Editorial

Generating Codebooks to Ensure the Independent Use of Research Data

Some Guidelines

Kai T. Horstmann[1], Ruben C. Arslan[2], and Samuel Greiff[3]

[1] Institute of Psychology, Humboldt-Universität zu Berlin, Germany
[2] Center for Adaptive Rationality, Max Planck Institute for Human Development, Germany
[3] Institute of Cognitive Science and Assessment, University of Luxembourg, Luxembourg

The publication of research data has become increasingly important. Recently, the *European Journal of Psychological Assessment (EJPA)* has implemented practices that strongly encourage and support data sharing for any article accepted for publication at *EJPA*. An essential element of data sharing is the documentation of research data, that is, providing a codebook that allows independent researchers to interpret the data file without additional assistance. However, general standards on how to document data have not yet been developed. Here, we define the purpose of a codebook and showcase how a codebook could be formatted and which information could be included. Given that different researchers have different data structures and use different formats, we argue that the general purpose of a codebook must be fulfilled rather than adhering to a specific format. We further highlight useful resources and provide a starting point for anyone wanting to release their documented data.

Generating Codebooks to Ensure the Independent Use of Research Data

Research on interindividual differences has a long tradition of data sharing. When constructing a psychological test using factor analysis, for example, providing a variance-covariance matrix and descriptive statistics of the items is usually sufficient to reproduce and evaluate key findings of a study. Yet, sharing only the minimal information required to reproduce the findings of a study is no longer sufficient, for two reasons: First, most studies nowadays, and especially at *EJPA*, report results from complex analyses, using multiple datasets (e.g., Rammstedt et al., 2020), longitudinal data (e.g., Blanke et al., 2020; Brandt et al., 2020), or data from different sources such as informant and peer-reports (e.g., Heintz, 2019). For those studies, it may simply be too difficult to share only the information that is required to reproduce the central findings. Second, and even more importantly, sharing the actual research data at the lowest possible level (e.g., item responses, reaction times, physiological recordings) allows other researchers to build on and extend the current work, to use the data for their own research questions, and to investigate the robustness of published research findings (Hendrick, 1988).

From our perspective, data are an integral part of a published study and should therefore always be shared, as long as legal and ethical requirements can be met (Wicherts et al., 2006).[1] Generally speaking, research data should be findable (e.g., via a search engine or a link in the manuscript), accessible (e.g., in a permanent open online repository), interoperable (e.g., be used with different programs), and re-useable (e.g., have a license that allows re-use of data), together referred to as the FAIR principle (Arslan, 2019; Borghi & Van Gulick, 2018; Hardwicke et al., 2018; Wilkinson et al., 2016). Besides, there are also numerous other positive side-effects of data sharing that we will not reiterate here (Kidwell et al., 2016; Munafò et al., 2017). For the current article, we will assume that sharing data is beneficial and that a specific researcher is already convinced to share their data. Thus, we will focus on only one but crucial aspect of data sharing: the codebook.

EJPA encourages and supports sharing data alongside with published manuscripts (Greiff et al., 2020). However,

[1] We are aware that both legal and ethical aspects of data sharing are far from clear in a globalized word, and that this can be a challenging issue, in particular for some kind of data (e.g., transcripts, videos).

https://doi.org/10.1027/1015-5759/a000620

simply sharing the data, for example, by uploading a raw csv-file to an online repository may not be sufficient to really enable the use of the dataset (Gilmore et al., 2018). Without an explanation of the dataset, it is nearly impossible to use the data for the two purposes described above. Making a dataset available almost always requires the publication of a codebook alongside the dataset. However, there has been a recent surge in open data practices, accompanied by a stark increase in publications that describe when, where, or how to share data. We acknowledge that diving into this literature can be overwhelming. Our goal was therefore to provide a comprehensive overview of issues around data sharing, with a particular focus on the codebook: First, we summarize different purposes of a codebook that should be understood as guidance for developing a codebook. Any codebook that enables data re-use for these purposes is already a good codebook. Second, we describe how a codebook could be structured. Here, we provide information on *how* a codebook should be structured and list all information that should be included in a codebook (such as items, their coding, etc.). Finally, we provide practical steps toward data sharing.[2]

Purpose of a Codebook

A codebook, sometimes also referred to as a data dictionary (Broman & Woo, 2018), for a dataset serves one main purpose: Combining it with the respective dataset allows any competent researcher to understand the data and make use of them. This means that the litmus test for the quality of a codebook is the answer to the following question:

> Can an independent researcher analyze and validly interpret the data without any further information about the data apart from what is provided in the codebook?

