Handbook on Conducting Reproduction and Replication Studies

Lukas Röseler* Lukas Wallrich* Helena Hartmann Luisa Altegoer

Veronica Boyce Sarahanne M. Field Janik Goltermann Joachim Hüffmeier

Charlotte R. Pennington Merle-Marie Pittelkow Priya Silverstein

Don van Ravenzwaaij Flavio Azevedo

2025-08-01

Handbook on Conducting Reproduction and Replication Studies

Lukas Röseler*, Lukas Wallrich*, Helena Hartmann, Luisa Altegoer, Veronica Boyce, Sarahanne M. Field, Janik Goltermann, Joachim Hüffmeier, Charlotte R. Pennington, Merle-Marie Pittelkow, Priya Silverstein, Don van Ravenzwaaij, Flavio Azevedo

FORRT

MüCOS
Münster Center
for Open Science

Table of contents

1 Summary

The practice of repeatedly testing published results with the same data (reproduction) or new data (replication) is currently gaining traction in the social sciences, owing to multiple failures to reproduce and replicate published findings. Along with increased skepticism have come guidelines for the repeated testing of hypotheses from various disciplines and fields. This guide aims to enable researchers to conduct high-quality reproductions and replications across social science disciplines. First we summarize recent developments, then provide a comprehensive guide to carrying out reproductions and replications, and finally present an example for how guidance needs to be tailored for specific fields. Our guide covers the entire research process: choosing a target study, deciding between different types of reproductions and replications, planning and running the new study, analyzing the results, discussing outcomes in the light of potential differences, and publishing a report.

Keywords: replication, repetitive research, reproducibility, meta-science, meta-research, open science, open research, open scholarship

FORRT

MüCOS
Münster Center
for Open Science

Part I

Foundations

2 Background

"The proof established by the test must have a specific form, namely, repeatability. The issue of the experiment must be a statement of the hypothesis, the conditions of test, and the results, in such form that another experimenter, from the description alone, may be able to repeat the experiment. Nothing is accepted as proof, in psychology or in any other science, which does not conform to this requirement." – (Dunlap 1926)

Repeatability is the cornerstone of many sciences: A majority of the scientific progress rests on the successful accumulation of evidence for claims through reproduction and replications to establish robust discoveries. Reproductions and replications, that is repeated testing of a hypothesis with the same (reproduction) or different (replication) data, are necessary.

Cumulative science without repetition is costly. The aim of this guide is to empower researchers to conduct high-quality reproductions and replications and thereby contribute to making their fields of research more cumulative and robust. Issues of replicability have been discussed across many disciplines, such as psychology ((Open Science Collaboration 2015)), economics ((Dreber and Johannesson 2024)), biology ((Errington et al. 2021)), marketing ((Urminsky and Dietvorst 2024)), linguistics ((McManus 2024)), computer science ((Hummel and Manner 2024)) and epidemiology ((Lash, Collin, and Van Dyke 2018)) and the number of replications has been rising sharply (see Figure 1).

Figure 1

Number of replication studies by year of publication based on the FORRT Replication Database (FReD, (Röseler et al. 2024)) based on the version from July 16, 2025. Code to reproduce the figure: https://osf.io/dznrb.

While the number of replication and reproduction studies has increased, the overall proportion of them is still very small, with reviews finding yearly replication rates of up to 1% ((Perry, Morris, and Lea 2022)). Moreover, much of the guidance on replications is being developed actively ((Clarke et al. 2024)) and in narrow parts of science, which leads to fragmentation, siloing, and potentially inconsistent information.

Here we attempt to integrate useful guidelines (e.g., (Block and Kuckertz 2018; Jekel et al. 2020)) into a comprehensive overview that allows diverse fields to profit from each other. In sum, this guide provides information about the entire process of research allowing researchers at all career stages to plan, conduct, and publish reproduction and replication studies. We limit our scope to quantitative research, given that the concept of reproducibility and replicability itself is highly contested among qualitative researchers (see Makel, Plucker, and Hegarty (2012); Cole et al. (2024); Pownall (2022); Bennett (2021)).

3 Understanding Replications and Reproductions

In this guide, we focus on studies that re-examine a previously tested hypothesis and refer to them as repetitions (i.e., reproductions and replications) with the general field being called repetitive research as suggested by Schöch (2023). However, it is important to note from the outset, that there is no overarching terminology or consensus (e.g., Voelkl et al., 2025), as the formal development of replication methods has begun relatively late in the social, behavioral, and cognitive sciences. For example, empirical psychology is more than 100 years old, but until the advent of the replication/reproducibility crisis in the early 2010s, replication methods have been rarely discussed (e.g., King, 1995). Different fields of research seem to tackle the task differently and independently, which has led to multiple overlapping terminologies across psychology (Schmidt, 2009; Hüffmeier et al., 2016), management (Tsang & Kwan, 1999), marketing (Urminsky & Dietvorst, 2024), organizational sciences (Köhler & Cortina, 2021), computer sciences (Heroux et al., 2018), language learning (MacManus, 2024), and the humanities (Schöch, 2023).

3.1 Reproduction and Replication

The terms reproduction and replication are used in different ways between disciplines; for example, in psychology, studies using different data are commonly referred to as replications and studies using the same data are referred to as reproductions, whereas in other fields, such as computational science or economics, these terms may be used in the opposite manner or treated interchangeably (see Milkowski et al., 2018; Ankel-Peters et

al., 2023). In this paper, replication is used to refer to efforts involving the analysis of different data, and reproduction to efforts involving the same data. The different data do not necessarily need to be from a different sample but can also constitute distinct (non-overlapping) subsets from the same sample (e.g., incidental or panel data; Huang & Huang, 2024).

Reproduction and replication should always be considered together and if possible, reproduction should come before replication. This is because, at the early stages of research, reproduction is much more cost efficient; first confirming whether the findings are reproducible can clarify whether a replication is worthwhile. Furthermore, if the research procedure consists of "moving away" from a specific finding in terms of changing the analysis code, materials, and dataset to test its generalizability or boundary conditions, a numerical reproduction (using the same data and same code) is the closest possible repetition of a finding and a useful foundation for further steps. We discuss multiple cases to illustrate the relationship between reproduction and replication in Table 1:

Table 1

Possible combinations of reproduction and replication outcomes.

Case	Reproducible?	Replicable?	Possible interpretation
A	Yes	Yes	The original finding is reproducible and generalizable.
B	Yes	No	The original finding is reproducible but not generalizable.
C	No	Yes	The original finding is not reproducible but replicators could determine a scenario where it holds.

Case	Reproducible?	Replicable?	Possible interpretation
D	No	No	The original finding is neither reproducible nor generalizable.

Note. A similar distinction is made by The Turing Way (XXXX) but uses a less specific terminology for reproductions.

3.2 Outcome

Common language often conflates outcome and study descriptions: researchers typically use the phrase "has been replicated" to refer to a replication attempt that has corroborated the findings of the original study, whereas "failed to replicate" or "could not be replicated" is used to refer to circumstances where a replication attempt has not corroborated the original results or has led to a different interpretation or conclusions (see also Patil et al., 2016a).

In this article, when we state that a "study was reproduced/replicated" we mean that there has been a replication attempt, irrespective of its outcome. With "replicable" and "reproducible" we express that there was support for the original hypothesis. Note that the outcome of a replication/reproduction study is often not straightforward, but may depend on the success criteria applied. This is discussed in section Defining and Determining Replication Success.

3.3 Types of replication

We heavily rely on the typology provided by Hüffmeier et al. (2016) where different types of replications are defined by the closeness or similarity between original and replication study. Similarity cannot be evaluated without a theory about the concepts involved. For example, the concept of age can differ strongly between replications of historical,

psychological, or biological studies, leading to different measures of the concept itself and thus different judgments about the similarity of an object's age.

Under the assumption of a stable world and constant laws or regularities that are investigated by the social, behavioral, and cognitive sciences, a reproduction and replication study's closeness to an original study is associated with replication 'success' (Hüffmeier et al., 2016, Lebel et al., 2018; see the discussion section for an in-depth discussion of success criteria). The argument can be made from two different philosophical perspectives that we call inductive (phenomenon-focused, effects application, bottom-up) and deductive (theory-focused, theory application, top down; e.g., Calder et al., 1981; Borgstede & Scholz, 2021). From an inductive perspective, a replication that is very similar to an original study should lead to the same result whereas one that differs with respect to any criterion may lead to different results.[^1] This is a stance often taken by proponents of findings that failed to replicate (e.g., Baumeister & Vohs, 2016; Syed, 2023), arguing that characteristics such as time or place are different and can be valid reasons for different results. From a deductive (theory-focused) view, the only changes that matter are those that affect the underlying theory. Consider for example a replication experiment that is identical in every aspect except for the season (summer instead of winter). If the theory that is tested is about color perception, the replication is likely judged to be close to the original study but if it is about participants' current tea preferences, it is likely judged to be different from the original study in a theoretically relevant aspect.[^2]

A related dimension of closeness concerns contextual sensitivity—the extent to which the meaning of a questionnaire or the effect of a manipulation depends on time, culture, or population. As Van Bavel et al. (2016) demonstrate, studies on contextually sensitive topics were significantly less likely to replicate successfully in Open Science Collaboration (2015), even though methodological fidelity was high. This raises important questions about what constitutes a "close" replication: Should a study on celebrity attitudes, for example, use the same examples (which may be outdated and thus psychologically inert), or should it adapt to locally and temporarily salient figures to trigger the same cognitive or emotional responses? In such cases, strict methodological similarity might paradoxically undermine theoretical closeness, and thus the validity of the replication attempt. This tension highlights that procedural fidelity does not always equate to theoretical

equivalence—particularly for studies involving social meaning, identity, or temporally anchored norms. Lebel et al. (2018, Figure 2) provide a taxonomy for classifying a replication study's closeness for psychological research.

Figure 2

Replication Continuum

Highly Similar ← → Highly Dissimilar

	Direct Replication			Conceptual Replication	
Design Facet	Exact Replication (All facets under researcher control are the same)	Very Close Replication (Procedure or physical setting is different)	Close Replication (IV or DV stimuli are different)	Far Replication (IV or DV operationalization or population is different)	Very Far Replication (IV or DV constructs are different)
Effect, Hypothesis	Same	Same	Same	Same	Same
IV Construct	Same	Same	Same	Same	Different
DV Construct	Same	Same	Same	Same	Different
IV Operationalization	Same	Same	Same	Different	
DV Operationalization	Same	Same	Same	Different	
Population (e.g., age)	Same	Same	Same	Different	
IV Stimuli	Same	Same	Different		
DV Stimuli	Same	Same	Different		
Procedural Details	Same	Different			
Physical Setting	Same	Different			
Contextual Variables	Different				
⋮	⋮				

Figure 3.1: Taxonomy for classifying a replication study's methodological similarity to an original study. Reprinted from LeBel et al. (2018) with permission.

Support for the view that methodological features that are theoretically irrelevant such as the use of text versus image stimuli or the type of sample can have a strong impact on the results is provided by Landy et al. (2020), who let different groups of researchers test identical hypotheses using different study designs. The groups arrived at entirely different and even opposite conclusions for similar hypotheses. The differences in the study designs were not predicted by the theories involved in the respective studies: A priori, none of the differences (e.g., within- vs. between-subjects design, picture vs. text stimuli) "should" have affected the outcomes. Note that other theories such as demand characteristics (Orne, 2017) could help in these cases. Moreover, Note that this does not disconfirm the deductive perspective but may be a demonstration of the lack of specification of theories - as well as a reminder that statistical choices affect statistical power by changing the variance, and thus standardised effect sizes. In line with deviations from original studies mostly having

uncertain consequences, close replications more directly test the credibility of original results, while conceptual replications that vary features of the design are concerned with generalizability.

Note that Nosek and Errington (2020) define replication as a study "for which any outcome would be considered diagnostic evidence about a claim from prior research". This can lead to issues when the original claim is not clear on its boundary conditions. Conceptual replications that highlight limitations to the claim made clearly count, e.g. when the original claim was about a universal effect, and the replication shows that it does not hold in a specific country. Conversely, "replications" that go beyond the claim made, and test the transferability of a claim explicitly made about, e.g., maths education to science education may indeed serve to be framed differently, as they do not directly speak to the claim made originally. Where original authors' failed to specify the scope of their claim, we would understand that they imply a broad/universally applicable relationship, which any attempts at generalisation help to corroborate or specify.

In terms of Schöch (2023), who defines an overarching type of repetitive research based on multiple dimensions, replications are concerned with the same question as a previous study, use the same (close replication) or a similar (conceptual replication) method and use different data (otherwise they are reproductions).

3.4 Types of reproduction

Reproductions can be numerical reproductions, testing whether the same data, code and software lead to the same results, or robustness reproductions, extending the original analysis and exploring the central finding's limits (Dreber & Johannesson, 2024). Most reproductions would include both a numerical reproduction as baseline and then a robustness reproduction, unless the numerical reproduction is not possible due to a lack of code or software.

