue room” in the fourth column, your code for “classical music” in the fifth column, and 5 in the last column.
4. Follow these same steps for the rest of the participants.

### Saving the Dataset

Once you have entered all the data, double-check that you’ve entered them correctly! This step can seem tedious or silly but is critical to ensuring the integrity of your data. After you’ve double (or triple!)-checked the data, save/export the dataset as a Comma Separated Values (.csv) file. You may receive a warning that saving a spreadsheet as this type of file will result in loss of information – ignore this message 😊 Save the dataset as Lab1.csv in the *AnalysisData* subfolder you created in Step 2.

## 5. Create a data dictionary for your analysis data

The next step is to create a dictionary for your analysis data. In an electronic document (Word, Google Docs, Pages, etc.), create a list of variables included in your analysis data csv. The variables you list should correspond to the names in the top row of your data. For each one, provide any information that another person might need to understand what the data mean. This should include, for each variable:

* a definition/description, including units and level of measurement;
* an explanation of the coding scheme, if it is not obvious.

You’ll notice that some of the information you are providing in the analysis data dictionary is the same as you provided in the original data dictionary. That’s fine. The two data sets are for different purposes, and a person who reads one data dictionary might not read the other one.

Save/export your file as a PDF file. Save this PDF file with the name *AnalysisDataDictionary* to the *AnalysisData* subfolder you created in Step 2. There will now be two items in this folder: the analysis data (the CSV file) and the data dictionary.

## 6. Zip your folders & submit the assignment

DETAILS OF SUBMISSION ARE TBD BY THE INSTRUCTOR.

# Exercise: Organizing and Documenting Data for Reproducibility with R

One of our big goals in this class is *reproducibility*. When we say an analysis is reproducible, we mean that another researcher can find our data, repeat the analyses exactly, and produce the same results.

The way to make our analyses reproducible is to organize and document our files carefully. The purpose of this assignment is to help you practice doing this. In this exercise, you will:

1. Learn the basics about data formatting.
2. Create a set of folders to hold your data and analyses.
3. Download a file of raw original data and create a data dictionary for it.
4. Create the analysis data.
5. Create the analysis data dictionary.
6. Zip up all your files/folders to submit them.

## 1. Understand data structure

### Start from the Very Beginning

For analyses to be reproducible, the data need to be organized and coded in an easy-to-understand way. Although most modern research is conducted electronically, there may still be instances in which data are collected on paper and then need to be put into a file that can be analyzed. Here, you are going to practice coding survey responses into an electronic spreadsheet. Going through this process one step at a time will help you understand how data should be coded and organized for easy and reproducible analysis.

### Data Preparation (from [Codebook Cookbook](http://www.medicine.mcgill.ca/epidemiology/joseph/pbelisle/CodebookCookbook.html) and [Research Methods in Psychology](https://opentextbc.ca/researchmethods/))

There is much freedom in the way data can be entered. A few rules, however, should be followed, to make both the data entry and subsequent data analysis as smooth as possible. The most common format is for **each row to represent a participant** and for **each column to represent a variable** (with the variable name at the top of each column). A sample data file is shown in the table below. The first column contains participant identification numbers. This is followed by columns containing demographic information (age), independent variables (mood, four ratings of self-esteem, and the total of the four self-esteem items), and finally two dependent variables (intentions and attitudes).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **age** | **mood** | **se1** | **se2** | **se3** | **se4** | **total** | **int** | **att** |
| 1 | 20 | 1 | 2 | 3 | 2 | 3 | 10 | 6 | 5 |
| 2 | 22 | 1 | 1 | 0 | 2 | 1 | 4 | 4 | 4 |
| 3 | 19 | 0 | 2 | 2 | 2 | 2 | 8 | 2 | 3 |
| 4 | 24 | 0 | 3 | 3 | 2 | 3 | 11 | 5 | 6 |

#### Variable Names

A unique, unambiguous name should be given to each variable. Variables names MUST consist of a single string that beings with a letter, and they should only include letters, numbers, and underscores (\_). Spaces and symbols are not allowed in variables names in most statistical programs, even if data entry programs like Excel do allow them. In this exercise, we will use a convention called Pascal case to name variables. In Pascal case, all spaces are eliminated, and the first letter of each word in the variable name is capitalized for readability.

It is good practice to enter variables names on top of each column. Variable names should be long enough to be meaningful, but short enough to be easy to type. A more detailed description of what the variable names mean is documented separately, in a document called the “data dictionary” (other names that are sometimes used for this kind of document are “metadata” and “codebook”). We’ll talk more about data dictionaries below.

#### Variable Codes

Each nominal (or sometimes called categorical) variable has a set of exhaustive, mutually exclusive values. Although these variables aren’t actually numeric, you can usually code the values of a nominal variable with numbers. In the table above, for example, we have coded the variable mood using 0 for negative mood and 1 for positive mood. Nominal variables can also be entered as text labels, if you prefer. For instance, we could have coded mood here using “P” and “N” for positive and negative moods, respectively.

In some cases, you might also have a coding scheme for an ordinal variable. For example, imagine a scale with three options: *Disagree*, *Neither agree/disagree*, *Agree*. In this case, it would be natural to code the three options, respectively, with the values 1–3.