This means that if the answer to this question is "yes", then anyone interested in the data would simply need to obtain the codebook and could start with their own analyses and interpretation in line with the idea of open science and transparency. For example, if contacted to share data for a meta-analysis, the author of the requested dataset would only need to send the data with codebook, without having to add any further information.

Potential Reasons to Share or Ask for Data

There are numerous reasons that someone might ask for data from a published study (Hendrick, 1988). Data that are shared alongside articles published in *EJPA* should fulfill all of these reasons sufficiently. When preparing a submission their data for release in *EJPA*, authors could therefore make sure that their data facilitate the following purposes:

(1) reproduction of results,
(2) finding alternative explanations,
(3) computing additional indices/statistics,
(4) systematic review and meta-analysis,
(5) combining datasets from different studies,
(6) testing new methods,
(7) teaching purposes, and finally
(8) cumulative science.

(1) Nullius in verba – reproduction of results. One of the most well-known phrases in scientific publishing is the phrase and motto of the Royal Society "nullius in verba", translated as "take no one's word for it" – meaning that science is built on the possibility to verify, rather than trust. Being able to verify the claims made in a scientific contribution should be an integral part of scientific publications (Munafò et al., 2017; Nosek et al., 2015). This means that not only the data need to be shared, but that they need to be shared in a way that allows for direct and straightforward verification.

(2) Alternative explanations and robustness checks. Secondly, others (most definitely Reviewer 2) may want to search for alternative explanations of an effect, or to conduct robustness checks for a claim previously published. This must not be to discredit the claims of the original publication, but simply to assess the validity of the claims made. This is especially true for the review process, and the ability for reviewers to assess data on which claims are based.

(3) Additional indices. Especially in the realm of psychological assessment, numerous indices and cut-offs to evaluate the fit of a model or the appropriateness of a statistic (co-)exist and continue to be developed. Such indices inform the claims that rest on the data. Authors may justifiably choose not to report all of them, but it may be relevant for others to compute additional indices to thereby better evaluate a scale; for example, before revising a scale, an author may choose to get additional item-statistics for said scale. Publications in *EJPA* should generally allow for such additional analyses.

(4) Systematic review and meta-analysis. An additional reason to share (or request) data is to conduct meta-analyses over previously publishes studies. Oftentimes, studies lack some crucial information about a certain effect (simply because it was not the focus of the study), while this effect

[2]Note that, throughout the article, we reference specific software or websites. All of these should be understood as suggestions, and *EJPA* does not require authors to use any of the mentioned resources, especially if alternative resources are or become available.

is paramount for a meta-analysis. Providing open data and a codebook allows others to compute whatever statistic they require and can also make it easier to find relevant datasets, whereas keyword searches will often fail to turn up studies that collected data on variables that were not the focus of the study. Note that in some instances, samples may be so rare or specific, that robust conclusions about an effect may only be drawn after years of publishing the respective data. A dataset that is usable and interpretable without any further assistance from the original author may therefore display its true worth years after its original publication.

(5) Combining and comparing data from different studies. Studies on psychological assessment often publish data from similar instruments that have either been revised over time or adapted for different purposes (e.g., different cultures or different age groups). Being able to combine datasets from different studies allows for a better examination and validation of measures or instruments. This is also true for detecting and avoiding jingle-jangle fallacies: If the original wording of the items is shared, it is easier (or even made possible at all) to identify domains named differently that refer to the same construct, and to identify domains named similarly while referring to different constructs.

(6) Testing new methods. The increasing complexity in study designs, larger datasets, more elaborate theories, and statistical models is likely to lead to more specialized disciplines: Some will become experts in study design, others in data collection, some in data analysis, and some in developing new methods for data analysis. Those who develop new models and analysis techniques may benefit greatly from the availability of open (and interpretable) research data.

