Table of contents

Summary

Many academic disciplines have changed considerably in recent years. New journals, methods, initiatives, societies, structures, and much more have emerged. There is talk of a "crisis" (Mede et al., 2020), a revolution (Nosek et al., 2018), or a renaissance (L. D. Nelson et al., 2018). Together, all these changes have led to a greater degree of openness and transparency and have substantially transformed the academic system.

This book provides an accessible explanation of the concept of openness, the origins of this change, and the Open Science movement. It explains the background of the academic system, newly developed methods, and problems with currently established methods, drawing mainly on examples and developments from psychology.

> **About this translation**
>
> While this entire book was written without any AI assistance, this English translation was produced with the help of Claude Sonnet 5 and has not yet been fully reviewed by the author. If you notice any translation errors or awkward phrasing, please refer to the German original or get in touch.
>
> Two things to know about this translation: First, German *Wissenschaft* has no exact English equivalent — it covers the natural sciences, social sciences, and humanities alike. Where the original says *Wissenschaft*, this translation generally uses "academia," "scholarship," or "research" rather than "science," and *Wissenschaftler*in* is rendered as "researcher" or "scholar" rather than "scientist." Second, the figures throughout this English edition

> still carry their original German labels — the diagrams have not yet been redrawn in English.

About the Book

Academic research is funded to a significant extent by the members of a society, for example through tax revenue. This means that a community supports individual people so that they can gain a deeper understanding of themselves and the world, and in return receives help in solving various problems. As a result, researchers owe a certain degree of accountability to society. This does not mean that every scholarly finding necessarily has to bear fruit, but rather that it meets scholarly quality criteria and is, at the very least, comprehensible and verifiable for other researchers. To what extent this is the case in various disciplines, and how people are working to make this accountability possible, is what I aim to lay out in this book.

This book itself is meant to be part of open science. In the context of the problems and solutions, I explain what the profession of *academic researcher* looks like, how the academic system works, and discuss, using selected examples, scholarly debates. Openness also means that the book is freely accessible.

Acknowledgments

This book would not have been possible without the support of my wife, Jessica. My optimism and my strength to advocate for an open, transparent, and fair science draw on her.

I thank Lena Röseler, Celina Wagner, Helga Röseler, and Anna Wentzel for proofreading and their countless suggestions for improvement, Wibke Fellermann for advice on style and layout, and Victoria Moser for reworking the figures.

In the examples, notes, and remarks, I have also drawn on numerous examples that I discussed with colleagues and friends. I thank Daniel Wolf, Johannes Leder, Thomas Lösch, Matthias Borgstede, Astrid Schütz, Ulrike Starker, Georg Felser, Viola Voß, Mitja Back, Jan H. Höffler, Helene Richter, Janik Goltermann, Gilad Feldman, Flavio Azevedo, Lukas Wallrich, and many others for the engaging conversations!

Use of AI

Claude Sonnet 5 was used to adjust the layout and identify spelling errors. An English translation currently in progress was also produced using Claude Sonnet 5.

Introduction: Traceable Science

Let us begin with a hypothetical example: in Germany, the default rule for organ donation could be changed. At present, no one is an organ donor unless they explicitly *consent* to organ donation (for example, by ordering an organ donor card). Under the proposed new system, everyone would still be free to decide, even after the change, whether or not they want to donate their organs. As in Austria, the alternative would be an *opt-out solution*, meaning that everyone is by default an organ donor[1] and must actively *object* to that status in order to opt out.

The basis for such a hypothetical political decision would be a scientific finding: under the opt-out solution, the proportion of organ donors is higher than when people have to actively pursue the status of "organ donor." The original finding was reported by Johnson and Goldstein (2003). Almost 20 years later, Chandrashekar et al. (2023) were able to demonstrate the finding again using newly collected data – in other words, to *replicate* it. All information about the replication can be read online, and its results are fully accessible online. Using free software, anyone can recalculate and check everything that an independent person has already done on behalf of the journal in which the results were published. Moreover, the effect holds not only for organ donation but for all kinds of other human decisions. The general phenomenon is called the *default effect*. The *scientific basis* for the political decision is, in this case, verifiable and traceable.

1. In the German-language original of this book, the decision to use gender-inclusive asterisk notation (e.g., *Organspender*in*) instead of the generic masculine form is based on the scientific finding that this better achieves the goal of addressing people of different genders equally (Brohmer et al., 2024). English does not carry the same grammatical gender distinctions as German, so this consideration does not apply in the same way to the English text; throughout this translation, gender-neutral terms such as "organ donor" are used instead.

The transparency of the scientific finding is clearly evident in this example. Unfortunately, though, this is not representative of science as a whole. Research findings often cannot simply be looked up; instead, research reports must be purchased from journals (or requested from the authors by email), data must be requested, and results can only be recalculated using expensive software. This kind of openness plays a particularly important role at present: there is widespread skepticism toward science (Sultan et al., 2024), and political decisions about heating, vaccination, nutrition, or mobility rely on science while simultaneously being perceived as wrong, costly, or threatening. Misleading or simply false information is spread, and people use artificial intelligence (large language models) that are based on potentially false information and are not able to indicate the origin of that information.

How the openness of science is changing, what researchers around the world are doing to open up science, and what proposed solutions are being discussed and implemented in this regard is the subject of this book. An overview of the dimensions of openness, based on Silveira et al. (2023), is given in the table below. Many academic disciplines lie somewhere between the extremes of "classical scholarship" and "open scholarship." The latter represents the ideal – that is, the state that would be most beneficial for scientific progress.

Table 1: Dimensions of openness in science.

Dimension	**Classical Science**	**Open Science**
Open Access	Research articles and books must be purchased or borrowed.	Research articles and books are freely available online.

Dimension	Classical Science	Open Science
Open materials and data	Materials and data underlying reported scientific findings are not shared, shared only on request, or shared only after considerable delay.	Materials and data are published together with the research report.
Open evaluation of researchers, universities, and journals	Research is evaluated on the basis of opaque and manipulable metrics.	Research is evaluated on the basis of objective criteria closely tied to scholarly quality assurance.
Open teaching	Acquiring knowledge is costly and requires being present at specific places at specific times.	Teaching materials and courses are freely available online, teaching is (also) conducted online, and care is taken to make materials low-barrier and accessible.
Open innovation, "Big Team Science"	Groups of researchers work in strong competition, isolated from one another, on similar problems.	Researchers cooperate internationally and across disciplines and pool resources.
Open infrastructure	In everyday research, researchers depend on funding for software or publication fees.	The processes of research and quality assurance are independent of commercial publishers and are organized independently and sustainably by researchers and universities themselves.
Citizen science / participatory science	Public participation in science is of secondary importance.	Members of the public are actively involved in the research process and enrich science.

Dimension	Classical Science	Open Science
Open scientific discourse	Scientific criticism is taken as personal criticism, harms one's academic career, and results in personal retaliation. Peer reviews of scientific articles are anonymous, are not credited as scientific output, and are visible only to a few people. Comments on research articles are rare and published only after a long delay.	A culture of appreciative and constructive engagement with error prevails; reviews are public and anonymous only on request. Scientific articles can be commented on within minutes.
Replication research	New studies must investigate new questions. Studies that re-examine things already demonstrated are not published: failed replication attempts are assumed to reflect methodological problems, while successful replications are considered uninformative.	Past findings are questioned, and independent researchers attempt to demonstrate them again using new data – that is, to replicate them. Journals publish these replication attempts regardless of their outcome.

What Is Open Science?

Research is often about details: What exactly happened in the study? What images did participants see, and on what kind of screen? What was the screen's refresh rate, and were the colors calibrated using a machine (e.g., a spectrophotometer)? Research studies and their findings are nowadays described in journal articles. It is not uncommon for one of the details listed above to be missing, unavailable in the supplementary materials, or for the authors to be unreachable because their e-mail address is no longer current. But if a finding depends on exactly such a detail, future researchers will have difficulty bringing it to light. At this concrete level, Open Science means the possibility, for anyone, within ethical and legal limits (e.g., anonymity), to gain detailed insight into the entire scientific process of knowledge generation. It should therefore be described precisely how the study was conducted, what materials were used, what data resulted from it, how that data was processed and analyzed, and finally, what can be learned from it.

At a more abstract level, Open Science refers to a greater degree of openness and transparency in all facets of academia. As described, this includes study materials (e.g., questionnaires, images, or videos), the research report, and the discourse involved in the peer review of research reports by colleagues. But the degree of transparency can also increase at the level of the academic system itself (that is, institutions, journals, and the laws governing research): for example, there are lists of journals whose articles can be read online free of charge (https://doaj.org). Finally, openness can also mean holding a conference in a "hybrid format" – that is, at a specific location, but with the option of online participation – so as not to exclude people who cannot travel there, thereby being *open*

toward them. The aim of greater transparency is to increase scientific quality: only what *can* be checked, can actually be checked. Ultimately, there is also the hope of transfer, that is, that scientific findings will be better translated into practice (Cole et al., 2024). Open Science is thus an umbrella term for a wide range of efforts to make science and research comprehensible and traceable. The logic behind this is that one need not simply take a researcher's word for it, but can instead verify the facts oneself without much effort.

Overall, Open Science is understood as a solution to numerous current, often interrelated problems – first and foremost among them the problem that roughly 50% of all psychological studies cannot be replicated (Open Science Collaboration, 2015), meaning that repeating a study produces different results. Fecher and Friesike (2014, p. 19) speak of five "schools of thought," that is, communities with different understandings that pursue different goals: infrastructure (e.g., platforms for storing or publishing research materials), public engagement and the involvement of society in the scientific process (e.g., by having citizens participate in conducting a study; *citizen science*), measurability of scientific success (e.g., alternatives to commonly used metrics such as the *impact factor*), democracy (e.g., access to knowledge), and pragmatism (e.g., greater efficiency of science through publicly available research data). An overview of the facets of Open Science is shown below. Precise estimates are difficult, but roughly speaking: if one picks a random social science study from an academic journal, reruns it according to the scientific gold standard, and tests the hypothesis examined in it once more, the probability of arriving at the same result is about as high as getting "tails" (vs. heads) on a coin flip. Problems partly associated with this include fraud in scientific publi-

cations (Gopalakrishna, Wicherts, et al., 2021) and mental health problems among early-career researchers (Satinsky et al., 2021).

