A practical guide for transparency in psychological science

Olivier Klein, Université Libre de Bruxelles, Belgium

Tom E. Hardwicke, Stanford University, USA

Frederik Aust, University of Cologne, Germany

Johannes Breuer, GESIS Leibniz Institute for the Social Sciences, Germany

Henrik Danielsson, Linköping University, Sweden and The Swedish Institute for

Disability Research, Sweden

Alicia Hofelich Mohr, University of Minnesota, USA

Hans IJzerman, Université Grenoble Alpes, France

Gustav Nilsonne, Karolinska Institutet and Stockholm University, Sweden, and

Stanford University, USA

Wolf Vanpaemel, University of Leuven, Belgium

Michael C. Frank, Stanford University, USA

Author note: Olivier Klein, Center for Social and Cultural Psychology, Brussels,

Belgium. Tom E. Hardwicke, Meta-Research Innovation Center at Stanford

(METRICS), Stanford University, Stanford, California, USA; Frederik Aust,

Department of Psychology, University of Cologne, Cologne, Germany; Alicia Hofelich

Mohr, Liberal Arts Technologies and Innovation Services, College of Liberal Arts,

University of Minnesota, Minneapolis, Minnesota, USA; Hans IJzerman, LIP/PC2S,

Université Grenoble Alpes, Grenoble, Isère, France; Gustav Nilsonne, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm Sweden, Stress Research Institute, Stockholm University, Stockholm, Sweden, and Department of Psychology, Stanford University, Stanford, USA; Michael C. Frank, Department of Psychology, Stanford University, Stanford, California, USA; Johannes Breuer, Data Archive for the Social Sciences, GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany;

Abstract

The credibility of scientific claims depends upon the transparency of the research products upon which they are based (e.g., study protocols, data, materials, and analysis scripts). As psychology navigates a period of unprecedented introspection, user-friendly tools and services that support open science have flourished. However, the plethora of decisions and choices involved can be bewildering. Here we provide a practical guide to help researchers navigate the process of preparing and sharing the products of their research (e.g., choosing a repository, preparing their research products for sharing, structuring folders, etc.). Being an open scientist means adopting a few straightforward research management practices, which lead to less error prone, reproducible research workflows. Further, this adoption can be piecemeal – each incremental step towards complete transparency adds positive value. Transparent research practices not only improve the efficiency of individual researchers, they enhance the credibility of the knowledge generated by the scientific community.

… until recently I was an open-data hypocrite. Although I was committed to open data, I was not implementing it in practice. ... Some of it was a lack of effort. It was a pain to document the data; it was a pain to format the data; it was a pain to contact the library personnel; it was a pain to figure out which data were indeed published as part of which experiments. Some of it was forgetfulness. I had neither a routine nor any daily incentive to archive data. (Rouder, 2016, p. 1063)

Introduction

Science is a cumulative and self-corrective enterprise; over time the veracity of the scientific literature should gradually increase as falsehoods are refuted and credible claims are preserved (Merton, 1973; Popper, 1963). These processes can optimally occur when the scientific community is able to access and examine the key products of research (materials, data, analyses, and protocols), enabling a tradition where results can be truly cumulative (Ioannidis, 2012). Recently, there has been growing concern that self-correction in psychological science (and scientific disciplines more broadly) has not been operating as effectively as assumed, and a substantial proportion of the literature may therefore consist of false or misleading evidence (Ioannidis, 2005; Johnson, Payne, Wang, Asher, and Mandal, 2016; Klein et al., 2014; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011; Swiatkowski & Dompnier, 2017). Many solutions have been proposed; we focus here on the adoption of transparent research practices as an essential way to improve the credibility and cumulativity of psychological science.

There has never been an easier time to embrace transparent research practices. A growing number of journals, including *Science*, *Nature*, and

4

Psychological Science, have indicated a preference for transparent research practices by adopting the Transparency and Openness Promotion guidelines (Nosek et al., 2015). Similarly, a number of major funders have begun to mandate open practices such as data sharing (Houtkoop et al., 2018). But how should individuals and labs make the move to transparency?

The level of effort and technical knowledge required for transparent practices is rapidly decreasing with the exponential growth of tools and services tailored to supporting open science (Spellman, 2015). While a greater diversity of tools is advantageous, researchers are also faced with a paradox of choice. The goal of this paper is thus to provide a practical guide to help researchers navigate the process of preparing and sharing the products of research, including materials, data, analysis scripts, and study protocols. In the supplementary material, readers can find concrete procedures and resources for integrating the principles we outline in their own research[1]. Our view is that being an open scientist means adopting a few straightforward research management practices, which lead to less error prone, reproducible research workflows. Further, this adoption can be piecemeal – each incremental step towards complete transparency adds positive value. These steps not only improve the efficiency of individual researchers, they enhance the credibility of the knowledge generated by the scientific community.

---

[1] One of the major burdens facing scientists is keeping up with the evolution in standards and resources. That's why the SM will be updated regularly and collaboratively. This "live" version (available at http://open-science-guide.uni-koeln.de) will therefore differ from the publisher's version.

Why Share?

Science is based on verifiability, rather than trust. Imagine an empirical paper with a Results section that claimed that "statistical analyses, not reported for reasons of brevity, supported our findings (details are available upon request)". Such opaque reporting would be unacceptable, because readers lack essential information to assess or reproduce the findings, namely the analysis methods and their results. Although publication norms for print journals previously supported only sharing verbal descriptions, rather than a broader array of research products, this same logic applies equally.

When study data and analysis scripts are openly available, a study's *analytic reproducibility* can be established by re-running the reported statistical analyses, facilitating the detection and correction of any unintended errors in the analysis pipeline (Hardwicke et al., 2018; Peng, 2006; Stodden, 2015; Stodden, Seiler, & Ma, 2018; see supplementary material [SM]: Promoting analytic reproducibility). Once analytic reproducibility has been established, researchers can examine the *analytic robustness* of the reported findings, by employing alternative analysis specifications (Silberzahn et al., 2017; Simonsohn, Simmons, & Nelson, 2015; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), highlighting how conclusions depend on particular choices in data processing and analysis. When stimuli and other research materials are openly available, researchers can conduct replication studies where new data are collected and analyzed using the same procedures to assess the *replicability* of the finding (Simons, 2014). And once a finding has been shown to be replicable, researchers can investigate its *generalisability*: how it varies across different contexts and methodologies (Brandt et al., 2014).

Transparency also enhances trust in the validity of statistical inference. Across statistical frameworks, conducting multiple tests and then selectively reporting only a subset may lead to improper and ungeneralizable conclusions (Goodman et al., 2016; Wasserstein & Lazar, 2016). Even if only a single analysis is conducted, selecting it based on post-hoc examination of the data can undermine the validity of inferences (Gelman & Loken, 2014). Transparency regarding analytic planning is thus critical for assessing the status of a particular statistical test on the continuum between exploratory and confirmatory analysis (De Groot, 2014; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Such transparency can be readily achieved by publicly documenting one's hypotheses, research design, and analysis plan before conducting a study in a process called *pre-registration* (De Angelis et al. 2004; Nosek, Ebersole, DeHaven, & Mellor, 2018; see SM: Preregistration).

Besides increasing the credibility of scientific findings, transparency also boosts the efficiency of scientific discovery. When information is not shared, the value of a study is limited because its products cannot be reused. In contrast, when research products are shared, subsequent researchers can avoid duplication of effort in data collection and decrease the expense involved in creating stimulus materials and analytic code. Further, sharing research products allows researchers to explore related hypotheses and can inspire new research questions. Shared research products can also be important models for researchers, especially trainees, to use in the development of their own materials and analyses. And, in the case of publicly funded research it should also be considered an ethical impetus to make the results of this publicly funded work available to the public.

Finally, there are practical reasons to embrace transparency. First, public sharing is probably the best protection against data loss, since – as we will discuss – best practices require sharing in durable repositories. Second, open research practices increase visibility and facilitate access to unique opportunities for collaboration, jobs, and funding (McKiernan et al., 2016). Third, data sharing has been associated with a citation benefit (Piwowar, & Vision, 2013). Fourth, and perhaps most importantly: in our own anecdotal experience (cf. Lowndes et al., 2017), a research workflow designed at its core to be shared with others is far more efficient and sustainable to use oneself. Accessing an old project to find data, code, or materials need not trigger nightmares. A useful saying to keep in mind is that "your closest collaborator is you six months ago, but you don't reply to emails" (Broman, 2016). Research is a finite enterprise for everyone: collaborators leave projects, change jobs, and even die. If work is not shared, it is often lost.

What to Share?

In this section, we review the different parts of the scientific process that can be shared. Our primary recommendations are:

1. Make transparency a default: If possible, share *all* products of the research process for which there are no negative constraints (due to e.g., funder, IRB/ethics, copyright, or other contract requirements). While attributes of the data, such as disclosure risk, sensitivity, or size, may limit sharing, there are many options for granting partial and restricted access to the data and associated materials.

2. If negative constraints prohibit transparency, explicitly declare and justify these decisions in the manuscript (Morey et al., 2016).

3. Any shared material incrementally advances the goals of increasing verifiability and reuse. Authors need not wait to resolve uncertainty about sharing all products before beginning the process: Bearing in mind any negative constraints (e.g., privacy of participants), *any* product that is shared is a positive step.

Navigating this space can be difficult (see Figure 1). For this reason, we recommend that lab groups discuss and develop a set of "Standard Operating Procedures" (SOP) to guide the adoption of transparent research practices in a manner that is well-calibrated to their own unique circumstances.[2] One part of that organization scheme is a consistent set of naming conventions and consistent project structure (see SM: Folder structure); an example of project created in accordance with our recommendations is available on The Open Science Framework (http://doig.org/10.17605/OSF.IO/XF6UG). Below, we review each of the different products that can be shared.

Study Protocol. A study protocol consists of a detailed written specification of hypotheses, methods, and analysis. For relatively straightforward studies, it may be reasonable to include all of this information in the main body of the primary report of the study. Alternatively, you may wish to share a separate protocol document and provide a higher-level verbal summary in the main report. For certain experimental procedures it may also be beneficial to include instructive video footage. One way to view the study protocol is as a verbal layer that collates, describes, and organises more specific research products, such as materials, software, and analysis code, and informs the reader how they were implemented during your study. Either way,

---

[2] For example, see the SOP for the Nosek group (https://osf.io/mv8pj/) and the Green group (https://github.com/acoppock/Green-Lab-SOP)

the level of detail should be sufficient to allow others to replicate your work without direct instruction from you.

Figure 1: Decision flowchart outlining important considerations when when sharing research products

1. Scope your sharing

Can you share?

Check with…

1. Funders
2. Ethics Review Board (IRB)
3. Contracts, Data Use Agreements, or Policies
4. Collaborators
5. Journal

What to share?

Study Protocol
Materials
Raw and/or analyzed data
Data Documentation
Analysis scripts
Research Reports

When to share?

Before data collection
As data are collected
When submitting a paper
When the paper is published
At the end of a project
After an embargo period

2. Assess your research products: Are your files...

Really big? *May need to find specialized repository or share compressed/aggregated version*

Sensitive? *Remove sensitive variables/information or share through restricted methods*

In a proprietary format? *If possible, also share a copy in a free/open format (e.g., csv/dat/txt)*

Identifying? *Remove or recode identifying variables or share through restricted methods*

Qualitative? *Redact/recode identifying text or share through restricted methods*

Clearly documented? *Ensure that an independent researcher, or future you, can make sense of your files*

3. Decide how to share

Any requirements?

Of funder or journal *Check requirements or recommendations*

Of your institution *Check sharing or ownership policies*

Of the data *Specializations for large, sensitive, or identifying data (see above)*

What are the options?