(7) Teaching purposes. Another, albeit not central reason to share data in a reusable way is that it can serve as the basis for teaching. One of the most famous and probably over-analyzed datasets is the Holzinger and Swineford (1939) dataset that is used to teach factor analysis. Open and usable data could supplement such data and introduce students to new techniques, especially if increasingly complex methods enter the curricula of psychology students.

(8) Cumulative knowledge. Finally, and most importantly, open and reusable data allow for a cumulative science. The data could technically be used for a different research purpose, a different question. This does not mean that the original authors have their idea "stolen", but an interested person, upon inspecting the codebook, could suggest a new research idea for a collaboration. Similarly, such data can be used to inform the next research project, for example, by generating estimates that can then be used for power-analysis, to test a new version of a questionnaire, or to extend and inspect the nomological network of a construct. Furthermore, open data can be used for exploratory analyses, which are then confirmed in a new sample. Although this reason is not the central purpose of good data sharing practices, it is arguably the most important one with regard to scientific advancement.

The Codebook

The content of a codebook and all the aspects related to the content such as which information is included, how it should look like, or which structure it should have should be guided in a way that all of the above purposes of using the dataset can be fulfilled.

Layout of a Codebook

A codebook can come in many different formats. In its simplest version, a codebook could be a plain text document or be integrated in the analysis code of the article. Other options are a word-document or a plain PDF file that includes all the relevant information. Although all of these versions of a codebook technically fulfill the purposes described earlier (i.e., anyone can use the data without having to obtain additional information from the author), useful codebooks follow three criteria:

(1) Human readable. A good codebook is readable and understandable by humans. That is, the information is presented in a way that allows anyone with slightly below average software skills to open the codebook and make use of the information provided;
(2) Machine readable. A good codebook should also be readable by machines. This means that the information provided can be accessed and summarized using automated approaches. This allows others to use the information in the codebook during data analyses itself. For example, the information in Table 1 could be used to recode variables from the data frame or to form scale composites by merging all items that belong to one scale. The codebook thereby becomes an element in the data analyses and data wrangling pipeline, thereby increasing transparency and reducing the possibilities for human error;
(3) Consistent. There are many ways codebooks can be structured, while it is still readable for humans and machines. It is important that codebooks follow a consistent structure. Above all other things, this means that each variable is referenced in the codebook, and that the same structure is used for all variables, and that there are no empty or undefined cells in the codebook.

Table 1. Overview of possible information to include in a codebook

Information	Description	Individual item	Scale score	Demographic	Technical
Name of the variable	The name of the variable, exactly as it is displayed in the dataset	BFI_extra_1	CFT-total	Gender	Time
Type of the variable	The type of the variable, usually numeric, factor, date, or character	Numeric	Numeric	Factor	Date
For questionnaire items					
Wording of the item	The original wording of the item, in its original language	„Ich gehe aus mir heraus, bin gesellig."	–	„Bitte geben Sie Ihr Geschlecht an."	–
English translation of the item	An English translation of the item (note, this does not need to be a validated one)	"I get out of myself, I'm sociable."	–	"Please indicate your gender."	–
Questionnaire/Source	The source of the item	BFI-2	CFT-20-R	Generated ad hoc	–
Dimension the item belongs to	If the item is part of a scale, the scale the item belongs to	Extraversion	–	–	–
Response format	The format on which the response was given.	5-point rating scale	–	Single choice	–
Response labels	The labels of the response scale. If this is too long, it can be added on an additional sheet in the codebook	response_list_bfi	–	response_list_gender	–
Theoretical minimum	Lowest theoretical value the variable could take	1	55	–	–
Theoretical maximum	Highest possible value the variable could take	5	160	–	–
Coding of item	If the item is coded in the direction of the construct or not	No	–	–	–
Optional item	Indicates if the item was optional or mandatory	Yes	–	No	–
Coding of missing data	Indicates how missing data are coded	−77	NA	–	NA
Description	A brief, verbal description of the variable, especially if not an item from a survey	–	Score of the dimension	–	The time the survey was taken by the participant
Concerning study design (Examples)					
Rating source	Indicates who responded to the survey	Self	–	Self	–
Assessment wave	In studies with multiple waves, when was the item taken	First assessment	Laboratory assessment	First assessment	–
Part of survey	In larger studies, indicate block or part of survey	Personality	Cognitive ability	Demographics	–
Additional information					
Link	If available, a link to an online source of the survey/item/measure	https://search.gesis.org/instruments_tools/zis247	–	–	–
Reference	Source of the survey or assessment tool	Danner et al. (2016)	Weiß (2019)	–	–
Comment	An additional comment in plain text, if required	–	–	–	–

