Draft for a chapter in the *Handbook of Language and Cognition* (Bloomsbury), edited by Stefan Hartmann and Laura J. Speed.

Reproducibility, replication, and preregistration

Lukas Sönning (University of Bamberg; lukas.soenning@uni-bamberg.de)
Timo Roettger (University of Oslo)

Abstract. The replication crisis has raised our awareness that empirical findings may be much less robust than they seem; indeed, some highly influential psycholinguistic studies have seen failures to replicate. This shows that there are fundamental flaws in the way we (used to) go about doing empirical work. The current chapter discusses three methodological hallmarks that have emerged from the critical discourse surrounding the replication crisis: reproducibility, replication, and preregistration. It has become clear that they deserve to be regarded as integral elements of the scientific process, and the aim of this chapter is to help authors improve their work in the light of these insights. Following some initial terminological clarifications, we first turn to various aspects related to reproducibility, ranging from the FAIR principles for data sharing and data publication to computational tools and workflows ensuring the recoverability of statistical results. The focus then shifts to replication and replicability, with an outline of the continuum from direct to conceptual replication as well as advice on how to facilitate attempts by others to replicate our work. We also consider design, analysis and interpretation strategies that may be used to increase the replicability of our statistical conclusions and the linguistic claims they inspire. Finally, we discuss the essential role played by preregistration, where the researcher specifies their linguistic hypotheses and data analysis protocol prior to data acquisition; this form of academic self-discipline has culminated in the emergence of Registered Reports, a new publication format that is beginning to gain ground across the linguistic and cognitive sciences.

Keywords: replication, reproducibility, preregistration, registered report, open science

1. Introduction

Research on language and cognition is, and has perhaps always been, strongly empirical. As a result, the methodological discussions surrounding what has come to be called the replication crisis (e.g., Ioannidis 2005; Open Science Collaboration 2015) are particularly relevant to the target audience of this handbook. A positive outcome of this crisis is that it has caused us to reflect on and revise our methodological practices. The present chapter deals with three themes that have received particular attention: reproducibility, replication, and preregistration. A careful engagement with these topics is desirable, since they offer important guideposts in our search for valid and generalizable insights about how language is processed and represented in the mind. Further, each of these keywords has concrete implications for the way we go about doing empirical work, and the past ten years have seen an emergence of several particularly helpful resources and platforms.

The current chapter primarily addresses the needs of cognitive linguists who are interested in how the scientific virtues of reproducibility and replicability as well as the methodological cornerstone of preregistration may inform their work. This means that we provide practical insights into various tools and strategies that help us establish sustainable research output with pointers to further reading where appropriate.

The chapter starts with some notes on terminology. Section 3 then deals with reproducibility. It sets out with some fundamentals such as the FAIR principles for data sharing and essentials for reproducible workflows and works its way up to more elaborate infrastructures such as data publication, standardized reproduction documentation, and version control. In Section 4, we turn to replication, starting with a distinction between different types of replication studies. This is followed

by some thoughts on how to enable others to replicate our work and how to increase the probability that such a replication study will be successful. The topic of preregistration is covered in Section 5, which introduces this newly adopted practice, its benefits for researchers, as well as practical advice on how to use existing resources to create and archive preregistrations. Section 6 concludes this chapter.

2. Terminology

Let us start by clearly delineating the way in which we use the terms 'reproducibility' and 'replicability' in the present chapter.[1] Reproducibility means that the findings of a study can be repeated using the *same* data and the *same* analysis procedure(s). Replicability, on the other hand, is more ambitious: It means that the findings can be repeated with *new* data. As we will see in Section 4, various types of replication studies can be distinguished, depending on how much the new set of data differs from the original one.

Unfortunately, there is terminological variation across disciplines, which means that the meaning of these labels (reproducibility vs. replicability, and also reproducible vs. replicable) is sometimes swapped. Our usage reflects the original distinction (see Barba 2018 for a review). It should be noted that the term 'replication' also appears in the literature on the design and analysis of experiments; there, it has a completely different meaning and refers to the number of repetitions that are collected for a specific condition in an experimental plan. The term "pseudoreplication" (e.g., Hurlbert 1984) is tied to this specific usage.

With these terminological remarks in place, let us now turn to the first and most basic issue, the reproducibility of statistical results.

3. Reproducibility

A finding is reproducible if it can be duplicated using the original data, either by the original researcher(s) or third parties. Section 3.1 discusses the importance of reproducibility in scholarly work. One requirement is the availability of the data, and Sections 3.2 and 3.3 will discuss principles and tools for data sharing. Reproducibility checks are greatly facilitated if analysis procedures are documented appropriately, ideally in the form of code. Sections 3.4, 3.5, and 3.6 deal with various aspects related to writing, organizing, and publishing analysis code.

3.1. The relevance of reproducibility

Professional ethics directs us to prevent the scientific record from being cluttered with unsubstantiated claims. Reproducibility is therefore among the minimal requirements for empirical work – if published results cannot be recovered from the data, linguistic conclusions may be invalid. We know from our own experience that human error creeps in all too easily, which has confronted us forcefully with the need to introduce reproducibility and error checks into our workflow. Perhaps as a natural consequence of this, we have also found ourselves giving higher credibility to scholarly work that is based on practices outlined in the following sections.