Part II

The Replication Process

4 Choosing the Target Study

Reproduction and replication studies can serve different goals and depending on the goal, the way of choosing a target study differs (see Pittelkow et al. 2023). In large-scale reproduction and replication projects, such as Brodeur et al. (2024), the Reproducibility Project: Psychology (Open Science Collaboration, 2015) or the Reproducibility Project: Cancer Biology (Errington et al., 2021), the primary aim is to assess the overall reliability of a field or a set of findings, leading to a top-down approach in which the decision to replicate comes first, followed by the selection of specific replication targets. This is often done in a way aimed to be representative of a field, ideally through random sampling, though this is generally constrained by practicalities. Here, individual studies are not the primary focus in the decision to repeat; instead, choices are guided by broader methodological or theoretical considerations. In contrast, individual researchers frequently adopt a bottom-up approach, where the decision to replicate is driven by engagement with a specific study (or theoretically related set of studies, e.g., Röseler, et al., 2021). This may occur when a researcher wishes to build upon an existing finding and seeks to verify its robustness before doing so, or when they harbor doubts about a claim and aim to test its validity. Since reproductions and replications can serve multiple purposes—from assessing theoretical frameworks to correcting potential errors—there is no singular correct way to decide what to repeat. The choice of targets ultimately depends on the overarching goals and methodological approach of the replication effort, as well as on practical constraints. However, what does matter is that the selection of reproduction and replication targets is well justified and transparently communicated. For instance, researchers can use structured frameworks such as the replication target selection checklist to ensure clarity and consistency in their decision-making process (Pittelkow et al. 2023). For a comment on what empirical reasons for replications are, see Kamermans et al. (2025).

4.1 Determining Reproduction and Replication Value

Whether a target study is "worth reproducing" or "worth replicating" is highly debated and is suggested to depend on several overlapping factors, including value (sometimes also referred to as impact or relevance), uncertainty, and feasibility (Isager et al. 2023). Below, different suggestions for operationalizing these factors are discussed systematically.

Note that there is also ongoing discussion about whether or not all studies are generally 'worth replicating'. One perspective is what is worthy of publication is worthy of replication (Feldman, 2025) or on a different note, what is worthy of publication should be worthy of replication - though this perspective is becoming complicated through the rise of influential preprints and a public-review-curate model to publications. Naturally, a public report of a study is necessary for other researchers to attempt a replication and an available dataset is needed for a reproduction. To take a more fine-grained look at the publication status, several different types of research emerge. An article can be retracted, that is, there is no confidence anymore in its findings due to research misconduct or severe errors. When the data of a study were fabricated and it was thus retracted, a reproduction will not be informative but a replication may inform researchers about the correctness of the hypothesis unlike the original report. Other reasons (or unclear reasons) for retraction may conversely increase the replication and reproduction value, as the source of a true claim may have become untrustworthy (and not easily citeable) due to issues unrelated to its truth (e.g. plagiarism).

Replicating and reproducing every finding that was ever published appears impossible to achieve, which is why researchers need to make decisions about prioritization. In the following, we discuss criteria by which such a prioritization can occur - restricted to quantitative research.

4.2 Value

The original study should be somehow relevant for the replication to have value (e.g., Karhulahti et al., 2024). It may have started a research stream. For example, Jacowitz and

Kahneman's (1995) studies on anchoring and adjustment were fundamental for how anchoring effects are investigated today, and were therefore replicated by Klein et al. (2014). Field et al. (2019) propose a method for the selection of replication studies that features the theoretical importance of the original study result. Relevance may be evidenced by many citations as they show that many studies are building on the finding, testing similar hypotheses, or criticizing the study. Note that a study could also be cited as a negative example or study that has not been replicated or retracted for some reason. Isager et al. (2021, 2023) suggest deciding what to replicate based on sample size and citation count (but see Pittelkow et al., 2025). In a Delphi study examining consensus among psychologists that had conducted empirical replications on what should influence the decision of what study to replicate, elements that came up were the importance of the original study for research (as indicated by citations, the phenomenon being over- or understudied, and the impact factor of the journal), the theoretical relevance of the study, and the implications of the original study for practice, policy, or clinical work (Pittelkow et al., 2023). The relevance of societal impact was also stressed by Bekker (2024), as a study may have a high value for a societal problem (e.g., a new vaccine or a repeated test of a claim that is relevant in the political discourse such as criminality among immigrants).

For scientists reproducing or replicating a study because they are interested in building on its findings (including if they wish to build upon their own original findings), their interest to build on it may be a sufficient indicator of its relevance to their research program.

4.3 Uncertainty

The more uncertain the original study's outcome is, the higher the potential of knowledge gained from reproduction and replication. Although no findings are definitive, research reports differ in the strength of the evidence they present (e.g., Registered Reports[^3] are typically more convincing than non preregistered studies; Soderberg et al., 2021). Similarly, sample size (within a given field) has been proposed as an indicator of evidence strength (Isager et al., 2021). Pittelkow et al. (2021, 2023) and Field et al. (2019) both argued for using the current strength of evidence in favour of the original claim as an

important element that features into the choosing a replication target. However, the degree of uncertainty can be uncertain or misjudged: In some areas of research a hypothesis had been claimed to be confirmed hundreds of times and yet, large-scale replication effort could not support the original hypothesis so that after hundreds of studies the existence of the phenomenon was still unknown (e.g., Friese et al., 2019). Meta-analyses allow some tests for uncertainty (e.g. via correction of bias, evaluation of risk of bias, or estimates of heterogeneity). Although there are numerous ways to meta-analytically evaluate the expected replicability of a set of claims, none of them is as solid as a well-designed replication attempt (Carter et al., 2019). Other heuristics to estimate robustness reproducibility and replicability of sets of findings have been proposed:. They include the caliper test, relative proximity, or z-curve (Bartoš & Schimmack, 2022; see Adler et al., 2023, for an overview and a ShinyApp that combines these tools). Individual findings can be assessed through forensic meta-science tests (for an overview, see Heathers, 2025), and through the assessment of papers for reporting issues, such as those identified by statcheck (Nuijten et al., 2017; DeBruine & Lakens, 2025). Moreover, methods such as sum of p-values (Held et al., 2024) and Bayesian re-analysis can be applied to help determine the degree of evidence for a given effect an original study might contain (Field et al., 2019, Pittelkow et al., 2021).

If the original paper reports multiple studies for the same phenomenon, researchers should check the proportion of significant studies and whether all of them confirm the hypothesis. More studies reduce the overall statistical power (power deflation). Provided the hypothesis is correct, a single study may test it with 90% power, that is, the statistical analysis will indicate the correctness of the hypothesis with a probability of .9. Now, if 10 studies are run with 90% power each, the chances of all of them supporting the hypothesis (even if it is true) are 0.9^10 = .35. For 80%, even finding five significant findings in a row is fairly highly unlikely (0.8^5 = .33). Thus, studies reporting a set of many and only significant findings when each of the studies does not have very high power should be taken with caution (see also Francis, 2012; Schimmack, 2012; Lakens & Etz, 2017).

For large parts of the literature and given the overall low replicability rate in many fields (though not all, e.g., Soto, 2019), the mere lack of a reproduction or close replication by independent researchers can be used as an argument for uncertainty (e.g., Pittelkow et al.,

2023). If a study has only been replicated by the original authors, it can be indicative of nobody else being interested in the phenomenon (i.e., low replication value) or nobody else being able to provide evidence for it (i.e., high uncertainty). For example, it is possible that reports of failed replications are held back by reviewers due to an aversion to null findings, replications, or findings criticizing their own work.

As replications can also be used to probe a phenomenon's generalizability, a lack of variety in study designs can motivate a replication attempt. If there is reason to assume that a phenomenon is highly dependent on context (e.g., works only for graduate students, with English-speaking people, when people are incentivized, for the chosen stimuli, …), it can be replicated and extended in other contexts. More generally, when background factors are introduced to a study (e.g., there was a positive correlation in study X but researchers suspect it to vanish under condition M), the original finding needs to be replicated in a part of the new study for the argument to work. An added benefit of this is to help avoid later claims of 'hidden moderators' in original studies; an argument which has been used previously to refute the validity of replication study results (Zwaan, et al., 2018).

Finally, uncertainty can be the result of a lack of specificity in the original report: If there are details missing that cannot be retrieved anymore (e.g., researchers involved in the original study cannot be reached), a replication can develop, test and share a comprehensive set of materials. For example, Chartrand and Bargh's (1999) seminal study on the chameleon effect requires many materials but none of them are openly available. Accordingly, Pittelkow et al. (2023) identified the clarity of the original study protocol as an important element that features into the decision of replication study selection. Reconstructing these materials and documenting a procedure would, thus, be a valuable contribution of a replication study.

Theoretical contribution

In some cases, theories are so vague that a failed replication would likely be criticized for misunderstanding the theory (e.g., Baumeister & Vohs, 2016). This suggests that the target theory was not well specified. If accepted as a reason not to replicate, it can discourage any form of replication despite the target finding being relevant. Instead, replication researchers can ask original authors for feedback on the study protocol before

collecting data to try to ensure that it tests (and then articulates) the intended theory. They can also engage in adversarial collaboration or "red teaming" (e.g., Cowan et al., 2020), that is work together with the original authors to design a study that they agree would be able to corroborate the original claim, or to call it into doubt.

Nevertheless, it has been argued that because so many original studies are flawed, the theories built upon them are weak, or contaminated. This, in turn, can lead to flawed replication studies, especially in the case of theory that aims to explain phenomena (Field et al., Volz, Kaznatcheev & van Dongen, 2024), risking a vicious cycle in which successful replications potentially perpetuate flaws across studies.

Availability of reproductions and replications

While a single replication (or robustness reproduction) cannot provide conclusive evidence in regard to the veracity of original claims, the first numerical reproduction, and arguably also the first robustness reproduction and replication adds the greatest value in terms of reducing uncertainty. Therefore, the search for existing reproductions and replications is a key part of the selection of a target study.

Although there is no comprehensive database with reproductions yet, researchers can check resources such as the Institute for Replication's discussion paper series (https://i4replication.org/discussion_paper.html; Brodeur et al., 2024), the Replication-Wiki (Höffler, 2017), the CODECHECK register (https://codecheck.org.uk/register/, Nüst & Eglen, 2021), or the Social Science Reproduction Platform (https://www.socialsciencereproduction.org).

With regard to replications, researchers can browse the FORRT Replication Database (https://forrt-replications.shinyapps.io/fred_explorer/; Röseler et al., 2024), though this does not (yet) provide a replacement for manual searches.

4.4 (Potential) Researcher Bias {#(potential)-researcher-bias}

Researchers typically work in relatively small communities to investigate the same phenomenon. These researchers are invested in their work and can be influenced by certain

researcher biases, such as confirmation bias (the tendency to preferentially seek out, evaluate and recall information that supports one's existing beliefs; e.g., Mahoney, 1977) and motivated reasoning (generating post-hoc rationalizations that frame previous decisions in a favourable light; see Hardwicke & Wagenmakers, 2024; Munafò et al., 2020). In some cases, researchers profit off their work and the (perceived) replicability of their findings may be associated with personal financial gain. Such conflicts of interest should be disclosed, but this is not always the case (see Heirene et al., 2024).

However, replication researchers are just as prone to bias as original authors can be. Certain studies are more likely to be chosen for replication than others (see Pennington, 2023; Yarkoni, 2013), and there may be a publication bias in replication studies in favor of nonsignificant findings (Berinsky et al., 2020), though there is no empirical evidence for this yet. Nevertheless, greater interest in failed replications seems very likely, incentivizing replication researchers to apply questionable research practices (QRPs) so that the results are nonsignificant ("null hacking", Protzko, 2018; Baumeister et al., 2022). The problems of p-hacking and null-hacking can mostly be solved through preregistration and the use of Registered (Replication) Reports (e.g., Brodeur et al., 2024; Soderberg et al., 2021). Another type of bias is that researchers may be interested in replicating specific studies because of personal admiration towards a study or personal doubts or envy towards a colleague.

4.5 Feasibility

Reproductions require the original dataset. We recommend that researchers check whether the journal that published the original study has a data editor or reproducibility manager who has done a reproducibility check or provides a replication package. A replication package is a collection of materials to allow reproduction of the original results. Ideally, the dataset in the replication package, or shared separately, adheres to the FAIR criteria (Wilkinson et al., 2016), that is, it should be findable, accessible, interoperable, and reusable. Otherwise, the reproduction author would need to send a data sharing request to the original authors. In any case, they may need to consult with the original authors

regarding software versions and code that does not work anymore due to changes in the software.