Note. Information = Information that is presented in the codebook; Description = a description of that information; Individual Item = an example for an individual item from a survey; Scale Score = an example for a scale score from a questionnaire or test; Demographic = example for a demographic item; Technical = example for an item that is not responded to by the participant.

Table 2. Example on how to provide information about response options of Rating Scale Items

Response list	Coded response	Label	Translation
response_list_bfi	1	Stimme überhaupt nicht zu	Disagree strongly
	2	Stimme eher nicht zu	Disagree a little
	3	Teils, teils	Neutral
	4	Stimme eher zu	Agree a little
	5	Stimme voll und ganz zu	Agree strongly
response_list_gender	1	Männlich	Male
	2	Weiblich	Female
	3	Divers	Non-binary
	4	Keine Angabe	Prefer not to say

Note. response_list = the name of the list, which can then be referenced in each item that makes use of this scale; coded response = the numerical value in the data frame that corresponds to the "label"; translation = an English translation of the label.

Information That Must Be Provided in a Codebook

The codebook should *at least* list all variables that are included in the dataset, but, for some purposes, it can also include other variables. It should make a reference to all variables that are included in the published data of the article and that are analyzed in the study. Usually, if the codebook is a table, one row in the codebook refers to one column in the data frame and ideally both are in the same order. In a data frame, rows typically contain participants, and columns contain variables. There may be a few exceptions, but the principles laid out here should be adaptable to most, if not all, other cases. In Table 1, we provide a list of entries that a codebook could have including different examples. Note that not all of this information needs to be provided, especially if additional materials are shared alongside the manuscript (e.g., if one can reference the surveys or tools, then not all of the information about the items need to be shared in the codebook). At the same time, it can be possible that some items or stimuli of a study cannot be shared due to legal restrictions, especially copyright. Our examples given in Table 1 therefore serve as an inspiration – as each case is unique, authors will have to decide themselves which information they can share and which they cannot share.

Response labels. One of the most commonly used item formats in psychological assessment is probably the rating scale (Baumeister et al., 2007; Wetzel & Greiff, 2018), which consists of an item and response labels (e.g., 1 = *agree*, 5 = *disagree*). This information should also be included. One practical and efficient way to include this information has been implemented by Arslan and colleagues (2020). In their platform, https://formr.org, response labels are defined once and then re-used throughout the survey. For codebooks, a similar approach could be feasible. This information can be displayed in a separate spreadsheet as exemplified in Table 2.

Additional information that can be provided in a codebook. A codebook must not only contain information about the data that were used in the study, but can also contain information about the data that were not used in the study. When designing a study and collecting data, it could be useful to have one codebook for the overall dataset. Whenever a subset of data from this dataset is published, the entire codebook is shared alongside the publication. This does not only save resources (when preparing the data for publication) but it also shows other researchers which additional information would in principle be available. This can be useful for reviewers if they would like to suggest additional robustness checks or for other researchers searching for alternative explanations for specific results. It can furthermore generate interest for collaboration, while the authors keep control over their yet unpublished data.

Formatting Principles

Most codebooks, and the one included here as an example, are Excel files or comparable spreadsheets. Both for the dataset and the codebook, some general rules apply. Broman and Woo (2018) lay out general principles on how to store information in a spreadsheet, and we strongly recommend this article for anyone starting to curate their data in a spreadsheet. However, some principles are especially relevant for the preparation of a codebook and the dataset.