While implementing these additions may slow our progress initially, it will save time and resources in the long run. This is not only true for editors and reviewers, whose precious time will not be

[1] This usage is in line with that documented in the FORRT glossary (*Framework for Open and Reproducible Research Training*; Parsons et al. 2022).

wasted on processing unsupported papers or evaluating poorly documented analysis procedures, but also for our future selves: It is much easier to return to and modify a carefully documented, bug-free analysis plan in response to new insights or reviewer feedback.[2]

3.2. Sharing data: The FAIR principles

Essential guidance for sharing data is provided by the FAIR principles (Wilkinson et al. 2016), which is short for *findable*, *accessible*, *interoperable*, and *reusable*. From a practical perspective, the process of data sharing usually involves three steps. The first is to ascertain whether there are legal or ethical issues that prohibit the sharing of the data at hand in a certain form. Next, the data need to be brought into an appropriate shape and finally we need to decide on a platform for dissemination. These steps will be dealt with in turn.

It is important to consider legal[3] and research-ethical concerns at the design stage of a study, and to consult the research data management service and ethics board at your institution for advice and approval. Proper planning will greatly facilitate data sharing and publication. For instance, the participant consent form should standardly include a tick box where the person actively grants permission to the publication of their data in an anonymized form. For examples illustrating how this may be phrased, please refer to Web appendix 1 (https://osf.io/8v7wf). Even if you have obtained consent for data publication, it is nevertheless important that they are anonymized if possible.[4] If raw or even derived data cannot be shared, creating synthetic datasets can be an option. These preserve the quantitative relationships between variables without including any identifying information (e.g., Quintana 2020).

Once legal and research-ethical concerns are taken care of, the data need to be converted into a format that ensures long-term reusability on a wide range of platforms; they should be *interoperable*. This is achieved by using persistent and open-source file types instead of proprietary formats that are tied to specific software. For data tables, a sensible default is a UTF-8 encoded tab-separated raw text file. For instructions on how to arrive at this format, please refer to Web Appendix 2 (https://osf.io/jchgy). An advantage of tab separation over comma separation (.csv) is the fact that the latter can invite errors when switching between operating systems that differ in their marking of decimal places (point vs. comma). Further, the organization of data in a spreadsheet should abide by certain principles: (i) one data table = one file, (ii) one row = one observation, (iii) one column = one variable, and (iv) one cell = one value (for more details, see The Touring Way Community 2025[5]).

Of course, data tables are not the only file type that is necessary to provide a thorough documentation of an empirical study. For PDF files, for instance, it is important to use the archivable PDF/A format. Table 1 provides examples of preferred (and undesirable) formats for a number of different file types.

---

[2] See https://book.the-turing-way.org/reproducible-research/overview for a more thorough discussion of the advantages of and barriers to reproducible research.

[3] Since legal guardrails will vary across countries, it is essential to consult applicable laws (see, e.g., https://gdpr.eu/ for the European *General Data Protection Regulation*) and official support services (e.g. https://forschungsdaten.info for the German context).

[4] There are various tools available that allow for automatic anonymization of quantitative and qualitative data, e.g. amnesia (https://amnesia.openaire.eu/) or QualiAnon (https://www.qualiservice.org/de/helpdesk/webinar/tools.html).

[5] https://book.the-turing-way.org/reproducible-research/rdm/rdm-spreadsheets

Table 1. Preferred file types for sharing data, code, and other materials. Based on the TROLLing deposit guidelines.[6]

File type	Preferred format	Undesirable format
Image	.tif .tiff .png .jpg	.psd .pct .gif .raw .bmp
Slide/illustration	.pdf (PDF/A)	.pptx
Spreadsheet	.txt .csv .tsv	.xlsx
Text	.txt .md .pdf (PDF/A)	.docx HTML
Statistical analysis	.R .RData (R) .dat/.sps (SPSS) .dat/.DO (Stata)	.por (SPSS Portable) .sav (SPSS) .dta (Stata) .7dat, .sd2, .tpt (SAS)

Once the data are in a form that ensures interoperability across systems, additional steps need to be taken to also make them *reusable*. The first is to provide detailed documentation of the contents of the data table in a separate text file, which is often referred to as a codebook or data dictionary. This includes a description of the columns (i.e. variables) in the table and a key to the labels or abbreviations that are used to represent them. This will also involve explanatory notes on technical terms, references to external sources, or a transparent documentation of the methods and materials used to collect the data. Finally, an appropriate license needs to be specified to inform others about the terms for reuse.

To ensure that the data are *accessible* for other scholars, a plain-text ReadMe file should describe the format of the data tables, how to access and open files and which software to use. Finally, in order for data to be *findable*, they should be hosted on a suitable platform and tagged with rich metadata to make them discoverable (e.g. in the form of keywords, information about the author(s) of the materials, alongside permanent contact information). A detailed comparison of different repositories is provided by Stall et al. (2025). Straightforward options are the Open Science Framework (OSF, osf.io) or Zenodo (zenodo.org), which allow for the specification of a number of metadata fields.

Aligning data with the FAIR principles and making them openly available is a major contribution to the sustainability and reproducibility of your work. In the next section, we will go one step further and consider the peer-reviewed publication of linguistic datasets.