While original data is not necessary for replications, thorough documentation of the original study is highly beneficial. Moreover, replication researchers should evaluate whether they can achieve the target sample size, which is often a multiple of the original sample size (see section Sample Size Determination). Pittelkow et al. (2023) identified the resources available to the replicating team in terms of funding, time, equipment, and (if relevant) previous experience and expertise as important elements that feature into the replication study selection. When choosing a target study, researchers should try to anticipate practical problems, and should restrict their choice of replication target to align with their lab resources in order to prevent 'secondary' research waste (Field et al., 2019). Specifically, some studies may be difficult to replicate (e.g., longitudinal studies). Other studies, such as those conducted with the use of highly technical, restricted, or expensive equipment, such as studies involving MRI scanning, might require expertise and knowledge that is not represented in all potential replication research teams (Field et al., 2019).

Moreover, there are no established standards for replications in some fields yet. In that case, replications may add less to the reduction of uncertainty and replicators need to propose methods. For example, replications with response-surface-analyses are not as established as those with t-tests for two-group study designs. Furthermore, the complexity of the data types can pose challenges for definitions of successful replications, such as in neuroimaging research (e.g., MRI studies) which often implicates outcome variables with an additional spatial component.

5 Planning and Conducting Reproductions and Replications

Planning depends on whether the focus is on a certain method or a theory, that is whether the replication will be close or conceptual. Table 2 provides an overview of reproduction and replication types, or more generally "repetitive research" (Schöch, 2023), drawn from different resources (e.g., Dreber & Johannesson, 2024; Hüffmeier et al., 2016; for an alternative taxonomy see also Cortina, Köhler & Aulisi, 2023) . The decision between these types is the first step in planning.

In addition, the formation of the replication team is important, as replications can take substantial resources. Notably, repetitive research has successfully been conducted collaboratively with graduate and undergraduate students (e.g., Boyce et al., 2024; Hawkins et al., 2018; Jekel et al., 2020; Moreau & Wiebels, 2023) and we recommend the use of replication studies to engage students of different levels in conducting and publishing research.

Table 2

Types of repetitive research ordered by reproduction and replication and respective closeness to the original study.

Type	Description	Goals
Computational Reproduction	Reanalysis of the same data with the same code	Correctness of original report
Recoding reproduction	Reanalysis of the same data, with new (equivalent) code	Correctness of original report

Type	Description	Goals
Robustness Reproduction	Reanalysis of the same data with new coding choices; can vary in closeness	Robustness of original finding and sensitivity to different analytical decisions or software
Multiverse analysis	Analyze data in all sensible ways (i.e., a large number of different robustness reproductions)	Robustness and generalizability of original finding, identification of potential moderators or sources for effect variability
Internal replication	Replicate one of your own studies as closely as possible	Demonstrate one's findings' generalizability across studies and rule out fear of false-positives (e.g., for new discoveries)
Close / direct / exact replication	Conduct a new study (based on work by other researchers) that is as close as possible to the original study	Rule out fear of the original finding being a false-positive, validate original materials or design, check generalizability/external validity for theoretically irrelevant variables (e.g., population, year of data collection)
Close replication with extension	Add a variable or procedure to a close replication	Rule out fear of the original finding being a false-positive, test generalizability of original finding

Conceptual / constructive replication	Conduct a study with changes that may be theoretically relevant but that tests the same hypothesis (e.g., different operationalization)	Generalizability of original finding, validity of theory

5.1 Post Publication Conversations

When planning the replication study, additional knowledge should be taken into account such as any discussions of the original finding. There can be other studies citing the original studies, criticizing them, disconfirming their underlying theory, identifying errors, reinterpreting the finding, or making suggestions for replications. All of these might highlight considerations that need to be taken into account when designing a replication study that robustly tests the original claim or its generalisability.

Thus, replication researchers should look for post-publication discussions on the target study such as published comments and reviews, blog posts, or discussions on social media. These can often be found via Altmetric (https://www.altmetric.com) or other tools that allow researchers to quickly identify discussions on social media or news outlets beyond scientific journals (PubPeer [https://pubpeer.com], Hypothes.is [https://web.hypothes.is]), or the in-development platform Alphaxiv.org [https://www.alphaxiv.org/]; for a review see Henriques et al., 2023).

5.2 Reproduction before Replication

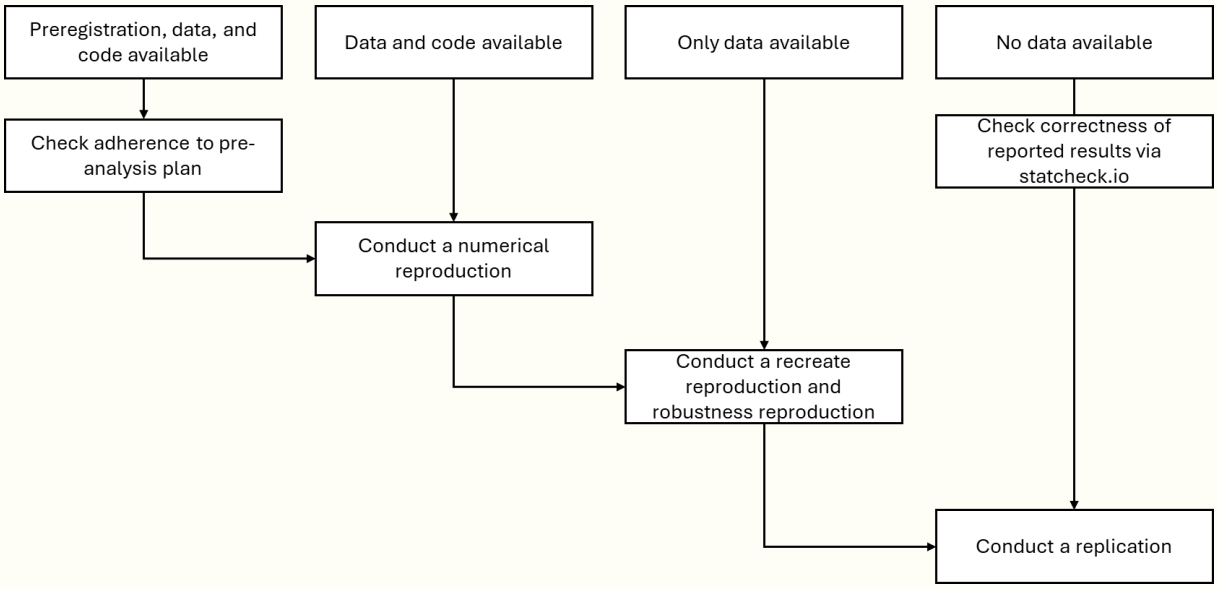
Many features of a replication study rest on the correctness of the original report. A reproduction allows researchers to investigate this by being able to uncover coding errors, fraud, robustness to analytical decisions, and generalizability. To make efficient use of resources, we encourage researchers to investigate the original finding's reproducibility and robustness first. In other words, ideally, reproductions should take place before

planning and conducting a replication study. Depending on the availability of the code and data, these can take several minutes to weeks.

If the original code and dataset are available, researchers can try to numerically reproduce the results. Beware, however, that differences in software versions or default settings may lead to slight deviations or require corrections in some cases (for a large-scale test of reproducibility see Brodeur et al., 2024). Similarly, the lack of a set seed for random number generators can mean that analyses relying on random numbers (e.g., bootstrapping) cannot be exactly reproduced. If no analysis script is available, analyses need to be recreated from the descriptions in the report (recoding reproduction). In this case, special attention should be paid to processing steps such as exclusion of outliers, transformation of variables, and handling of missing data. However, in many research areas information on these steps is often incomplete (Field et al., 2019); older research tends to be especially limited in terms of the methodological details they provide. In addition, we recommend testing the robustness of the original finding by making small alterations to the data processing and analyses procedure (robustness reproductions). For example, if the analyses were run for a subset of the data (e.g., participants aged 21 to 30 or without outliers ± 3 standard deviations), this subset can be changed (e.g., participants aged 18 to 30 or without outliers ± 2 standard deviations). Here, the initial focus should be on choices that are not determined by the theory that is presented, though this can also be used to explore the generalisability of some aspects of theory. Finally, if the original study was preregistered and the original code is available, reproduction researchers can check whether the original analyses adhere to the preregistered analysis plan.

If neither code nor data are available (or shared by the authors), no reproduction is possible. Researchers can still use automated tools to compare reported p-values with those that can be computed from test statistics via the website statcheck.io (where documents may be uploaded), the corresponding R package (Nuijten & Polanin, 2020), or papercheck (DeBruine & Lakens, 2025), which is still actively maintained.

Figure 3

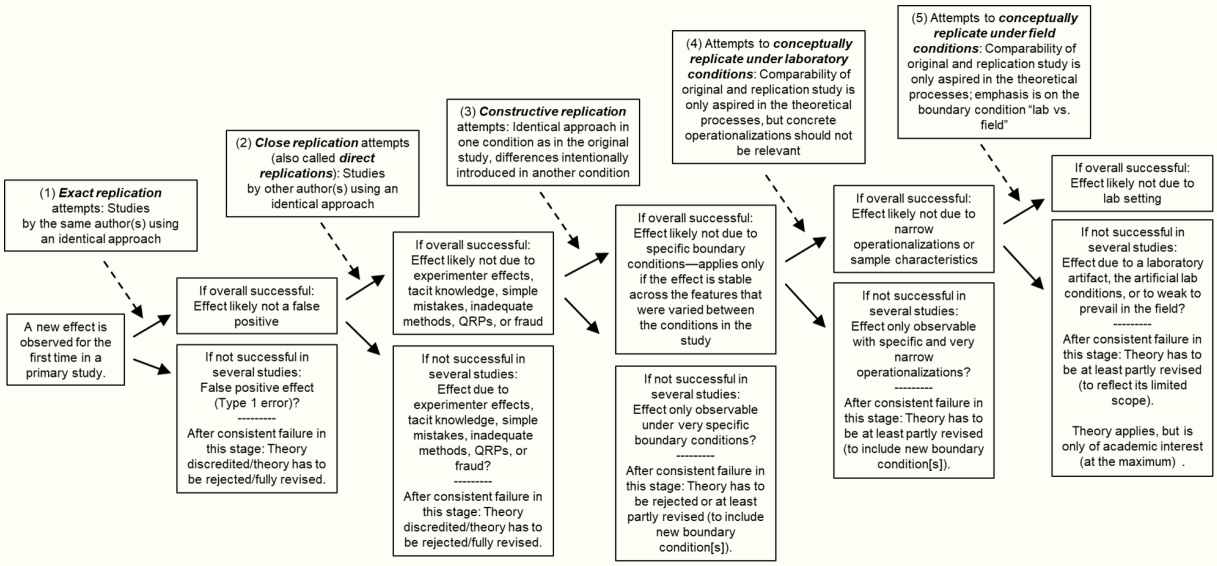


5.3 Close replication before conceptual replication

If the goal is to increase the generalizability of a specific finding, we also suggest starting with replications that adhere as close as possible to the original study (e.g., close replications) and only later conduct conceptual replications. Based on Hüffmeier, Mazei, and Schultze (2016), we propose the typology and order of replication attempts in Figure 4. Importantly, replications at any stage should not compromise any aspects of an original study, but rather (at the latest from the third study stage [constructive replications] onwards) try to improve one or more aspects of the original study, such as "[…] more valid measures, more critical control variables, a more realistic task, a more representative sample, or a design that allows for stronger conclusions regarding causality", see Köhler & Cortina, 2021, p. 494). Köhler and Cortina term such replications "constructive replications" and caution against the conduct of "quasi-random" replications that vary features without clear rationale.

Finally, there may be cases where the sequence of replications is not necessary, or where the context of the replication team requires a focus on generalisability to a specific context (see section The Role of Differences for the Interpretation of Findings).

Figure 4



Note: This an adaptation and update of the typology of replication studies by Hüffmeier, Mazei, and Schultze (2016). The typology is conceptualized as a hierarchy of studies that together help to (i) establish the validity and replicability of new effects, (ii) exclude alternative explanations, (iii) test relevant boundary conditions, and (iv) test generalizability.

6 Planning and Conducting Reproductions and Replications

Planning depends on whether the focus is on a certain method or a theory, that is whether the replication will be close or conceptual. Table 6.1 provides an overview of reproduction and replication types, or more generally "repetitive research" ((Schöch 2023)), drawn from different resources (e.g., (Dreber and Johannesson 2024; Hüffmeier, Mazei, and Schultze 2016); for an alternative taxonomy see also (Cortina, Köhler, and Aulisi 2023)). The decision between these types is the first step in planning.

In addition, the formation of the replication team is important, as replications can take substantial resources. Notably, repetitive research has successfully been conducted collaboratively with graduate and undergraduate students (e.g., (Boyce et al. 2024; Hawkins et al. 2018; Jekel et al. 2020; Moreau and Wiebels 2023)) and we recommend the use of replication studies to engage students of different levels in conducting and publishing research.