Naming variables. For mainly historical reasons, some researchers still tend to use highly abbreviated and sometimes obscure variable names (e.g., to fit the arbitrary 8 character limit that Mplus used to require). The current state-of-the-art, however, is to go for clarity rather than brevity, for instance *participant_id*, *start_date_time, bfi10_extraversion_2R, education_level_highest_obtained, occupational_status*, and so on. Unfortunately, there are no commonly agreed upon standards, although a collaborative initiative (https://github.com/psych-ds/psych-DS) aims to develop them. Most commonly used data collection software allows researchers to define meaningful variable names already while setting up the study, which frontloads the effort and can avoid confusion on part of the original team as well as other researchers who seek to reuse the data. Research published in the *EJPA* should use English variable names.

Researchers differ in whether they use an R at the end of an item to denote whether an items *should be* reverse-coded or already *has been*, we suggest an explicit approach to avoid confusion: When sharing data with reverse codings, reverse-code the values and add an R at the end of the variable name to show that this has been done, then add value labels that show that the verbal anchors for the numeric representations are reversed (e.g., 1 = *strongly agree*, 5 = *strongly disagree*).

Avoid color coding. Coding cells, rows, or columns in certain colors can be used to enhance the organization of the spreadsheet. However, it should not be used as additional information. Colors should only be used to give structure and guidance, and no essential information should be lost when stripping all colors off the spreadsheet. The same is true for any additional data that are not stored in a cell of the spreadsheet, such as a comment.

No empty undefined cells. All cells in a codebook should be filled and contain explicit information. Even if a certain statement does not apply (see, e.g., Table 1), then this should be made explicit either by writing "n/a" or specifying what empty cells signify. Users will thereby be informed that the information is not missing due to a potential oversight, but that the information is not applicable to a certain type of variable.

Just one information per cell. Each cell of a codebook should only contain one piece of information. For example, the item and the number of the item in the survey are oftentimes presented in the same cell, whereas these should be split in two cells. The same is true for the anchors of rating scales that should be presented in two or more columns. Merging information from two cells is usually straightforward, whereas splitting information is oftentimes difficult.

Be clear about missing values. Certain formats, such as SPSS or Stata files, allow for labeled missing values, yet we often see data where the typical missing values (−1, 999, etc.) are not properly marked up. These are the exact pitfalls that make data hard to reuse without prior knowledge. When labeled missing values are required, researchers should ensure they are used correctly. In many cases, simply coding missing values explicitly is preferable. Datasets exported in .R as .csv files distinguish between empty strings and missing values, representing the former as "" and the latter as simply nothing. Researchers should try "round-tripping" their datasets (i.e., exporting it and re-importing it) to ensure this important information is not lost (or that no new information is added, which can also be the case). Here, more complex formats (such as .rds or .dta) can help preserve metadata more fully. As there is no reason to share a dataset only in one format, it can be smart to share both a simple .csv file and a more complex format with richer metadata.

Long versus wide format. Data that contain multiple observations per participant can be provided in two different forms, in a wide or a long format. Wide format means that the data frame contains a variable for each observation (e.g., extraversion at time 1, extraversion at time 2, etc.). In long format, each participant would have several rows. The first row of participant 1 would contain their extraversion score at time 1, the second row their extraversion at time 2, and so forth. Here, any additional variable would then indicate the measurement occasion, and another variable would indicate the identifier (id-variable) of the participant. Whenever data from repeated assessments are shared, the long format is generally preferable, as each variable (e.g., extraversion) must only be referenced once in the codebook.

Additional Information

Not all information that enables independent researchers to re-use data must necessarily be provided in a codebook. Additional information can be supplied in a ReadMe file or as metadata for a codebook.