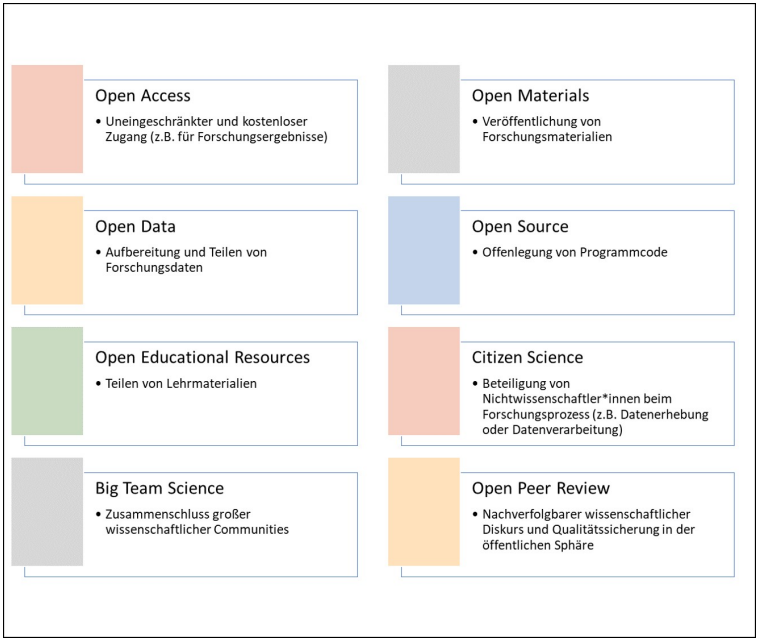

Figure 1: Facets of Open Science.

How Should We Deal with Open Science?

The topic of Open Science and the replication crisis[1]—or crisis of confidence—is difficult to communicate for several reasons. First, it amounts to harsh criticism of science and the scientific system, which can be misused to downplay scientific findings and which endangers trust in science as a whole. The arguments and findings presented here can, for example, be used to criticize political or individual decisions that are based on science. Against this, it should be said: by no means are all scientific findings "wrong" or unreplicable. On many points there is still broad consensus—that is, scientists agree on the truth of the findings.

On the other hand, the problems discussed here go beyond what naturally occurs in almost every scientific field. The more closely one looks, the more likely it becomes that any given scientific finding can be relativized. Contradictions or conflicts can be found everywhere, arise naturally, and are the target of scientific scrutiny. Every researcher can attest: on closer inspection, nothing remains black or white but dissolves into many shades of gray. The replication failures discussed below amount to more than the uncertainty inherent to science: they endanger entire lines of scientific inquiry.

Before we proceed to the details, it is also important to note that, at present (autumn 2023), most of the problems discussed in this book have not yet been solved, or have only been partially solved, and are hotly debated on a weekly basis. After a decade of the Open Science movement, a point has

1. This refers to the fact that a large proportion of published replication studies (that is, studies that re-test what is already known) arrive at different results, meaning that original studies often cannot be replicated (see also the following chapters on history and stocktaking).

been reached at which most scientists have developed an awareness of the problem. Some are familiar with proposed solutions, but these have neither become fully established, nor is it clear which problems have already been solved.

I understand Open Science as a large trigger warning that should precede the social sciences and that reads: Attention, we currently have very serious problems here. It is entirely possible that half of everything we believe we know is simply wrong. Opinions on the Open Science movement are by no means uniform, however: one can span a spectrum here between "Should we really spend the next 20 years finding out which findings of the last 100 years are correct? Let's just start over from zero," and "Actually, this is all quite normal; I see no reason to speak of a 'crisis' here."

Figure 2: Spectrum of reactions to the replication crisis.

The problem is especially large for textbooks. Imagine you have written a book about a branch of psychology (social psychology is often the showcase example here) that rests on decades of research, hundreds of publications, and thousands of studies. Here it hardly helps to ignore the book entirely. But replicating all the studies yourself is not practicable either. What is clear to most scientists is this: things cannot continue the way they have so far. A fifty percent guarantee for scientific findings should not be the standard for something that calls itself science. But we do not need to start from scratch either.

Open Science vs. Open Scholarship

Unlike German *Wissenschaft*, which covers all scholarly disciplines, English "science" is usually taken to mean only the natural sciences, so that, strictly speaking, "Open Science" excludes the humanities. Because English therefore has to address all fields jointly as "Science and Humanities," the term "Open Scholarship" is often used instead (Parsons et al., 2022).

Further information: reactions from various members of the German Psychological Society (DGPs) to the board's official statement, taken from a discussion forum, can be read online: https://replicationindex.com/wp-content/uploads/2022/12/DGPs_Diskussionsforum.pdf.

The History of the Open Science Movement

In this book, the open science movement is understood as a reaction to identified problems, and thus as a self-correcting process within science. At its center is a lack of trust in findings published in scientific journals. Some of these problems have been known in similar form for decades. However, it took decisive events before they were publicly discussed and proposed solutions were developed. Because of difficulties in *replicating* past results that had been considered reliable (that is, arriving at the same result in repeated investigations), this is often referred to as a *replication crisis*. Due to the uncertainty about the exact causes of low replication rates, the term "crisis of confidence" is sometimes preferred instead (Feest, 2019).

Further reading:

- The history of the replication crisis is described by L. D. Nelson et al. (2018) and Schimmack (2020). A more positive perspective is taken by Korbmacher et al. (2023).
- The high value of replication studies was already being discussed in political science in 1995 (King, 1995).
- A case of data manipulation at a highly regarded physics research institution, which came to light in 2002, is described in an online documentary (https://www.youtube.com/watch?v=nfDoml-Db64) and a book (Sarachik, 2009).
- The history of the open-access movement has been documented by the Open Access Network: https://open-access.network/informieren/open-access-grundlagen/geschichte-des-open-access

Beginnings of a Revolution

Around the year 2010, a series of events accumulated in psychology that, taken individually, could have been dismissed as isolated incidents, but that together painted a negative picture of the science. These cases fall into two categories: the Stapel affair was a clear case of fraud through the "invention" of data or the reporting of studies that were never conducted. The other cases were not clear-cut fraud; instead, they involved methods that were, at the time, accepted within the field, used to produce results.

The Stapel Affair

By chance, in 2011 early-career researchers discovered that the data for a study by their colleague, Diederik Stapel, had never been collected by anyone. It had been invented, or fabricated. As of July 2024, 58 of Stapel's journal articles have been identified and retracted for containing fabricated or manipulated data.[1] Scientific institutions such as peer review, whose purpose is quality assurance, had failed. Since the incident, several further cases have come to light, in part through renewed analysis of the data from the studies in question (O'Grady, 2021), and often through whistleblowers—that is, people with knowledge of the wrongdoing who wish to remain anonymous for their own protection. Surveys of researchers in the Netherlands have found that up to 10% of all researchers have engaged in the fabrication or manipulation of data (Gopalakrishna, Riet, et al., 2021). It should be noted here that fabricated data tend to make studies especially innovative, surprising,

1. Retracted articles can be searched by topic, author, journal, etc. using the Retraction Database: http://retractiondatabase.org/.

or clear-cut—properties that increase the likelihood of publication in a journal.

Bem: Feeling the Future

A short time later, Daryl Bem—known for foundational psychological and philosophical theories such as self-perception theory (Bem, 1967)—published the finding that people can predict the future (Bem, 2011). More precisely, some people, under certain conditions, can predict the future. Across nine studies, the research group found that people could make predictions about erotic images. The results were published in the highly regarded *Journal of Personality and Social Psychology*. For many psychologists, it was immediately clear: either fundamental assumptions of their worldview were wrong ("people cannot predict the future"), or something was wrong with the results. Several researchers attempted to explain how these results had come about. Analyses of the published data using alternative statistical methods arrived at the same conclusion (Wagenmakers et al., 2011), yet replications by independent researchers failed (Muhmenthaler et al., 2022; Robinson, 2011; Roe et al., 2012).

Bargh: Influence through *Priming*

His studies were celebrated in marketing and sold as neuroscientific insights: people who drink a hot beverage (compared to a cold one) rate other people as "warmer" (more generous, more considerate) (Williams & Bargh, 2008). People who solve anagrams related to old age (e.g., "TREEMTIRNE" instead of "RETIREMENT," or "YARG" instead of "GRAY," or, to try it yourself, "NOISER") subsequently walk more

slowly (Bargh et al., 1996). Many of these studies were not replicated: researchers noticed, in the anagram study, that Bargh and colleagues had measured walking time with stopwatches while knowing which participants had solved the "old age" words and which had solved the neutral anagrams—even though every first-year psychology student learns that this should not be the case, and that experimenters should be "blind" to the purpose of the study and to which condition each participant was assigned. In their replication (Doyen et al., 2012), Doyen and colleagues first measured time themselves, just as Bargh et al. had in the original study: this problematic measurement approach produced the same result. When they instead had time measured using light barriers, without the experimenters knowing which hypothesis was being tested or which participant was solving which anagrams, the effect could not be replicated.

! Important Events beyond Psychology

In behavioral biology, Crabbe et al. (1999) showed much earlier that findings on the influence of genes on behavior depended strongly on which laboratory the mice were tested in. Although calls for improvements to the research process have recurred ever since (Kafkafi et al., 2018), a revolution comparable to that in psychology never materialized.

A somewhat different problem arose in the course of sequencing human DNA: there, companies filed for intellectual property rights while the research process was still underway. The negative social and scientific consequences were discussed early on (Moore, 2000) and remain part of scientific and political debate to this day.

> Problems with reproducibility—that is, confirming published results based on the same data—are also known from preclinical animal studies of various drugs (biology), cancer research (medicine), genetics (biology), treatments for ALS (neurology), and widely cited medical interventions (medicine) (Royal Netherlands Academy of Arts and Sciences, 2018, p. 22, Table 1).

Stocktaking

A similar crisis of confidence occurred in social psychology in the 1960s (Lakens, 2023). A crucial difference between the old crisis and the current one is the stocktaking that has taken place: parallel to the peculiar findings on Bem's precognition studies, psychologists around Brian Nosek formed an international network and investigated the replicability of 100 studies from prominent psychology journals (Open Science Collaboration, 2015). They found that only 39 of the 100 original findings could be replicated. For all other studies, the replication results differed from the original results. Many further large-scale projects followed, all with similar results: replication rates fell far below what had been hoped for.

> **i** A critical look at the Open Science Collaboration, 2015
>
> Although this "Reproducibility Project: Psychology" shook the entire field to its core and paved the way for a paradigm shift, some researchers also point to negative effects on subsequent replication research. By publishing 100 studies simultaneously through a group of more than 100 researchers, the project set unrealistic standards for replication research. At the same time, quality control was less rigorous, since the individual studies could not all be reviewed to the extent that would have been the case for a traditional publication, e.g. (Röseler et al., 2022). Some equally ambitious projects have since been published, such as the Many Labs studies, e.g. (Klein et al., 2014; Klein et al., 2018), or attempts in which independent groups tested and replicated the same hypotheses (Landy et al., 2020). Such projects are often limited to studies that can be replicated within an

> online survey. Formats such as longitudinal studies or behavioral observations are underrepresented due to the greater practical difficulty involved.