Table 6.1: Types of repetitive research ordered by reproduction and replication and respective closeness to the original study.

Type	Description	Goals
Computational Reproduction	Reanalysis of the same data with the same code	Correctness of original report
Recoding reproduction	Reanalysis of the same data, with new (equivalent) code	Correctness of original report

Type	Description	Goals
Robustness Reproduction	Reanalysis of the same data with new coding choices; can vary in closeness	Robustness of original finding and sensitivity to different analytical decisions or software
Multiverse analysis	Analyze data in all sensible ways (i.e., a large number of different robustness reproductions)	Robustness and generalizability of original finding, identification of potential moderators or sources for effect variability
Internal replication	Replicate one of your own studies as closely as possible	Demonstrate one's findings' generalizability across studies and rule out fear of false-positives (e.g., for new discoveries)
Close / direct / exact replication	Conduct a new study (based on work by other researchers) that is as close as possible to the original study	Rule out fear of the original finding being a false-positive, validate original materials or design, check generalizability/external validity for theoretically irrelevant variables (e.g., population, year of data collection)
Close replication with extension	Add a variable or procedure to a close replication	Rule out fear of the original finding being a false-positive, test generalizability of original finding
Conceptual / constructive replication	Conduct a study with changes that may be theoretically relevant but that tests the same hypothesis (e.g., different operationalization)	Generalizability of original finding, validity of theory
Multilab / multisite replications	Conduct studies that adhere to a predetermined replication protocol in many different locations at the same time	Robustness, generalizability over different locations, and receive a precise estimate of effect size and heterogeneity

6.1 Post Publication Conversations

When planning the replication study, additional knowledge should be taken into account such as any discussions of the original finding. There can be other studies citing the original studies, criticizing them, disconfirming their underlying theory, identifying errors, reinterpreting the finding, or making suggestions for replications. All of these might highlight considerations that need to be taken into account when designing a replication study that robustly tests the original claim or its generalisability.

Thus, replication researchers should look for post-publication discussions on the target study such as published comments and reviews, blog posts, or discussions on social media. These can often be found via Altmetric (https://www.altmetric.com) or other tools that allow researchers to quickly identify discussions on social media or news outlets beyond scientific journals (PubPeer [https://pubpeer.com], Hypothes.is [https://web.hypothes.is]), or the in-development platform Alphaxiv.org [https://www.alphaxiv.org/]; for a review see (Henriques et al. 2023)).

6.2 Reproduction before Replication

Many features of a replication study rest on the correctness of the original report. A reproduction allows researchers to investigate this by being able to uncover coding errors, fraud, robustness to analytical decisions, and generalizability. To make efficient use of resources, we encourage researchers to investigate the original finding's reproducibility and robustness first. In other words, ideally, reproductions should take place before planning and conducting a replication study. Depending on the availability of the code and data, these can take several minutes to weeks.

If the original code and dataset are available, researchers can try to numerically reproduce the results. Beware, however, that differences in software versions or default settings may lead to slight deviations or require corrections in some cases (for a large-scale test of reproducibility see Brodeur et al., 2024). Similarly, the lack of a set seed for random

number generators can mean that analyses relying on random numbers (e.g., bootstrapping) cannot be exactly reproduced. If no analysis script is available, analyses need to be recreated from the descriptions in the report (recoding reproduction). In this case, special attention should be paid to processing steps such as exclusion of outliers, transformation of variables, and handling of missing data. However, in many research areas information on these steps is often incomplete (Field et al., 2019); older research tends to be especially limited in terms of the methodological details they provide. In addition, we recommend testing the robustness of the original finding by making small alterations to the data processing and analyses procedure (robustness reproductions). For example, if the analyses were run for a subset of the data (e.g., participants aged 21 to 30 or without outliers ± 3 standard deviations), this subset can be changed (e.g., participants aged 18 to 30 or without outliers ± 2 standard deviations). Here, the initial focus should be on choices that are not determined by the theory that is presented, though this can also be used to explore the generalisability of some aspects of theory. Finally, if the original study was preregistered and the original code is available, reproduction researchers can check whether the original analyses adhere to the preregistered analysis plan.

If neither code nor data are available (or shared by the authors), no reproduction is possible. Researchers can still use automated tools to compare reported p-values with those that can be computed from test statistics via the website statcheck.io (where documents may be uploaded) or the corresponding R package ((Nuijten and Polanin 2020)).

6.3 Close replication before conceptual replication

If the goal is to increase the generalizability of a specific finding, we also suggest starting with replications that adhere as close as possible to the original study (e.g., close replications) and only later conduct conceptual replications. Based on (Hüffmeier, Mazei, and Schultze 2016), we propose the typology and order of replication attempts in Figure 6.1. Note that there may be cases where the sequence of replications is not necessary, or where the context of the replication team requires a focus on generalisability to a specific context (see ?@sec-interpretation).

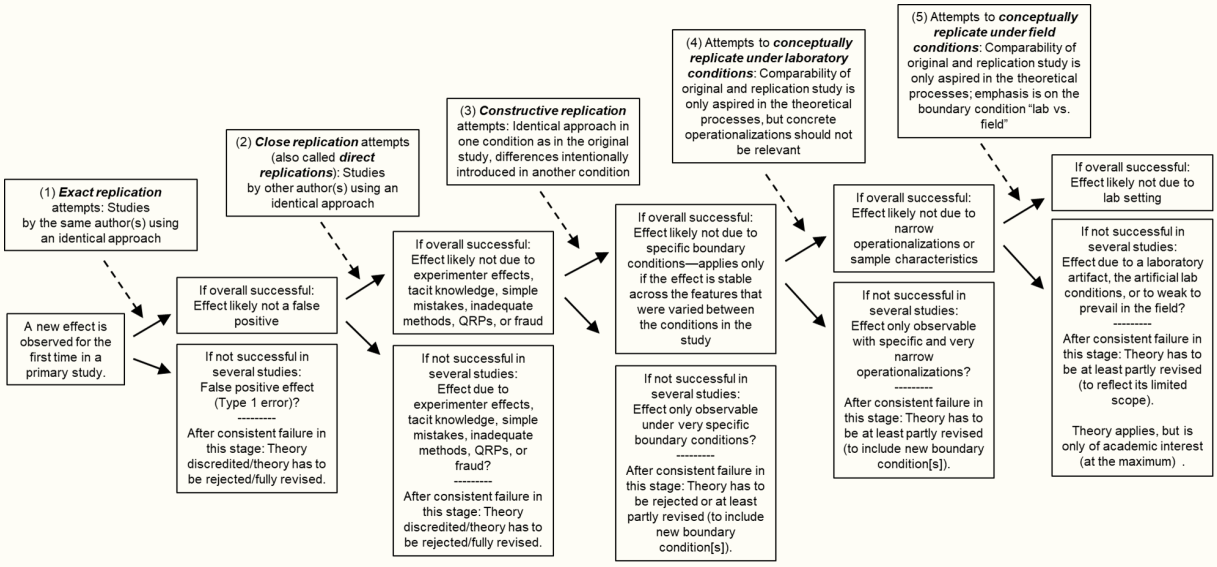


Figure 6.1: Sequence of replications from exact replications to conceptual replications under field conditions. Note: This an adaptation and update of the typology of replication studies by (Hüffmeier, Mazei, and Schultze 2016). The typology is conceptualized as a hierarchy of studies that together help to (i) establish the validity and replicability of new effects, (ii) exclude alternative explanations, (iii) test relevant boundary conditions, and (iv) test generalizability. Importantly, replications at any stage should not compromise any aspects of an original study, but rather (at the latest from the third study stage [constructive replications] onwards) try to improve one or more aspects of the original study, such as "[…] more valid measures, more critical control variables, a more realistic task, a more representative sample, or a design that allows for stronger conclusions regarding causality", see (Köhler and Cortina 2021, 494)). (Köhler and Cortina 2021) term such replications "constructive replications" and caution against the conduct of "quasi-random" replications that vary features without clear rationale.

At the start, you will need to collect all available materials and data, and contact the original authors if there is something missing. It will often be necessary to contact the authors more than once because missing descriptions of details of the original study only become apparent once the replication study is planned. In most cases, the original paper identifies one of the authors as "corresponding author" with an e-mail address. We recommend a quick web search to check if this is the current email address, as researchers frequently change institutions and thus e-mail addresses. Sometimes, it may be most helpful to write to the last authors instead, who tend to have more stable e-mail addresses, or to copy all authors into the email Templates for asking for materials and sharing replication results are in the appendix. Note that original authors may not respond due to institutional changes or not being active in academia anymore.

7 Execution of Reproductions

7.0.1 Gathering resources

Prerequisites for reproduction studies are available data and ideally also code. These are usually linked within the manuscript and shared via repositories (e.g., Zenodo, OSF.io, github.com, gitlab.com) or they are part of the supplemental materials that are listed on the article's website. In special cases, an entire original manuscript may be reproducible and written in Markdown language. Researchers searching for target studies to reproduce can check topfactor.org and filter for Data transparency level 3 (https://topfactor.org/journals?factor=Data+Transparency; will no longer be updated. They can also use the extensive database of economics studies with available data compiled by Sebastian Kranz (https://ejd.econ.mathematik.uni-ulm.de/).

If data are not publicly available, researchers can contact the authors of the original study. In this case, we recommend them to adhere to Guide for Accelerating Computational Reproducibility in the Social Sciences (ACRE) guidelines for constructive communication (Berkeley Initiative for Transparency in the Social Sciences, 2020; https://bitss.github.io/ACRE/comunications.html).

When re-using data, researchers need to respect licenses. Generally, research data should be licensed openly, that is re-use and alteration should be permitted, likely requiring citation of the original resource (e.g., CC-BY 4.0 Attribution). Note, however, that non-derivative licenses may prohibit reproductions; in that case separate approval would be required from the copyright holder.

When it comes to reporting, Ankel-Peters et al. (2025) provide a table for reporting results from the computational reproduction that includes resource availability (e.g., raw data, cleaning code, analysis code).

7.0.2 Contacting Authors

Reproduction authors may have to contact the original authors if there is something missing. It will often be necessary to contact the authors more than once because missing descriptions of details of the original study only become apparent once the replication study is planned. In most cases, the original paper identifies one of the authors as "corresponding author" with an e-mail address. We recommend a quick web search to check if this is the current email address, as researchers frequently change institutions and thus e-mail addresses. Sometimes, it may be most helpful to write to the last authors instead, who tend to have more stable e-mail addresses, or to copy all authors into the email Templates for asking for materials and sharing replication results in the appendix. Note that original authors may not respond due to institutional changes or not being active in academia anymore.

7.0.3 Identification of Claims

Statistical analyses and their results are always used as a way to evaluate a certain claim. While Ankel-Peters et al. (2025) recommend reproductions to identify "results [that] are essential for the paper's main argument to hold", we acknowledge that a reproduction can also focus on secondary results if they are relevant in some other context. In either case, reproduction researchers need to justify the choice of the claim in their report

7.0.4 Preregistration

Preregistrations contain a description of the planned study or analysis prior to their execution. This way, they can reduce researchers' 'degrees of freedom'. In the case of reproductions, they can prevent QRPs (e.g., "null hacking", Bryan et al., 2019; "gotcha

bias", Berinsky, Druckman, & Yamamoto, 2021) as long as the entire analysis plan is preregistered (Brodeur et al., 2024) and the data have not yet been accessed. While a numerical reproduction with available code does not require preregistration, we recommend a priori specification of all further planned analyses.

It should be noted that a preregistered analysis plan or analysis script is much easier to create with access to data and reproductions are impossible with unavailable data, pre-registration cannot exclude the risk of authors having already looked at the data, yet making fraudulent claims regarding data access in a pre-registration is evidently academic misconduct. How much weight readers and reviewers will give to a pre-registration based on data that could have been accessed already will differ, but generating it is a way to keep ourselves accountable and produce robust reproductions.

7.0.5 Deviations

To increase trust in the reported results, reproduction researchers need to report them in a transparent way, in a possible pre-registration and the final report. Ideally, all changes to the original procedure are explained, justified, and hypotheses about their expected effect on the outcomes are reported. Note that some journals' publishing reproductions require adherence to special requirements such a Registered Report format (e.g., Journal of Open Psychology Data) or including a minimum of two independent reproductions (e.g., Journal of Robustness Reports).

7.0.6 Analysis

The main part of the reproduction is the analysis. Factors that are potentially relevant for reproduction success include the software of the machine that is running the code as well as versions of the software and additional packages or plug-ins. For example, users of the open source software R can get a comprehensive overview of the program version and their machine using the function sessionInfo(), which should be included in supplementary materials. For python users, a package has been developed to run a similar function session_info.show() (https://gitlab.com/joelostblom/session_info).