3.3. Data publication: The TROLLing archive

Data can alternatively be published in a dedicated data repository.[7] While this involves some extra effort, it also holds out many benefits both for the individual researcher as well as other stakeholders such as journal editors and scholars building on the findings of your study. For illustration, we will consider the Tromsø Repository of Language and Linguistics (TROLLing), a linguistic data archive associated with the Dataverse project[8] and based at UiT The Arctic University of Norway. The main advantage of using TROLLing is that a dataset undergoes a proper review and curation process,

[6] https://site.uit.no/dataverseno/deposit/prepare/

[7] See https://www.re3data.org/ for a registry of research data repositories.

[8] https://dataverse.org/

which ensures that it is aligned with the FAIR principles and documented thoroughly. Let us consider the publication process in some detail.

To publish a dataset, the researcher creates what is referred to as a TROLLing post, which includes a bundle of files and metadata. The heart of a post is its ReadMe file, which is prepared based on a standardized template.[9] This raw text file provides orientation and background information. It includes, among other things, a dataset abstract, details on the methods used, and a codebook for each data table, i.e. a key to its contents. The TROLLing website provides researchers with detailed guides on how to prepare and deposit their data.[10]

Upon submission, the dataset is subject to peer review by a trained data curator at TROLLing. The researcher receives feedback in the form of a curation report, which draws their attention to various features that may require specification, revision, or elaboration. This includes checks with an eye to the FAIR principles (e.g. the use of proper file formats, see Table 1), feedback on the form, consistency and transparency of the documentation provided in the ReadMe file, and advice on relevant legal and research-ethical aspects. Since TROLLing is dedicated to the language sciences, the curation service is able to offer substantive feedback on the documentation of data and methods. Finally, the dataset receives a license that clearly states the terms for (re)use as well as a DOI for persistence and proper citation.

We can tell from our own experience that the additional effort involved in data publication, as well as the review process it undergoes, has the advantage that everything undergoes a thorough quality check prior to publication, and indeed, prior to running the final analyses reported in a paper. We now move on to tools and resources that help with establishing a reproducible data analysis pipeline.

3.4. Reproducible workflows: Some basic principles

In this section, we will only concentrate on the use of statistical programming languages such as R or Python. On the one hand, this is because there is a clear trend toward the use of these environments for data analysis; on the other, code-based software is the ideal vehicle for making an analysis reproducible. In this section, we will cover a few fundamentals for ensuring reproducibility.

The first principle is that a data analysis project for a specific study should always be self-contained. This means that there is a single folder on your computer which includes everything your analysis relies on (typically data and analysis scripts) and everything it generates (tables, figures, and sometimes complete reports). If you are using an RStudio project, for instance, it is likewise located in this folder and operates from there. This makes it easier to keep things together and in order, and it facilitates sharing and archiving (and potentially version-controlling) the analysis project.

The second principle is that every step of the analysis should be preserved in code form. Most importantly, this means that the script includes code that loads the data, and code generating and exporting graphics. Novices sometimes make use of the point-and-click facilities in RStudio to load their data; we do not recommend this practice. Since a code-form paraphrase of such commands will appear in the console, researchers can quickly learn to spell out and preserve these instructions in their scripts.

Third, each step in the data analysis pipeline should be documented and explained. This means that an analysis script will consist of both code as well as explanations in the form of comments. At a

---

[9] See Conzett & Dijkstra Haugstvedt (2024): https://doi.org/10.5281/zenodo.10849096

[10] https://site.uit.no/dataverseno/deposit/

minimal level, this includes brief notes on what the following code chunk does. Dynamic documents such as Jupyter or Quarto/RMarkdown notebooks allow for a closer integration of code and prose, which makes it much easier to document analyses.

The final principle we will mention is that everything should always be reproducible from data and code. This means that at any point during the time you spend working on your project, it is set up in a way that the whole analysis pipeline runs through automatically, starting with setting up R (loading relevant packages) and loading the data, followed by initial data analysis using tables and graphs, and then (possibly) statistical modeling. This avoids certain opportunities for human error and it ensures that the output of your analysis is reproducible at all times. In RStudio, for instance, this means that you do not rely on anything that is loaded into your workspace. The workspace should always be cleared automatically when ending an R session.

3.5. Analysis code: Two formats

The code that is necessary for running and reproducing an analysis can be written, maintained, and shared in two formats: as a raw script or a dynamic document. In this section, we will discuss these basic formats and mention some of the benefits of dynamic documents.

Analysis scripts are raw text files that contain code as well as notes that must be marked explicitly as non-code; in R, this is done by starting a line in the script with a pound sign (#). The raw-text format of scripts does not provide much room for structuring the analysis and makes it cumbersome to include detailed explanatory notes. Nevertheless, from a reproducibility point of view raw scripts are a perfectly appropriate format.

Dynamic documents, which are also called (computational) notebooks and represent what Knuth (1984) refers to as literate programming, offer many advantages. Examples of this format are Jupyter notebooks or Quarto (formerly RMarkdown) documents. Both allow you to interweave explanatory notes, code, as well as its output (tables and graphs) into a single document. A direct benefit is that this makes it easier for the researcher to keep track of their analysis. Further, headers and sub-headers as well as various other formatting options (use of italics, bold print, and bullet points) help generate clearly structured documentation. Dynamic documents may then be exported into a range of formats, including a PDF or an html file, which are more reader-friendly compared to a raw-text file. For these (and several other) reasons, computational notebooks are becoming the de facto standard for the documentation of statistical analyses.