## 2. Create a set of folders

The assignment will require you to create a folder that contains four sub-folders. The top folder should be given a name that represents the assignment (e.g., *ReproducibilityAssignment*). The four subfolders should be named as follows:

1. *OriginalData*
2. *AnalysisData*
3. *Scripts*
4. *Outputs*

Schematic illustration of the recommended folder structure.

Throughout the rest of this assignment, you will be creating files to go in the subfolders. (The *Scripts* subfolder will remain empty. This is where we would normally store the R scripts we use for analysis.)

## 3. Download the original data and create data dictionary

### About the Study

Researchers were interested in whether different kinds of music, listened to in differently colored rooms, would influence participants’ relaxation levels. Forty participants were randomly assigned to listen to either classical or heavy metal music for 15 minutes while seated in a room with either blue or red walls. After the 15 minutes, participants were asked to rate their level of relaxation on the following 1-7 scale:

1. Not at all relaxed
2. More anxious than relaxed
3. Somewhat anxious
4. Neutral
5. Somewhat relaxed
6. More relaxed than anxious
7. Very relaxed

The researchers predicted that listening to classical music would be more relaxing than listening to heavy metal music, and that this effect would be even bigger for participants in a blue room.

### About the Data

For every participant, the researchers recorded a participant number, their age, their year in college, the color of the room they were in, the genre of music they listened to, and their mood score. You’ll need to download and inspect these data. You can download the data from a repository we have created on the Open Science Framework (osf.io) website. The page holding the data is [here ](https://osf.io/4qkc7/?view_only=eec8506bf0084236afc488b4380f8761). The data are stored in a PDF. After you download the PDF, move it to the *OriginalData* folder that you created in Step 2.

### Create a data dictionary for the original data

The next step is to create a data dictionary file to accompany the original data file. The data dictionary file provides information like variable definitions and coding, sampling methods, and anything else a user would need to know to work with and interpret the data appropriately. In an electronic document (Word, Google Docs, Pages, etc.), provide the following information:

* The source from which you obtained the data. In this case, that would be URL from which the data were downloaded.
* The date on which you downloaded the original data file.
* A brief (1-2 sentence) summary of the research that produced the data.
* The following information for each variable:
  + name;
  + a definition/description, including units and level of measurement;
  + an explanation of the coding scheme, if it is not obvious.

Save/export your data dictionary file as a PDF with the name *OriginalDataDictionary*. Save this PDF file to the *OriginalData* subfolder you created in Step 2. There will now be two items in this folder: the original data (the PDF file) and the data dictionary.

## 4. Create the analysis data by transferring the original data into a spreadsheet

Most software programs require that the data be organized inside a table or spreadsheet. The most compatible file type across different programs is a Comma Separated Values file (or .csv). These can be created and saved in software like Excel or in an online app like Google Sheets.

### Familiarize Yourself with the Data

Open the PDF of the *OriginalData*. Each participant’s data is separated from the next by a line. Ask yourself – how many variables are there for each participant? This is the number of columns you will have in your dataset. Which variables are nominal and will need to be coded? What codes will you use?

### Put the Data into a Spreadsheet

1. Open a new spreadsheet in your preferred spreadsheet software (e.g., Excel, Google Sheets, Numbers).
2. In your spreadsheet file, create a column for each variable with a short, easy-to-understand variable name in the top row. Remember the best practices for naming and coding variables described above.
3. Under the top row, every row of data will represent one participant. Carefully type in the data for each variable, one participant at a time. In other words, for Participant 1, you would enter 1 in the first column, 22 in the second column, 4 in the third column, your code for “blue room” in the fourth column, your code for “classical music” in the fifth column, and 5 in the last column.
4. Follow these same steps for the rest of the participants.

### Saving the Dataset

Once you have entered all the data, double-check that you’ve entered them correctly! This step can seem tedious or silly, but is critical to ensuring the integrity of your data. After you’ve double (or triple!)-checked the data, save/export the dataset as a Comma Separated Values (.csv) file. You may receive a warning that saving a spreadsheet as this type of file will result in loss of information – ignore this message 😊 Save the dataset as Lab1.csv in the *AnalysisData* subfolder you created in Step 2.

## 5. Create a data dictionary for your analysis data

The next step is to create a dictionary for your analysis data. In an electronic document (Word, Google Docs, Pages, etc.), create a list of variables included in your analysis data csv. The variables you list should correspond to the names in the top row of your data. For each one, provide any information that another person might need to understand what the data mean. This should include, for each variable:

* a definition/description, including units and level of measurement;
* an explanation of the coding scheme, if it is not obvious.

You’ll notice that some of the information you are providing in the analysis data dictionary is the same as you provided in the original data dictionary. That’s fine. The two data sets are for different purposes, and a person who reads one data dictionary might not read the other one.

Save/export your file as a PDF file. Save this PDF file with the name *AnalysisDataDictionary* to the *AnalysisData* subfolder you created in Step 2. There will now be two items in this folder: the analysis data (the CSV file) and the data dictionary.

## 6. Zip your folders & submit the assignment

DETAILS OF SUBMISSION ARE TBD BY THE INSTRUCTOR.