Apart from a numerical reproduction where the same code is used, reproduction researchers can explore alternative ways that should and should not affect the results, test new hypotheses or theories, and run exploratory analyses. Their report should be clearly structured to discern these methods. Finally, for statistical analyses, the reproduction report should include reproducibility indicators (Dreber & Johannesson, 2024) that summarize statistical significance and relative effect sizes across the original and reproduction results. Ankle-Peters et al. (2025) recommend a visual summary of these indicators in the form of a reproducibility dashboard and specification curves (e.g., Simonsohn et al., 2020, see also Mazei et al., 2025). We strongly recommend reproduction researchers to consult the respective resources for further details.

7.0.7 Discussion

The discussion section should include a clear evaluation of the reproduction success on different levels (Ankel-Peters et al., 2025). Researchers should report possible reasons for failure (e.g., objective coding errors, changes in software packages) and the role of differences between the original and the reproduction studies' results with respect to their conclusions. Finally, if the original authors provided comments, the reproduction report should include a discussion of them.

8 Execution of Replications

8.0.1 Preregistration and Registered (Replication) Reports {#{#preregistration-and-registered-(replication)-reports}}

Due to the replications being met with skepticism, we encourage researchers to adhere to the highest standards of openness and transparency. This includes preregistering the replication including the analysis plan (ideally with an analysis code that was tested beforehand using data from test runs or simulations), and criteria for the results to distinguish between a replication success and failure. A preregistration without an analysis plan provides no safeguard against p-hacking (Brodeur et al., 2024). Beware that these criteria can be structured sequentially. For example, if there is a manipulation check, it can be defined that it has to work for the replicability to actually be evaluated. Boyce et al. (2024) also found that repeating unsuccessful replications did not change the outcomes unless obvious weaknesses were fixed.

There is a specific preregistration template by Brandt et al. (2014) but it may not fit the structure of some studies beyond social psychology (e.g., personality science or cognitive psychology; for a list of preregistration templates see https://osf.io/7xrn9 and https://osf.io/zab38/wiki/home). To facilitate publication of the replication, we furthermore encourage submitting it as a Registered Report. A rejection due to the results is not possible at this point. A list of journals offering Registered Reports (irrespective of replications) is available online (https://docs.google.com/spreadsheets/d/1D4_k-8C_UENTRtbPzXfhjEyu3BfLxdOsn9j-otrO870/edit#gid=0).

A special review platform for Registered Reports is Peer Community in Registered Reports (PCI-RR; https://rr.peercommunityin.org) where a community reviews pre-prints. Once

accepted by PCI-RR, authors can decide to publish their paper in participating journals (PCI friendly journals) without another editorial round.

Finally, replication researchers need to deal with deviations from their preregistration in a transparent way. In principle, there is nothing wrong with deviating from what one had planned but most importantly, all changes should be listed, discussed, and it should be made transparent how the changes affected the results (for recommendations on changes and documentation, see Heine et al., 2024; Lakens, 2024; Willroth & Atherton, 2024). If changes are noticed during the data collection, many platforms also allow the upload of amendments with preserved version history.

8.0.2 Sample Size Determination

For replication studies, power analyses or other types of sample size justification can be simpler than for studies testing entirely new hypotheses because there already is a study that did what one is planning, with a result that one can refer to. However, we advise against simply using the original study's sample size. While the maxim for most decisions is "stay as close as possible to the original study", sample sizes of replication studies usually need to be larger. To be informative, replication failure should provide evidence for a null hypothesis or a substantially smaller effect size, which requires a larger sample. While a general tutorial for sample size justification is provided by Lakens (2022b), we briefly present approaches that are fit for replication studies.

As a pair of original and replication studies is usually concerned with multiple effect sizes (e.g., for different scales/items/groups/hypotheses), their number and individual power need to be considered carefully. If the interpretation will rely on the significance of all effect sizes, the total power will be smaller than the power for each individual test. To get along with limited resources, researchers may choose one single effect size and argue that it is central, or clearly specify other methods for aggregation across results (e.g., testing multivariate models).

8.0.2.1 Small Telescopes Approach

The idea behind the small telescopes approach (Simonsohn, 2015) is that a replication study should be precise but how far this precision exceeds the original study should be limited. Specifically, the replication study should be able to detect an effect size for which the original study had insufficient power (usually 33%). If that effect size can be ruled out, the original study can be treated as uninformative, as with such low power, the result becomes more likely to have been a false positive.

This approach is based on the notion that replications should assess the evidentiary value of the original study, and that the 'burden of proof' shifts back to proponents of a hypothesis if their evidence is shown to be very weak. It is particularly appropriate when original studies are very imprecise. In that case, a replication that finds a much smaller effect may well still be compatible with the (wide) confidence interval of the original study, and it might be impossible to reject the original claim on that basis.

As an example, Schultze et al. (2017, Figure 4) found an effect in three studies with an average effect size of r = -.11, 95% CI [-.22, -.01].

If we wanted to achieve high power to rule out an effect of -.01, and thus show that the true effect does not fall into their confidence interval, we would need a sample size of 108,218 participants (alpha = 5%, one-tailed test[^4]). Conversely, with the small telescopes approach, we would aim to test whether the replication effect is smaller than the effect which the original study had 33% power to detect, r = -.043 (alpha = 5%, one-tailed test). Simonsohn (2015) showed that this requires a sample 2.5 times as large as the original for 80% power. However, we deem that level of power insufficient for replications, and instead suggest aiming for 95% power (given that a false negative in a replication leads to a wrong claim regarding the absence of an effect). This requires a multiple of 4.5 (rather than 2.5), so a sample is in this case of 4.5 * 793 = 3,569 participants. If this replication then results in an estimate that is significantly smaller than the effect the original study had 33% power to detect, the small telescopes approach would suggest treating the original study as unable to provide reliable evidence for its claim.

8.0.2.2 Equivalence Testing

If statements can be made about the smallest effect size of interest (SESOI), researchers can aim to test whether the replication effect is smaller than that. Given that the direction is fixed by the original, this simply requires running a one-sided test, e.g., a t-test in the case of a two group design, in the "lesser" direction. If the replication effect size is significantly smaller than the SESOI, the original claim is taken to be refuted in this instance by those who accept that this is really the smallest effect of interest. Lakens et al. (2018) provide a practical tutorial on equivalence testing, though they focus on cases where observations in either direction would falsify the null hypothesis.

8.0.2.3 Bayesian Approach

External knowledge can be incorporated into sample size planning (uninformative / flat priors; heterogeneity; shrinkage) using the R package BayesRepDesign (Pawel et al., 2023). Moreover, Micheloud and Held (2022) provide a method for incorporating an original study's uncertainty into power calculations. With interim analyses (e.g., sequential testing) , a replication study can also be stopped early and prevent wasting resources (Wagenmakers, Gronau, & Vandekerckhove, 2019). However, when planning to use Bayes Factors to make inferences about replication success, it is important to plan to use plausibly narrow priors. Priors that assign substantial likelihood to effects rarely observed (e.g., N(0,1) priors for standardized mean differences in the social sciences) may be taken to unfairly privilege the null hypothesis, which is inappropriate for a study setting out to find support for it.

8.0.2.4 Meta-Analytical Estimates

If the replication study is part of a larger research programme, it is possible to include other studies in the estimate of the (minimum) effect size one wishes to detect/rule out. The target study may be part of a multistudy paper with at least one other study that includes an effect size for the hypothesis of interest. Researchers can compare the effect

sizes and possibly pool them to get a more precise estimate (for a related Shiny App, see for instance McShane & Böckenholt, 2017).

Metrics such as average effect sizes, heterogeneity, or the confidence interval width are valuable estimates needed for the replication's sample size justification. If there is a meta-analysis on the general topic, researchers can also use that to inform sample size planning, but should prioritise estimates that aim to correct for publication bias and other QRPs (for an overview see Nagy et al., 2024). They should also choose effect sizes from a set of studies that resembles the planned replication study as closely as possible. For correlational effects, researchers can check metabus.org (Bosco et al., 2017) to identify similar studies.

8.0.2.5 Multilab Replications

Multilab replications, that is replications that are conducted by different groups of researchers in different locations adhering to the same protocol, allow researchers to investigate heterogeneity of effects and estimate effect sizes with high precision. There are currently no standards for planning sample sizes for multilab replications. Depending on the specific goals, a power analysis needs to account for possible moderator hypotheses and the desired precision of effect size, heterogeneity estimates, or cultural variables. Note that this often requires large sample sizes for any level of the moderator (e.g., culture, profession). Usually, the different labs are required to collect data from a minimum number of participants. Each lab's study and all analysis scripts should be preregistered to prevent local and global QRPs such as optional stopping or ad hoc exclusions of single labs.

8.0.3 Changes in the Methods

A replication study should closely resemble the original study, in the case of conducting a direct/close replication. However, this is difficult for multiple reasons: First, original studies may not include sufficient detail to allow for a replication (see Aguinis & Solarino, 2019; Errington et al., 2021). Second, scientific progress in the form of new methods

and insights and cultural changes might require replication researchers to make changes or additions to their study. Third, obvious errors must be corrected. We elaborate on a number of reasons to deviate from an original study. In the replication report, all deviations should be reported and justified exhaustively.

- Unspecific original materials: If the original study does not specify a key element that is needed for the replication, replication researchers can reach out to the original study's authors and ask for the details. If this is not possible because authors cannot be reached or they are unwilling/unable to share the materials, new materials must be created. In this case, special attention should be paid to the theory, so that the new materials exhibit both face and construct validity.

- Deprecated materials: If a psychological study about person perception published in the 1980s used celebrities, the examples used may no longer have the same status. For example, Mussweiler et al. (2000) used "a 10-year old car (1987 Opel Kadett E)" to be evaluated in German Marks. For a new study, car and currency would have to be replaced as a car's age is strongly associated with price. Like most studies, the original provides no details about the conditions that a new stimulus would have to meet. Ideally, the theoretical requirements for stimuli should be specified in primary research, where they are not, replication authors need to make their own assumptions and report them explicitly (see Simons et al., 2017).

- Translation: Most published original studies are in English. If the replication sample's mother tongue is not English, translation may be necessary. Standards for translation differ strongly even between subfields. For example, when a personality scale is translated, the translated version will usually be validated and tests of invariance will be required. In social psychology, such procedures are less common, and often merely a back-translation is conducted. However, in any field, measurement invariance is required if one wants to compare effect sizes across samples, so that this should be tested rather than assumed where possible.

- Necessity of a special sample: Many large-scale replication projects (e.g., Chang & Feldman, 2024) made use of click workers (e.g., via MTurk) or use student samples. Replicators should consider if using such samples satisfy their needs and evaluate

which platform to use (for best practices and ethical considerations, see Kapitány & Kavanagh, 2024). Even if the original study used such a convenience population, changing to a different convenience population may require tweaks to maintain comparability, e.g. with regard to participant attentiveness and engagement with the paradigm.

- Quality of methods and apparatus: Replicating old studies often faces the problem that something new has been discovered that should be taken into account. If a specific tool or method is used, there may be another recent method that is more reliable. For example, software for eye tracking studies from the early 2000s is now deprecated; there is new hardware and software that researchers will use. This might also apply to analysis methods, yet where possible, both the results from the original methods as well as state-of-the-art methods should be reported; where a choice has to be made, it is essential that invalid methods are avoided while comparability is maintained as far as possible. Finally, if the original finding's generalizability is tested, new items or tasks that vary more or less systematically can be added to compare results for the original parts versus these extensions (though order effects have to be carefully considered, as a second manipulation might affect participants differently from a first manipulation)

- Adding checks: Doing a replication often implies some uncertainty in the results, so it is wise to include checks that will be helpful to interpret the results, especially if they are negative. For example, if there are occurrences that would make the results meaningless, it is good to have a way to measure them and incorporate that into the study. This could include positive or negative controls (items that are diagnostic of the method rather than the question of interest), manipulation checks (generally placed after the critical parts of the experiment), or attention checks. See Frank et al. (2025, chapter 12.3) for further discussion.

8.0.4 Piloting

If considerable resources are linked to the full execution of a replication (e.g., in a Registered Replication Report), or when new materials are used, researchers may want to

consider piloting it (or parts of it) first. For multi-lab replications, researchers may want to consider a sequential study order in contrast to a simultaneous design: As Buttliere (2024) put it: "Who gets better results, 39 people doing it the first time or one person doing it 39 times?" (p.4) Beware that piloting may not be of value if it is simply an underpowered version of the study; instead it may be used to identify flaws in the methodology or test assumptions about the distribution of values or participants' qualitative responses. Importantly, small pilot studies should never be used to derive effect sizes for power analyses as their results are too imprecise.

For instance, researchers should follow general best practices for their replications including piloting their study on a few participants to ensure that the instructions are clear, that the procedure works smoothly (e.g., website loads appropriately), and that all necessary data are recorded. A debriefing survey where pilot participants are asked about their experience, the clarity of instructions, and the clarity of any user interface, can help to identify some issues that could undermine the replication. See Frank et al. (2025, chapter 12.3.1) for further discussion on piloting studies.