Irrespective of the specific format you use, it often makes sense to compartmentalize an analysis into different files, e.g. one for data cleaning and preparation, one for running the statistical analysis, one for producing figures and tables, etc. If the structure of a data analysis project becomes more elaborate, it may be worthwhile to consider investing additional effort into the documentation of the code and the general workflow. Two rather advanced options are discussed in the next section.

3.6. Advanced reproducibility: The TIER protocol and version control

This section briefly presents two state-of-the-art resources for reproducible research: The TIER protocol, which is a unified guide for reproduction documentation, and the use of Github for version control.

Anyone who has turned to the supplementary materials of another study to try to reproduce (parts of) the analysis will very likely have spent quite some time finding their way around the archived

files. The TIER protocol (current version: 4.0)[11] was conceptualized and written with the needs of the reproducer in mind. It is a detailed scheme for documenting a statistical data analysis project in a standardized way, to allow others to easily find their way around a project directory. The proposed protocol suggests a default directory structure as well as a standard set of components such as a data cleaning pipeline, a data dictionary, and scripts running the analysis. Importantly, the documentation includes a ReadMe file, which is meant to be the first file opened by the person who wants to repeat the analyses. This file provides instructions on how to run the code and actively helps others to navigate the directory. The second general component is what is referred to as a master script, which in some sense represents the command center of the project. It calls and runs the set of scripts in the appropriate order, which means that reproducing the analysis (ideally) simply involves running this script. Overall, the TIER protocol includes many useful tips and ideas, and may serve as a source of inspiration when developing your own workflow.

Another tool that has been adopted for reproducible research is version control using Github. While this platform was developed with the needs of software programmers in mind, it is beginning to play an increasingly important role in academia. In general, the version-control features offered by Github allow you to keep an external backup of the analysis code, which also includes the equivalent of tracked changes. These features provide an additional safety net, and they allow you to return to an earlier stage of your code. While this comes at the cost of relatively counterintuitive operation, the RStudio environment makes it relatively straightforward to use Github for version control.

3.7. Recommendations for further reading

For more background on establishing reproducibility, we can make the following recommendations:

- The Turing Way handbook[12] and Schweinberger (2025)[13] provide information on various aspects related to reproducibility.
- Strand (2025) gives many valuable tips for avoiding errors in laboratory research.
- Andreassen (2022) has more background on data archiving, and practical advice may be found on the TROLLing website.[14]
- The Tier protocol 4.0 is documented in detail online.[15]
- Numerous online materials illustrate the use of Quarto/RMarkdown and Jupyter notebooks.

The reproducibility of quantitative findings is essentially a prerequisite for the hallmark of replicability, to which we now turn.

4. Replication

We will start by delineating different types of replication study, and will then ask what researchers can do (i) to make it possible for others to attempt to replicate their research (Section 4.2), and (ii) to increase the probability that their findings are replicable (Section 4.3).

---

[11] https://www.projecttier.org/tier-protocol/protocol-4-0/
[12] https://book.the-turing-way.org/
[13] https://ladal.edu.au/tutorials/repro/repro.html
[14] https://site.uit.no/trolling/
[15] https://www.projecttier.org/tier-protocol/protocol-4-0/

4.1. Terminology

Often, a distinction is made between two forms of replication study.[16] A direct replication (also called exact/precise/literal replication) makes an effort to duplicate the original study as closely as possible. This means that apart from a different team of investigators and a different sample of human subjects (but from the same target population), everything is aligned as closely as possible with the original study. The basic goal of a direct replication study is to probe the generality of a finding beyond the original sample of participants, to other members of the same target population.

The other major form that is commonly recognized is a conceptual replication (also called construct replication). This type deliberately deviates from the original in one or several additional ways, i.e. beyond the sampling of new subjects. For instance, the repetition of a cognitive-linguistic study using speakers of a different language would classify as a conceptual replication. While this type of replication study represents a more severe probing of the generality of the original finding, it is clear that the two forms are best considered as endpoints on a continuum. To be able to interpret a conceptual replication (whether it is successful or not), an (ideally) exhaustive listing of the differences between the replication and the original study must be provided. Aspects in which the two studies diverge (e.g. differences in method, target (sub)population, or materials) then point to dimensions across which the original findings do or do not generalize. An unsuccessful conceptual replication becomes difficult (some would argue: impossible) to interpret, however, if the studies differ on multiple dimensions. Initially, direct replications should therefore take precedence over conceptual replications.

4.2. Enabling others to replicate your work

Any study should be carried out and documented in a way that allows others to carry out a direct or conceptual replication. In general, this involves three things: documentation of (i) data collection, (ii) data analysis, and (iii) the target of inference.

The first requirement is a thorough statement of the methods used for data collection. Depending on the specific field of study, this may involve details about participant recruitment (or sampling), the materials used, and measurement procedures. If the data are archived in accordance with the principles outlined in Section 3, the ReadMe file or the data dictionary may host additional notes to complement and extend the information provided in the methods section of a publication. The general goal is to enable others to arrive at a set of data that are as nearly comparable to yours as possible, apart from covering a different sample of participants.