Numerous consortia followed. Some projects focused on individual phenomena. For example, 17 research groups joined forces to replicate the *facial feedback* finding (Strack et al., 1988) in (Wagenmakers et al., 2016). In this experiment, participants hold a pen in their teeth, and depending on the pen's orientation, they either tense the muscles used for smiling or do not. In the "smiling" condition, participants subsequently rated comics as funnier. The replication failed. In 2022, a further study with more than 3,000 participants from 19 countries was published – this time with direct involvement from Fritz Strack, who had conducted the original study (Coles et al., 2022). Again, it was shown that the position of a pen in the mouth has no effect on the evaluation of stimuli. Beyond social-psychological findings, researchers also focused on areas such as research with infants (Byers-Heinlein et al., 2020) or studies from specific journals (Camerer et al., 2018).

Defining replicability

Determining what could and could not be replicated is only possible once *replication success* has been defined. In common usage among researchers, "was replicated" means that a study investigating the same research question arrived at results similar to those of an original study. Replication failures are described as "could not be replicated." Subtly different from this, "was not replicated" can mean that no replication attempts exist, or that they could not even be conducted because the

original study failed to adequately describe key details (Errington et al., 2021). For a science that is over 100 years old, it seems surprising that there is still no clear definition of key concepts surrounding replication, let alone that replicating studies has become routine. While different fields have settled on divergent taxonomies – that is, models for classifying different types of replication – terms related to *replication* are used in this book as described in the table below. Depending on whether the same or different data and the same or different analyses are used, we speak of reproducibility, replicability, robustness, or generalizability.

Table 2: Replication taxonomy following the Turing Way (The Turing Way Community & Scriberia, 2024).

		Data	
		same	different
Analysis	same	reproducible	replicable
	different	robust	generalizable

Statistical comparison of original and replication findings

Every measurement carries some degree of imprecision. In the social sciences, measuring properties or phenomena such as decision-making heuristics is extremely difficult. When original results are compared with replication results, both results carry this imprecision. This makes it difficult to say whether differences arose from random fluctuation or from problems with the original study. There are numerous statistical methods for comparing the two results. They typically differ in how much they account for the imprecision of the original study. Because of older methodological standards, original results tend to be extremely imprecise, and the more heavily

> they are weighted, the more favorable the comparison turns out – that is, the higher the estimated replication rate. Depending on the method used, replication rates can then range between 40% and 80%. A comparison of criteria against the FORRT replication database is available online.

Further reading

For a more systematic taxonomy of replication types, grounded in information science, see Plesser (2018). LeBel et al. (LeBel et al., 2019) have proposed a taxonomy for replication study outcomes based on statistical methods. The closeness of replications is discussed philosophically, for example, by Choi (2023) and Leonelli (2023).

Table 3: Replication taxonomy.

Distinguishing criterion	Categories
Outcome of a replication study	Successful
	Failed
	Unclear or mixed

Distinguishing criterion	Categories
Closeness of a replication study to the original study (following LeBel et al. [2019] and Hüffmeier et al. [2016])	**Direct replication** (same experimenters, same experimental materials, new participants) **Close replication** (different experimenters, experimental materials as similar as possible, new participants) **Conceptual or constructive replication** (different experimenters, different experimental materials, new participants)
Goal of the replication	**Reproduction** Arriving at the same results using the same data and the same code **Replication** Arriving at the same results using different data

"One swallow does not make a summer"

Whether a scientific finding "holds water" – that is, whether it has a valid claim to truth – depends, in replication research, on many factors beyond the way it was originally established. What were the results of the replication study? How many studies were conducted, and how varied were they? What exactly did the methods look like? What were the differences between the replications and the original study? While individual studies always yield some gain in knowledge, at the very least whether a particular

method is feasible (Sikorski & Andreoletti, 2023), they can vary considerably depending on the field of research (Landy et al., 2020; McShane et al., 2022). To see the bigger picture, more is needed – for example, a statistical summary of many individual studies combined into one overall study (a *meta-analysis*). An example using fictitious data is shown in the following figure.

If one looks at many studies that have examined the relationship between two things – for instance, income and educational attainment – the studies will differ in their details: what exactly was counted as income (net, gross, social benefits, family members' income, values over a period of time or from a specific point in the past, etc.), or which people were surveyed (students, working professionals, whether respondents were paid, etc.). All these differences may affect the relationship in question, and even when they do not, relationships are often subject to fluctuations arising from the measurement methods themselves. In this example, the strength of the relationship can be reduced to a single number along with an associated precision. Here, the number is called the "correlation" and the precision the "error bar." The forest plot (also called a *blobbogram*) displays possible correlations from various studies. Within a meta-analysis, study results can then be combined and differences examined.

Phenomenon-centered replication projects

In contrast to the broadly scoped *Reproducibility Project: Psychology* (Open Science Collaboration, 2015) and other attempts to estimate replication rates for entire fields (Brodeur, Mikola, et al., 2024; Camerer et al., 2016; Feldman, 2021), other efforts have focused on fundamental phenomena. In such cases, dozens of groups around the world have

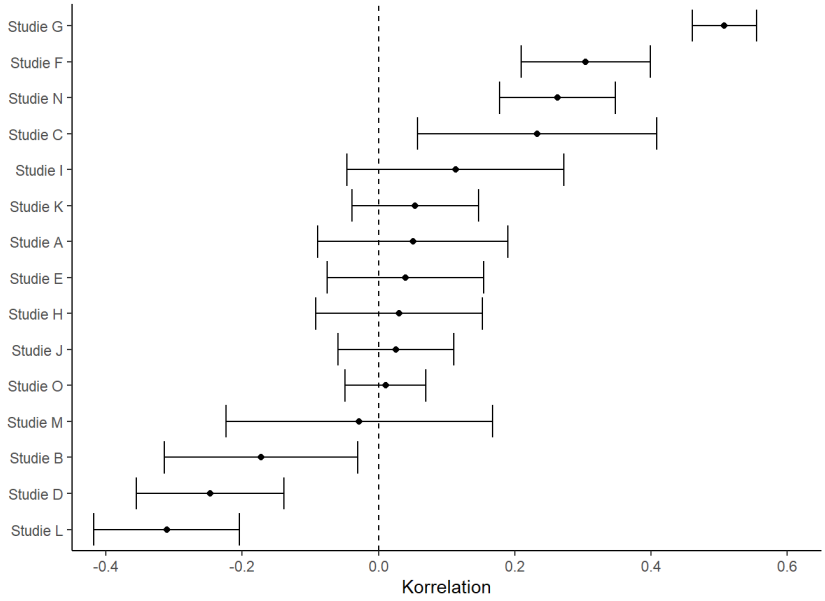

Figure 3: Forest plot with simulated correlations from 15 fictitious studies.

joined forces, agreed on a study design, and carried out the studies with an enormous number of participants. Most of these projects originate from psychology. While the effect sizes found in them – that is, roughly speaking, the magnitude of a relationship or finding – were in almost all cases far below those of previous studies (Kvarven et al., 2020), for the majority of the studies they were also null, meaning the phenomena were "not detectable" at all (Alogna et al., 2014; Bouwmeester et al., 2017; Cheung et al., 2016; Eerland et al., 2016; O'Donnell et al., 2018; Rife et al., 2024; Vaidis et al., 2024; Wagenmakers et al., 2016). For instance, it was shown with enormous precision that a story about a professor does *not* make participants perform better on a subsequent test of intelligence (O'Donnell et al., 2018).

Efficient use of resources?

How should resources be handled in replications? This question inevitably arises whenever many researchers join forces. Does each group create the study independently? Does everyone adhere to a jointly agreed protocol? Do they run the study sequentially, so as to learn from one another? In *Registered Replication Reports*, a study design is typically agreed upon in advance with other researchers (e.g. the authors of the original study). In other cases, a study design is developed collaboratively that should be ideal for testing the theory (the *creative destruction approach*, Tierney et al., 2020). Teams in different countries then translate the protocol and adhere to it closely during implementation. These protocols are sometimes not pre-tested (Buttliere, 2024), but they are often based on successful, well-known studies. This has the advantage that differences between groups cannot be attributed to differences in implementation, and that cultures can be compared (Kakinohana et al., 2022). A disadvantage, however, is that if the experiment already fails to work at one, two, or five sites, it becomes questionable whether the remaining 30 groups should even attempt it. In the words of Buttliere (2024): "Who gets better results? Thirty-nine people doing something for the first time, or one person doing something 39 times?"

Discipline-centered replication projects

Fewer than half of all psychological findings are replicable, then. Does this mean that all social-science textbooks across all disciplines are half wrong? The clear answer is *no*. The accurate answer is *it depends*.

Beyond psychology

The extent to which this depends on the discipline within the social sciences has so far been examined primarily in psychology. Current trends suggest that replication rates in personality psychology and cognitive psychology (Soto, 2019) are higher than in social psychology (Open Science Collaboration, 2015) or in marketing (Charlton, 2022). While hundreds of replication attempts for social-psychological studies have already been published, there are currently far fewer in other fields such as marketing – as of October 2022, only nine. Fields outside psychology are likewise affected by replication problems. Almost all disciplines are affected by problems of replicability, reproducibility, and traceability. New approaches to addressing these issues are being discussed in medicine, biology, chemistry, physics, history, political science, education, computer science, and many other fields.

Change in the System

Since the low replicability of psychological studies became widely known, the scientific system has changed in many respects with regard to its openness and transparency. Some journals now require, as a condition for publishing research articles, that the data be made publicly available or that an explanation be given for why this is not the case (e.g., for data that are difficult to anonymize). In rare cases, such as at the journal Meta-Psychology, researchers recompute all the results. As a counter-proposal to the "impact factor" – a metric that indicates how often a journal is cited and that has been shown to have nothing to do with the quality of the research it contains (Brembs, 2018) – the TOP Factor was introduced (TOP: Transparency and Openness Promotion). For a list of criteria, it indicates the degree to which they are met by different journals. In other fields, such as business administration or marketing, journals are even rated by a small number of selected researchers using rating scales that are published annually as rankings. TOP Factors, by contrast, are objective, are continually updated, and can be viewed and verified publicly at topfactor.org. For each factor there are four levels: no mention, level 1, level 2, and level 3. Level 3 represents the ideal; with regard to data transparency, for instance, this would mean that an article is published only once the data are publicly available and the analyses have been successfully recomputed (*reproduced*) by an independent person. To score a journal, corresponding points are assigned for levels 1-3 and summed.[1]

1. Social scientists know that summing these values in this way is nonsensical, since level 2 is not "twice as good" as level 1, and the different factors should not always be weighted equally – yet that is exactly what the arithmetic assumes. Still, as a heuristic

Table 4: Overview of the TOP guidelines.