8.0.5 Collaborating and Consulting with the Original Authors

To reduce the chance that a failure to replicate is dismissed by the original study's authors afterwards by pointing out a flaw in the method, researchers can consult with the original authors before running the study. However, in the past, this still has not kept the original authors from dismissing a replication as an inadequate test of a hypothesis (e.g., Baumeister & Vohs, 2016). Note that replication researchers have even been accused of "null hacking" (Protzko, 2018) although little evidence exists for this claim (Berinsky et al., 2020). While involving original authors can help in creating a good study when reporting is poor, ideally original studies should be reported in sufficient detail for others to replicate them without further involving the original authors. Historically, the relationship between involvement of original authors and the average replication effect size is not clear (although there have been lab effects in some cases, Power et al., 2013). This is showcased here in a few examples:

- Ten effects from Open Science Collaboration (2015) were replicated in Many Labs 5 (Ebersole et al., 2020), where the original authors commented on the study protocols of the planned replication before these replications were conducted, and "the revised protocols produced effect sizes similar to those of the RP:P protocols (Δr = .002 or .014, depending on analytic approach)."

- McCarthy et al. (2021) conducted a multisite replication of hostile priming where one of the original authors was involved. Each laboratory conducted a close and a conceptual replication and found no difference and recommended that "researchers should not invest more resources into trying to detect a hostile priming effect using methods like those described in Srull and Wyer (1979)".

- After Baumeister and Vohs (2016) criticized the failed registered replication report by Hagger et al. (2016) for their methods, Vohs et al. (2021) conducted another registered replication report and also found a null effect.

- After no effect of the pen-in-mouth task was found in the facial feedback Registered Replication Report by Wagenmaker et al. (2016), another multilab test, which included one of the original authors, arrived at the same results (Coles et al., 2022).

- The Many Labs 4 project set out to test the effect of author involvement on replication success but found an overall null effect for the group of studies that did and that did not include original findings' authors (Klein et al., 2022).

- For social priming studies' replication success, "the strongest predictor of replication success was whether or not the replication team included at least one of the authors of the original paper" (Mac Giolla et al., 2022, Abstract).

8.0.6 Adversarial Collaborations

Although they are not specific to replication projects, researchers have often issued calls for adversarial collaborations (e.g., Clark et al., 2022, Cowan et al., 2020, Corcoran et al., 2023). Thereby, groups of researchers can collaborate and try to settle conflicting views by designing and conducting a study designed to settle a debate. A related idea are "red

teams" where experts are invited to critique the analysis plan, without becoming authors and thus without a conflict of interest in terms of desired results (Lakens & Tiokhin, 2020).

8.0.7 Analysis

Analyses of replication results are often a compromise or a combination of the original analysis and the current state-of-the-art. Generally, replication studies should follow the original analysis plan as closely as possible. That does not only concern statistical procedures but also data processing (e.g., exclusion of outliers, transformation and computation of variables). Even when following the original analysis plan for their confirmatory analysis, researchers should still follow best practices and examine their raw data to check for distributional anomalies to detect whether participants might be inattentive, guessing or speeding, and report relevant sensitivity checks where helpful. Some things to check for include theory-agnostic condition/manipulation checks (e.g., were participants faster in the condition focused on speed?) and the results of attention checks or control trials. Generally, it is advisable not to remove participants from the main analysis on that basis, but instead to confirm that the rates of non-compliance are acceptably low and to report robustness to the exclusion of these participants. See Ward and Meade (2023) for a comprehensive review of strategies for assessing and responding to careless responding.

At times, methodological advances may suggest that the original statistical tests are not robust. In such cases, researchers may want to run both the test that the original study used, as well as the statistical approach that is most appropriate by today's standards (for instance, both the t-test that can be compared with the original, and the mixed-effect model that is justified by the study design). Where original data is available, or can be obtained from the original authors, researchers might be able to also update the analyses in the original study, which facilitates interpretation.

Where original statistical analyses are fundamentally flawed, replication researchers are faced with a difficult choice. For instance, it has been convincingly demonstrated that the famous Dunning-Kruger effect (1999) is based on analyses strongly influenced by a statistical artifact, namely regression to the mean (Gignac & Zajenkowski, 2020). In such

a context, one may want to report results based on the original methods alongside more robust test, yet needs to be very careful to frame them in a way that "replication success" cannot be claimed in the absence of evidence for the original claim.

. Exclusion criteria are another area where there may be tension between the original study and current best practices. Typically, it makes sense to run the analysis both ways to check for robustness, yet one analysis choice should be preregistered as the central analysis.

Naturally, original and replication results should be compared. Unstandardized values can be informative with respect to sample characteristics (e.g., overall reaction times). How to do this analytically depends on the choice of success criteria discussed in the next section.

9 Discussion

9.0.0.1 Defining and Determining Replication Success

There is no strong consensus yet on what constitutes a replication success and some approaches can be biased (e.g., Schauer & Hedges, 2021) or imprecise (Patil et al., 2016b). Like in classical null hypothesis significance testing (NHST), replication researchers face the trade-off between dichotomizing something that is not dichotomous (success vs. failure) and making a clear decision about the outcome. On the one hand this is a question about statistical choices and their interpretation, namely how to compare original and replication effect sizes (or p-values) and how to interpret differences. On the other hand, it is a more complex question about how to interpret a mixed pattern of results, where some results are consistent across original and replication, while others are not. Here, it is important for replication researchers to specify which effects are of primary interest in their pre-registration, and how they will aggregate results, noting that requiring multiple effects to yield the same result will reduce statistical power.

Below, we briefly present different approaches to assessing replication success. For a review and a computational implementation of these and other replication success criteria, see also Heyard et al., 2025; Muradchanian et al., 2021; Röseler & Wallrich, 2024; Errington et al., 2021, Table 1).

9.0.0.1.1 Qualitative criteria

Researchers often face the difficulty of making an overarching decision on replication success, failure, or inconclusiveness due to the decision relying on a complex interplay of differences between original study and replication, different original and replication

results, and vague theories. We term cases where the decision has not yet been formalized qualitative. It is inferior to quantitative criteria in that it is not formally reproducible but it is superior to quantitative criteria in that it allows researchers to consider a large number of details.

9.0.0.1.2 Quantitative Criteria

With quantitative criteria we refer to those that can be automatically computed based on study details such as effect sizes and statistical tests. They are often a combination of whether the replication effect is different from the expected effect under the assumption of the null hypothesis, different from the original effect, and whether the aggregated effect is different from zero. We list different approaches and brief descriptions in Table 3.

Table 3

Quantitative criteria to operationalize replication success

Reference	Type	Description
Brandt et al., 2014	NHST	Comparison of replication effect to 0 and to original effect Success: different from the null (i.e. statistically significant), and similar to the original (i.e. in its 95% confidence interval) or larger and in the same direction Informative failure to replicate: either not different from null, or in the opposite direction from the original, and significantly different from original (i.e. outside its 95% confidence interval) Practical failure to replicate: both significantly different from the null and significantly smaller than the original Inconclusive: neither significantly different from null nor the original

Reference	Type	Description
Anderson & Maxwell, 2016	NHST (Bayesian)	Consider 6 distinct goals To infer an effect Conduct statistical test on replication effect, and conclude success when it is significant and in same direction as original To infer a null effect NHST: Conduct an equivalence test, showing that the hypothesis that the effect is very small is significant (e.g., one-sided t-test against d = .1) Bayesian: Test whether posterior is in Region of Practical Equivalence (better for multivariate hypotheses) Bayesian: Bayes Factors quantify the relative support for alternative hypothesis and null hypothesis, and can thus provide evidence for the null Quantify the size of an effect Focus on increasing precision of effect size estimate, primary success criterion is to achieve a confidence interval width that allows for decision-making (e.g. to assess value for money) To infer an effect across the two studies Combine original and replication effects in meta-analysis to estimate a more precise and robust effect Accounts for the fact that a non-significant replication can

Reference	Type	Description
Steiner et al., 2023	NHST	Test difference and equivalence of original and replication effects Depending on the combination of outcomes, there is equivalence, difference, a trivial difference, or indeterminacy
Bonett, 2020	NHST	Compute confidence intervals for original and replication effect as well as for the difference of the two effects Combination of the three confidence intervals determines the type of evidence across 9 categories (Bonett, 2020, Table 2)
Verhagen and Wagenmakers, 2014	Bayesian	Test the likelihood of the replication finding between two competing priors: a skeptic hypothesis suggesting and effect size of zero, and a proponent hypothesis, based on the posterior probability from the original study (i.e. a prior centred on the effect size obtained in the original study) Results in a weighted-likelihood ratio between the two hypotheses (i.e. a Bayes Factor)

Reference	Type	Description
Held et al., 2022	Bayesian	Establish a "sceptical prior" to be just strong enough to render the original study's finding non-significant, and test the replication effect against it. Replication success is then declared if the replication study's data conflicts with this sceptical prior, which is equivalent to the relative effect size exceeding a minimum required threshold.

If researchers are unsure about how to compare results for their replication study with originals, they can also browse the FORRT Replication Database (Röseler et al., 2024) and look for replication studies from their area of research, though they might note that current practice is highly variable and often not sufficiently justified.

9.0.0.2 Interpreting Divergent Results (Replication Failures) {#interpreting-divergent-results-(replication-failures)}

When replications succeed, the original claim gains further credence (as long as the methods are sound). However, when replications fail, many explanations and interpretations can be advanced, which need to be carefully considered and discussed in a report. While replication failure can highlight issues with statistical conclusion validity in the original studies (John et al., 2012; Nelson et al., 2018; Simmons et al., 2011), other explanations need to be considered, including issues with internal, external, and construct validity in both original and replication studies (Fabrigar et al., 2020; Vazire et al., 2022). For example, internal validity is threatened when attrition rates differ between experimental

conditions in original or replication studies, creating potential confounds in the interpretation of treatment effects (Zhou & Fishbach, 2016). Construct validity is threatened when original or replication studies use unvalidated ad-hoc measures, fail to employ validated manipulations of the target construct, or when differences in sample characteristics between original and replication studies mean that manipulations and measures do not work as intended (Fabrigar et al., 2020; Fiedler et al., 2021; Flake & Fried, 2020). External validity is threatened when original findings do not generalize to the specifics of the replication study due to person and context differences between studies that moderate the effect. Thus, before making statements about the original finding's robustness and generalizability, replication researchers need to critically discuss potential methodological shortcomings in both original studies and replication attempts that limit statistical conclusion, internal, external, and construct validity.

9.0.0.2.1 Hidden Moderator Account

One challenge for replication researchers is the identification of hidden/unknown confounds that may influence or bias the phenomenon under study. Each study has a set of potential extraneous or unknown moderator variables that is unique to it. These may seem trivial, such as the brightness of an experimental laboratory, or important, such as a cultural difference between the replicating and original studies. Yet even seemingly trivial differences could potentially change results. Often statistical and methodological choices are made to circumvent or attenuate these issues. However, for some paradigms, these variables could be unknown to the original researcher (Fiedler, 2011). These are referred to in the literature as unknown moderators, background variables, hidden moderators or fringe variables. While they are always a way to reject unpleasant replication results, they can potentially bias replications, which highlights that a single replication is never entirely conclusive (though it might raise enough doubts that researchers do not see the value in addressing the remaining uncertainty). It should be noted that the same argument could be applied to raise doubts about any original study, questioning whether the effect is really due to the hypothesised cause or due to some hidden moderator or background variable. Clearly a skeptic who stops at that level would not be taken very seriously, so that it is important to move conversations about replication failure beyond general suspicion of hidden moderators.

Bargh (2006) suggested that the evidence generated by empirical findings far outweighs the resources of (social) psychology to conceptualize and understand the mechanisms underlying their effects. Therefore, boundary conditions are not easily specified, which can impact both direct and conceptual replication success. Replication failure indicates that the original claim does not generalise to the setting of the replication. Whether that generalises to the setting of the original study needs to be considered in light of theory, and might be a legitimate matter of contention.

9.0.0.3 The Role of Differences for the Interpretation of Findings

Each replication outcome should be evaluated in the light of its closeness, which is why all deviations with the respective reasons and, if possible, their potential impact on the results should be discussed. Existing theories may help assess whether a deviation should affect the outcomes. For example, most psychological theories are agnostic towards age so that a different distribution of participants' age will be unproblematic in most cases. Researchers may choose to evaluate replications from both phenomenon-focused / inductive and theory-focused / deductive views. Different types of interpretations are listed in Figure 5 and integrated from previous accounts by Borgstede and Scholz (2021) and Freese and Peterson (2017, Figure 3).

Figure 5

Interpretation of replication outcomes depend on similarity of closeness and results as well as the view (inductive vs. deductive).