How open *Public or restricted?*

Where to share *With the paper or in a third-party repository?*

Who mediates *Who is responsible for access or file maintenance?*

How preserved *Who is responsible for long-term preservation?*

How discoverable *Is sufficient meta-data provided so files be found through the article, a website, data repository, or search engine?*

Materials. What constitutes materials differs widely from application to application, even within psychology. In simpler studies, the material may be a list of questionnaire items or stimuli presented to participants manually (videos, images, sounds, etc.). In other studies, materials may include elaborate video stimuli, (video-taped) procedures for an interaction with a confederate or participants (Grahe, Brandt, & IJzerman, 2015), or computer code to present stimuli and collect responses. For clinical studies, materials may include case report forms and

materials for informed consent. Sharing these materials is valuable for both interpretation of research results and for future investigators. A detailed examination of stimulus materials can lead to insights about a particular phenomenon or paradigm by readers or reviewers. In addition, since these materials are often costly and difficult to produce, lack of sharing will be a barrier for the replication and extension of findings.

Data and Metadata. Sharing data is a critical part of transparency and openness, but investigators must make decisions regarding what data to share. "Raw data" are the data as originally recorded, whether by software, an experimenter, a video camera, or other instrument (Ellis & Leek, 2017). Sharing such data can raise privacy concerns due to the presence of identifying or personal information. In some cases, anonymization may be possible, while in others (e.g., video data), anonymity may be impossible to preserve and permission for sharing may not be granted. Regardless of the privacy concerns surrounding raw data, it should almost always be possible to share anonymized data in tabular format as they are used in statistical analyses (see SM: Anonymization). Such data should typically be shared in an easily readable format that does not rely on proprietary software (e.g., comma-separated values, or CSV). Ideally, the script for generating these processed data from the raw data should be made available as well to ensure full transparency (see SM: Automate or thoroughly document all analyses).

One critical ingredient of data sharing is often overlooked: the need for *metadata*. Metadata is a term describing documentation that accompanies and explains a dataset (see SM: Data documentation). In psychology, metadata typically include information on who collected the data, how and when they were collected, the number of variables and cases in each data file, and dataset-level information

such as verbose variable and value labels. Although they can also be shared in standardized, highly structured, machine-readable formats, often metadata are simply a separate document (called a "codebook" or "data dictionary"; see SM: data documentation) that gives verbal descriptions of variables in the dataset. Researchers do not need to be experts in metadata standards: Knowing the structure of metadata formats is less important than making sure the information is recorded and shared. Machine-readable structure can always be added to documentation by metadata experts after the information is shared.

Analysis Procedure. To ensure full transparency and reproducibility of research findings it is critical to share detailed documentation of how the analytic results reported in a research project were obtained (see SM: Analytic reproducibility). Researchers analyze their data in many different ways, and so the precise object(s) to be shared will vary. Nevertheless, the aim is to provide an exact specification of how to move from raw data to final descriptive and statistical analyses, ensuring complete documentation of any cleaning or transformation of data. For some researchers, documenting analyses will mean sharing, for example, R scripts or SPSS syntax; for others it may mean writing a step-by-step description of analyses performed in non-scriptable software programs such as spreadsheets. In all cases, however, the goal is to provide a recipe for reproducing the precise values in the research report.

One challenge for sharing analyses is the rapid pace of change in hardware and software (SM: Avoid "works on my machine" errors). Some researchers may find it discouraging to try and create a fully-reproducible analytic ecosystem with all software dependencies completely specified (e.g., Boettiger, 2015; SM: Sharing software environments). We have several recommendations. First, do not let the
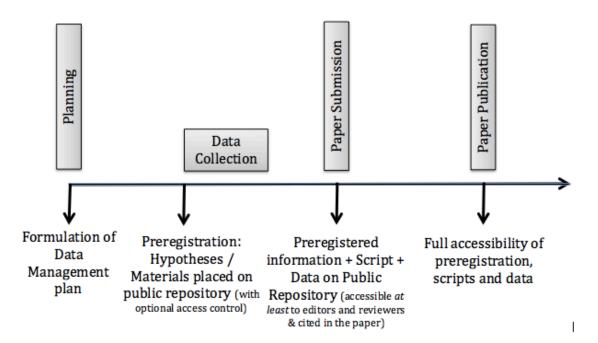
perfect be the enemy of the good. Share and document what you can, as it will provide a benefit compared with not sharing. Second, document the specific versions of the analysis software and packages/add-ons that were used (American Psychological Association, 2010; Eubank, 2016). And finally, when possible, consider using open source software (e.g., R, Python) as accessing and executing code is more likely to be possible in the future compared with commercial packages (Huff, 2017; Morin et al., 2012).

  Research Reports. While we primarily focus on research products beyond the standard written report in this guide, research reports themselves (i.e., published papers) also provide important information about how materials were used, how data were collected, and the myriad other details that are required to understand other products. Making research reports publicly available (through "Open Access") greatly facilitates the use of shared research products. Two main options exist to publish Open Access: Green (posting research online through preprint repositories, like PsyArxiv; https://psyarxiv.com) or Gold (full open access via the publisher, most of which currently still charge Article Processing Costs). For the Green route, preprints do not typically affect the traditional publication process as most journals do not consider them a 'prior publication' (Bourne, Polka, Vale, & Kiley, 2017). A further discussion of Open Access is beyond the scope of this article. However, you can always check a particular journal's stance on open access by typing its name into the SHERPA/ROMEO database (http://www.sherpa.ac.uk/romeo/index.php). This will also tell you whether the journal makes the final article publicly available on its website, and whether this will require you to pay a fee.

When to Share

When it comes to the question of when to share, any time is better than never. However, benefits are maximised when sharing occurs as soon as possible. We consider the possibilities for sharing 1) before data collection, 2) during data collection, 3) when submitting a paper, 4) when the paper is published, 5) at the end of a project, 6) after a specified embargo period. Figure 2 presents a typical workflow.

Figure 2: Typical workflow indicating when to share research products at different stages of the research process.

Planning to share. Sharing research products is easier when you have planned for it in advance. For example, it makes sense to store and structure your data files in a systematic manner throughout the data collection period (i.e., to have a basic "data management plan"). Sharing then only requires a few clicks to upload the files. Many researchers justify not sharing their data because of the time and effort it takes (Borgman, 2012; Houtkoop, Chambers, Macleod, Bishop, Nichols, & Wagenmakers,

2018) – starting early helps avoid this problem. Ideally, researchers should create a data management plan at the beginning of their study (for information on how create one, see, e.g., Jones, 2011).

Before data collection. Sharing key study design and analysis information prior to data collection can confer a number of significant benefits, such as mitigating selective reporting bias or "p-hacking" (Nosek et al., 2018; Simmons et al., 2011). Your study protocol – hypotheses, methods, and analysis plan – can be formally "registered" by creating a time-stamped, read-only copy in a public registry (e.g., The Open Science Framework), such that they can be viewed by the scientific community (a "pre-registration"; see SM: Pre-registration). If you wish, it is possible to pre-register the study protocol, but keep it private under an "embargo" for a specified period of time (e.g., until after your study is published).

While embargoes on preregistrations can mitigate the fear of being scooped, flexibility in the release of pre-registered documents limits transparency. For example, researchers may strategically release only those documents that fit the narrative they wish to convey once the results are in. It is therefore preferable to encourage transparency from the outset. At the very least, the scientific community should be able to check whether a study was preregistered and, preferably, have access to the content of this preregistration, regardless of whether it is communicated in the final paper.

"Registered Reports" (Chambers, 2013; Hardwicke & Ioannidis, 2018; also see https://cos.io/rr/) address this concern by embedding the pre-registration process directly within the publication pipeline. Researchers submit their study protocol to a journal where it undergoes peer-review, and may be offered *in principle acceptance* for publication before the study has even begun. This practice could yield additional

advantages beyond standard pre-registration, such as mitigation of publication bias (because publication decisions are not based on study outcomes), and improved study quality (because authors receive expert feedback before studies begin).

The central purpose of pre-registration is transparency with respect to which aspects of the study were pre-planned (confirmatory) and which were not (exploratory). Viewed from this perspective, pre-registration does not prevent researchers from making changes to their protocol as they go along, or from running exploratory analyses, but simply maintains the exploratory-confirmatory distinction (Wagenmakers et al., 2012). When used appropriately, pre-registration has the potential to reduce bias in hypothesis-testing (confirmatory) aspects of research. This ambition, however, does not preclude opportunities for exploratory research when it is explicitly presented as such (Ioannidis, 2014; Nosek et al., 2018).

During data collection. Study protocol and materials can be readily shared once data collection commences. Rouder (2016) has, additionally, advocated sharing data while they are being collected, a concept he calls "born-open data" (see SM: Born-open data). Born-open-data are automatically uploaded to a public repository, for example, after every day of data collection. Besides the obvious advantages of greater transparency and immediate accessibility, born-open-data can simplify data management (e.g., the published data constitute an off-site backup to a professionally managed data storage). Because of technical and privacy issues, this approach may not be right for every project. However, once the system is set up, sharing data requires minimal effort (other than appropriate maintenance and periodic checking).

Upon paper submission or publication. A more common practice is to share research products when submitting a paper to a journal or when the paper is

published. Of these two possibilities, we recommend sharing on submission. First, editors/reviewers may need access to this information in order to properly evaluate your study. Second, sharing on submission adds value to your paper by demonstrating that your research products will be made available to the scientific community. Finally, sharing on submission allows for errors to be caught before publication, reducing the possibility of later public correction. If 'blind-reviewing' is important and author names are displayed alongside shared research products, some repositories, such as The Open Science Framework, offer a "view-only" link option, that (partially) circumvents this problem.

After an embargo period. Finally, there may be reasons why researchers cannot or do not want to share all research products immediately. It is possible to archive products in an accessible repository right away, and temporarily delay their release by placing them under an embargo.

How to Share?

Once a researcher has decided to share research products, one of the most important decisions to make is where and how to share. Journals and professional societies often recommend that data and other research products be made "available upon request" (Vasilevsky et al., 2017). However, a number of studies suggest that data requests typically do not result in data access (Alsheikh et al., 2011; Dehnhard et al., 2013; Stodden et al., 2018; Wicherts et al., 2006; Vanpaemel et al., 2015; Vines et al., 2014). Similarly, sharing via personal websites is very flexible and increases accessibility and discoverability compared to sharing on request, but is also not a sustainable solution: Research products may become

inaccessible when the personal website is deleted, overhauled, or moved to a different hosting service. Thus we do not recommend either of these options.

Instead, we recommend the use of independent public repositories for sharing research products. When choosing a repository, researchers should consider whether the repository:

1. Uses persistent and unique identifiers for products (such as DOIs).

2. Accommodates structured metadata to maximize discoverability and reuse.

3. Tracks data re-use (e.g., citations, download counts).

4. Accommodates licensing (e.g., provides the ability to place legal restrictions on data reuse or signal there are no restrictions).

5. Features access controls (e.g., allows restriction of access to a particular set of users).

6. Has some persistence guarantees for long-term access.

7. Stores data in accordance with local legislation (e.g., the new General Data Protection Regulation for the EU, http://www.eugdpr.org/).

Within this category, we highlight the Open Science Framework (http://osf.io). This repository satisfies the first six criteria above (the last one being dependent on the exact location of the researchers[3]), is easy to use, and provides for sharing the variety of the products listed above (for a detailed tutorial on using the Open Science Framework to share research products, see Soderberg, 2018). Note that some research communities make use of specialized repositories, for example, brain imaging data (https://openeuro.org/) or video and audio recordings (https://nyu.databrary.org/). Such repositories are more likely to have metadata

---

[3] OSF, which is located in the United States, satisfies US legislation. Besides, the "privacy shield" agreement between the EU and the US (https://ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/eu-us-privacy-shield_en) guarantees strong protection for data that are transferred from the EU to the US. Note, however, that compliance with these legislations depends on the use of proper anonymization procedures by the researchers (see Supplementary Material).

standards and storage capacity calibrated to specific data types. For an overview of other public repositories, see Table 1.

Insert Table 1 about here

We recommend to share on a platform, such as the OSF, that make it possible to attribute a unique and persistent URL (such as a DOI) to the project. Several studies have indicated that regular URLs used by journals to link to supplementary files can often break over time, severing access to research products (Gertler & Bullock, 2017; Evangelou, Trikalinos, & Ioannidis, 2005). Using persistent URLs increases the chances that research products will be accessible for the long term.

Sharing can raise a number of legal and ethical issues, and these vary between countries and between institutions. Handling these is vastly simplified by addressing them ahead of time. For example, consent forms (see SM: Informed consent) can explicitly request participant consent for public data sharing, as it can be hard or impossible to obtain retroactive consent. Additionally, researchers should always clarify any requirements of their institution, granting agency, and intended publication venue. Below we review issues related to privacy and licensing.