The second requirement is to provide details about the analysis of the data, including any intermediate steps that were taken. These could involve data transformation, the exclusion (or "correction") of outlier values, the imputation of missing data, or other operations that will affect the data that inform the results of a study. Ideally, the analysis pipeline is documented step by step in the form of reusable code or a dynamic document, as described in Section 3.

Finally, the original study must clearly identify the target of inference, i.e. the population to which its results are meant to generalize. This will help a direct replication study to decide on how to sample new subjects, and it will allow a conceptual replication study to recognize relevant differences from the original work. We will return to this point in the next section.

Enabling others to duplicate your study as precisely as possible ensures that a failure to replicate is not due to insignificant procedural details. In some sense, it is a first step toward increasing the

[16] See, e.g. the FORRT glossary (https://forrt.org/glossary/).

chances of a successful replication. In the next section, we look at additional strategies that may be pursued to this end.

4.3. Increasing the replicability of your work

The replication crisis has been a forceful reminder that the inferential procedures used in a specific study can produce over-optimistic views of the accuracy and generalizability of statistical conclusions. While methodological reform movements have been primarily concerned with the reasons for replication failures, we can change perspective and ask what we can do to increase the chances that our findings will replicate. In this section, we discuss four aspects that contribute to this goal. We will use the shorthand label "effect" to refer to the association, difference, or pattern of interest, without wanting to imply that it reflects a causal relationship.

The first is to increase the a-priori likelihood of observing the effect that is reported – Ioannidis (2005: 699) refers to this as the "pre-study odds". There are broadly two lines of evidence that can be leveraged to increase the probability of observing an effect. The first is empirical, and it refers to an existing body of findings that may conspire to provide a basis for relatively firm expectations. This is to say that a study is ideally embedded in an incremental, cumulative research program. The second line of evidence is theoretical, and it refers to the expectations that are formulated based on a discipline's formalized state of knowledge. If a confirmatory study is grounded in strong theory that generates solid directional hypotheses, this will also increase the probability of finding such an effect. Taken together, the credibility and robustness of an empirical finding is amplified if it rests on convergent arguments (data at hand, prior research, and theory), an inductive constellation Mayo (2018: 15) refers to as "lift-off".

The second strategy is to actively embrace relevant sources of variation in study design and data analysis. This piece of advice reflects the insight that statistical inferences often do not hold at the intended level of generality (Winter & Grice 2021, Yarkoni 2022). In fact, error probabilities in the form of p-values or confidence intervals always require specification of a target population. Contrary to the everyday meaning of the term, a population need not only refer to a collection of human beings. In Generalizability Theory, the different dimensions across which we generalize are referred to as facets (e.g., Shavelson & Webb 1991). Each facet is a potential source of variation, and therefore a threat to the generality of an effect. Often, the extrapolation from a sample of subjects to a population of human beings is but one particular source of statistical uncertainty. Depending on the focus of the study, other dimensions may include the lexical items and materials or stimuli used in the study, or the real-world contexts they represent. Any of these facets could be a source of non-negligible variation, meaning that the magnitude (and possibly the direction) of an effect could vary across units (e.g. items, materials, or contexts). When designing an experiment, consideration should be given to potentially relevant sources of variation. The statistical conclusions of a study will then be strengthened if such built-in sources of variation can be demonstrated to be largely inert.

A third strategy is to subject findings to fiercely critical scrutiny; as Baayen et al. (2017: 227) state: "An effect is worth taking seriously only if it withstands truly serious attempts to bring it down." Designing a study in a way that gives different sources of variation the opportunity to call into question the generality of an effect is a key step in this regard. Apart from this, the statistical interrogation of the data may launch such attacks. In particular, the generality of an effect is qualified by the presence of non-negligible interactions of the factor of interest with other variables. Put differently, the credibility of an effect is strengthened if it can be demonstrated to be stable across subgroups in the data; Kish (1987: 51-57) refers to this as internal replication. The statistical analysis should therefore generally be allowed to speak to the presence of such interactions. A particular case in point is the examination of effect variation across subjects (and possibly also

items) by including into the analysis what – in regression parlance – are called random slopes (see Barr et al. 2013). Further, when working within a Bayesian paradigm, an effect can be challenged with skeptical priors.

Finally, limitations on the generality of an effect should be stated clearly. This is best done in the form of constraints-on-generality statements prior to the presentation of the results (Simons et al. 2017, Roettger 2021b). Essentially, this is the logical counterpart to embracing sources of variation at the design stage. On the one hand, this means that an explicit nod is given to those facets that are not represented in the design, and which could therefore not be addressed statistically. The findings are then specific to the particular units (e.g. items, materials, or contexts) covered by the study. Further, limited generality may also be due to a failure of a design to represent the breadth of variation underlying a particular facet. This is to say that a convincing demonstration of stability across units should be based on a set of conditions that approximate the existing level of variation. Related to this is a reflection on the extent to which the units in your study permit generalization to the underlying population of units; for instance, whether the lexical items in the experiment are representative of the population of lexical items to which the study (and any following replication attempt) intend to extrapolate. Effectively, constraints on generality provide guardrails for a conceptual replication study.

4.4. Recommendations for further reading

The following resources may be consulted for more background on the topics covered in this section:

- Replication and replicability are discussed in more detail in King (1995) and Nosek & Errington (2020)
- Yarkoni (2022) offers many thoughts on the problem of generalization; see also Chapter 7 in Abelson (1995) and Chapter 1 and Section 15.2 in Anderson (2001)

A particularly promising way to enable others to replicate your work (and arguably also increase its replicability) is the concept of preregistration, which the next section introduces.