![][image5]

9.0.0.4 Comments from the Original Study's Authors {#comments-from-the-original-study's-authors}

If the replication results do not converge with the original results, replication researchers can reach out to the original study's authors and ask for a comment that they can publish together with the replication report. A template for asking for a comment is in

the appendix. Note that some journals (e.g., Journal of Replications and Comments in Economics) require such statements at the time of submission.

Part III

Advanced Topics and Applications

10 Publishing and Communicating

The final step of replication research is publishing and communicating the results. Researchers should adhere to best practices of transparency and openness promotion guidelines (TOP, 2025; Grant et al., 2024) and to the reporting standards of their respective field (e.g., JARS standards for reporting psychology replications, https://apastyle.apa.org/jars/quant-table-6.pdf). For example, they should report a link to the preregistration, analysis plan, and analysis script, share all materials and data (if possible in light of ethical and legal limitations) under an open license (see also Janz & Freese, 2021), and report methods and results comprehensively.

We list several options for writing and publishing the report in Table 4. These are non-exclusive, that is, researchers can choose multiple of them. An overview of active journals that exclusively publish replications is in Table 5.

Table 4

Reporting and communicating reproductions and replications.

Type	Description
FORRT Replication Database	This open and collaborative database contains thousands of replication findings and makes them visible. Anyone can enter results using a guided survey (https://t1p.de/fred_submit).

Type	Description
PubPeer	Researchers can comment on the original study and say that there is a replication attempt, describe the outcome, and provide links/references/DOIs to the replication(s). Researchers checking pubpeer.com or using the browser plug-in that automatically highlights studies for which there are comments will see your comment.
Manuscript (required for Preprint and Journal Article)	Manuscripts are mostly used as they are the traditional form of a research article. For judgment and decision making, there are useful examples by Feldman (2024). For reproducibility analyses the I4R Replication Report Template (https://osf.io/j2qrx) can be used. Moreover, Röseler et al. (2025, https://osf.io/brxtd) provide general templates for reproductions and replications.
Preprint	We recommend publishing a report in the form of a traditional or standardized manuscript as a preprint. This secures open access and makes the report visible, citable, and commentable. There are many preprint servers across the social sciences (e.g., PsyArxiv, SOCARXIV, SportRxiv, MediArXiv, MindRxiv, EdArXiv, AfricArXiv, or MetaArXiv). In some countries, researchers have a legal right for a secondary publication of their research (green open access). Be aware that preprints are faster in terms of publication than journal articles, but are usually not peer-reviewed.

Type	Description
Journal article	Most researchers have to "play by the rules", that is, publish or perish (Bakker et al., 2012; Koole & Lakens, 2012). While some have argued for a pottery barn rule (https://thehardestscience.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/) where journals that published the original finding have to publish respective replication attempts, many journals are not (yet) interested in replications. Notable exceptions are listed in the appendix. This is why journals dedicated to replications have emerged (see Table 5). Moreover, researchers can submit their preprint to a PCI community (see https://peercommunityin.org/current-pcis/), which is a preprint peer-review service. Several journals are PCI-friendly, which means that they publish articles recommended by the respective PCI. Many institutions and libraries recommend adding a CC-BY disclaimer on journal submissions that give the researchers the right to use the accepted manuscript as they like or choosing Diamond Open Access journals that are defined by no fees for publishing and reading research.

Table 5

Active journals dedicated to reproductions and replications.

Journal name	Commercial status	Owner(s)	Discipline(s)	Article types	Website
Journal of Comments and Replications in Economics	Non-commercial, diamond OA	ZBW	Economics	Replications, Reproductions and comments research	https://jcr-econ.org
Replication Research	Non-commercial, diamond OA	Münster Center for Open Science and FORRT	Multidisciplinary	Reproductions, Replications, Conceptual articles	https://replicationresearch.o
Journal of Open Psychology Data	Commercial, Gold OA (APCs: 450 pounds)	Ubiquity Press	Psychology	Reproductions (only as Registered Reports)	https://openpsychologydata.metajnl.com
Journal of Robustness Reports	Non-commercial, diamond OA	SciPost	Multidisciplinary	At least two independent reproductions are required, limited to 500 words	https://scipost.org/JRobustRep
Rescience C	Non-commercial, diamond OA	Olivia Guest, Benoît Girard, Konrad Hinsen, Nicolas P. Rougier	Multidisciplinary	Reproductions	https://rescience.github.io
Journal of Management Scientific Reports	Commercial (subscription based)	Sage	Management	Replications, reproductions, related methods	https://smgmt.org/jomsr/

Journal name	Commercial status	Owner(s)	Discipline(s)	Article types	Website
Journal of Reproducibility in Neuroscience	Non-commercial, diamond OA	Center of Trial and Error	Neuroscience	Replications, Comments, Reviews, conceptual articles	https://jrn.trialanderror.org
Rescience X	Non-commercial, diamond OA	Etienne B. Roesch	Multidisciplinary	Replications (Experiments)	http://rescience.org/x
AIS Transactions on Replication Research	Non-commercial, diamond OA	Association for Information Systems (?)	Information Systems	Exact, Methodological, Conceptual Replications	https://aisel.aisnet.org/trr/

11 Field-Specific Replication Challenges: An example from MRI research

11.0.1 Introduction

While the principles of reproducibility and replication apply across scientific disciplines, certain fields face distinct methodological and practical challenges. Neuroimaging research, particularly MRI-based studies, is one example where field-specific complexities cause specific challenges for data sharing, reproducibility and replicability. Other fields may have different specialized requirements on these topics. Generally, false-positive findings are likely driven by a combination of low statistical power, a high number of researcher degrees of freedom and statistical tests, and biased motivation towards obtaining positive (i.e., significant) results (Ioannidis, 2005). Most of these factors are arguably aggravated in MRI studies, making replication research in this field particularly relevant albeit challenging. In addition, the analyzed data and obtained findings are characterized by a three-dimensional spatial component (or four dimensions in case of functional MRI studies (fMRI) in combination with time series data), which further complicates the matter. In the following we summarize the inherent peculiarities of replication research in the field of neuroimaging.

11.0.2 Researcher Degrees of Freedom

Brain imaging comes with a massive number of researcher degrees of freedom along the preprocessing and analysis pipelines. Preprocessing steps include for example motion correction procedures, spatial normalization and smoothing, with additional steps necessary

for some imaging modalities, such as temporal signal filtering for fMRI. For each of these steps a multitude of parameter options and toolboxes are available. It has been shown that different preprocessing toolboxes can lead to fundamentally different results, even when aiming to harmonize all parameters (Zhou et al., 2002), and that different teams analyzing the same dataset can arrive at different final conclusions dependent on the used pipeline (Botvinik-Nezer et al., 2020). Furthermore, a large variety of operationalizations of neurobiological targets is available. For example, cerebral gray matter structure could be investigated as voxel-wise gray matter, segmentation-based regional cortical surface, thickness or gyrification.

Analysis-wise, the high number of researcher degrees of freedom is mainly a consequence of the multidimensional data structure. Basically, the central question is where in the brain to look for effects and how to define significance in the face of a large number of tests. There is an immensely high number of single data points represented by spatial units in the obtained individual images (e.g., two-dimensional pixels or three-dimensional voxels). Analysis is often done utilizing mass-univariate approaches where a statistical model is calculated separately for each of these spatial units. For example, in cerebral MRI research the analysis of 400k voxels is common. To avoid false-positive findings, region-of-interests (ROIs) are often defined or the analysis is restricted to a smaller region in the brain (i.e., small volume correction) to narrow down the search space and unique methods to correct for multiple testing are applied (Han et al., 2019). This again results in a multitude of options, such as the anatomical vs. functional definition of a ROI based on several different atlases and a variety of voxel-based or cluster-based inference methods to choose from. Botvinik-Nezer et al. (2020) gave the same fMRI dataset (raw data and preprocessed data), along with predefined hypotheses to 70 independent analysis teams and observed substantial variation in obtained results, attributable to variability in the analysis pipelines (in fact, none of the 70 teams used the same pipeline). Even when the same code and data is available the reproducibility of MRI analysis can be challenging (Leehr et al., 2024).

11.0.3 Sample Size Justification

The gold standard for sample size justification is a power analysis. In neuroimaging this is complicated by the outlined mass-univariate three-dimensional data structure. Any power analysis would need to incorporate assumptions about the covariance structure of all data points, as well as the spatial extent and distribution of statistical effects, and the method to correct for multiple tests. While these numerous tests are not independent from another, the extent of their spatial covariance structure is difficult to assess and depends on preprocessing steps, such as image smoothing but is also on the data and the specific research question. Due to the high number of single data points, the obtained result is not a single statistical estimate with an effect size but rather a highly individual three-dimensional distribution of effect sizes around a peak localization. Simulation-based power analysis approaches have been previously suggested to address this problem. However, valid simulations require assumptions about valid spatial distributions of effects (contingent on regional anatomical peculiarities and on the specific research question), often difficult to assess and many developed power analysis tools have been discontinued. To date the utilization of power analysis is extremely rare in MRI research.

Without proper power estimation, justifying sample size becomes challenging. As in other fields of research the statistical power ultimately depends on the expected effect size. Recent large-scale investigations in the domain of mental health neuroimaging suggest that maximum underlying effect sizes are very small across various neuroimaging modalities (below 2% explained variance; Marek et al., 2022; Winter et al., 2022) and could require thousands of individuals to obtain robust and replicable statistical estimates (Marek et al., 2022). In contrast, given the labor-intensive and costly nature of MRI assessments, most MRI studies tend to have small sample sizes, making them likely underpowered (Button et al., 2013). Smaller samples may be suitable however, for research questions where the neurobiological effect sizes are expected to be larger, such as in psychosis research or when using highly individually tailored or within-subject designs (Lynch et al, 2024; Marek et al., 2022; Rosenberg & Finn, 2022; Spisak et al., 2023).

11.0.4 Criteria of Replication Success

Regarding the definition of replication success, the three-dimensional data structure requires special attention when defining replication success. In addition to other possible definitions, it has to be defined where in the brain the criteria of replication success should be met. As discussed above, there is not only one effect size but rather a 3D map with an effect size for each spatial unit (e.g., voxel). Goltermann & Altegoer (2025) describe a variety of potential criteria focusing on statistical significance in accordance with different spatial definitions revolving around the original finding. These include significance either at the peak voxel location (where the effect in the original study had the largest effect size), or in a ROI that can be defined in terms of spatial proximity to this peak voxel (for example a 15mm sphere with the peak voxel as a center) or in terms of an anatomically defined region where the original effect was found (for example anywhere in the hippocampus). Another possibility is the definition of a ROI directly deducted from the original results mask, if available (i.e., the original thresholded mask). Each of these spatial definitions comes with important limitations. For example, the meaning of proximity could be judged very different in different locations in the brain, as some anatomically or functionally defined structures may vary in size and distinctiveness (e.g., comparing the small and clearly-defined amygdala with a large and difficult to define dorsolateral prefrontal cortex). Thus, it may be necessary to combine several criteria in a systematic and/or subjective manner.

It should be noted that these criteria apply to voxel-based analyses. For other neuroimaging techniques, such as segmentation-based MRI analysis, diffusion tensor imaging (white matter integrity), or functional connectivity metrics, other criteria for replication success may be necessary.

11.0.5 Open Science Practices in Neuroimaging

While suggestions on open science practices and replication studies are not fundamentally different from other research areas, their necessity for neuroimaging studies could be even more pressing and there are some peculiarities to consider. Due to the high number of researcher degrees of freedom the utilization of automated preprocessing pipelines is

highly advisable (e.g., Esteban et al., 2019), ideally in combination with containerized toolbox environments for preprocessing and analysis (Renton et al., 2024). In face of reproducibility challenges the transparent publication of preprocessing and analysis scripts becomes even more vital. While the publication of data is advised whenever possible, this can be difficult when sensitive patient data is included and whenever anonymization is difficult. For example, while this is currently subject of debate, MRI-derived brain scans may retain fingerprint-like identifiable features, even when removing the face from the image (Jwa et al., 2024, Abramian & Eklund, 2019). When the publication of raw data is not possible, comprehensive statistical brain maps (i.e., the statistical results in each voxel) should be made publicly available in non-thresholded form (Taylor et al., 2023) and/or data can be published in aggregated form (e.g., summarized for one brain region). Preregistrations can and should be used to make the exploitation of researcher degrees of freedom more transparent. To facilitate preregistrations in neuroimaging, there are multiple templates available. To incorporate all the specifics coming with MRI studies Beyer, Flannery et al. (2021) developed a fMRI specific template, which can be assessed here: https://doi.org/10.23668/psycharchives.5121. For replication research, preregistrations should contain a definition of replication success criteria that take into consideration the spatial dimension of results. Overall, open science practices and replications are still extremely rare in neuroimaging research despite their pressing relevance. Finally, there are also unique tensions to be navigated between open science practices in neuroimaging and the ongoing climate crisis, for example the sustainability of data sharing (see Puhlmann et al., 2025 for a perspective).