Considering participants' privacy can be both an ethical issue and a legal requirement (for example, The United States' Health Insurance Portability and Accountability Act and The European Union's General Data Protection Regulation, see SM: EU Data Protection Guidelines). In short, researchers must take appropriate precautions to protect participant privacy prior to sharing data. Fortunately, many datasets generated during psychological research either do not contain identifying information, or can be anonymized ("de-identified") relatively straightforwardly (see SM: Anonymization). However, some forms of data can be quite difficult to

anonymize (e.g., genetic information, video data, or structural neuroimaging data; Gymrek et al., 2013; Sarwate et al., 2014), and require special considerations beyond the scope of this article. Because it is often possible to identify individuals based on minimal demographic information (e.g., postal code, profession, age; Sweeney, 2000), researchers should consult with their ethics board to find out the appropriate legal standard for anonymization.

One further legal concern for sharing research products is their ownership. Researchers often assume that publicly available research products have been placed in the public domain, that is, that the authors waive all property rights. However, by default, researchers retain full copyright of the products they create. Unless they are published with a permissive license, the products technically cannot be used or redistributed without approval from the authors – despite scientific norms to the contrary. Thus, to reduce uncertainty about copyright, shared products should ideally be licensed using an easy-to-understand standard license such as a Creative Commons (CC; https://creativecommons.org/licenses/) or Open Data Commons (ODC; https://opendatacommons.org/licenses/). In the spirit of openness we recommend to release research products into the public domain by using maximally permissive licenses, such as CC0 and ODC-PDDL, or to condition re-use only on attribution (e.g., CC-BY and ODC-BY). Licensing research products is as easy as including a text file alongside your research products containing a statement such as "All files in this repository are licensed under a Creative Commons Zero License (CC0 1.0)".

So, why not share?

Given all of the arguments we and others have presented, why would researchers still not share their data? Beyond the concerns described above

(privacy, etc.), one commonly heard worry is that researchers will make use of shared resources to gain academic precedence ("scooping"; Houtkoop et al. 2018). In our view, this worry is usually unwarranted. Most subfields of psychology are not so competitive as to be populated with investigators who are looking to race to publish a particular finding. In addition, in many cases the possibility of being scooped is likely outweighed by the benefits of increased exposure, as noted by Gary King[4]: "The thing that matters the least is being scooped. The thing that matters the most is being ignored". Researchers who are truly concerned about being scooped – whether justifiably or not – can simply wait to share their materials, code, and data until after they publish, or release research products under a temporary embargo. Such embargoes slow verification and reuse, but they are far better than not sharing at all.

Another worry is that errors will be revealed by others checking original data, or original conclusions will be challenged by alternative analyses (Houtkoop et al., 2018). Indeed, it seems likely that errors *will* be discovered and conclusions *will* be challenged as widespread adoption of transparent research practices adds fuel to the idling engines of scientific self-correction and quality control, such as replication and peer review. It is understandable that researchers worry about errors being discovered in their own work, but such errors are inevitable - we are after all, only human. A rise in informed critical discourse will be healthy for science and make discovery of such errors normative. We believe that more, rather than less, transparency is the best response. Honesty and transparency are likely to enhance – rather than diminish – one's standing as a scholar (Fetterman & Sassenberg, 2015).

---

[4] https://www.youtube.com/watch?v=jD6CcFxRelY

Researchers may also be concerned that learning and then implementing transparent research practices will be too time-consuming (Houtkoop et al., 2018). In our experience, there is indeed a significant time-cost to learning such practices. Nonetheless, these should not necessarily be embraced and mastered at once. It is often through "baby steps", via trial and error, that the practice of open science can become natural and habitual. It helps to include, for example, "research milestones" in one's workflow. Adding in milestones also contributes to an optimal teaching strategy, with students learning how to engage in open science practices in small steps. Besides, there are major benefits that make this time well spent for the individual researcher. First, transparent research practices are often synonymous with good research management practices, and therefore increase efficiency in the longer term. For example, it is much easier to locate stimuli from an old project or re-use analysis code when it is well-documented and available in a persistent online repository. Second, transparent practices can lead to benefits in terms of citation and reuse of one's work (see SM: Incentivizing Sharing). Finally, transparent research practices inspire confidence in one's own research findings, allowing one to more readily identify fertile avenues for future studies that are truly worth investing resources in.

Conclusion

The field of psychology is engaged in an urgent conversation about the credibility of the extant literature. Numerous research funders, institutions, and scientific journals have endorsed transparent and reproducible research practices

through the TOP guidelines (Nosek et al., 2015)[5] and major psychology journals have begun implementing policy changes that encourage or mandate sharing (see e.g., Kidwell et al., 2016; Nuijten et al., 2017). Meanwhile, the scientific ecosystem is shifting and evolving. A new open science frontier has opened, and flourishes with a plethora of potential tools and services to help researchers adopt transparent research practices.

Here we have sketched out a map to help researchers navigate this exciting new terrain. Like any map, some aspects will become outdated as the landscape evolves over time: Exciting new tools to make research transparency even more user-friendly and efficient are already on the horizon. Nevertheless, many of the core principles will remain the same, and we have aimed to capture them here. Our view is that being an open scientist means adopting a few straightforward research management practices, which lead to less error-prone, reproducible research workflows with each incremental step adding positive value. Doing so will improve the efficiency of individual researchers and it will enhance the credibility of the knowledge generated by the scientific community.

References

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th edition). Washington, DC: American Psychological Association.

---

[5] See also guidelines for specific fields and types of research: the CONSORT statement (randomized controlled trial: http://www.consort-statement.org/), the ARRIVE guidelines (animal research: https://www.nc3rs.org.uk/arrive-guidelines) or the PRISMA statement (meta-analysis: http://prisma-statement.org/). The EQUATOR website (http://www.equator-network.org/) lists the main reporting guidelines.

Boettiger, C. (2015). An Introduction to Docker for Reproducible Research. *ACM SIGOPS Operating Systems Review*, 49(1), 71–79. DOI: https://doi.org/10.1145/2723872.2723882

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the Association for Information Science and Technology*, 63(6), 1059-1078. DOI: https://doi.org/10.1002/asi.22634

Bourne, P. E., Polka, J. K., Vale, R. D., & Kiley, R. (2017). Ten simple rules to consider regarding preprint submission. *PLoS Computational Biology*, 13(5), e1005473–6. http://doi.org/10.1371/journal.pcbi.1005473

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?, *Journal of Experimental Social Psychology, 50*, 217–224. DOI: https://doi.org/10.1016/j.jesp.2013.10.005

Broman, K. (2016). Steps toward reproducible research [Slides]. Retrieved from bit.ly/jsm2016

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. Cortex, 49(3), 609-610. DOI: https://doi.org/10.1016/j.cortex.2012.12.016

De Angelis, C., Drazen, J. M., Frizelle, F. A. P., Haug, C., Hoey, J., Horton, R., … Weyden, M. B. V. D. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine, 351*(12), 1250–1251. DOI: https://doi.org/10.1056/NEJMe048225

De Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta psychologica*, *148*, 188-194. DOI: https://doi.org/10.1016/j.actpsy.2014.02.001

Dehnhard, I., Weichselgartner, E., & Krampen, G. (2013). Researcher's willingness to submit data for data sharing: A case study on a data archive for psychology. *Data Science Journal, 12*, 172-180. DOI: https://doi.org/10.2481/dsj.12-037

Ellis SE, Leek JT. (2017) How to share data for collaboration. PeerJ Preprints 5:e3139v5 DOI: https://doi.org/10.7287/peerj.preprints.3139v5

Evangelou, E., Trikalinos, T. A., & Ioannidis, J. P. A. (2005). Unavailability of online supplementary scientific information from articles published in major journals. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, 19(14), 1943–1944. http://doi.org/10.1096/fj.05-4784lsf

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

Fetterman AK, Sassenberg K (2015). The reputational consequences of failed replications and wrongness admission among scientists. PLoS ONE, 10(12): e0143723. DOI: https://doi.org/10.1371/journal.pone.0143723

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460-465. DOI: https://doi.org/10.1511/2014.111.460

Gertler, A. L., & Bullock, J. G. (2017). Reference Rot: An Emerging Threat to Transparency in Political Science. *PS: Political Science & Politics*, 50(01), 166–171. http://doi.org/10.1017/S1049096516002353

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*, 1–6. DOI: https://doi.org/10.1126/scitranslmed.aaf5027

Grahe, J., Brandt, M. J., IJzerman, H. (2015). The Collaborative Education and Replication Project. Retrieved from http://osf. io/wfc6u. DOI: https://doi.org/10.1371/journal.pone.0038234

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339(6117), 321-324. DOI: https://doi.org/10.1126/science.1229566

Hardwicke, T. E., & Ioannidis, J. P. A. (2018, April 16). Mapping the universe of Registered Reports. Retrieved from http://doi.org/10.17605/OSF.IO/FZPCY

Hardwicke, T. E., Mathur, M. B., MacDonald, K. E., Nilsonne, G., Banks, G. C.,... Frank, M. C. (2018, March 19). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. Retrieved from https://osf.io/preprints/bitss/39cfb/

Heycke, T., Aust, F., & Stahl, C. (2017). Subliminal influence on preferences? A test of evaluative conditioning for brief visual conditioned stimuli using auditory unconditioned stimuli. *Royal Society Open Science, 4*, 160935. doi: http://doi.org/10.1098/rsos.160935

Houtkoop, B., Chambers, C., Macleod, M., Bishop, D., Nichols, T., &
Wagenmakers, E. J. (2018). Data sharing in psychology: A survey on barriers
and preconditions. *Advances in Methods and Practices in Psychological
Science.* Advance online publication. DOI:
https://doi.org/10.1177/2515245917751886

Huff, K. (2017). Lessons Learned. In Kitzes, J., Turek, D., & Deniz, F. (Eds.). *The
Practice of Reproducible Research: Case Studies and Lessons from the
Data-Intensive Sciences*. Oakland, CA: University of California Press.
Retrieved from https://www.gitbook.com/book/bids/the-practice-of-
reproducible-research

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS
Medicine*, *2*(8), e124. DOI: http://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting.
Perspectives on Psychological Science, *7*(6), 645–654. DOI:
http://doi.org/10.1177/1745691612464056

Ioannidis, J. P. (2014). Clinical trials: what a waste. *British Medical Journal,* 349.
DOI: 10.1136/bmj.g7089

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2016). On the
reproducibility of psychological science. *Journal of the American Statistical
Association*, *5*(4), 0–0. DOI: http://doi.org/10.1080/01621459.2016.1240079

Jones, S. (2011). *How to Develop a Data Management and Sharing Plan*.
Retrieved from http://www.dcc.ac.uk/resources/how-guides

Kidwell, M. C. Lazarevic, L. B. Baranski, E. Hardwicke, T. E. Piechowski, S. Falkenberg, L.-S. Nosek, B. A. et al. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. PLoS biology, : 1–15, DOI: https://doi.org/10.1371/journal.pbio.1002456

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A Many Labs replication project. *Social Psychology*, *45*, 142–152. DOI: http://doi.org/10.1027/1864-9335/a000178

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., et al. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, *1*(6), 0160–7. DOI: http://doi.org/10.1038/s41559-017-0160

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., et al. (2016). How open science helps researchers succeed. *eLife*, *5*, 1–19. DOI: http://doi.org/10.7554/eLife.16800

Merton, R. K. (1973). *The Sociology of Science. Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., Lewandowsky, S., Morey, C. C., Newman, D. P., Schönbrodt, F., Vanpaemel, W., Wagenmakers, E.-J., & Zwaan, R. A. (2016). The peer reviewers' openness initiative: Incentivising open research practices through

peer review. *Royal Society Open Science, 3*, 150547, 1-7. DOI:
https://doi.org/10.1098/rsos.150547

Morin, A., Urban, J., Adams, P. D., Foster, I., Sali, A., Baker, D., & Sliz, P.
(2012). Shining Light into Black Boxes. *Science*, 336(6078), 159–160. DOI:
https://doi.org/10.1126/science.1218263

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S.
J., et al. (2015). Promoting an open research culture. *Science*, *348*(6242),
1422–1425. DOI: http://doi.org/10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A., & Mellor, D. (2018). The
preregistration revolution. *PNAS*. Advance online publication. DOI:
https://doi.org/10.1073/pnas.1708274114

Nuijten, M. B., Borghuis, J., Veldkamp, C. L., Dominguez-Alvarez, L., Van Assen,
M. A., & Wicherts, J. M. (2017). Journal Data Sharing Policies and Statistical
Reporting Inconsistencies in Psychology. *Collabra: Psychology,* 3(1). DOI:
http://doi.org/10.1525/collabra.102

Open Science Collaboration (2015). Estimating the reproducibility of
psychological science. *Science*, *349*(6251), 1–8. DOI:
http://doi.org/10.1126/science.aac4716

Peng, R. D. (2006). Reproducible epidemiologic research. *American Journal of
Epidemiology*, *163*(9), 783–789. DOI: http://doi.org/10.1093/aje/kwj093

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation
advantage. *Peer J*, *1*, e175. DOI:10.7717/peerj.175

Popper, K. (1963). Conjectures and refutations: The growth of scientific knowledge. London: Routledge and Kegan Paul.