ReadMe

One file that, from our experience, is often forgotten although helpful, is a ReadMe file.[3] A ReadMe file is a plain text file named *ReadMe.txt* that can be added to a project (i.e., a set of combined folders that are related to one publication, such as data, materials, analyses scripts, etc.) and contains plain text information on how to navigate a project. The idea is that when someone opens a project folder, the first file they are supposed to look into is the ReadMe file. The readme could contain information about which software to use, how to load the data into the software, a brief verbal description of each item (i.e., data file) in the project folder, or a license. Even if a ReadMe only contains basic information, it still lets the user know that they have not overlooked anything. Finally, ReadMe files do not have to be plain text files, the same information can also be included elsewhere (e.g., in the wiki of the project or on a separate website). However, we generally recommend adding a plane text ReadMe so that everything is kept in one place.

Generating Metadata

Basic metadata for a dataset makes it (a) uniquely identifiable, for example, using a document object identifier (DOI)

[3]Further information and a generic example can be found here https://data.research.cornell.edu/content/readme

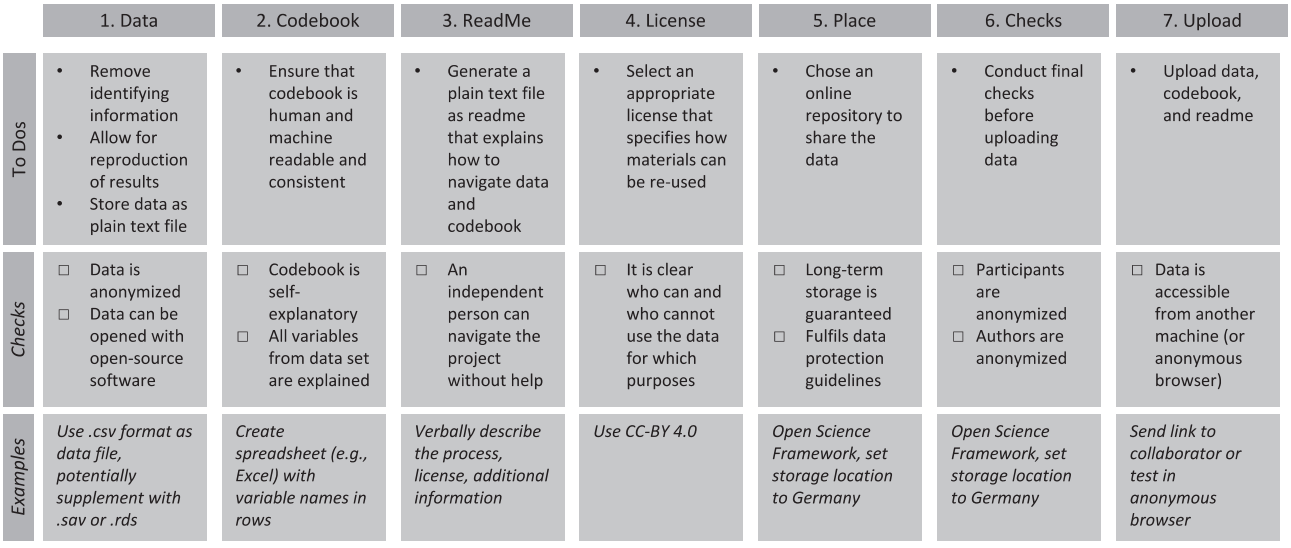

Figure 1. The process of data preparation. The process of data preparation for online release, including checks and examples.

or another permanent address, so that multiple copies of a dataset can be identified, (b) specifies the terms of use, such as a license, a way to get access, and a preferred citation (c) specifies the time and location the data were collected (e.g., to enable post-hoc aggregation of timeseries during the first months of the SARS-CoV-2 outbreak) and published. Oftentimes, people additionally name and describe the dataset and variables collected.

Some data repositories, such as IPUMS (Integrated Public Use Microdata Series, https://www.ipums.org/), allow researchers to search inside the variable names and labels of the data stored in their repository, but the Open Science Framework (OSF), which has recently been favored by psychologists, does not possess this functionality yet. In addition, Google has recently rolled out its Dataset Search (https://datasetsearch.research.google.com/). Datasets are spread across the internet and this new search engine aggregates them wherever they are found and properly marked up using a specific standard, documented on https://schema.org/Dataset. Although a JSON-LD document describing a dataset in a machine-readable fashion can, in principle, be written by hand, most of the requested information is already part of the documentation in tabular form that researchers tend to write for human readers. The codebook R package (Arslan, 2019) and the accompanying website (https://codebook.formr.org) make it easy to add variable and value labels once, and then use them to generate documentation for humans and search engines.