5. Preregistration

Introduced to improve the robustness of randomized controlled trials in clinical research more than two decades ago, preregistrations have been adopted quickly across the behavioral sciences within the last five to ten years. After a short introduction to what preregistrations are (Section 5.1), we turn toward their purpose and what is in it for individual researchers (Section 5.2). Finally, we will give practical advice on how to preregister a study and where to create and archive preregistrations (Section 5.3).

5.1. Preregistrations and Registered Reports

A preregistration is, simply put, a time-stamped document that describes hypotheses, data collection procedure, and analysis plan; it is stored on an independent repository before data collection and/or analysis commences. Preregistrations can differ in the level of detailed they provide, ranging from basic descriptions of the study design to very thorough descriptions of the procedure and statistical analysis. In the most transparent version of a preregistration, all relevant materials, experimental protocols, and statistical procedures are published alongside the preregistration prior to data collection (ideally documented according to FAIR principles as outlined in Section 3).

When writing a preregistration, the researcher should keep a skeptical reader in mind. The goal of the preregistration is to reassure the skeptic that all necessary decisions have been made in advance. Importantly, preregistration is not restricted to studies that collect data but can also be used for secondary data analysis such as corpus-based work (Van den Akker et al. 2021). In preregistrations of already existing datasets, researchers specify their planned quantitative analyses and disclose their prior knowledge about the data. More recently, pre-registration formats for qualitative research have also been proposed (e.g., Haven & van Grootel 2019).

Peer-reviewed preregistrations, called Registered Reports (Nosek & Lakens 2014), include the theoretical rationale and research question(s) of the study as well as a detailed methodological protocol. Registered Reports can therefore be conceived of as full-fledged manuscripts, just without results. They can be submitted to an increasing number of journals, where they receive critical feedback from peer reviewers on how well the proposed method addresses the research question. After addressing the reviewers' concerns and refining the study design, the manuscript may be in-principle accepted, in which case the results of the study will be published irrespective of whether they confirm the researchers' predictions or not.

5.2. Reasons to preregister a study

Preregistering your work comes with several advantages. We discuss three of the most important ones here: First, preregistrations allow others to critically assess your research methods. Potential weaknesses can then be detected not only prior to publication, but before resources are invested into data collection. Allowing for such critical evaluations arguably improves the quality of empirical work.

Second, preregistrations draw an explicit line between post-hoc, exploratory analyses and initially planned, confirmatory analyses (Wagenmakers et al. 2012, Nosek et al. 2018). Since the conduct of a study commits to certain decisions prior to collecting or analyzing data, preregistrations reduce researcher degrees of freedom – i.e. the strategic exploitation of the flexibility that is inherent to empirical work. Constraints on this flexibility are important because the academic incentive system is arguably skewed toward favoring certain research outcomes over others; for instance, results that corroborate prespecified hypotheses are easier to publish than null results and exploratory analyses (Sterling 1959, Franco 2014). This can provoke motivated reasoning, which can further feed into human biases such as confirmation and hindsight bias (Tversky & Kahneman 1974), or lead to questionable research practices such as selective reporting and hypothesizing after results are known (John et al. 2012). As a result, the probability of statistically supported claims being false positives drastically increases (Simmons et al. 2011, Roettger 2019).

Such after-the-fact reasoning is particularly tempting in research on language, i.e. an intricate, multimodal communication system. Theoretical phenomena can be approached empirically from different angles, opening up many alternative paths to answer the same research question. For example, Coretta et al. (2023) gave the same speech-production dataset to 46 teams of researchers and asked them to answer the same research question, resulting in substantial variability in their analytical approaches, the reported effect sizes and their conclusions.

Registered Reports can additionally protect you from receiving methodological criticism from reviewers after the results are known. Thus, reviewers might identify reasons for why your results are different from what they expected and potentially make the publication of your study difficult (Nosek & Lakens 2014). During a Registered Report workflow however, reviewers only assess whether your report is a sound attempt to answer the research question. This precludes (or at least severely constrains) methodological criticism after results are known.

Finally, publicly stored preregistrations can help to reduce publication bias, as the number of failed attempts to reject a hypothesis can be tracked transparently. Recent work has shown that Registered Reports produce a more realistic number of null results than regular publication routes (Scheel et al. 2021).

5.3. Storing, sharing and deviating from preregistrations

Since preregistrations in their simplest form are time-stamped methodological protocols, any such document can be uploaded to a publicly available repository. However, there are many templates and tools out there that help with identifying the most important pieces of information that a preregistration should contain.

The OSF (osf.io) and AsPredicted (aspredicted.org) are the most commonly used templates and include examples of information that should typically be included in a pre-registered study plan. The idea behind these generic templates is to provide an accessible point of entry. AsPredicted covers fewer methodological details than the OSF, which in turn also allows for an "Open-Ended Registration", which is less restricted in its structure. PROSPERO (crd.york.ac.uk/prospero/) is a registry to preregister systematic reviews.