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior research methods*, *48*(3), 1062-1069. DOI: https://doi.org/10.3758/s13428-015-0630-z

Sarwate, A. D., Plis, S. M., Turner, J. A., Arbabshirani, M. R., & Calhoun, V. D. (2014). Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. *Frontiers in neuroinformatics, 8,* 35. DOI: https://doi.org/10.3389/fninf.2014.00035

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. C., ... & Carlsson, R. (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. DOI: http://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. DOI: http://doi.org/10.1177/1745691613514755

Soderberg, C. K. (2018). Using OSF to Share Data: A Step-by-Step Guide. *Advances in Methods and Practices in Psychological Science*. Advance online publication. DOI: https://doi.org/10.1177/2515245918757689

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. DOI: http://doi.org/10.1177/1745691615609918

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. DOI: http://doi.org/10.1177/1745691616658637

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

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

Sweeney L. (2000). Simple demographics often identify people uniquely. http://impcenter.org/wp-content /uploads/2013/09/Simple-Demographics-Often -Identify-People-Uniquely.pdf.

Świątkowski, W. & Dompnier, B., (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology*. 30(1), pp.111–124. DOI: http://doi.org/10.5334/irsp.66

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*, 1-5. DOI: https://doi.org/10.1525/collabra.13

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: Are journal data sharing policies

meeting the mark?. *PeerJ*, 5, e3208. DOI:

https://doi.org/10.7287/peerj.preprints.2588

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M.

T., ... & Rennison, D. J. (2014). The availability of research data declines

rapidly with article age. *Current Biology, 24*(1), 94-97. DOI:

https://doi.org/10.1016/j.cub.2013.11.014

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., &

Kievit, R. A. (2012). An agenda for purely confirmatory research.

Perspectives on Psychological Science, 7(6), 632–638. DOI:

http://doi.org/10.1177/1745691612463078

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values:

Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

DOI: http://doi.org/10.1080/00031305.2016.1154108

Wicherts, J., Borsboom, D., Kats, J. & Molenaar, D. (2006). The poor availability

of psychological research data for reanalysis. *American Psychologist*, *61*,

726–728. DOI: https://doi.org/10.1037/0003-066x.61.7.726

Contributions

OK initiated the project.

He coordinated it with MF.

All authors contributed to the writing and commented on previous versions.

FA designed the example project.

Table 1

Features of selected public repositories that hold psychological data

	Operator(s)	For-profit/Non-profit	Country /jurisdiction	Focus/specialization	Costs for data sharer	Self-deposit[1]	Private (i.e., non-public) storage/projects possible	Restrictions of access possible (for published/public projects)	Embargo period possible	Content types[2]
Code Ocean	Code Ocean	Non-profit	USA	None	50 GB of storage, 10 hrs/month cloud computing time (with up to 5 concurrent runs) free for academic users	Yes	Only before publication of the project	No	No	Software applications, Source code, Structured graphics, Configuration data
DANS EASY	Netherlands Organisation for Scientific Research (NOW) & Royal Netherlands Academy of Arts and Sciences (KNAW)	Non-profit	Netherlands	None	Free for up to 50GB	Yes	No	Yes	Yes	Scientific and statistical data formats, Standard office documents, Plain text, Images, Audiovisual data, Raw data, Structured text, Structured graphics, Databases
Dryad	Dryad Digital Repositor	Non-profit	USA	Medical and life sciences	$120 for every data package up to 20GB (+	Yes	No	No	Yes	Scientific and statistical data formats,

y					$50 for each additional 10 GB); no costs for data sharers if charges are covered by a journal or the submitter is based in a country that qualifies for a fee waiver					Standard office documents, Plain text, Software applications, Source code, Structured text, other
figshare	Digital Science	For-profit	UK	None	Free for up to 100GB	Yes	Yes (up to 20GB for free accounts)	Yes[3]	Yes[3]	Scientific and statistical data formats, Standard office documents, Plain text, Images, Audiovisual data, Raw data, Archived data, Source code, Structured graphics
GESIS datorium	GESIS - Leibniz Institute for the Social Sciences	Non-profit	Germany	Social sciences	Free	Yes	No	Yes	Yes	Scientific and statistical data formats, Standard office documents, Plain text, Raw data, Structured graphics, other

GESIS standard archiving	GESIS - Leibniz Institute for the Social Sciences	Non-profit	Germany	Social sciences, survey data (esp. from large or repeated cross-sectional or longitudinal studies)	Free	No	No	Yes	Yes	Scientific and statistical data formats, Standard office documents, Plain text, Archived data
Harvard Dataverse	Harvard University, Institute for Quantitative Social Sciences	Non-profit	USA[4]	None	Free	Yes	Yes	Yes	Yes	Scientific and statistical data, formats, Standard office documents, Raw data, Archived data, Software applications, Source code, Databases
Mendeley Data	Elsevier (in cooperation with DANS)	For-profit	Netherlands	None	Free[5]	Yes	Yes	No	Yes	Scientific and statistical data formats, Standard office documents, Plain text, Software applications, Structured text, Configuration data, other
openICPSR	Inter-University Consortium of	Non-profit	USA	Political and social research, social and	Free up to 2GB[6]	Yes	Yes	Yes	Yes	Scientific and statistical data formats, Standard

	Political and Social Science			behavioral sciences						office documents, Plain text, Archived data, Structured text, Structured graphics
Open Science Framework	Center for Open Science	Non-profit	USA	None	Free	Yes	Yes	No	Yes	Scientific and statistical data formats,Standard office documents, Plain text, other
PsychData	ZPID - Leibniz Institute for Psychology Information	Non-profit	Germany	Psychology, data for peer-reviewed publications	Free	No	No	Yes	Yes	Scientific and statistical data formats, Standard office documents, Plain text
UK Data Service standard archiving	UK Data Archive & Economic and Social Research Council (ESRC)	Non-profit	UK	Social research, esp. large-scale surveys, longitudinal, and qualitative studies	Free	No	No	Yes	Yes	Scientific and statistical data formats, Standard office documents, Plain text, Images, Audiovisual data, Raw data, Structured graphics
UK Data	UK Data	Non-	UK	Social	Free	Yes	No	Yes	Yes	Scientific and

Service ReShare	Archive & Economic and Social Research Council (ESRC)	profit		sciences						statistical data, formats, Standard office documents, Plain text, Images, Audiovisual data, Raw data, Structured graphics
Zenodo	European Organization for Nuclear Research & Open Access Infrastructure for Research in Europe (OpenAIRE)	Non-profit	EU	None	Free[7]	Yes	No	Yes	Yes	Scientific and statistical data formats, Standard office documents, Plain text, Images, Audiovisual data, Raw data, Archived data, Source code, Structured text, Structured graphics, Networkbased data, other

[1] If self-deposit (i.e., researchers can directly upload their own materials) is not possible, this means that the repository is curated (or at least more strongly curated than the others). The advantage of these repositories is that they offer additional help and services by professional archiving staff (e.g., in the creation of study- and variable-level documentation or the conversion of files to nonproprietary formats).

[2] We used the content type category from the re3data schema v3.0 here. Descriptions of these types can be found in the documentation which is available at http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1397899:6/component/escidoc:1398549/re3data_schema_documentation_v3_0.pdf

[3] Individual files can be embargoed or made confidential.

[4] Dataverse is a special case in several regards. There is the overall Dataverse Project (https://dataverse.org/), then there are different Dataverse repositories (e.g., the Harvard Dataverse or DataverseNL by the Dutch Data Archiving and Networked Services) which host multiple individual Dataverses (e.g., by individual universities, research

groups or researchers). If the institution a researcher is affiliated with does not have its own Dataverse repository or Dataverse, it is possible to create a Dataverse within the Harvard Dataverse repository. For a more detailed description of Dataverse and its organizational architecture, see King (2007) and Leeper (2014).

[5] The FAQ on the Mendeley Data website states that they may introduce a freemium model in the future "for instance charging for storing and posting data, above a certain dataset size threshold" (see https://data.mendeley.com/faq).

[6] If more storage space or additional services are needed the researchers or their institutions branded OpenICPSR hosting or the "Professional Curation Package" to access all of the ICPSR (curation) services (see https://www.openicpsr.org/openicpsr/pricing).

[7] The Zenodo terms of use state that "content may be uploaded free of charge by those without ready access to an organized data centre".

A practical guide for transparency in psychological science

Supplementary Materials[1]

---

[1] We intend to update this document on (at least) an annual basis as standards and resources evolve. The current version is available here: http://psych-transparency-guide.uni-koeln.de/.

Anonymization

Researchers need to ensure that the privacy of human participants is properly protected in line with national and/or international law. One way to achieve this goal is to anonymize[2] the data, rendering identification of participants nearly impossible. There are two ways in which participants can be identified: 1) through direct identifiers, such as names, addresses or photos, 2) through combinations of indirect identifiers (e.g., date of birth + job title + name of employer). Below we detail ways of minimizing risks, but often the risk of re-identification can never be eliminated completely. Researchers must weigh risks and benefits, bearing in mind that the research participants also have a legitimate interest in the realisation of benefits due to their participation.

First, researchers are advised to consider the legal standards that apply to them in particular. The United States Department of Health and Human Resources has developed a "de-identification standard" (http://bit.ly/2Dxkvfo) to comply with the HIPAA (Health Information Portability and Accountability Act) Privacy rule. Readers may also refer to the guide to de-identification (http://bit.ly/2IxEo9Q) developed by the Australian National Data Services and the accompanying decision tree (http://bit.ly/2FJob3i). Finally, a subsection below deals with new EU data protection laws.

In general, since a relatively limited set of basic demographic information may suffice to identify individual persons (Sweeney, 2000), researchers should try to limit the number of recorded identifiers as much as possible. If the collection of direct or many indirect identifiers is necessary, researchers should consider whether these need to be shared. If directly identifying variables are only recorded for practical or logistic purposes, e.g., to contact participants over the course of a longitudinal study, the identifying variables should simply be deleted from the publicly shared dataset, in which case the data set will be anonymized.

A special case of this situation is the use of participant ID codes to refer to individual participants in an anonymous manner. ID codes should be completely distinct from real names (e.g., do not use initials). Participant codes should also never be based on indirectly identifying information, such as date of birth or postal codes. These ID codes can be matched with identifying information that is stored in a separate and secure, non-shared location.