Factor	Explanation
Data citation	Most research articles are based on data. Specifically, this concerns the possibility of citing data independently of the articles (e.g., via its own identifier, such as a *Digital Object Identifier* [DOI]).
Data transparency	Whenever possible, data should be made publicly available. This can be done directly through journals or through so-called *research data repositories*.
Analysis code transparency	The analysis code is used to evaluate the research data. Other researchers should have the opportunity to run the code and *reproduce* or check the results.
Research materials transparency	Research materials can include questionnaires, images or films shown to participants, but also odors presented, fruits tasted, or software used. Whenever possible, these too should be deposited (in digital form) in a repository. This is naturally not possible for physical objects. For genes, for instance, there are *alphanumeric codes* that researchers use to communicate with one another.
Guidelines for describing the study design and analyses	To understand, but also to replicate, a study, it is important to describe the study design in detail. In recent years, for example, some journals have lifted the word limit for the corresponding sections of research articles.

estimate of a journal's openness, this approach is more sensible than measures such as the impact factor and the Hirsch index (i.e., *bibliometric measures*).

Factor	Explanation
Preregistration of studies	See also the section on preregistration in this book: if a study tests hypotheses (i.e., expectations or predictions for the outcome), these should be specified in advance and not adjusted after seeing the results. Ideally, journals require preregistrations for such studies and require researchers to provide a link to them.
Preregistration of the analysis plan	The path from data to results is a long one. The results depend on many decisions. To strip themselves of this flexibility, researchers should describe the analysis plan in the preregistration.
Replication	Despite their value, there are still many journals that do not publish replication studies. Ideally, journals encourage researchers to submit replication studies to them.

This change has been driven above all by early-career researchers (ECRs). At many universities in recent years, Open Science initiatives or regular meetings ("Reproducibili-Tea") have emerged, consisting almost exclusively of postdocs (researchers who have completed their PhD and hold a fixed-term research position; in Germany this covers essentially everyone below professor level), doctoral students, and students. In Germany, these have networked together to form NOSI (the German Network of Open Science Initiatives) (Schönbrodt, Baumert, et al., 2022).

Has the replicability rate increased?

The replication crisis has now been ongoing for more than a decade, and quite a bit has changed. Has this also solved the problem of replicability? For one thing, in many areas it is still not even clear what can be replicated. In marketing, the journal *Journal of Business Research* briefly introduced a replication section (Easley & Madden, 2013), which was later moved to a different journal. Projects are currently underway to estimate replicability for various disciplines. Their results are, for the most part, still preliminary and unclear. Another project is collecting replication results in order to track, over the long term, how replication rates have changed by discipline and over time. Up-to-date figures are available online (https://forrt-replications.shinyapps.io/fred_explorer/, Röseler, Kaiser, et al., 2024). An evaluation spanning multiple disciplines and years will still require hundreds of additional replication studies.

The Open Science Revolution as a Paradigm Shift

Historians and theorists of science describe the development of science as discontinuous. Textbooks do not simply keep getting thicker; instead, some chapters get shorter because the knowledge they describe is discarded, while others get thicker because new findings are added. At times, chapters even disappear entirely. One of the best-known models of science comes from Thomas Kuhn (1970/1996), who was originally a physicist and adopted and further developed ideas from the physician and sociologist Ludwik Fleck (1935/2015). According to this model, knowledge grows for a time, but findings accumulate that are incompatible with the rest of established knowledge. At some point, these so-called anomalies can no longer be ignored. From that point

on, the scientific worldview tips over: new theories are devised that can explain the anomalies, and old knowledge is either discarded or integrated into the new theories. This tipping is referred to as a paradigm shift or scientific revolution. Classic examples of such a paradigm shift include the transition from the geocentric to the heliocentric worldview in astronomy, prospect theory for decisions under uncertainty (Kahneman & Tversky, 1979), which linked psychological and economic models, or – as is argued here in this book and elsewhere (Sönning & Werner, 2021) – the replication crisis in psychology. In this case, the anomalies are the findings of Bem (2011) or Bargh et al. (1996), or isolated replication failures. These were incompatible with existing knowledge, and once many such findings had accumulated, they could no longer be ignored or dismissed. The claim that replication researchers had done something wrong, were unqualified, or had bad luck was no longer a convincing explanation. Unlike classic revolutions in the sense of Kuhn (1970/1996), which involve a transition from one theoretical perspective to another, the Open Science Revolution does not center on any particular theory or research discipline being discarded, but rather on the scientific method and the scientific system of the social sciences itself. Moreover, social science disciplines such as economics or psychology do not each have only a single paradigm; they can have multiple independent paradigms (Hoyningen-Huene & Kincaid, 2023) – that is, several strands with little in common that are researched independently of one another. Following Kuhn's philosophy-of-science terminology, alongside "replication crisis" the term *credibility revolution* (Korbmacher et al., 2023) is also used. For philosophical perspectives, see also Rubin (2023).

A paradigm shift is comparable to an ambiguous image such as the duck-rabbit figure (Figure 5), as drawn, for example, by Wittgenstein

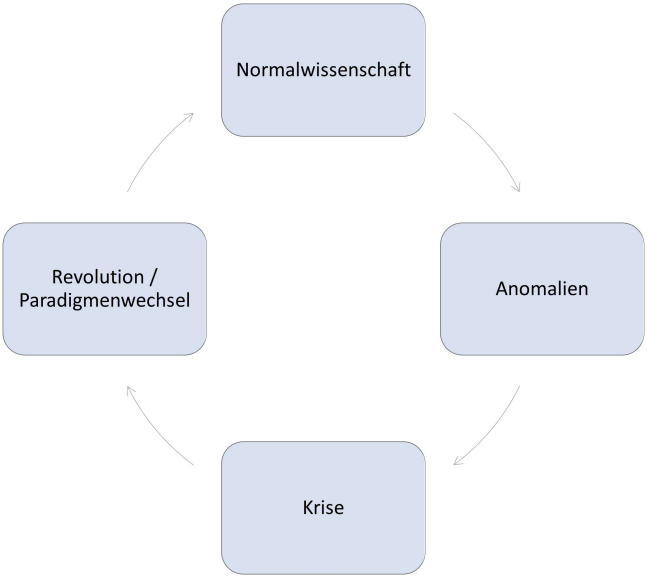

Figure 4: Scientific revolution according to Kuhn (1970/1996), figure adapted from Fiorentino & Montana Hoyos (2014).

(1968). Up to a certain point, everyone agrees that it is a rabbit. But gradually, new insights and perspectives are added. The tipping point is crossed, the duck becomes the accepted interpretation, and no one would take it for a rabbit any longer. With the duck-rabbit image, of course, one can jump back and forth between interpretations at will. In scientific *progress*, however, new knowledge is added and a return to the earlier view becomes very difficult.

From the very beginning of their studies or doctoral training, researchers are trained to deal with findings in accordance with the prevailing paradigm.

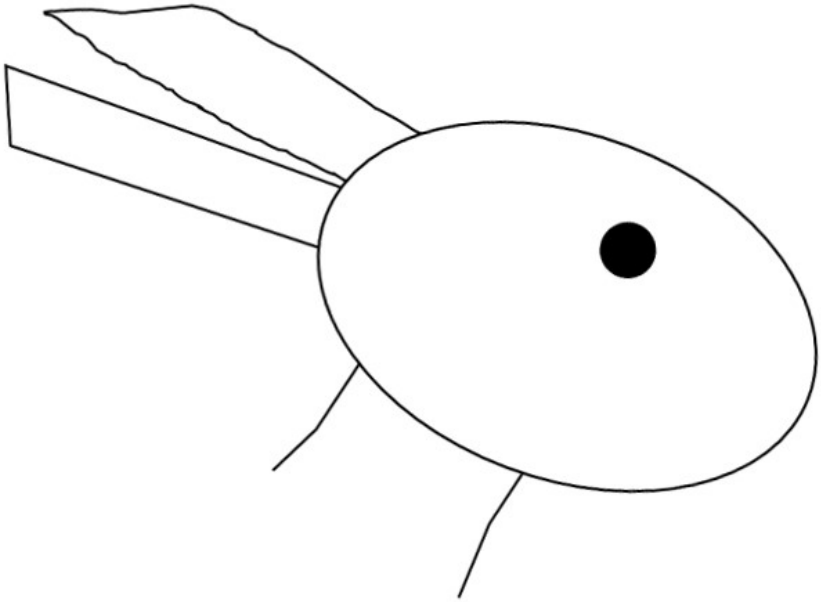

Figure 5: Duck-rabbit ambiguous image, freely adapted from Wittgenstein: the protrusions on the left side can be interpreted either as the beak of a duck facing left or as the ears of a rabbit facing right.

It never works on the first try: a paradigm shift in psychology

Already during my studies, I was trained to deal with failed replications in a manner consistent with the prevailing paradigm. This was before the replication crisis had crystallized. In the first study I took part in, in 2012, we tried, for example, to replicate the finding that colorful sets of objects appear less numerous than single-colored sets of the same size. In terms of consumer psychology, this is somewhat counterintuitive — a long-running German TV jingle for Smarties (in Germany, colourful chocolate lentils similar to US M&M's, rather than the tart pressed-sugar discs sold as "Smarties" in the US) promised "lots and lots of colourful Smarties" ("viele viele bunte Smarties") — but it can be plausibly explained from a Gestalt-psychological perspective (Redden & Hoch, 2009). After a group of fellow students found the opposite – namely, that *colorful Smarties actually appeared more numerous than single-colored Smarties* – we were unable to demonstrate either effect in two follow-up studies of our own. Regardless of whether the Smarties were presented on plates, in cups, or in bowls, regardless of whether the candies were blue, red, yellow, or mixed colors, and regardless of whether quantities were estimated or Smarties were poured from large bottles into containers, our participants were simply not influenced by "colorfulness." Over the following years, I was able, as a teaching assistant, to run further replication attempts. The supervising professor, who was also my mentor, explained: I have honestly never seen a hypothesis confirmed on the very first attempt. After every experiment you get a bit smarter and learn what to do better next time. It's completely natural that it takes a few attempts

before you figure out how to confirm the hypothesis. After we had run six studies with a total of 1383 participants, asked the authors of the original study for advice, eaten piles of candy, and still failed to confirm the hypothesis across all the studies, I had lost confidence in the finding.

About six years later, during my doctoral studies, I thought back on those studies and discussed them with the professor. In light of the replication crisis, it had become clear: if you run multiple experiments and the hypothesis is actually false, chance alone will occasionally still produce a confirmation of the hypothesis. This is comparable to the fact that even a fair coin can land on the same side six times in a row. But if you flip 10 coins six times each, it is not at all unusual for one of the 10 coins to land on the same side all six times. Seen from this perspective, the remark that results never seem to come out the way you want them to on the first attempt has a bitter aftertaste: if the hypothesis is false, it is indeed unlikely that it will nevertheless be confirmed – in a single study. But not if many studies are conducted. In that case, it is actually to be expected that, sooner or later, some study will confirm the hypothesis – even though it is actually false (!). We eventually published these studies together with several of the people involved (Röseler, Felser, et al., 2020).