After completing and submitting a template, a time-stamped version of the document is recorded, which can either be made public right away, kept private, or automatically published after an embargo (OSF only). After the study is conducted and submitted to a journal, a link to the pre-registration should be provided in the manuscript so that readers can verify whether the final study followed the registered plan. Both OSF and AsPredicted offer anonymized view-only links that can be shared during peer review.

When it comes to Registered Reports, there are over 300 documented journals that have adopted Registered Reports, including outlets relevant to cognitive linguists: *Language and Cognition, Cognitive Linguistics, Journal of Memory and Language,* and *Glossa: Psycholinguistics* (see https://cos.io/rr for an exhaustive list).

Of course, things don't always go as planned – the complexity of behavioral studies cannot always be foreseen, and statistical properties of data cannot always be anticipated. This means that deviations from preregistrations are normal. It is possible to change the preregistration and document these changes alongside the reasons as to why and when changes were made. This procedure still provides substantially lower risk of cognitive biases impacting the conclusions compared to a situation without any a-priori specifications. It also makes these changes to the analysis transparent and detectable. However, deviations from the preregistration should be recorded prominently in the manuscript. One way of documenting deviations and unregistered steps in a structured way is the following table template (Willroth & Atherton 2024), designed to ensure transparency for readers. For Registered Reports, deviations must be documented and will be examined during a stage-2 review process, where reviewers assess, among other things, whether these deviations jeopardize the originally planned link between theoretical predictions and data (analysis).

5.4. Recommendations for further reading

For a deeper engagement with preregistration and its benefits we suggest the following resources and articles:

- A detailed introduction to preregistration for language scientists can be found in Roettger (2021a), Mertzen et al. (2021), and Brown & Strand (2023).

- Wicherts et al. (2016) provides a list of relevant researcher degrees of freedom for behavioral scientists.
- A concise and practical primer to preregistrations and Registered Reports is given by the UK Reproducibility Network (Stewart et al. 2020).

6. Conclusion

The replication crisis has put the methodological spotlight firmly on three hallmarks of empirical work: reproducibility, replicability and preregistration. We have reviewed many resources that can help us improve our empirical work, and indeed the list of tools we have touched upon may at first seem overwhelming. Importantly, however, each dimension shows a crescendo, which allows us to start small and work our way upwards: from the FAIR principles to proper data archiving using TROLLing, from reproducible workflows and dynamic documents to version control, from constraints-on-generality statements to variability-centered study design, and from a basic preregistration to a Registered Report. While we cannot expect studies (and scholars) to excel on all of these dimensions, it should have become clear that there are many opportunities for researchers to update their workflow and enhance their toolkit. Step by step, this will enable us to contribute to the continuing emergence of a new methodological paradigm and cultivate a more transparent, cautious, and self-critical approach to the empirical study of the relation between language and the mind.

Acknowledgements

Funding for the present work was provided by the German Research Foundation (DFG; grant number 548274092) and by the Research Council of Norway (NFR; grant number 350026)

References

Abelson, Robert P. 1995. *Statistics as principled argument*. New York: Taylor & Francis.

Anderson, Norman H. 2001. *Empirical direction in design and analysis*. Mahwah, NJ: Lawrence Erlbaum.

Andreassen, Helene N. 2022. Archiving research data. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *The open handbook of linguistic data management*, 89-100. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12200.003.0011

Baayen, Harald, Shravan Vasishth, Reinhold Kliegl & Douglas Bates. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94. 206-234. https://doi.org/10.1016/j.jml.2016.11.006

Barba, Lorena A. 2018. Terminologies for reproducible research. https://arxiv.org/abs/1802.03311. (27 September, 2025.) https://doi.org/10.48550/arXiv.1802.03311

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Brown, Violet A. & Julia F. Strand. 2023. Preregistration: Practical considerations for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research* 66(6), 1889-1898. https://doi.org/10.1044/2022_JSLHR-22-00317

Collister, Lauren B. 2022. Copyright and sharing linguistic data. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *The open handbook of linguistic data management*, 117-128. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12200.003.0013

Conzett, Philipp & Noortje Dijkstra Haugstvedt. 2024. DataverseNO README File Template - General (v2.4). Zenodo. https://zenodo.org/records/10849096. (27 September, 2025.) https://doi.org/10.5281/zenodo.10849096

Coretta, Stefano, Joseph V. Casillas, Simon Roessig, et al. 2023. Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science* 6(3). 1-29. doi:10.1177/25152459231162567

Franco, Annie, Niel Malhotra & Gabor Simonovits. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science* 345(6203). 1502-1505. https://doi.org/10.1126/science.1255484

Haven, Tamarinde L. & Leonie Van Grootel, Dr. 2019. Preregistering qualitative research. *Accountability in Research* 26(3). 229-244. https://doi.org/10.1080/08989621.2019.1580147

Hurlbert, Stuart H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54(2). 187-211. https://doi.org/10.2307/1942661

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

John, Leslie K., George Loewenstein & Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23(5). 524-532. http://doi.org/10.1177/0956797611430953

King, Gary. 1995. Replication, replication. *PS: Political Science and Politics* 28(3). 444-452. https://doi.org/10.2307/420301

Kish, Leslie. 1987. *Statistical design for research.* New York: Wiley.