In the case that indirect identifiers are an important part of the dataset, researchers should carefully consider the risks of re-identification. For some variables it may be advisable or even required to restrict or transform the data. For example, for income information, a simple

---

[2] The terms "anonymize" and "de-identify" are used differently in various privacy laws, but typically refer to the process of minimizing risk of re-identification using current best statistical practices (El Emam, 2013).

step is to restrict the upper and lower range (using top- and/or bottom-coding). Similarly location information such as US zip codes may need to be aggregated so as to provide greater protection (especially in the case of low-population areas in which a city or US zip code might be identifying information in conjunction with a variable like age). To analyze these risks more generally for a dataset, it may be useful to consider the degree to which each participant is unique in the dataset and in the reference population against which it may be compared. The nature of the reference population is usually described by the sampling procedure. For instance, the reference population may consist of students at the university where the research was conducted, or of patients at a hospital clinic where a study was performed, or of the adult population of the town where the research was done. Another potentially useful method is to consider threat models, i.e. how reidentification could be performed by different actors with different motives. Such a thought exercise can help uncover weaknesses in data protection. For example, one threat model is that the participant tries to reidentify themselves. In this case, one needs to consider what potentially identifying variables the participants has access to, and what harm may result from successful reidentification in view of what the participant already knows about themselves. Another threat model could be that a third party tries to identify a specific participant based on publicly available information. In this case, it is necessary to consider what publicly available information, if any, would permit reidentification by matching to the original dataset. Such threat assessments have the purpose of determining the risk of (re-)identification and should be used by researchers (ideally with the help of data archiving specialists from libraries, institutional or public repositories) to choose appropriate technical and/or organizational measures to protect participants' privacy (e.g., by removing or aggregating data or restricting access).

Finally, in case anonymization is impossible, researchers can obtain informed consent for using and sharing non-anonymized data (see below for example templates for consent) or place strict controls on the access to the data.

EU Data Protection Guidelines

Many researchers will be required to follow new EU data protection guidelines. The European Parliament, the Council of the European Union, and the European Commission have implemented the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679), a regulation that aims at strengthening and unifying data protection for all individuals within the European Union (EU). It is effective as of May 25, 2018. This new regulation makes a distinction between pseudonymisation and anonymisation.

Pseudonymisation refers to the processing of personal data in such a way that it can no

longer be associated with a specific data subject unless additional information is provided. It typically involves replacing identifying information with codes[3]. The key must then be kept separately. The GDPR promotes the use of pseudonymisation as a standard data protection practice for scientific research purposes. *Anonymous* data are defined as information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable by any means. This regulation does not concern the processing of such anonymous information, including for statistical or research purposes. More information on this regulation can be found on the European Commission's website (http://bit.ly/2rnv0RA). Chassang (2017) also discusses its implications for scientific research in more detail.

The EU-funded project OpenAIRE (https://www.openaire.eu/) offers the free-to-use data anonymization tool Amnesia "that allows to remove identifying information from data" and "not only removes direct identifiers like names, SSNs etc but also transforms secondary identifiers like birth date and zip code so that individuals cannot be identified in the data" (https://amnesia.openaire.eu/index.html).

Informed consent

When asking study participants for informed consent, it is important to also inform them about the data sharing plans for the study. ICPSR offers some recommendations for informed consent language for data sharing (http://bit.ly/2tWFAQK) and the data management guidelines of the German Psychological Association (http://bit.ly/2ulBgt5) provide an example informed consent in Appendix B. Based on these two resources and the informed consent forms we have used in our own labs, we created two informed consent templates that researchers can use and adapt to their needs: one for when no personal data is being collected (https://osf.io/sfjw9/), and one for when personal data is being collected (https://osf.io/kxbva/). For further recommendations on how to formulate an informed consent that is compatible with open science practices, see Meyer (2018).

Born-open data

The most radical form of data sharing involves publishing data as they are being collected. Rouder (2016) implements this "born-open" approach using the publicly hosted version control system *Github*. Data can similarly be "born open" with other tools that may be more familiar to a wider range or researchers and are easier to set up. For example, a born-open

---

[3] It should therefore be distinguished from a practice sometimes called "pseudo-anonymization", which involves only partial anonymization (e.g., "Michael Johnson" becoming "Michael J.").

data workflow can be set up using Dropbox[4] and the Open Science Framework (OSF; see http://help.osf.io/m/addons/l/524148-connect-add-ons). Once the connection is set up, the Dropbox storage is available in the files widget. If a file is changed in the Dropbox, all previous versions can be viewed and downloaded in the OSF repository. Currently, a drawback of this approach, compared to using a hosted version control system, is that OSF does not log and display changes made to files as Recent Activities. Hence, if files are deleted, they vanish without a trace, putting a serious limit on transparency.

Version control software, on the other hand, automatically tracks changes to a repository and allow users to access previous versions. Such platforms (e.g., github.com, gitlab.com, or bitbucket.org) have both advantages and disadvantages. They can be used to facilitate collaboration and tracking changes as well as to share research products: They have greatest potential when used for the complete research "pipeline from data collection to final manuscript submission" (Rouder, 2016, p. 1066; Gandrud, 2013b). But for researchers with no previous experience with version control systems, such platforms can have a steep learning curve. In addition, services that host version control systems may have a different commitment to preserve resources than repositories that are explicitly designed to archive research products. However, note that, for example, GitHub repositories can be archived using the publicly funded research data repository Zenodo (https://guides.github.com/activities/citable-code/).

Folder structure

Typically a "project" on the OSF, or on any other repository, will be associated with one or more studies as reported in a paper. The folder structure will naturally depend on what you wish to share. There is no commonly accepted standard. The folders can, e.g., be organized by study, by file type (analysis scripts, data, materials, paper), or data type (raw vs. processed). However, different structures may be justified as a function of the nature of the study. Some archives may also require a specific structure. One example is the BIDS format for openneuro/openfmri (https://doi.org/10.1038/sdata.2016.44). The structure we suggest here is inspired by The DRESS Protocol of The Tier project (http://www.projecttier.org/tier-protocol/dress-protocol/). See Long (2009) for other examples of folder and file structures.

---

[4] Alternatively, the same workflow could also be set up with other cloud storage providers that are integrated with OSF, such as ownCloud, Google Drive, or Box, though funders or institutions may have restrictions on the use of these providers.

Root folder

The root folder contains a general readme file providing general information on the studies and on the folder structure (see below):

- Short description of the study
- A description of the folder structure
- Time and location of data collection for the studies reported
- Software required to open or run any of the shared files
- Under which license(s) the files are shared (see section on licenses in the main paper)
- Information on the publication status of the studies
- Contact information for the authors
- A list of all the shared files

Study Protocol or Preregistration

The repository should contain a description of the study protocol. This can coincide with the preregistration document or the method section of the research report. In the example project (https://osf.io/xf6ug/), which is a registered report, we provide the full document as accepted in-principle at stage 1. If the study protocol or the preregistration consist of multiple files (e.g., analysis scripts or protocols of power analyses) these documents can placed in a Study protocol-folder together with the description of the study protocol.

Materials

If possible, this folder includes all the material presented to the participants (or as-close-as-possible reproductions thereof) as well as, e.g., the software used to present the stimuli and user documentation. The source of this material should be documented, and any licensing restrictions should be noted in a the readme file. In the example project (https://osf.io/xf6ug/), we provide the experimental software used for stimulus presentation and response collection, and the stimuli that we are legally able to share. License information on reuse is included in the README file.

Raw data

This folder includes the original data, in the "rawest" possible form. These could, for example, be individual e-prime files, databases extracted from online survey software, or scans of questionnaires. If this form is not directly exploitable, a processed version (e.g., in

CSV format) that can be imported by any user should be included, in an appropriately labeled folder. For example, raw questionnaire responses as encoded could be made available in this format. Ideally, both versions of the data (i.e., before and after being made "importable") are included. In the example project (https://osf.io/xf6ug/), we provide raw text files saved by the experimental software for each participant. A file containing a description of each dataset should also be included (see section on data documentation).

Processed data

This folder contains the cleaned and processed data files used to generate the results reported in the paper as well as descriptions of the datasets. If data processing is extensive and complex, this can be the most efficient way to enable data re-use by other researchers. Nevertheless, in order to ensure full analytic reproducibility, it is always important to provide raw data in addition to processed data if there are no negative constraints (e.g., identifiable information embedded in the raw data). In the example project (http://doig.org/10.17605/OSF.IO/XF6UG), we provide the processed datasets in the native R Data format. A file containing a description of each dataset should also be included (see section on data documentation).

Analysis

This folder includes detailed descriptions of analysis procedures or scripts used for transforming the raw data into processed data, for running the analyses, and for creating figures and tables. Instructions for reproducing all analyses in the report can be included in the README or in a separate instruction document inthis folder. If parts of the analyses are computationally expensive, this folder can also contain intermediate ("cached") results if this facilitates fast (partial) reproduction of the original results. In the example project (http://doi.org/10.17605/OSF.IO/XF6UG), we provide the R Markdown file used to create the research report (including the appendix), and cached results in the native R Data format. For convenience we also provideR-script versions of the R Markdown files, which can be executed in R without rendering the manuscript. The folder also contains a subfolder "Analysis functions", which contains custom R functions that are loaded and used in the R Markdown files.

Research Report

A write-up of the results, in the form of a preprint/postprint, or the published paper is included here. In our example project, the data and analysis folder contains an R Markdown document that includes the text of the paper interleaved with the R code to process the raw

data and perform all reported analyses. When rendered, it generates the research report (in APA manuscript style) using a dedicated package, papaja (https://github.com/crsh/papaja; Aust & Barth, 2017). The advantage of this approach is that all values presented in the research report can be directly traced back to their origin, creating a fully reproducible analysis pipeline, and helping to avoid copy and paste errors.

Data documentation

Simply making data available is not sufficient to ensure that it is re-usable (see e.g., Kidwell et al., 2016). Providing documentation (often referred to as 'metadata', 'codebooks', or 'data dictionaries') alongside data files will ensure that other researchers, and future you, can understand what values the data files contain and how the values correspond to findings presented in the research report. This documentation should describe the variables in each data file in both human- and machine-readable formats (e.g., csv, rather than docx or pdf).[5] Ideally, codebooks are organized in such a way that each line represents one variable and each information relative to a variable represents a column. Extraneous information, that cannot be read (e.g., colors, formatting), should be be included in the codebook as well. For an example of a codebook based on survey data, see this example by Kai Horstmann (https://osf.io/e4tqy/); for an example based on experimental data see the codebook in our example OSF project (https://osf.io/up4xq/).

Codebooks should include the following information for each variable: the name, description of the variable, units of measurement, coding of values (e.g., "1 = Female","2 = Male"), possible options or range in which the data points can fall (e.g., "1 = not at all to 7 = Very much"), value(s) used for missing values, and information on whether and how the variable was derived from other variables in the dataset (e.g., "bmi was derived from body_weight m and body_height l as $= \frac{}{2}$."). Other relevant information in a codebook entry can include the source of a measure, instructions for a questionnaire item, information about translation, or scale that an item belongs to.[6]

---

[5] Codebooks can be generated from the data set metadata in popular statistical software, including SPSS (http://libguides.library.kent.edu/SPSS/Codebooks), Stata (http://www.stata.com/manuals13/dcodebook.pdf), or R (http://www.martin-elff.net/knitr/memisc/codebook.html; https://cran.r-project.org/web/packages/codebook/index.html), or with data publishing tools (e.g., http://www.nesstar.com/software/publisher.html). The author of the *codebook* (Arslan, 2018) package for R has also created a web app for creating codebooks for SPSS, Stata or RDS files: https://rubenarslan.ocpu.io/codebook/www/

[6] For those interested in metadata and codebooks, the Digital Curation Centre provides a helpful overview (see http://www.dcc.ac.uk/resources/metadata-standards). Common metadata standards are the basic and general-purpose Dublin Core and the more social-science-focused Data Documentation Initiative (DDI) that was originally developed for survey data (Leeper, 2014).

Analytic Reproducibility

Below we provide more detailed guidance on a number of topics in analytic reproducibility.

Document hardware and software used for analyses

The more detailed the documentation of analyses, the more likely they are to be fully reproducible. The hardware, the operating system, and the software compiler used during the installation of some statistical software packages can affect analytical results (e.g., Glatard et al., 2015; Gronenschild et al., 2012). Any nonstandard hardware requirements, such as large amounts of RAM or support for parallelized or distributed computing, should be noted.

Similarly, analysis software is subject to change. Software updates may introduce algorithmic changes or modifications to input and output formats and produce diverging results. Hence, it is crucial to document the analysis software that was used including version numbers (American Psychological Association, 2010; Eubank, 2016; Gronenschild et al., 2012; Keeling & Pavur, 2007; Piccolo & Frampton, 2016; Rokem et al., 2017; Sandve, Nekrutenko, Taylor, & Hovig, 2013; Xie, 2015). If analyses involve any add-ons to the base software they, too, should be documented including version numbers.