Structure of the Crisis of Confidence

In a conversation about preregistration – a method for increasing the replicability of a finding – a colleague once replied: "That's all well and good, but as long as the system doesn't change, replication rates won't change either." How can we make sense of this objection? Let us approach this from the other end: at the latest since the Reproducibility Project Psychology (Open Science Collaboration, 2015), most psychologists have understood that there are serious problems with the replicability of findings. Where exactly do these come from? An important problem here is *publication bias*, which refers to the fact that when scientific reports are published, a selection has to be made, so that only certain results end up being published (e.g., results that support a particular theory). This creates a distorted picture of reality. Researchers often even compile reports using only certain findings. "Failed experiments" – that is, ones in which a hypothesis was not confirmed or a theory was not supported – end up in the drawer (the *file-drawer problem*, Rosenthal, 1979; Sterling, 1959). In more extreme cases, scientists make use of various, largely accepted methods to present the data in a way that supports the hypothesis, or act as if what can be read from the data was the hypothesis from the very beginning (HARKing, Parsons et al., 2022). But what drives people who chose an academic career mainly out of interest in how the world works to unconsciously embellish the truth, or even to deliberately manipulate it? At the beginning of the complex process that has led to the replication crisis stands the current academic system: a large proportion of people employed in academia operate under extreme pressure and precarious working conditions. In order to have their contract renewed after one or two years, they must demonstrate publications in journals that are

as prestigious as possible. These are obtained through especially striking and exciting results. The incentive system of academia therefore rewards not truth, accuracy, modesty, or transparency, but above all the things that are not within a researcher's control: exciting and unambiguous results (Bakker et al., 2012).

The process leading from precarious working conditions to low replication rates is illustrated here. In the following chapters, the problems and possible solutions are discussed in detail.

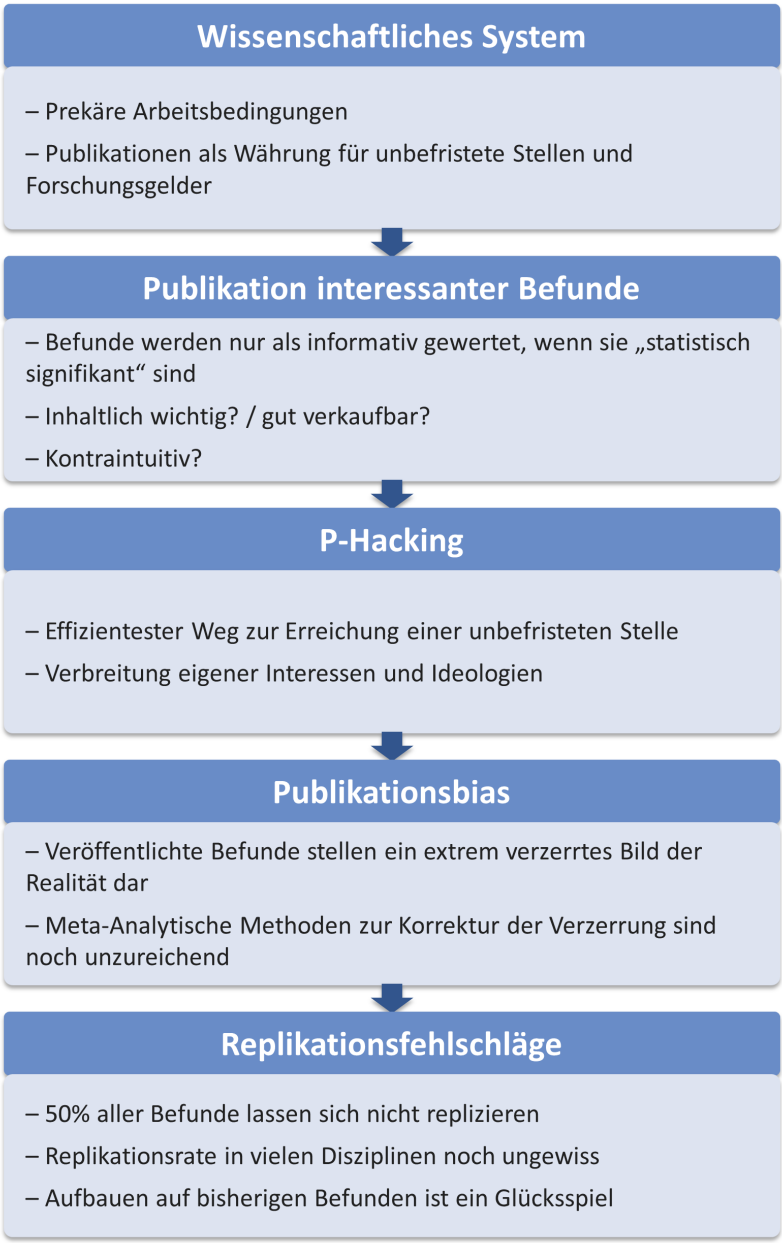

Figure 6: The incentive system of academia as a cause of replication failures.

Problems

In the following, numerous problems of the scientific system are explained and listed, ones that researchers have identified while grappling with science itself. Some of these problems have been known for many decades, others only for a few years. It should be noted while reading that for *all* of these problems, solutions have already been developed and, in some cases, already implemented. What these solutions look like, and which solutions are tailored to which problems, is explained in detail in the chapter *Solutions*.

Problems and solutions are each divided according to different levels, which are shown in the following table.

Table 5: Levels of problems and solutions.

Level	Explanation
System	For people in a society to be able to pursue research as a profession, a framework is needed through which this activity is organized. This includes, for example, the constitutional protection of academic freedom that exists in some countries (in Germany, Art. 5(3) of the Basic Law, *Grundgesetz*, guarantees academic freedom, *Wissenschaftsfreiheit*), the funding of research through tax revenue, universities, degree programs, the link between research and teaching, but also scholarly journals and books.

Level	Explanation
Method	How researchers arrive at knowledge is described by the method they use. Methods are tools that are adapted to the problems at hand, change over time, and are, in themselves, neither good nor bad. Examples of methods include telescopes, X-ray machines, questionnaires, but also computer models, algorithms, and statistical or mathematical models.
Theory	A scientific theory is understood as a well-founded assumption about how a particular relationship or phenomenon works. "Well-founded" means that theories are only used for as long as they agree with observations (or as long as there is no better explanation for them). They are therefore never final. Examples of theories include the theory of evolution regarding the adaptation of organisms to their environment, or dissonance theory from psychology.
World / Epistemology	Epistemology, or the theory of knowledge, discusses what people can know. It is relevant to science insofar as things that cannot be known also cannot be researched.

The System

Alongside the ideal of what academia should be or how it should work, there is academia as it is actually integrated into our social system. One particularity here is that researchers pursue this activity as a profession — that is, they earn money doing it. Committees decide how this money, which comes largely from taxpayers, is distributed, and these committees themselves consist of researchers (*academic self-governance*). Because of the heavy workload of simultaneously conducting and administering research, decision-makers simplify their work by using shortcuts to select highly qualified individuals. As a result, the currency of academia has largely become the number of research articles (*papers*) published in academic journals. Another commonly used indicator is the number of other scholarly articles that cite a person's research — in other words, citation counts. Role models such as Charles Darwin or William James, and even some current Nobel laureates, would stand no chance of obtaining a permanent position in academia by today's standards — they simply did not write enough papers. Many of the problems discussed here have been known in psychology, for example, for several decades, making a replication crisis all but inevitable (M. Baker, 2016; Cronbach et al., 1991; Greenwald, 1976).

Scholarship versus the Academic Business

While some natural and social sciences relied heavily on book publications in their early days, with a substantial body of work built up mostly by individual scholars (e.g., Galileo Galilei for physics or William James for psychology), today it is academic journals that take center stage. These jour-

nals are comparable to the magazines found at a newsstand or supermarket, except that they consist of articles (mostly written in English) authored by researchers, and in many cases are only available online or through university libraries. Researchers then download individual articles from these journals over the internet via their university network, and libraries hold contracts with publishers, paying money so that members of the university have access to the catalogs. Every submitted article addresses a research question defined by the researchers themselves (e.g., how many psychological studies replicate on average?). The articles usually report studies along with their results. Before publication, articles are reviewed (*review*) — not by journal staff, but by colleagues (*peers*). This *peer review* is meant to ensure the quality of research. Typically, reviewers check that the conclusions are justified by the data collected, that the research question is clearly answered, that the article is comprehensible, and that the findings are exciting or surprising. Journals differ in which topics they cover (e.g., social psychology, consumer behavior, applied sports science), how rigorous their peer review is, and how many researchers read and cite them. The academic disciplines are thus deeply embedded in a system that allows researchers and publishers to earn a steady income.

Precarious Working Conditions

So much for the general framework: researchers conduct studies and share their results mostly in the form of *publications*, such as journal articles or books. As for jobs in academia (that is, primarily at universities), they are hierarchically structured. In Germany in particular, permanent positions are few and are almost exclusively professorships, with everyone below that level — doctoral candidates and postdoctoral researchers (*post-*

docs) — on fixed-term positions. (Systems with permanent lectureships or tenure-track assistant professorships, as in the UK and US, distribute security somewhat differently, but the underlying competition for scarce permanent posts is comparable.) Anyone working in academia usually finds themselves in fierce competition for one of the few permanent positions (Rahal et al., 2023). The underlying assumption is that competition among researchers boosts productivity. From the start of a doctorate[1] to an appointment to a professorship — one of the rare permanent positions — contracts typically last only one to three years, and positions are usually less than full-time. At the same time, in many disciplines it is uncommon to complete a doctorate within the typical three years while working fewer than 40 hours a week. While the majority of scientific publications are based on studies conducted as part of doctoral projects, doctoral candidates are simultaneously the people in the system considered to have the least value — that is, whose labor is cheapest. Researchers in fixed-term positions are thus exposed to enormous performance pressure. Mental health problems such as burnout or depression are widespread among doctoral candidates (Jaremka et al., 2020; Liu et al., 2019). The path to a professorship is achieved above all by those who publish many articles in prestigious journals. Given the immense workload and the sheer number of articles submitted to journals, there is no longer time to carefully check and recompute results (Nuijten et al., 2017); instead, what matters most is how clearly the results answer the research question — or, more precisely, how clearly they confirm it (Giner-Sorolla, 2012; Mynatt et al., 1977). In other words, the very part of scientific work that gets rewarded is the one that is not actually in the researchers' hands: the outcome of an investigation. From that

1. The process of obtaining a doctoral degree/title, which — depending on the discipline and institution — involves writing either a single book (monograph) or several journal articles (cumulative dissertation).

point on, published articles and prestige — rather than quality (Brembs, 2018) — become the currency of academia: they determine who receives research funding, and research funding and publications in turn determine who is appointed to professorships. In the hiring committees that make these decisions, members usually do not read the applicants' articles; they merely count how many are listed and in which journals. Applicants are sometimes asked to report their citation counts. Some of these figures (e.g., the impact factor) are owned by corporations, and access to them must be purchased through the institution.