I Want to Share My Data – and Now What?

When preparing a dataset for release, the following, practical steps can be taken. Note that there are numerous other descriptions, some more sophisticated (e.g., Arslan, 2019), but the steps described here will serve the purposes described earlier, that is, enabling independent researchers to understand the provided data. Figure 1 gives an overview of all steps.

Check Your Starting Position

Much of what has been said here may cause the impression that one may have to "downgrade" their existing files, maybe from an SPSS file (.sav) or Stata file (.dta), which may already contain labels, and split it into a .csv and an excel sheet. This is not necessarily our intention. Instead, these files may already fulfill the purposes described earlier. Furthermore, it is possible to transform such files into human and machine readable codebooks using the codebook package (Arslan, 2019) and the accompanying web-app.

Preparing Data

The dataset that is used in the publication should be edited such that (a) all identifying information of the participants is removed, and that (b) all key analyses can be reproduced. Distinguishing identifying information from non-identifying information can sometimes be challenging, but in most cases, it is not. E-mail addresses, IP-addresses, Names, unique participant ids, and so forth should be removed. Test-scores on psychological tests can generally be retained, and demographic information (age, gender) as well. Preparing such a shareable version of the dataset right at the beginning of a project (i.e., even before the first substantial data analyses) is highly recommended.

The dataset should then be transformed in a plain text format, such as .txt or .csv. Providing the data in a plain text format is the minimum, but there is no reason not to

provide additional formats (e.g., .rds or .sav), if you think they will make it easier to re-use the data. If an analyses script for the data analysis is shared, it should start with loading the same dataset that is shared alongside the manuscript.

Generating a Codebook

Based on the so-obtained dataset, one can then generate the structure of a codebook. To get started, all variable names of that dataset can be saved as a new variable in a spreadsheet. The columns can then be named based on the suggestions in Table 1. Afterwards, all cells in the so-created spreadsheet should be filled out. We provide an excel sheet with a general template that can be adapted (https://osf.io/nerpa/).

Generating a ReadMe

Next, a readme should be generated, describing the process of how to obtain the data and how to get started. We provide an example ReadMe (https://osf.io/nerpa/).

License

One important aspect when releasing a dataset is choosing a license that indicates how the data can be re-used and distributed. This license can be added to the readme file. A widely used license is, for example, the CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/), and data from publications in *EJPA* must at least be as open as specified under CC BY.

Place to Share

Before sharing their data, authors need to decide where to share their data. PsychArchive (https://www.psycharchives.org/) is a preferred repository for the *EJPA*, run by a German research institute. Other alternatives for sharing research data have been summarized by Gilmore et al., (2018). Refer also to Table 1 for some examples. Of course, other platforms are also acceptable, as long as the data are licensed such that they can be re-used. If possible, a digital object identifier (DOI) should also be created at this point. This can then be referenced in the article and in the ReadMe.

Check Anonymity

Just before uploading the files to an openly accessible platform, authors should conduct two final checks. First, they should check if participants in their data are no longer identifiable. Second, they should check – for blind review purposes – if they themselves are not identifiable. This information can be hidden in the meta-data of the files, in the ReadMe or sometimes even in the briefing and debriefing that are given to the participants and may be included in the codebook or elsewhere.

Upload

Finally, the authors can upload the data to their chosen repository. Here it is once again advisable to check if the data are really accessible, for example, by asking a collaborator to use the provided link and see if they can access, download, and use the data.