The utility of a detailed documentation of the employed software is limited to a large extent by the availability of the software and its previous versions. An interested reader may not have the license for a given commercial software package or may be unable to obtain the specific version used in the reported analysis from the distributor. In contrast to commercial software, open source software is usually free of charge, can be included in shared software environments, and previous versions are often much easier to obtain. For these and other reasons open source software should be prefered to commercial closed source solutions (Huff, 2017; Ince, Hatton, & Graham-Cumming, 2012; Morin et al., 2012; Rokem et al., 2017; Vihinen, 2015).

Consider sharing software environments

Beyond a list of software, there are convenient technical solutions that allow researchers to share the software environment they used to conduct their analyses. The shared environments may consist of the analysis software and any addons but can even include the operating system (e.g., Piccolo & Frampton, 2016).

A software environment is organized hierarchically with the operating system at its base. The operating system can be extended by operating system libraries and hosts the analysis software. In addition some analysis software can be extended by add-ons that are specific to that software. Technical solutions for sharing software environments are available at each level of the hierarchy. Moving from the top to the base of the hierarchy the number of

obstacles for reproducibility decreases but the technical solutions become more complex and less convenient. Choosing between dependency management systems, software containers, and virtual machines involves a trade-off between convenient implementation and degree of computational reproducibility.

Open source analysis software, such as R and Python, support rich ecosystems of add-ons (so-called packages or libraries) that enable users to perform a large variety of statistical analyses. Typically multiple add-ons are used for a research project. Because the needed add-ons often depend on several other add-ons recreating such software environments to reproduce an analysis can be cumbersome. Dependency management systems, such as packrat (Ushey, McPherson, Cheng, Atkins, & Allaire, 2016) and checkpoint (Microsoft Corporation, 2017) for R, address this issue by tracking which versions of which packages the analyst used. Critically, reproducers can use this information and the dependency management systems to automatically install the correct versions of all packages from the the Comprehensive R Archive Network (CRAN).

Software containers, such as Docker (Boettiger, 2015) or ReproZip (Chirigati, Rampin, Shasha, & Freire, 2016), are a more comprehensive solution to sharing software environments compared to add-on dependency management system. Software containers can bundle operating system libraries, analysis software, including add-ons, as well as analysis scripts and data into a single package that can be shared (Huff, 2017; Piccolo & Frampton, 2016). Because the operating system is not included these packages are of manageable size and require only limited computational resources to execute. With Docker, software containers can be set up automatically using a configuration script—the so-called Docker file. These Docker files constitute an explicit documentation of the software environment and can be shared along with data and analysis scripts instead of packaging them into a single but comparably large file (as ReproZip does). A drawback of software containers is that they are not independent of the hosting operating system and may not support all needed analysis software.

Virtual machines allow sharing the complete software environments including the operating system. This approach eliminates most technical obstacles for computational reproducibility. Common virtualization software, such as VirtualBox (https://www.virtualbox.org/), bundle an entire operating system with analysis software, scripts, and data into a single package (Piccolo & Frampton, 2016). This file can be shared but is of considerable size. Moreover, execution of a virtual machine requires more computational resources than a software container. Similar to Docker, workflow tools, such as Vagrant (https://www.vagrantup.com/), can set up virtual machines including the operating system automatically based on a configuration script, which constitutes an explicit documentation of the environment and facilitates sharing the software environment.

Automate or thoroughly document all analyses

Most importantly, analytic reproducibility requires that all steps necessary to produce a result are documented (Hardwicke et al., 2018; Sandve et al., 2013) and, hence, documentation of analyses should be considered from the outset of a research project (p. 386, Donoho, 2010). The documentation could be a narrative guide that details each analytical step including parameters of the analysis (e.g., variable coding or types of sums of squares; Piccolo & Frampton, 2016). However, ideally an interested reader can reproduce the results in an automated way by executing a shared analysis script. Hence, if possible the entire analysis should be automated (Huff, 2017; Kitzes, 2017; Piccolo & Frampton, 2016). Any manual execution of analyses via graphical user interfaces should be documented by saving the corresponding analysis script or by using workflow management systems (Piccolo & Frampton, 2016; Sandve et al., 2013).

If possible the shared documentation should encompass the entire analytic process. Complete documentation ideally begins with the raw data and ends with the reported results. If possible, steps taken to visualize results should be included in the documentation. All data manipulation, such as merging, restructuring, and transforming data should be documented. Manual manipulation of raw data should be avoided because errors introduced at this stage are irreversible (e.g., Sandve et al., 2013).

Use UTF-8 character encoding

Character encodings are systems used to represent symbols such as numbers and text usually in a numeral system, such as binary (zeros and ones) or hexadecimal. Not all character encoding systems are compatible and these incompatibilities are a common cause of error and nuisance. Text files contain no information about the underlying character encoding and, hence, the software either makes an assumption or guesses. If an incorrect character encoding is assumed characters are displayed incorrectly and the contents of the text file may be (partly) indecipherable. UTF-8 is a widely used character encoding system that implements the established Unicode standard. It can represent symbols from most of the world's writing systems and maintains backward compatibility with the previously dominant ASCII encoding scheme. Its standardization, wide adoption, and symbol richness make UTF-8 suitable for sharing and long-term archiving. When storing text files, researchers should ensure that UTF-8 character encoding is applied.

Avoid "works on my machine" errors

When a fully automated analysis fails to execute on the computer of someone who wants to reproduce it although the original analyst can execute it flawlessly, the reproducer may be

experiencing a so-called "works on my machine" error (WOMME). In the political sciences the rate of WOMME has been estimated to be as high as 54% (Eubank, 2016). Trivially, the replicator may be missing files necessary to run the analysis. As discussed above, WOMME can also be caused by hardware and software incompatibilities. Moreover, the file locations specified in analysis scripts are a common source of WOMME. Space and other special characters in file and directory names can cause errors on some operating systems and should be avoided. Similarly, absolute file paths to a specific location (including hard drive and user directory) are a likely source of WOMME. Hence, researchers should use file paths to a location relative to the current working directory if possible (e.g., Eubank, 2016; Gandrud, 2013a; Xie, 2015) or load files from a permanent online source. To guard against WOMME, researchers should verify that their analyses work on a computer other than their own, prefer open source analytical software that is available on all major operating systems, and ideally share the entire software environment used to conduct their analyses (see the *Sharing software environments* section). Another option to avoid WOMME is to share data and code via cloud-based platforms, such as Code Ocean (https://codeocean.com/) or RStudio Cloud (https://rstudio.cloud/), that ensure computational reproducibility by running the analysis code in a cloud environment instead of locally on a user's computer.

Share intermediate results for complex analyses

Some analyses can be costly to reproduce due to non-standard hardware requirements, because they are computationally expensive, or both. Besides pointing out the costliness of such analyses, researchers can facilitate reproducibility of the simpler analysis steps by sharing intermediate results. For example, when performing simulations, such as the simulation of a statistical models' joint posterior distribution in Bayesian analyses, it can be helpful to store and share the simulation results. This way interested readers can reproduce all analyses that rely on the simulated data without having to rerun a computationally expensive simulation.

Set and record seeds for pseudorandom number generators

Some statistical methods require generation of random numbers, such as the calculation of bootstrap statistics, permutation tests in large samples, Maximum likelihood estimation using optimization algorithms, Monte Carlo simulations, Bayesian methods that rely on Markov Chain Monte Carlo sampling, or jittering of data points in plots. Many statistical applications employ algorithmic pseudorandom number generators (PRNG). These methods are called pseudorandom because the underlying algorithms are deterministic but produce sequences of numbers, which have similar statistical properties as truly random sequences. PRNG

apply an algorithm to a numerical starting point (a number or a vector of numbers), the so-called seed. The resulting sequence of numbers is fully determined by the seed—every time the PRNG is initiated with the same seed it will produce the same sequence of pseudorandom numbers. Whenever an analysis involves statistical methods that rely on PRNG the seeds should be recorded and shared to ensure computational reproducibility of the results (Eubank, 2016; Sandve et al., 2013; Stodden & Miguez, 2014), ideally by setting it at the top of the analysis script.

Practical Implementation:

Note that the analysis software or add-ons to that software may provide more than one PRNG and each may require its own seed. In principle, any whole number is a valid seed for a PRNG but in practice larger numbers sometimes yield better sequences of pseudorandom numbers in the sense that they are harder to distinguish from truly random sequences. A good way to generate a PRNG seed value is to use a true random number generator, such as https://www.random.org/integers/.

SPSS

SPSS provides the multiplicative congruential (MC) generator, which is the default PRNG, and the Mersenne Twister (MT) generator, which was added in SPSS 13 and is considered to be a superior PRNG—it is the default in SAS, R, and Python. The MC generator can be selected and the seed value set as follows:

SET RNG=MC SEED=301455.

For the MC generator the seed value must be a any whole number between 0 and 2,000,000. The MT generator can be selected and the seed value set as follows:

SET RNG=MT MTINDEX=158237730.

For the MC generator the seed value can be any real number. To select the PRNG and set the seed value in the graphical user interface choose from the menus Transform > Random Number Generators.

SAS

SAS relies on the MT generator. The seed value can be set to any whole number between 1 and 2,147,483,647 as follows:

call streaminit(663562138);

R

R provides seven different PRNG but by default relies on the MT generator. The MT generator can be selected explicitly and the seed value set to any whole number as follows:

set.seed(seed = 923869253, kind = "Mersenne-Twister")

Note that some R packages may provide their own PRNG and rely on seed values other than the one set by set.seed().

Python

Python, too, relies on the MT generator. The seed value can be set to any whole number, a string of letters, or bytes as follows:

random.seed(a = 879005879)

Note that some Python libraries may provide their own PRNG and rely on seed values other than the one set by random.seed().

Make your analysis documentation easy to understand

It is important that readers of a narrative documentation or analysis scripts can easily connect the described analytical steps to interpretative statements, tables, and figures in a report (e.g., Gandrud, 2013; Sandve et al., 2013). Correspondence between analyses and reported results can be established by adding explicit references to section headings, figures, or tables in the documentation and by documenting analyses in the same order in which the results are reported. Additionally, it can be helpful to give an overview of the results produced by the documented analysis (see the Project Tier DRESS Protocol). Additional analyses that are not reported can be included in the documentation but should be discernible (e.g., by adding a comment "not reported in the paper"). A brief justification why the analyses were not reported should be added as a comment.

Best practices in programming discourage extensive commenting of analysis scripts because comments have to be diligently revised together with analysis code—failing to do so yields inaccurate and misleading comments (e.g., Martin, 2009). While excessive commenting can be useful during analysis, it is recommended to delete obscure or outdated comments once a script is finalized to reduce confusion (Long, 2009). Comments should explain the rationale or intent of an analysis, provide additional information (e.g., preregistration documents or standard operating procedures, Lin & Green, 2016), or warn that, for example, particular analyses may take a long time (Martin, 2009). If comments are needed to explain how a script works, researchers should check whether they can instead rewrite the code to be clearer. Researchers can facilitate the understanding of their analysis

scripts by adhering to other common best practices in programming, such as using consistent, descriptive, and unambiguous names for variables, labels, and functions (e.g. Kernighan & Plauger, 1978; Martin, 2009) or avoiding to rely on defaults by explicitly setting optional analysis parameters. Extensive narrative documentation is not necessary in a script file (Eglen et al., 2017), and is better suited to dynamic documents (see below).

As a final note, it can be beneficial to split the analysis documentation into parts (i.e., files and directories) in a way that suits the research project. A basic distinction applicable to most cases is between processing of raw data—transforming original data files into restructured and cleaned data—and data analysis and visualization (see, e.g., http://www.projecttier.org/tier-protocol/specifications/).

Dynamic documents

It is important that readers of a narrative documentation or analysis scripts can easily connect the described analytical steps to interpretative statements, tables, and figures in a report (e.g., Gandrud, 2013a; Sandve et al., 2013).