Part IV

Conclusion and Checklist

12 Conclusion

As replication researchers from multiple disciplines, we have discussed current standards, best-practices, and open debates surrounding the planning and execution of reproductions and replications. We have also highlighted the need for field-specific guidance and debate by presenting the special case of replications with MRI data. Our recommendations are summarized in the checklist below. With decades of research waiting to be reproduced and replicated, we hope to provide a starting point for interdisciplinary discussions and support researchers in embracing the essential and exciting element of repetitive research.

12.1 Reproductions and Replications Checklist

- ☒ Justify choice of target study and claims
- ☒ Choose a reproduction/replication type that aligns with your aims
- ☒ Gather and review all relevant materials
- ☒ Reproduce before you replicate, where possible
- ☒ Discuss all updates, changes, and extensions of the original materials (as close as possible, as updated as necessary)
- ☒ Preregister your study and analysis plan

☒ Predetermine conditions for success and failure

☒ Use balanced language when describing the outcomes

☒ Carefully evaluate outcomes and potential reasons for divergences

☒ Report your research comprehensively and openly accessible

References

Bennett, E. A. 2021. "Open Science from a Qualitative, Feminist Perspective: Epistemological Dogmas and a Call for Critical Examination." Psychology of Women Quarterly 45 (4): 448–56. https://doi.org/10.1177/03616843211036460.

Block, J., and A. Kuckertz. 2018. "Seven Principles of Effective Replication Studies: Strengthening the Evidence Base of Management Research." Management Review Quarterly 68 (4): 355–59. https://doi.org/10.1007/s11301-018-0149-3.

Boyce, V., B. Prystawski, A. B. Abutto, E. M. Chen, Z. Chen, H. Chiu, and M. C. and Frank. 2024. "Estimating the Replicability of Psychology Experiments After an Initial Failure to Replicate," May. https://doi.org/10.31234/osf.io/an3yb.

Clarke, B., P. Y. (K.) Lee, S. R. Schiavone, M. Rhemtulla, and S. Vazire. 2024. "The Prevalence of Direct Replication Articles in Top-Ranking Psychology Journals." American Psychologist. https://doi.org/10.1037/amp0001385.

Cole, N. L., S. Ulpts, A. Bochynska, E. Kormann, M. Good, B. Leitner, and T. Ross-Hellauer. 2024. "Reproducibility and Replicability of Qualitative Research: An Integrative Review of Concepts, Barriers and Enablers." https://doi.org/10.31222/osf.io/n5zkw_v1.

Cortina, J. M., T. Köhler, and L. C. Aulisi. 2023. "Current Reproducibility Practices in Management: What They Are Versus What They Could Be." Journal of Management Scientific Reports 1 (3-4): 171–205. https://doi.org/10.1177/27550311231202696.

Dreber, A., and M. Johannesson. 2024. "A Framework for Evaluating Reproducibility and Replicability in Economics." Economic Inquiry. https://doi.org/10.1111/ecin.13244.

Dunlap, K. 1926. "The Experimental Methods of Psychology." In Psychologies of 1925, edited by C. Murchison, 331–51. Clark University Press. https://doi.org/10.1037/11020-022.

Errington, T. M., M. Mathur, C. K. Soderberg, A. Denis, N. Perfito, E. Iorns, and B. A. Nosek. 2021. "Investigating the Replicability of Preclinical Cancer Biology." eLife 10: e71601. https://doi.org/10.7554/eLife.71601.

Hawkins, R. X., E. N. Smith, C. Au, J. M. Arias, R. Catapano, E. Hermann, and M. C. and Frank. 2018. "Improving the Replicability of Psychological Science Through Pedagogy." Advances in Methods and Practices in Psychological Science 1 (1): 7–18. https://doi.org/10.1177/2515245917740427.

Henriques, S. O., N. Rzayeva, S. Pinfield, and L. Waltman. 2023. "Preprint Review Services: Disrupting the Scholarly Communication Landscape?" https://doi.org/10.31235/osf.io/8c6xm.

Hüffmeier, J., J. Mazei, and T. Schultze. 2016. "Reconceptualizing Replication as a Sequence of Different Studies: A Replication Typology." Journal of Experimental Social Psychology 66: 81–92. https://doi.org/10.1016/j.jesp.2015.09.009.

Hummel, T., and J. Manner. 2024. "A Literature Review on Reproducibility Studies in Computer Science." In Proceedings of the 16th ZEUS Workshop on Services and Their Composition (ZEUS 2024)(CEUR). Vol. 3673.

Jekel, M., S. Fiedler, R. Allstadt Torras, D. Mischkowski, A. R. Dorrough, and A. Glöckner. 2020. "How to Teach Open Science Principles in the Undergraduate Curriculum—the Hagen Cumulative Science Project." Psychology Learning & Teaching 19 (1): 91–106. https://doi.org/10.1177/1475725719868149.

Köhler, T., and J. M. Cortina. 2021. "Play It Again, Sam! An Analysis of Constructive Replication in the Organizational Sciences." Journal of Management 47 (2): 488–518. https://doi.org/10.1177/0149206319843985.

Lash, T. L., L. J. Collin, and M. E. Van Dyke. 2018. "The Replication Crisis in Epidemiology: Snowball, Snow Job, or Winter Solstice?" Current Epidemiology Reports 5: 175–83.

Makel, M. C., J. A. Plucker, and B. Hegarty. 2012. "Replications in Psychology Research: How Often Do They Really Occur?" Perspectives on Psychological Science 7 (6): 537–42. https://doi.org/10.1177/1745691612460688.

McManus, K. 2024. "Replication Studies in Second Language Acquisition Research: Definitions, Issues, Resources, and Future Directions: Introduction to the Special Issue." Studies in Second Language Acquisition 46 (5): 1299–319. https://doi.org/10.1017/

S0272263124000652.

Moreau, D., and K. Wiebels. 2023. "Ten Simple Rules for Designing and Conducting Undergraduate Replication Projects." PLOS Computational Biology 19 (3): e1010957. https://doi.org/10.1371/journal.pcbi.1010957.

Nuijten, M. B., and J. R. Polanin. 2020. "'Statcheck': Automatically Detect Statistical Reporting Inconsistencies to Increase Reproducibility of Meta-analyses." Research Synthesis Methods 11 (5): 574–79. https://doi.org/10.1002/jrsm.1408.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." Science 349 (6251): aac4716. https://doi.org/10.1126/science.aac4716.

Perry, T., R. Morris, and R. Lea. 2022. "A Decade of Replication Study in Education? A Mapping Review (2011–2020)." Educational Research and Evaluation 27 (1-2): 12–34. https://doi.org/10.1080/13803611.2021.2022315.

Pownall, M. 2022. "Is Replication Possible for Qualitative Research?" https://doi.org/10.31234/osf.io/dwxeg.

Röseler, L., L. Kaiser, C. Doetsch, N. Klett, C. Seida, A. Schütz, and Y. and Zhang. 2024. "The Replication Database: Documenting the Replicability of Psychological Science." Journal of Open Psychology Data 12 (1): 8. https://doi.org/10.5334/jopd.101.

Schöch, C. 2023. "Repetitive Research: A Conceptual Space and Terminology of Replication, Reproduction, Revision, Reanalysis, Reinvestigation and Reuse in Digital Humanities." International Journal of Digital Humanities 5 (2): 373–403. https://doi.org/10.1007/s42803-023-00073-y.

Urminsky, O., and B. J. Dietvorst. 2024. "Taking the Full Measure: Integrating Replication into Research Practice to Assess Generalizability." Journal of Consumer Research 51 (1): 157–68. https://doi.org/10.1093/jcr/ucae007.

A Author Contributions

No.	Author (last name, first name)	ORCID-ID	Contribution (CRediT)	Affiliation
1	Röseler, Lukas*	https://orcid.org/0000-0002-6446-1901	Conceptualization, Project Administration, Writing – original draft, Writing – review & editing	Münster Center for Open Science, University of Münster
2	Wallrich, Lukas*	https://orcid.org/0000-0003-2121-5177	Writing – original draft, Writing – review & editing	Birkbeck Business School, University of London
3	Hartmann, Helena	https://orcid.org/0000-0002-1331-6683	Writing – original draft, Writing – review & editing	Department for Neurology and Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University Hospital Essen

No.	Author (last name, first name)	ORCID-ID	Contribution (CRediT)	Affiliation
4	Hüffmeier, Joachim	https://orcid.org/0000-0002-0490-7035	Writing - review & editing	TU Dortmund University
5	Goltermann, Janik	https://orcid.org/0000-0003-3087-1002	Writing - review & editing	University of Münster, Institute for Translational Psychiatry
6	Charlotte R. Pennington	https://orcid.org/0000-0002-5259-642X	Writing - review & editing	School of Psychology, Aston University, Birmingham
7	Veronica Boyce	https://orcid.org/0000-0002-8890-2775	Writing - review & editing	Department of Psychology, Stanford University
8	Sarahanne M. Field	https://orcid.org/0000-0001-7874-1261	Writing, review & editing	Department of pedagogy, University of Groningen
9	Merle-Marie Pittelkow	https://orcid.org/0000-0002-7487-7898	Writing, review & editing	Berlin Institute of Health at Charité - Universitätsmedizin Berlin
10	Don van Ravenzwaaij	https://orcid.org/0000-0002-5030-4091	Writing, review & editing	Department of psychology, University of Groningen

No.	Author (last name, first name)	ORCID-ID	Contribution (CRediT)	Affiliation
11	Priya Silverstein	https://orcid.org/0000-0003-0095-339X	Writing, review & editing	Center for Neuroscience and Cell Biology, University of Coimbra; Institute for Globally Distributed Open Research and Education
12	Luisa Altegoer	https://orcid.org/0000-0001-8466-7328	Writing - review & editing	University of Münster, Institute for Translational Psychiatry
13	Azevedo, Flavio	https://orcid.org/0000-0001-9000-8513	Writing – original draft, Writing – review & editing	University of Utrecht, Department of Interdisciplinary Social Science

*shared first authorship

B Potential Conflicts of Interest

A large proportion of the authors are members of FORRT, an organization dedicated to integrating open and reproducible science into higher education. LR, LW, FA, and JG are inaugural editors of the in-development journal Replication Research (https://replicationresearch.org). Besides their conviction of the value of replications, their current project's success relies on researchers conducting reproductions and replications. LR is the managing director of an institutional open science center and proponent of repetitive research. The authors declare that they have no further potential conflicts of interest.

C Funding

LR received funding from the University of Münster and the 'Landesinitiative opennaccess.nrw'. LW and LR received funding from 'UK Research and Innovation'. LW, HH, and FA received funding from the 'Nederlandse Organisatie voor Wetenschappelijk Onderzoek'. JG received funding by 'Innovative Medizinische Forschung' (IMF) of the medical faculty of the University of Münster (GO122301). HH was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 422744262 - TRR 289 (gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 422744262 - TRR 289).

D Acknowledgments

This work is an initiative from The Framework for Open and Reproducible Research Training (FORRT; https://forrt.org), and all core-team authors are active members of FORRT's Replication Hub (https://forrt.org/replication-hub).

We thank Patrick Smela for valuable ideas on defining replication success and Abel Brodeur for suggestions about definitions and the relationship between reproducibility and replicability.

E Asking for materials and data

Dear [name of author(s)], we are conducting replication research using some of your research. Specifically, we [brief name of the phenomenon and study that was replicated]. [We do this because … e.g., your research addresses a very important question.] Can you please send us the following materials to help us design a replication as close as possible to your original study?

- [list of required materials/data/code]
- [list of required materials/data/code]
- Citation of original study: [add citation] We are looking forward to your responses! Thank you [Your name] Asking for comments on an experimental paradigm Dear [name of authors], We are planning a replication of some of your research. Specifically we are aiming to replicate your study [study details and citation]. [we are interested in these findings because …] I'm writing to share a mock-up of the replication to get your feedback on whether this paradigm accurately captures the design of your study. Please let me know if you have any comments or concerns that you'd like to share. Here's a link to my paradigm. Any insights you have into details that differ from your own study would be much appreciated. I will be replicating your experiment on [planned recruitment sample]. [I know this is a deviation from the original population you tested, and I will note this sample decision prominently in any writeups.]

Thanks again,
Your name Signature

F Asking for comments on replication results

Dear [name of author(s)], we have conducted replication research using some of your research. Specifically, we [brief name of the phenomenon and study that was replicated]. In our study [description of results]. We want to provide you with an opportunity to comment on these findings. We plan to publish the replication report via [paper or publication platform, e.g., FORRT's Replication Hub], which asks replication studies to be submitted alongside comments from the authors of the original study. Your comment – if you choose to give one – will be part of the report.

- Citation of original study: [add citation]
- Replication study:[add link to document or attach it to the e-mail]

We are looking forward to your responses!

Thank you
[Your name]