An Ideal World of Data Sharing

The structure of the codebook and also the dataset described here fulfill the purposes of open data and are sufficient for *EJPA*. At the same time, this is only a bare minimum (comparable to Level 1 of the TOP guidelines; Nosek et al., 2015), and it would be desirable to have (at least psychology-wide) standards for data sharing. Similar to the standard format of a manuscript (i.e., title, abstract, introduction, methods, results, discussion, references, footnotes, tables, figures, appendix; see American Psychological Association, 2020), discipline wide standards would facilitate discovering, (re-)using, and interpreting datasets. Furthermore, this would allow the automatic integration of datasets across studies. However, these guidelines would have to be developed (which is a huge task in itself), and then they had to be accepted, implemented, and controlled (which is an even bigger task). As long as these common frameworks do not exist, it would be too much to ask authors at *EJPA* to prepare their data in a *(journal-)specific* format. We also acknowledge that many authors may have already curated their data in some form or other. *EJPA* has therefore decided to pursue a *your data your way* approach, as long as the minimum requirements outlined here are met. In an ideal world, however, the curation of research data should receive the same – or maybe even more – attention to detail as the preparation of a manuscript. We hope that this editorial can provide a first step toward such practices.

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association. https://doi.org/10.1037/0000165-000

Arslan, R. C. (2019). How to automatically document data with the codebook package to facilitate data reuse. *Advances in Methods and Practices in Psychological Science, 2*(2), 169–187. https://doi.org/10.1177/2515245919838783

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). formr: A study framework allowing for automated feedback generation and

complex longitudinal experience-sampling studies using R. *Behavior Research Methods, 52*(1), 376–387. https://doi.org/10.3758/s13428-019-01236-y

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Blanke, E. S., Kalokerinos, E. K., Riediger, M., & Brose, A. (2020). The Shape of Emotion Regulation. *European Journal of Psychological Assessment, 36*(3), 447–455. https://doi.org/10.1027/1015-5759/a000586

Borghi, J. A., & Van Gulick, A. E. (2018). Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLoS One, 13*(7), Article e0200562. https://doi.org/10.1371/journal.pone.0200562

Brandt, N. D., Becker, M., Tetzner, J., Brunner, M., Kuhl, P., & Maaz, K. (2020). Personality across the lifespan. *European Journal of Psychological Assessment, 36*(1), 162–173. https://doi.org/10.1027/1015-5759/a000490

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician, 72*(1), 2–10. https://doi.org/10.1080/00031305.2017.1375989

Danner, D., Rammstedt, B., Bluemke, M., Treiber, L., Berres, S., Soto, C., & John, O. P. (2016). Die deutsche Version des Big Five Inventory 2 (BFI-2) [The German version of the Big Five Inventory, BFI-2]. In *ZIS – Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis247

Gilmore, R. O., Kennedy, J. L., & Adolph, K. E. (2018). Practical solutions for sharing data and materials from psychological research. *Advances in Methods and Practices in Psychological Science, 1*(1), 121–130. https://doi.org/10.1177/2515245917746500

Greiff, S., van der Westhuizen, L., Mund, M., Rauthmann, J. F., & Wetzel, E. (2020). Introducing new open science practices at *EJPA*. *European Journal of Psychological Assessment, 36*(5), 717–720. https://doi.org/10.1027/1015-5759/a000628

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science, 5*(8), Article 180448. https://doi.org/10.1098/rsos.180448

Heintz, S. (2019). Do others judge my humor style as I do? *European Journal of Psychological Assessment, 35*(5), 625–632. https://doi.org/10.1027/1015-5759/a000440

Hendrick, T. E. (1988). Justifications for the sharing of social science data. *Law and Human Behavior, 12*(2), 163–171.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary Educational Monographs, 48*, xi + 91.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), Article e1002456. https://doi.org/10.1371/journal.pbio.1002456

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), Article e0021. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2020). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment, 36*(1), 149–161. https://doi.org/10.1027/1015-5759/a000481

Weiß, R. H. (2019). *CFT 20-R mit WS/ZF-R* [CFT 20-R with WS/ZF-R]. Hogrefe.

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales. *European Journal of Psychological Assessment, 34*(1), 1–5. https://doi.org/10.1027/1015-5759/a000469

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*(7), 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.-B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18

Published online November 30, 2020

Open Data

Data can be accessed on Open Science Framework at https://doi.org/10.17605/OSF.IO/NERPA.

Kai T. Horstmann
Institute of Psychology
Humboldt-Universität zu Berlin
Rudower Chaussee 18
12489 Berlin
Germany
kai.horstmann@hu-berlin.de