In addition to these devastating problems, there are systemic issues of sexual harassment (Hoebel et al., 2022) and abuse of power, which are difficult to resolve given the currently strictly hierarchical structure of the system (Forster & Lund (2018); see also https://www.netzwerk-mawi.de/ and various recent reports, [1; 2; 3]). According to reports from the German network against abuse of power in academia, cases are kept quiet and the perpetrators quietly move to another university, so the problem never actually gets resolved. M. M. Elsherif et al. (2022) illustrate vividly who has it particularly easy in this system with their *Academic Wheel of Privilege* (p. 85; see also https://www.psychologicalscience.org/observer/gs-navigating-academia-as-neurodivergent-researchers). For example, female doctoral candidates in the Netherlands received worse grades than their male counterparts specifically when the dissertation committee — the group of professors evaluating the doctorate — consisted entirely of men (Bol, 2023). On top of this, universities fall well short of the statutory quota for employing people with recognised severe disabilities that applies to German employers, and further below the rates achieved in other sectors (German-language source).

Too Much Research

During a doctorate, researchers must typically publish three scientific articles within a total of 3-6 years, plus parental leave (or, in fairer cases, produce three articles worthy of publication). For postdocs, the number required is even higher. Under Germany's Academic Fixed-Term Contract Act (*Wissenschaftszeitvertragsgesetz*, WissZeitVG), individuals may be employed at German universities for a maximum of six years before and six years after completing a doctorate. For the *Habilitation* — a second, post-doctoral qualification thesis that in Germany and several other continental European countries is the traditional prerequisite for a full professorship, with no direct equivalent in the UK or US system — a similar time frame is allotted; the rule of thumb in psychology, for instance, is around six articles. How extensive the articles are plays only a secondary role. For example, conducting a meta-analysis — in which existing findings on a particular topic are systematically collected and statistically summarized or compared — often takes several years. A longitudinal study can, depending on the research question, even take decades. By contrast, a cross-sectional survey using an online questionnaire can be completed within a few weeks. A doctoral candidate who conducts a single meta-analysis could not obtain a doctorate with it. Had she instead conducted three simple online studies and published them separately, it would have been easier.[2] These arbitrary requirements have led researchers to place an enormous burden on themselves merely by reviewing their colleagues' articles — a burden that hampers scientific progress (Hanson et al., 2023).

2. One could object here that some journals only publish articles that report multiple studies. In doing so, however, these journals clearly miss their actual goal, namely the *internal* replication of one's own findings. Instead, it tempts researchers to run several studies with few participants each, rather than one study with many respondents.

To illustrate the workload involved, consider a thought experiment: suppose there were 10 researchers who together published 10 articles a year — sometimes alone, sometimes as a group — and each article was reviewed by two people; then each person would need to review two articles. For the system to work, each person would need to review the average number of articles published per person, multiplied by the number of reviewers required. With 10 publications per person and three reviewers, that would be 10x3=30 reviews. But not all articles are published by the journal they are first submitted to, nor are they published immediately. Researchers often submit their articles to the "top-ranked" journals. After several reviewers there have assessed the article, it gets rejected (Jaremka et al., 2020). At best, revisions are requested, which often trigger another round of peer review, and even then the article is not always published afterward. So our calculation does not add up: let us assume, *conservatively*, that an article is reviewed nine times in total (e.g., three reviewers once, then rejection, then three reviewers again, revision, a second review, acceptance). 10x3 becomes 10x9 — accounting for some vacation time, that amounts to a bit more than two reviews per week, ideally taking up to two working days. Under this arithmetic, there is less time left for teaching, knowledge transfer, supervising students or doctoral candidates, acquiring research funding, university self-governance, and so on. Because of the sheer number of articles that must be published and the strict review system, academia piles up a mountain of work on itself — one that is not realistically manageable, and under which the quality of research ultimately suffers. For example, neither reviewers nor journal editors noticed that more than 30 articles contained the phrase "Regenerate response" right in the middle of the text — a phrase that, in OpenAI's ChatGPT, lets a user reword AI-generated text at the click of a button (https://retractionwatch.com/2023/10/06/signs-of-undeclared-chatgpt-

use-in-papers-mounting/). Some articles even stated, "As an AI language model, I …" (PubPeer search for this phrase"). In one case, the publisher Elsevier altered a published article — not through the recommended route[3] of a transparent *erratum* or *corrigendum*, i.e., a public notice explaining the change and its reasons, but without any explanation or the authors' consent (https://predatory-publishing.com/elsevier-changed-a-published-paper-without-any-explanation/).

Publish or Perish

Anyone working in academia should know the most important rule of the game: whoever wants to survive must publish articles. In short: publish or perish. For a doctorate, a *Habilitation* (see above), acquiring research funding, and being called to a professorship (in Germany, the formal *Berufung* procedure), publications are the top criterion. The defining feature of a currency is that it allows a value to be assigned to things. So what does value look like in research? To describe and select journal subscriptions (that is, explicitly *not* to evaluate) across different research fields, bibliometricians devised various metrics, such as the impact factor or the h-index. Both figures are calculated from how often articles are cited, broken down by journal or by person. For instance, the journal impact factor is calculated by dividing the total number of citations in a given year by the number of citable publications in the reference years (e.g., the three preceding years). Journals advertise high impact factors and sometimes do not allow citation of comments (or publish comments under the same reference as the original studies) in order to keep their impact factors as high as possible. They can, of course, also choose whether to report the 3-year

3. Ethical guidelines for the publication process are available, for example, from the Committee on Publication Ethics (https://publicationethics.org/guidance/Guidelines).

or 2-year impact factor. Calculation methods also differ in terms of which underlying data they draw on. The journal impact factor database currently belongs to the company Clarivate Analytics and is not publicly accessible. This means the exact figures cannot easily be recalculated and checked. Given the important yet opaque role of citation metrics, it is hardly surprising that they get manipulated, and that 3 of the top 10 journals in one field turned out not to be scientific journals at all. It is also clear that citation metrics are either unrelated to, or negatively related to, scientific quality (Brembs, 2018; Etzel et al., 2024, Table 3). For example, there was the problem that Microsoft Excel recognized gene names as dates, changed their format, and rendered the actual names unrecognizable. As a result, parts of the corresponding results became useless (Ziemann et al., 2016). This problem occurred especially in prestigious journals. They did nothing about it; instead, Excel itself was adjusted a few years later.

Besides journal rankings, rankings of individuals can also be created. Researchers from Stanford published one such list of the *top 2% most-cited researchers worldwide.* Aside from inviting misuse, it contains numerous errors (Abduh, 2023), such as researchers who supposedly published articles for hundreds of years — both before and after their death. Companies that own publishing houses and other tools used in research (e.g., reference-management software) also collect data about researchers (e.g., which articles are accessed for how long, which passages of text are highlighted). Libraries are sometimes asked to install publishers' tracking software. The publishers then sell the collected data back to the researchers. This practice endangers academic freedom, since states can require publishers to disclose the names of researchers working on politically sensitive topics. The initiative Stop Tracking Science campaigns against this practice.

Researchers often choose the journal in which to publish their results based on citation metrics, and in hiring committees, applicants are frequently selected on the same basis. In personal accounts, for instance, professors report that across countless hiring committees over several decades, they never once read a single one of the applicants' research articles. To get their articles into "high-impact journals," researchers are even willing to embellish their results — in order to increase their chances of publication (Bohorquez et al., 2024). Calls for the responsible use of these metrics have been made for some time now (Hicks et al., 2015), and numerous universities and researchers have joined forces to make the evaluation of research more meaningful (e.g., through the San Francisco Declaration on Research Assessment). How incentive structures and career status affect research misconduct has been studied with mixed results. The underlying problem: as soon as a system has a clear evaluation criterion, everything gets oriented toward it (*gaming the system*). According to Fanelli, Schleicher, et al. (2022), incentive structures that lead to poor research — in the form of image manipulation in biology — are especially prevalent in China, and considerably less so in the USA, the United Kingdom, and Canada.

Another product of the *publish-or-perish structure* is that most instruments developed in psychology to measure personality traits are used only a handful of times — and mainly by their own developers, at that (Elson et al., 2023). Elson et al. (2023) make the point: psychological measurement instruments are not toothbrushes! An even more extreme version of this problem can be seen with so-called *paper mills* (van Noorden, 2023): organizations that automatically generate large quantities of scientific articles without actually conducting the studies described in them. Researchers can then buy co-authorships to increase their number of published articles

and gain more citations. Depending on the journal, these articles are not caught during peer review, that is, evaluation by other researchers. There are concerns that the buyers of such articles can themselves become very successful, even becoming journal editors, thereby protecting themselves from being caught. The true extent of the paper-mill problem is unclear and remains largely unexplored (Byrne & Christopher, 2020). One telltale sign for detecting articles produced with the help of artificial intelligence is so-called *tortured phrases* (Cabanac et al., 2021) — phrases that are grammatically correct but rarely occur in ordinary usage or make little sense.

The Bottleneck Hypothesis and the Drive for Innovation

Renowned scientific journals receive countless submissions every day but publish only a limited number of articles. They must therefore select very strictly what gets reviewed and, ultimately, published. Because a journal's goal is to be widely read, journal editors choose the articles with the greatest potential to become well known and heavily cited (Giner-Sorolla, 2012). This applies, for example, to contributions with particular practical implications, surprising findings, or especially consistent findings. Studies whose results do not permit clear conclusions — or whose authors draw their conclusions too cautiously — are therefore out of the running. In the neurosciences, for instance, some journals openly state that they do not publish replication studies and that they select for novelty, while most other journals take no position on the matter at all (A. W. K. Yeung, 2017). In psychology, in 2017 only 33 out of 1,151 journals stated that they accept replications (Martin & Clarke, 2017). Innovative findings, or findings presented as innovative (Stavrova et al., 2024), are cited more often, so the journal gains more readers, more submissions,

and thus more money through subscriptions and publication fees — yet heavily cited articles tend to replicate worse than less-cited ones (Serra-Garcia & Gneezy, 2021), and prestigious journals are magnets for questionable research practices (Kepes et al., 2022) and for research that is demonstrably no better, or even of lower quality (Brembs, 2018).