Dynamic documents constitute a technically sophisticated approach to connect analytical steps and interpretative statements (e.g., Gandrud, 2013a; Knuth, 1984; Kluyver et al., 2016; Welty, Rasmussen, Baldridge, & Whitley, 2016; Xie, 2015). Dynamic documents intertwine automated analysis scripts and narrative reporting of results. When a document is compiled all embedded analysis scripts are executed and the results are inserted into the text. The mix of analysis code and prose creates explicit links between the reported results and the underlying analytical steps and makes dynamic documents well suited for documentation and sharing. It is possible to extend this approach to write entire research papers as dynamic documents (e.g., Aust & Barth, 2017; Allaire et al. 2017b). When sharing researchers should include both the source file, which contains the executable analysis code, and the compiled file, preferably in HTML or PDF format.

Below we provide a brief overview of three software solutions for creating dynamic documents: R Markdown (Allaire et al., 2017a), Jupyter (Kluyver et al., 2016), and StatTag (Welty, et al., 2016).

R Markdown

rmarkdown is an R package that provides comprehensive functionality to create dynamic documents. R Markdown files consist of a front matter that contains meta information as well as rendering options and is followed by prose in Markdown format mixed with R code chunks. Markdown is a formatting syntax that was designed to be easy-to-read and -write (e.g., *italic* yields *italic*) and has gained considerable popularity in a range of applications.

When the document is compiled the R code is executed sequentially and the resulting output (including figures and tables) is inserted into the document before it is rendered into a HTML, Word, and PDF document. Although R Markdown is primarily intended for R, other programming languages, such as Python or Scala, have limited support.

R Markdown uses customizable templates that control the formatting of the compiled document. The R package papaja (Aust & Barth, 2017) provides templates that are specifically designed to create manuscripts in APA style and functions format analysis results in accordance with APA guidelines. Additional document templates that conform to specific journal or publisher guidelines are available in the rticles package (Allaire et al., 2017b)

The freely available integrated development environment RStudio provides good support for R Markdown and can be extended to, e.g., count words (Marwick, n.d.) or search and insert citations from a BibTeX file or Zotero library (Aust, 2016).

Jupyter

Jupyter is a web-application for creating dynamic documents that support one or multiple programming languages, such as Python, R, Scala, and Julia. Like R Markdown, Jupyter relies on the Markdown formatting syntax for prose and while the primary output format for dynamic documents is HTML, Jupyter documents can be rendered to other formats with document templates, albeit less conveniently. Like in R Markdown, Jupyter can be extended, e.g., to search and insert citations from a Zotero library.

StatTag

StatTag can be used to create dynamic Word documents. It supports integration with R, SAS, and SPSS by inserting the contents of variables defined in the analysis scripts into the word document. Other document formats are not supported.

Comparison

StatTag may be the most beginner friendly but currently least flexible option and it is the only of the three presented options that supports SAS and SPSS. Jupyter is the recommended alternative for researchers using Python, Scala, and Julia, or for researchers whose workflows combine multiple programming languages including R. While Jupyter is well suited for data exploration, interactive analysis, and analysis documentation, R Markdown is better suited for writing PDF and Word documents including journal article manuscripts. In contrast to Jupyter, R Markdown relies entirely on text files, works well with any text editor or integrated development environment, and is better suited for version control systems such

as git. Technical requirements and personal preferences aside, R Markdown, Jupyter, and StatTag are all well suited for documenting and sharing analyses.

Preregistration

How should you pre-register your study? There has been growing awareness of pre-registration in recent years, but there are still few established guidelines to follow. In brief, an ideal pre-registration involves a written specification of your hypotheses, methods, and analyses, that you formally 'register' (create a time-stamped, read-only copy) on a public website, such that it can be viewed by the scientific community. Another form of pre-registration known as "Registered Reports" (Chambers, 2013; Hardwicke & Ioannidis, 2018), involves submitting your pre-registration to a journal where it undergoes peer-review, and may be offered *in principle acceptance* before you have even started the study, indicating that the article will be published pending successful completion of the study according to the methods and analytic procedures outlined, as well as a cogent interpretation of the results. This unique feature of Registered Reports may offer some remedy to the issue of publication bias because studies are accepted for publication based on the merits of the research question and the methodological quality of the design, rather than the outcomes (Chambers et al., 2014).

Really, it is up to you how much detail you put in your pre-registration and where you store it. But clearly, a more detailed (and reviewed) pre-registration will provide more constraint over the potential analytical flexibility, or 'researcher degrees of freedom', outlined above, and will therefore allow you and others to gain more confidence in the veracity of your findings. To get started, you may wish to use an established pre-registration template. The Open Science Framework (OSF) has several to choose from (for a brief tutorial on how to pre-register via the OSF, see https://osf.io/2vu7m/). In an OSF project, click on the "Registrations" tab and click "New Registration". You will see a list of options. For example, there is a template that has been developed specifically for social psychology (van 't Veer & Giner-Sorolla, 2016). For a simple and more general template you may wish to try the "AsPredicted preregistration". This template asks you 9 key questions about your study, for example, "Describe the key dependent variable(s) specifying how they will be measured." One downside of templates is that they do not always cover important aspects of your study that you think should be pre-registered but the template creators have not anticipated. Templates can also be limited if you want to specify detailed analysis code within your pre-registration document. As a result, you may quickly find that you prefer to create your own custom pre-registration document (either from scratch or adapted from a template). Such a

document can still be registered on the OSF, you just need to upload it to your OSF project as a regular file, and register it using the procedure outlined above, this time choosing the "OSF-Standard Pre-Data Collection Registration" option instead of one of the other templates.

After completing a template, or choosing to register a custom document, you will be asked if you would like to make the pre-registration public immediately, or set an embargo period of up to four years, after which the pre-registration will be made public. Note that the AsPredicted template mentioned above is actually based on a different website (https://aspredicted.org/) that provides its own registration service as an alternative to the OSF. If you use the AsPredicted service, all pre-registrations are private by default until they are explicitly made public by their owners. This may sound appealing, but it is potentially problematic: when registrations are private, the scientific community cannot monitor whether studies are being registered and not published (e.g., a file-drawer effect), or whether multiple, similar pre-registrations have been created. We would therefore recommend using the OSF, where all pre-registrations will eventually be made public after four years. Once the registration process is complete (you and your collaborators may need to first respond to a confirmation e-mail), you will be able to see the frozen, read-only, time-stamped version of your project containing your pre-registration. You may need to click on the green "view registration" button if you used a template, or click on your custom pre-registration document in the "files" window to see the content itself. The url displayed in the address bar is a unique, persistent link to your pre-registration that you can include in your final published article.

When you write up your study, you should explicitly indicate which aspects were pre-registered and which were not. It is likely that some deviations from your plan were necessary. This is not problematic, simply note them explicitly and clearly, providing a rationale where possible. Where you were able to stick to the plan, these aspects of your study retain their full confirmatory status. Where deviations were necessary, you and your readers have the information they need to judge whether the deviation was justified. Three additional tools may be helpful in such cases. Firstly, one can anticipate some potential issues, and plan for them in advance using a 'decision-tree'. For example, one might pre-specify that "if the data are normally distributed we will use a Student's t-test, but if the data are not normally distributed we will use a Mann-Whitney U test". Of course, the number of potential things that can "go wrong" and require deviation from the pre-specified plan are likely to multiply quite rapidly, and this approach can become untenable.

A more long-term solution is for an individual researcher or lab to write a "Standard Operating Procedures" (SOP) document, which specifies their default approach to handling various issues that may arise during the studies that they typically run (Lin & Green, 2016). For example, the document might specify which data points are considered "outliers" in reaction time data, and how those outliers are typically handled (e.g., excluded or retained). SOPs should also be registered, and either included along with your main pre-registration as an appendix or linked to directly. Of course, SOPs are only useful for issues that you have already anticipated and planned for, but it can be a valuable safety-net when you forget to include relevant information in your main pre-registration. SOPs can be continuously updated whenever new scenarios are encountered, such that there is a plan in place for future occasions.

Finally, a useful approach for handling unanticipated protocol deviations is to perform a *sensitivity analysis* (Thabane et al, 2013). Sensitivity analyses are employed when there are multiple reasonable ways of specifying an analysis. For example, how should one define exclusion criteria for outliers? In a sensitivity analysis, a researcher runs an analysis several times using different specifications (e.g., exclusion thresholds), and evaluates the impact of those specifications on the final outcome. An outcome is considered 'robust' if it remains stable under multiple reasonable analysis specifications. One might also consider running a *multiverse analysis*: a form of factorial sensitivity analysis where different specifications are simultaneously considered for multiple aspects of the analysis pipeline, giving a much more in depth picture of the robustness of the outcome under scrutiny (Steegen et al., 2016; also see Simonsohn et al., 2015). Indeed, multiverse analyses (and sensitivity analyses more broadly) are highly informative even when one has been able to stick to the pre-registered plan. To the extent that the pre-registered analysis plan included fairly arbitrary specifications, it is possible that that plan does not provide the most robust indication of the outcome under scrutiny. The gold standard here is to pre-register a plan for a multiverse analysis (Steegen et al., 2016).

Incentivising Sharing

When sharing data, code, and materials, when reusing resources shared by others, and when appraising research merits, scientists form part of an ecosystem where behaviour is guided by incentives. Scientists can help shape these incentives and promote sharing by making use of mechanisms to assign credit, and by recognizing the value of open resources published by others.

How to get credit for sharing

To make a published dataset citable, it is recommended to use a repository that provides a persistent identifier, such as a Digital Object Identifier (DOI). Others will then be able to cite the data set unambiguously.

A further mechanism that can help a researcher get credit for open data is the data article.The purpose of a data article is to describe an dataset in detail, thereby increasing the potential for reuse (Gorgolewski, Margulies, & Milham, 2013). Examples of journals that publish data articles and cover the field of psychology are *Scientific Data* (https://www.nature.com/sdata/), the *Journal of Open Psychology Data* (https://openpsychologydata.metajnl.com/), and the *Research Data Journal for the Humanities and Social Sciences* (http://www.brill.com/products/online-resources/research-data-journal-humanities-and-social-sciences). Data articles can be used to provide documentation going beyond metadata in a repository, e.g. by including technical validation. They can be a good means of enhancing the visibility and reusability of the data and are especially worthwhile for data with high reuse potential.

Initiatives to increase data sharing

Numerous research funders, universities/institutions, and scientific journals have adopted policies encouraging or mandating open data (reviewed e.g. in Chavan & Penev, 2011 and Houtkoop, Chambers, Macleod, Bishop, Nichols, & Wagenmakers, 2018). The Peer Reviewers' Openness (PRO) Initiative is seeking to encourage transparent reporting of data and materials availability via the peer review process (Morey et al., 2016). Signatories of the PRO Initiative commit to reviewing papers only if the authors either make the data and materials publically available, or explain in the manuscript why they chose not to share the data and materials.

A recent systematic review (Rowhani-Farid, Allen, & Barnett, 2017) found that only one incentive has been tested in health and medical research with data sharing as outcome measure: Badges to Acknowledge Open Practices (https://osf.io/tvyxz/wiki/home/). Kidwell et al. (2016) observed an almost 10-fold increase in data sharing after badges were introduced at the journal *Psychological Science*. However because this was an observational study, it is possible that other factors contributed to this trend. A follow-up study of badges at the journal *Biostatistics* found a more modest increase by about 7% on an absolute scale (Rowhani-Farid & Barnett, 2018).

Another strategy for incentivizing sharing comes from fellowships funding the expansion of transparent research practices in academic institutions, such as the rOpenSci fellowship program and the Mozilla Science Fellowship program.