Knuth, Donald E. 1984. Literate programming. *The Computer Journal* 27(2). 97-11. https://doi.org/10.1093/comjnl/27.2.97

Mayo, Deborah G. 2018. *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781107286184

Mertzen, Daniela, Sol Lago & Shravan Vasishth. 2021. The benefits of preregistration for hypothesis-driven bilingualism research. *Bilingualism: Language and Cognition* 24(5). 807-812. https://doi.org.10.1017/S1366728921000031

Nosek, Brian A. & Timothy M. Errington. 2020. What is replication? *PLOS Biology* 18(3): e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, Brian A. & Daniël Lakens. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology* 45(3). 137-141. https://doi.org/10.1027/1864-9335/a000192

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven & David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11). 2600-2606. https://doi.org/10.1073/pnas.1708274114

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

Parsons, Sam, Flávio Azevedo, Mahmoud M. Elsherif, et al. 2022. A community-sourced glossary of open scholarship terms. *Nature Human Behaviour* 6. 312-318. https://doi.org/10.1038/s41562-021-01269-4

Quintana, Daniel S. 2020. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* 9: e53275. https://doi.org/10.7554/eLife.53275

Roettger, Timo. 2021a. Preregistration in experimental linguistics: applications, challenges, and limitations. *Linguistics* 59(5). 1227-1249. https://doi.org/10.1515/ling-2019-0048

Roettger, Timo B. 2021b. Context sensitivity and failed replications in linguistics – a reply to Grieve. *Linguistics* 59(5). 1357-1358. https://doi.org/10.1515/ling-2020-0239

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1-27. http://doi.org/10.5334/labphon.147

Scheel, Anne M., Mitchel R. Schijen, & Daniël Lakens. 2021. An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science* 4(2). 1-12. https://doi.org/10.1177/25152459211007467

Schweinberger, Martin. 2025. The Language Technology and Data Analysis Laboratory (LADAL). Brisbane: The University of Queensland, School of Languages and Cultures. https://ladal.edu.au/ (Version 2025.04.01). (27 September, 2025.)

Shavelson, Richard J. & Noreen M. Webb. 1991. *Generalizability theory: A primer*. London: Sage.

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

Simons, Daniel J., Yuichi Shoda & D. Stephen Lindsay. 2017. Constraints on generality (CoG): Proposed addition to all empirical papers. *Perspectives on Psychological Science* 12(6). 1123-1128. https://doi.org/10.1177/1745691617708630

Sönning, Lukas & Valentin Werner. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59(5). 1179-1206. https://doi.org/10.1515/ling-2019-0045

Stall, Shelley, Marianne E. Martone, Ishwar Chandramouliswaran, Lisa Federer, Julian Gautier, Jennifer Gibson, Mark Hahnel, Jennie Larkin, Nicole Pfeiffer, Brian Sedora, Ida Sim, Tim Smith, Ana E. Van Gulick, Erin Walker, Julie Wood, Maryam Zaringhalam & Alberto Zigoni. 2023. Generalist repository comparison chart (Version 3.0). Zenodo. https://zenodo.org/records/7946938. (27 September 2025.) https://doi.org/10.5281/zenodo.7946938

Sterling, Theodore D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association* 54(285). 30-34. https://doi.org/10.1080/01621459.1959.10501497

Stewart, Suzanne L., Eike M. Rinke, Ronan McGarrigle, Dermot Lynott, D., Carole Lunny, Alexandra Lautarescu, Matteo M. Galizzi, Emily K. Farran & Zander Crook. 2020. Pre-registration and registered reports: A primer from UKRN. OSF preprint: https://osf.io/preprints/osf/8v2n7_v1. (27 September 2025.)

Strand, Julia F. 2025. Error tight: Exercises for lab groups to prevent research mistakes. *Psychological Methods* 30(2). 416-424. https://doi.org/10.1037/met0000547

The Turing Way Community. 2025. *The Turing Way: A handbook for reproducible, ethical and collaborative research*. Zenodo. https://zenodo.org/records/15213042 (Version 1.2.3). (27 Setptember 2025.) 10.5281/ZENODO.3233853

Tversky, Amos & Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185(4157). 1124-1131. https://www.science.org/doi/10.1126/science.185.4157.1124

Van den Akker, Olmo, Sara Weston, Lorne Campbell, Bill Chopik, Rodica Damian, Pamela Davis-Kean, Andrew Hall, Jessica Kosie, Elliott Kruse, Jerome Olsen, Stuart Ritchie, KD Valentine, Anna van 't Veer & Marjan Bakker. 2021. Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology* 5: 2625. https://doi.org/10.15626/MP.2020.2625

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas & Rogier A. Kievit. 2012. An agenda for purely confirmatory research. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 7(6). 632-638. https://doi.org/10.1177/1745691612463078

Wicherts, Jelte M., Coosje L. Veldkamp, Hilde E. Augusteijn, Marjan Bakker, Robbie C. Van Aert & Marcel A. Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7: 1832. https://doi.org/10.3389/fpsyg.2016.01832

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18

Willroth, Emily C. & Olivia E. Atherton. 2024. Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science* 7(1): 25152459231213802. https://doi.org/10.1177/25152459231213802

Winter, Bodo & Martine Grice. 2021. Independence and generalizability in linguistics. *Linguistics* 59(5). 1251-1277.

Yarkoni, Tal. 2022. The generalizability crisis. *Behavioral and Brain Sciences* 45(e1). https://doi.org/10.1017/S0140525X20001685