How does this come about? People speak of a *bottleneck* — many submissions but few publications. Combined with the incentive to publish in precisely these journals, this leads researchers to use every means available to gain a chance at publication. At some institutes, publishing in a particular journal automatically earns a doctoral candidate the highest possible grade. Other institutes already specify, in the job posting for a doctoral position, which journal the results of the research project must be published in. The fact that prestigious journals such as *Nature* or *Science* mainly publish articles with clear-cut messages — articles that also get read more often (Stavrova et al., 2024) — spurs researchers on to *manufacture* clear-cut results. If a finding happens not to fit the hypothesis being tested, it is either distorted through data-analysis practices that are currently tolerated in many places (questionable research practices) until it does fit, or it is not published at all and ends up in the file drawer (Gopalakrishna, Wicherts, et al., 2021; Schneider et al., 2024). Likewise, this discourages researchers from repeating an experiment that has already been conducted — that is, from running a replication study (Marefat et al., 2024).

The File-Drawer Problem

It has been known for several decades that researchers primarily publish those studies that support their theories (Rosenthal, 1979; Sterling, 1959). In an extreme case, someone might run five studies to test a theory, con-

firm the theory in only one of them, and publish only that one. Other researchers who then search the (published) literature see only the "successful" study. This creates the impression that the theory is correct, even though the majority of the studies do not actually support that conclusion. Because study results are subject to natural statistical fluctuation, it is likely that, among many studies, at least one will show the desired result purely by chance. The file-drawer problem allowed entire lines of research to develop that have since gone completely extinct once awareness of replication studies took hold (Brockman, 2022; Mac Giolla et al., 2022).

For *meta-analyses* — studies that summarize existing findings — a range of methods have already been developed to assess the severity of the file-drawer problem. Numerous methods for correcting this bias also already exist (Fisher et al., 2017; Hedges & Vevea, 1996; Schimmack, 2020; Simonsohn et al., 2014b; van Aert & van Assen, 2018). However, none of these methods works in every possible scenario (Carter et al., 2019). So researchers cannot avoid actually publishing their failed studies.

Medicine represents a special case: all studies conducted there must be publicly registered. Any subsequent publication must then cite a registration number. Public information on registered studies therefore makes it possible to track which individuals, institutions, or countries actually publish how many of the studies they conduct. Researchers in Berlin have developed a website with an interactive *dashboard* for exactly this purpose, allowing the collected data to be searched and displayed (Franzen et al., 2023; Riedel et al., 2022). As of the current state (July 2024), https://quest-cttd.bihealth.org/ shows that, of all registered studies, only 46% were published within the following two years and 74% within the

following five years. People who volunteer as participants in these studies, or third-party funders, thus gain insight into the size of the drawer in which failed studies and "not-so-exciting results" end up.

Access to Knowledge

Because research is embedded in the commercial publishing system, much of academia finds itself trapped in a social dilemma that results in severely restricted access to knowledge. The dominant model for academic journals — most of which belong to publishers such as Springer, Elsevier, Sage, or Taylor and Francis — is a subscription model. University libraries regularly pay money to publishers so that members of the university (students and staff) have access to the works published there. Anyone without a subscription can buy individual articles. If you try to download an article online without being on a university network, it is not free: the article sits behind a paywall. If an article is to be published so that everyone can access it freely (*open access*), this costs the authors extra — typically between €2,000 and €9,000 per article, and up to €20,000 for books (*article/book/chapter processing charges*). This also affects the article's own authors: upon publication, they sign away all their rights to the publisher, and can only access their own article through a university with a subscription, or by paying a fee (around 30 US dollars), unless they paid the enormous open-access fees in the first place. So in order to read research at all, universities must buy subscriptions or individual articles. As a result, individuals, institutions, countries, or even entire regions with fewer financial resources — such as the Global South — are systematically disadvantaged in their ability to take part in international scholarly discourse.

This social dilemma lies precisely in how difficult it is to change this system. Brembs et al. (2023) describe it as follows: libraries sign subscription contracts with publishers in order to make research available at their universities. If they canceled these contracts, that could slow down research and weaken a university's standing. Researchers cannot boycott the well-known journals of commercial publishers, because doing so would jeopardize their careers. They depend on publishing in prestigious journals. It is also extremely difficult to suddenly change the behavior of millions of researchers, most of whom have lived with the current system for many years. The journals, as the third party in this dilemma, profit from this dependency and can raise prices at will and worsen publication terms as they please (e.g., requiring authors to sign away all rights to their research). Aside from their brand value — that is, the prestige and trust mistakenly (Brembs, 2018) placed in them — they contribute almost nothing to this system. Universities and countries already have the option of managing research articles online in journals of their own; peer review and quality assurance are carried out by volunteer researchers rather than journal staff; and, as is evident from existing journals[4] (Carlsson et al., 2017), researchers can handle the formatting of articles themselves with little effort. Corresponding solutions are discussed in the chapter *Open Access Publications*.

i Wasting Taxpayer Money on Commercial Publishers?

The German-language science show MAITHINK X explains why taxpayer money is wasted under the current system. For instance, the total amount of APCs — that is, publication costs for articles in

4. This refers to journals that publish articles without requiring a subscription and without a paywall, and at no cost to the authors.

> open-access journals, not including library subscription costs, book publications, and much more — is estimated to have exceeded 2.5 billion US dollars worldwide, and more than tripled between 2019 and 2023 (Haustein et al., 2024). What the program falls short on is discussing possible solutions (see the chapter Solutions > The System).

Scientific Quality Control

The problem of quality control runs through all the problems of the scientific system described so far. During the review process before publication, many errors go undetected, and after publication, reports are so opaque that it is often not even possible to expose fabricated data. Reviews of scientific articles usually remain confidential, and when one journal rejects an article, it simply gets submitted to another. The criticisms already raised are then lost along the way (Aczel et al., 2021).

Scientific articles end with a paragraph on potential conflicts of interest. There, researchers must disclose whether they received benefits from companies for their research, or whether they profit in any way from the research (e.g., by owning stock in the company whose drug they tested favorably). In gambling research, such disclosures are frequently missing (Heirene et al., 2021). Some large companies systematically undermine scientific consensus by spreading false information: a well-known example is the tobacco industry and the well-established negative effect of smoking on health (Reed et al., 2021).

Some journals exploit this situation by publishing research in exchange for high publication fees (e.g., €4,500 per article) despite weak or entirely

absent quality control. This practice is called *predatory publishing* and refers to researchers effectively buying publications, with publishers profiting from it. Because the peer-review process is anonymous and kept confidential, it is sometimes unclear whether a given journal has any quality control in place at all. Paper mills go a step further still. There, people can buy co-authorships on articles. These articles are frequently produced by individuals or generated using algorithms. Cases in which such nonsensical articles are exposed are common, but the articles are rarely retracted or given a public notice explaining the error (Cabanac & Labbé, 2021). Further detection methods are being developed, checking, for example, whether the authors of journal articles have an institutional email address (that is, an address from a university or research institution) (Sabel et al., 2023).

Besides publications, researchers can also buy citations or artificially inflate their own (Singh Chawla, 2024) — for instance, according to Google Scholar, Larry the cat was briefly credited with 132 citations. Because reviews remain confidential, "review mills" have also emerged: there, authors are pressured into citing specific research articles in order to boost the reviewers' own citation counts. These reviews consist of vague, formulaic boilerplate text (Oviedo-García, 2024).

Further Information

- The free film "Paywall: The Business of Scholarship" covers the topic of paywalls and public access to scientific knowledge: https://paywallthemovie.com/paywall.

- M. M. Elsherif et al. (2022) discuss the topic of neurodiversity in research. Via FORRT.org, the group also maintains a database of neurodivergent researchers and other projects.
- Thériault and colleagues recently released the Missing Majority Dashboard (Thériault, 2023), which automatically uses the OpenAlex literature database to show which continents authors come from. North America and Europe are clearly overrepresented.
- TED Talk on inclusive science: https://www.youtube.com/watch?v=tB_HeqnonNM.
- Retractions from the Retraction Watch Database can be viewed visually by country and type in the Retraction Dashboard: https://retraction-dashboard.netlify.app.
- In an interview (in German), Jasmin Schmitz explains various types of research misconduct: www.youtube.com/watch?v=EqW7vP6dizU.

Career in Academia

A career in academia usually requires a relevant degree (bachelor's and master's) and begins with a doctorate. During the doctoral phase, one learns the craft of research and earns the doctoral degree: studies are conducted, data are analyzed, and research articles are published. In Germany, doctoral candidates are typically employed as *wissenschaftliche Mitarbeiter* ("research associates") on a 50–66% contract — a salaried university post, unlike the fully funded studentships common in the UK or US — and in that role they review articles and grant applications on behalf of their supervisors, write their own grant applications, spend several months abroad, create, supervise, and grade exams, supervise theses, and take part in university self-governance by attending meetings and serving on committees. Positions or scholarships are usually set for a term of three years. At the start of my own doctorate, I was told that with a 40-hour week you won't manage to finish your doctorate in the intended time – in other words, 50% pay for more than 100% of a full-time workload, on a fixed-term contract. Yet contracts are not always limited to the duration of the doctorate. Many people find out only a few months before their contract ends what percentage of a salary they will receive next, how extensive their teaching load will be, and how long the next contract will run. In short: working conditions are not optimal, and choosing a path into academia requires a high degree of flexibility.

Working conditions in academia are also often described as *precarious*, meaning delicate or problematic. The idea behind the strong competition is that pressure fosters productivity. Yet built into the rules of the system is also the recognition that pressure does not lead to good science. In the (pre-

sumably rare) extreme case, researchers commit fraud (e.g., by fabricating data, as in the Stapel affair or the Himalayan fossil hoax).

Double dependency

There are various ways to fund the doctoral phase: some pursue a doctorate alongside employment at a company, or scholarships that cover basic living costs for a limited period (in Germany, for example, €1,100 **per month** for 36 months under the state of Saxony-Anhalt's graduate funding scheme, with social security contributions on top — German doctoral stipends are typically in this range). The most common path, however, is a position as a research associate. In that case, the supervising person is also the person who evaluates the doctoral thesis. Being handed 120 first-year exam scripts to mark — cohort sizes at German universities are large — can, in extreme cases, come into conflict with the time needed to work on one's own research. Doctoral candidates therefore usually depend on their supervising professors both through their employment *and* through the evaluation of their work – a situation referred to as *double dependency*.

> Family-unfriendly scholarships
>
> Fixed-term positions, double dependency, meager scholarships, and performance pressure all have clear negative effects on the health of researchers, and thus on the quality of science. What about doctoral scholarships? Scholarships often run for one year with an option for extension, or for up to three years. Researchers who pursue a doctorate directly after school and their degree are about 24 years old, may already have a partner, and may wish to start a

family. Under such uncertain contractual conditions, that is difficult. Scholarships often provide no option for parental leave and must then simply be paused – or they keep running with no extra time added at the end. They forbid additional income, which means Germany's earnings-related parental benefit (*Elterngeld*, normally about two-thirds of prior net income) collapses to its statutory floor of €300 per month. On top of that, they stipulate that the funded person – *including their spouse* – must not exceed a certain maximum level of assets, so that funding through personal savings is not an option either.