Reusing others' research products

Citation of research products – software, data, and materials, not just papers – contributes to better incentives for sharing these products. Commonly cited barriers to data sharing include concerns of researchers who generate data that others will publish important findings based on these data before they do ("scooping"), duplication of efforts leading to inefficient use of resources, and that new analyses will lead to unpredictable and contradictory results (International Consortium of Investigators for Fairness in Trial Data Sharing et al., 2016; Smith & Roberts, 2016). While, at least to our knowledge, there exists no reported case of a scientist that has been scooped with their own data after publishing them openly, and while differences in results can be the topic of a fruitful scientific discourse, fears such as these can be allayed by consulting the researchers who published the data before conducting the (re)analysis. A further reason for consulting researchers who created data, code, or materials is that they are knowledgeable about the resource and may be able to anticipate pitfalls in reuse strategies and propose improvements (Lo & DeMets, 2016). While the publication of a resource such as data, code, or materials generally does not in itself merit consideration for co-authorship on subsequent independent reports, it may be valuable to invite the resource originators into a discussion about the proposed new work. If the resource originators make an important academic contribution in this discussion, it is reasonable to consider offering coauthorship. What constitutes an important contribution can only be determined in relation to the case at hand; development of hypotheses, analytical strategies, and interpretations of results are examples that may fall in this category. Approaching open resources with an openness towards collaboration may, thus, help to increase value, as well as promoting a sharing culture. Bear in mind that offering co-authorship for researchers whose only contribution was to share their previously collected data with you on request disincentivizes public sharing.

Recommendations:
- Contact the originators of the resource beforehand and tell them about your plans
- Consider including the original authors in a collaboration if they make an important academic contribution to your efforts
- Do not include researchers as co-authors if their only contribution was sharing-on-request of previously collected data.
- Cite both the resource and any associated papers
- Address open science practices explicitly in assessments of merits, and appraise the value of open science contribution

When appraising scientific merits

The merit value of open science practices may strongly influence scientists' behaviour. Therefore, when appraising merits, for example in decisions about hiring and awarding research grants, open science practices should be assessed. Some departments of psychology[7] encourage or even require job advertisement to include a statement expressing commitment to open science[8]. Examples of such job advertisements are available here: https://osf.io/b9zks/.

Several factors affect the value of open scientific resources.
Indicators of higher value may include:

- The resource is available in a trusted repository
- The resource is comprehensively described
- Data and metadata conform to discipline-specific standards
- The resource appears, on the face of it, to have high reuse potential
- There is evidence of reuse

---

[7] such as those at Ludwig Maximilian University (Munich, Germany) and at the University of Oregon.
[8] An example of such a statement is "such as "Our department embraces the values of open science and strives for replicable and reproducible research. For this goal we support transparent research with open data, open materials, and study pre-registration. Candidates are asked to describe in what way they already pursued and plan to pursue these goals." (http://www.fak11.lmu.de/dep_psychologie/osc/open-science-hiring-policy/index.html).

References

Allaire, J. J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., ... Arslan, R. C. (2017a). rmarkdown: Dynamic Documents for R [Computer software]. Retrieved from https://cran.r-project.org/web/packages/rmarkdown

Allaire, J. J., R Foundation, Wickham, H., Journal of Statistical Software, Xie, Y., Vaidyanathan R., ... Yu, M. (2017b). rticles: Article Formats for R Markdown [Computer software]. Retrieved from https://cran.r-project.org/web/packages/rticles

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th edition). Washington, DC: American Psychological Association.

Arslan, R. C. (2018). Automatic codebook generation with {codebook}. Retrieved from https://github.com/rubenarslan/codebook

Aust, F. (2016). citr: 'RStudio' Add-in to Insert Markdown Citations [Computer software]. Retrieved from https://cran.r-project.org/web/packages/citr

Aust, F., & Barth, M. (2017). papaja: Create APA manuscripts with R Markdown [Computer software]. Retrieved from https://github.com/crsh/papaja

Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review, 49*(1), 71-79. DOI: http://www.doi.org/10.1145/2723872.2723882.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. Cortex, 49(3), 609–610. DOI: http://doi.org/10.1016/j.cortex.2012.12.016

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. AIMS Neuroscience, 1(1), 4–17. DOI: http://doi.org/10.3934/Neuroscience.2014.1.4

Chassang, G. (2017). The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience*, 11: 709 . DOI: http://www.doi.org/10.3332/ecancer.2017.709

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics, 12,* Suppl 15, S2. DOI : https://doi.org/10.1186/1471-2105-12-S15-S2

Chirigati, F., Rampin, R., Shasha, D., & Freire, J. (2016). Reprozip: Computational reproducibility with ease. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2085-2088). ACM.

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, 11(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028

Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P., … Poline, J.-B. (2017). Toward standard practices for sharing computer code and programs in neuroscience. *Nature Neuroscience*, 20(6), 770–773. DOI: https://doi.org/10.1038/nn.4550

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

El Emam, K. (2013). Guide to the de-identification of personal health information. Boca Raton, FL: CRC Press.

Gandrud, C. (2013a). *Reproducible Research with R and Rstudio*. Boca Raton, FL: CRC Press. https://github.com/christophergandrud/Rep-Res-Book

Gandrud, C. (2013b). GitHub: A tool for social data set development and verification in the cloud. Available at SSRN: https://ssrn.com/abstract=2199367 or http://dx.doi.org/10.2139/ssrn.2199367

Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., … Evans, A. C. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, 9. DOI: https://doi.org/10.3389/fninf.2015.00012

Gorgolewski, K. J., Margulies, D. S., & Milham, M. P. (2013). Making data sharing count: a

publication-based solution. *Frontiers in Neuroscience, 7*, 9.

Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., Mengelers, R., Rozendaal, N., Os, J. van,

& Marcelis, M. (2012). The Effects of FreeSurfer Version, Workstation Type, and Macintosh

Operating System Version on Anatomical Volume and Cortical Thickness Measurements.

PLOS ONE, 7(6), e38234. DOI: https://doi.org/10.1371/journal.pone.0038234

Hardwicke, T. E., Mathur, M. B., MacDonald, K. E., Nilsonne, G., Banks, G. C.,... Frank, M. C.

(2018, March 19). Data availability, reusability, and analytic reproducibility: Evaluating the

impact of a mandatory open data policy at the journal Cognition. Retrieved from

https://osf.io/preprints/bitss/39cfb/

Houtkoop, B., Chambers, C., Macleod, M., Bishop, D., Nichols, T., & Wagenmakers, E. J. Data

sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and

Practices in Psychological Science.* Advance online publication. DOI:

https://doi.org/10.1177/2515245917751886

Huff, K. (2017). Lessons Learned. In Kitzes, J., Turek, D., & Deniz, F. (Eds.). *The Practice of

Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*.

Oakland, CA: University of California Press. Retrieved from

https://www.gitbook.com/book/bids/the-practice-of-reproducible-research

International Consortium of Investigators for Fairness in Trial Data Sharing, Devereaux, P. J.,

Guyatt, G., Gerstein, H., Connolly, S., & Yusuf, S. (2016). Toward Fairness in Data Sharing.

The New England Journal of Medicine, *375*(5), 405–407. DOI:

http://doi.org/10.1056/NEJMp1605654

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs.

Nature, 482(7386), 485–488. DOI: https://doi.org/10.1038/nature10836

Keeling, K. B., & Pavur, R. J. (2007). A comparative study of the reliability of nine statistical

software packages. *Computational Statistics & Data Analysis*, 51(8), 3811–3831. DOI:

https://doi.org/10.1016/j.csda.2006.02.013

Kernighan, B. W., & Plauger, P. J. (1978). *The Elements of Programming Style* (2nd edition). New York: McGraw-Hill.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS Biology*, *14*(5), e1002456. DOI: https://doi.org/10.1371/journal.pbio.1002456.

Kitzes, K. (2017). The Basic Reproducible Workflow Template. In Kitzes, J., Turek, D., & Deniz, F. (Eds.). *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*. Oakland, CA: University of California Press. Retrieved from https://www.gitbook.com/book/bids/the-practice-of-reproducible-research

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., … Jupyter Development Team. (2016). Jupyter Notebooks – A publishing format for reproducible computational workflows. In *Proceedings of the 20th International Conference on Electronic Publishing* (pp. 87–90). DOI: https://doi.org/10.3233/978-1-61499-649-1-87

Knuth, D. E. (1984). Literate Programming. *The Computer Journal*, 27(2), 97–111. DOI: https://doi.org/10.1093/comjnl/27.2.97

Leeper, T. J. (2014). Archiving Reproducible Research with R and Dataverse. *R Journal*, *6*(1).

Lo, B., & DeMets, D. L. (2016). Incentives for Clinical Trialists to Share Data. *The New England Journal of Medicine*, *375*(12), 1112–1115. DOI: http://doi.org/10.1056/NEJMp1608351

Long, J. S. (2009). *The workflow of data analysis using Stata*. College Station, TX: Stata Press.

Lin, W., & Green, D. P. (2016). Standard Operating Procedures: A Safety Net for Pre-Analysis Plans. *PS: Political Science & Politics*, 49(3), 495–500. DOI: https://doi.org/10.1017/S1049096516000810

Martin, R. C. (2009). *Clean Code - A Handbook of Agile Software Craftsmanship*. Upper Saddle River, NJ: Prentice Hall. Retrieved from http://ricardogeek.com/docs/clean_code.html

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science.* Advance online publication. DOI: https://doi.org/10.1177/2515245917747656

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., et al.

(2016). The peer reviewers' openness initiative: Incentivizing open research practices

through peer review. *Royal Society Open Science*, *3*(1), 150547–7. DOI:

http://doi.org/10.1098/rsos.150547

Morin, A., Urban, J., Adams, P. D., Foster, I., Sali, A., Baker, D., & Sliz, P. (2012). Shining Light

into Black Boxes. *Science*, 336(6078), 159–160.

DOI: https://doi.org/10.1126/science.1218263

Piccolo, S. R., & Frampton, M. B. (2016). Tools and techniques for computational reproducibility.

GigaScience, 5, 30. DOI: https://doi.org/10.1186/s13742-016-0135-4

Staneva, V. (2017). Assessing Reproducibility. In Kitzes, J., Turek, D., & Deniz, F. (Eds.). *The*

Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive

Sciences. Oakland, CA: University of California Press. Retrieved from

https://www.gitbook.com/book/bids/the-practice-of-reproducible-research

Rowhani-Farid, A., Allen, M., & Barnett, A. G. (2017). What incentives increase data sharing in

health and medical research? A systematic review. *Research Integrity and Peer Review*,

2(1), 4. DOI: https://doi.org/10.1186/s41073-017-0028-9

Rowhani-Farid, A., & Barnett, A. G. (2018). Badges for sharing data and code at Biostatistics: an

observational study. *F1000Research*, *7*, 90. https://doi.org/10.12688/f1000research.13477.1

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior research methods,*

48(3), 1062-1069. DOI: https://doi.org/10.3758/s13428-015-0630-z

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible

Computational Research. *PLOS Computational Biology*, 9(10), e1003285.

https://doi.org/10.1371/journal.pcbi.1003285

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results.

Psychological science, 26(5), 559-569. DOI: http://www.doi.org/10.1126/science.aab2374

Smith, R., & Roberts, I. (2016). Time for sharing data to become routine: the seven excuses for

not doing so are all invalid. *F1000Research*, *5*, 781. DOI:

http://doi.org/10.12688/f1000research.8422.1

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency

through a multiverse analysis. *Perspectives on Psychological Science*, 11, 702-712. DOI:

https://doi.org/10.1177/1745691616658637

Stodden, V., & Miguez, S. (2014). Best Practices for Computational Science: Software

Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open

Research Software*, 2(1). DOI: https://doi.org/10.5334/jors.ay

Sweeney L. (2000). Simple demographics often identify people uniquely.

http://impcenter.org/wp-content /uploads/2013/09/Simple-Demographics-Often -Identify-

People-Uniquely.pdf.

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., et al. (2013). A tutorial

on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical

Research Methodology*, 13(1), 92. DOI: http://doi.org/10.1186/1471-2288-13-92

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion

and suggested template. Journal of Experimental Social Psychology, 67(C), 2–12.

DOI: http://doi.org/10.1016/j.jesp.2016.03.004

Vihinen, M. (2015). No more hidden solutions in bioinformatics. *Nature News*, 521(7552), 261.

DOI: https://doi.org/10.1038/521261a

Welty, L.J., Rasmussen, L.V., Baldridge, A.S., & Whitley, E. (2016). StatTag [Computer

software]. Chicago, IL: Galter Health Sciences Library. DOI:

https://doi.org/10.18131/G36K76

Xie, Y. (2015). *Dynamic Documents with R and knitr* (2nd edition). Boca Raton, FL: CRC Press.