Depression, burnout, and #IchbinHanna

Across disciplines, a movement developed under the hashtag #IchbinHanna ("I am Hanna") in response to a since-deleted explainer video from the German Federal Ministry of Education and Research, in which a fictional doctoral candidate named "Hanna" explains the Academic Fixed-Term Contract Act (WissZeitVG). The movement sharply criticized both the video's matter-of-fact explanation (https://www.youtube.com/watch?v=PIq5GlY4h4E) and working conditions in academia more broadly. The German-language video stated (my translation): "so that up-and-coming researchers also have a chance to acquire this qualification, and so that one generation doesn't block all the positions, universities and research institutions are permitted to conclude fixed-term contracts under the special rules of the WissZeitVG. This creates turnover, which fosters innovative capacity." In Germany, roughly 90% of academic staff below professor level are employed on fixed-term contracts — far

higher than in most other national systems (German-language source: https://www.youtube.com/watch?v=H1wJmqpGhJc).

Does planning insecurity combined with massive competitive pressure really foster innovative capacity? That seems unlikely: mental illnesses such as burnout (Jaremka et al., 2020) and signs of anxiety disorders and depression are widespread among researchers. According to a review study by Satinsky et al. (2021), between 18 and 31% of doctoral candidates suffer from depression.

> **Who can help?**
>
> Quick help for psychological difficulties is available, for example, through the crisis hotline of TelefonSeelsorge (a German telephone counseling service). Some universities have their own dedicated support offices, and cities often have local psychotherapy outpatient clinics and counseling centers. Readers elsewhere can find local crisis lines via findahelpline.com; in the US and Canada, dial 988, and in the UK, Samaritans can be reached at 116 123.

Top-down change

Those who could change the system – that is, everyone with a permanent contract – no longer suffer under it. For those who do suffer under the system, it is unwise to try to change it by turning to professors, since these are precisely the people who, as members of appointment committees, decide whether someone stays in academia by awarding professorships. In extreme cases, this can lead someone to hold back criticism of the work of more senior researchers out of fear of diminishing their own chances at a professorship. On the other hand, professors have proven that

they are very good at playing by the rules of the game. Admitting that they obtained one of the rare and highly sought-after positions through the quantity rather than the quality of their research would mean devaluing themselves.

Further information

- The German registered association Respect Science advocates for improving the working conditions of researchers.
- TED Talk on publication culture and Open Science: https://www.youtube.com/watch?v=c-bemNZ-IqA.

Methods

In everyday thinking, the myth still often prevails that science is distinguished from non-science by *the scientific method*. That is false (Feyerabend, 1975/2002). It is true that scientific knowledge differs from everyday knowledge (and also from religion) through a higher degree of systematicity (Hoyningen-Huene, 2013), but there is neither a single nor a constant scientific method. Instead, methods have changed over time, and that is a good thing. New technologies enable, for example, highly precise measurements using electron lasers in physics, 3D scans of artifacts that otherwise only a few people would ever see in the historical sciences, or databases of voluntarily provided and anonymized chat logs in the social sciences (https://db.mocoda2.de/c/home).

Just as a hammer or other tool is neither good nor bad, methods too are neither good nor bad, neither right nor wrong – rather, they are used *appropriately and correctly*, or they are misused. Instead of "misuse," the social sciences speak of questionable research practices, or QRPs for short. They allow researchers to generate the findings they want. In what follows, widespread and frequently applied (John et al., 2012) techniques are presented (for an overview of the research on this topic over the last 50 years, see also Neoh et al., 2023).

Exploratory versus confirmatory research

Understanding questionable research practices (QRPs) requires an important methodological distinction: like a walk, a scientific investigation can be exploratory or goal-directed. Sometimes one strolls freely through the

area and makes new discoveries along the way; sometimes the destination and the route are clear and determined in advance. In a scientific context, this is referred to as exploratory and confirmatory research. In exploration, at most the research question and rough features of the method are fixed; in a confirmatory test, everything is worked out in advance: the procedure, the possible results, and explanatory approaches for every possible outcome. A specified hypothesis, together with its associated theory, is then either confirmed or not. Neither approach is superior to the other. Investigations in a largely unexplored area typically begin with exploration, while more prior research goes hand in hand with clearer expectations. It should be noted that these are extreme types of research that form a spectrum, and that only both approaches together allow for genuine gains in knowledge. Within the hermeneutic circle (simply put, "the circle of understanding"), a general regularity is formulated from individual observations (*induction*), and this regularity is subsequently tested against further individual observations (*deduction*). Depending on the regularity in question, deduction can be logically necessary, since a further statement is inferred from previously established statements. The prior assumptions are called *premises*, and the *conclusion* follows from them. If both premises are correct, the conclusion must also be correct. Induction, by contrast, is not a necessity (see the *problem of induction*, Hume, 1748/2011).

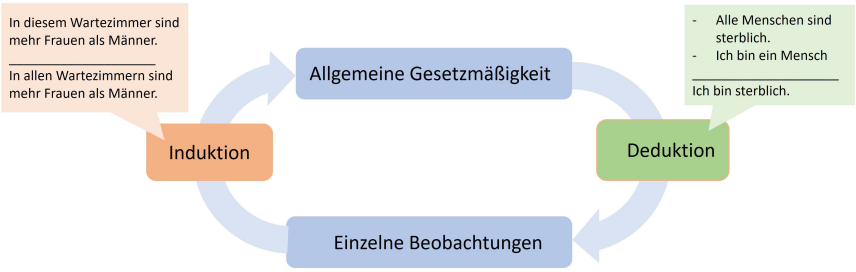

Figure 7: The hermeneutic circle.

Problems arise when exploratory research is communicated as if it were confirmatory – that is, when it is presented as though a single observation had confirmed an already-formulated regularity, rather than merely having inspired it. This kind of faulty logic is called circular reasoning: the regularity holds because of the observation, and the observation matches the regularity.

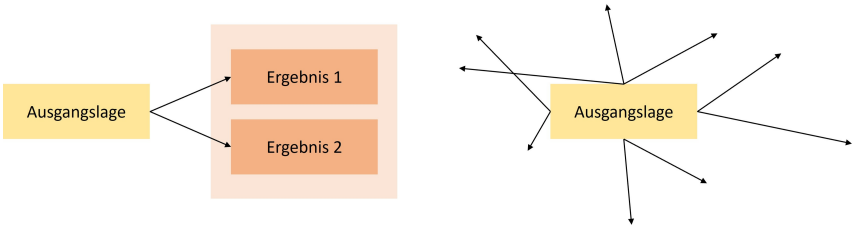

Figure 8: Sketched and simplified procedure for confirmatory (left) versus exploratory (right) research. Confirmatory approaches often depict a narrow and controlled slice of a phenomenon. Exploratory approaches are not goal-directed, the direction can change, and they are sometimes associated with unforeseen results.

Methods of data generation

Scientific disciplines typically draw on many different methods. Ideally, findings are independent of the method that led to their discovery, and different methods lead to the same finding. Typical methods in the social sciences include surveys using standardized questionnaires, behavioral observation via video recordings followed by coding of behaviors by multiple people who are unaware of the study's purpose, indirect methods (Schimmack, 2019) in which something other than the actual target construct is asked about, behavioral measures such as eye tracking, or simulation studies, which are used, for example, to compute traffic flows on the basis of predetermined principles or to predict panic attacks

(Robinaugh et al., 2021). These methods almost always generate data – for instance, a table in which data are recorded across several columns for each observational unit (e.g., each participant) and which are then almost always analyzed statistically. The need for such analysis arises because the observed regularities are not absolute laws in the sense of "all male babies weigh more than all female babies," but rather statistical regularities in the sense of "on average, male babies weigh a few hundred grams more than female babies shortly after birth, but not every male baby weighs more than every female baby." A comparison of height by sex (figures for Germany) can be seen on Statista (German-language page).

The following figure shows the frequencies of different values (a *histogram*). The further to the right a value is, the higher that value is (e.g., birth weight), and the higher the bar, the more often that value occurs. The yellow and purple distributions overlap, meaning that not all yellow values are lower than all purple values. On average, however, the yellow values are lower.

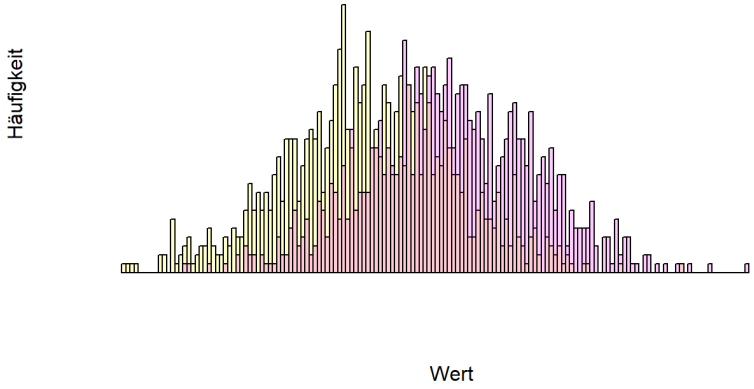

Figure 9: Histogram of two overlapping distributions (e.g., birth weight by sex).

Statistical significance

One of the most widely used methods in the social sciences (and beyond) is statistics, more precisely *inferential statistics*. Here, a limited set of observations (e.g., completed questionnaires from 100 people) is generalized to all possible observations (e.g., all people). The relationships under investigation are rarely clear-cut, but statistical regularities are common. Characteristic of this is a certain *element of chance*. If you weigh a recently born male and female baby, the probability is very high that the male baby weighs more. But it also frequently happens that this is not the case. The situation is similar with a fair coin – one that on average lands on heads and tails equally often: it is unlikely that it will land on heads on all four out

of four tosses, but landing on heads once or twice does happen fairly often (specifically, in 6.25% of all cases in which a fair coin is tossed four times in a row).

Inferential statistical tests now assume that, when looking at a statistical relationship (e.g., sex and birth weight, body weight and height, parents' education level and children's education level), "only chance is at work" (Röseler & Schütz, 2022). Under this assumption, it is calculated how often an observed relationship of the observed strength would occur if there *actually* were no relationship at all. For example: "that a fair coin lands on heads four times happens in 6.25% of all cases." For six tosses, it would be 1.5625%. The art of statistical inference lies in finding the point at which researchers conclude that chance was not at work, because the calculated probability is so low. Conventionally, this threshold lies at 5%, for new findings sometimes at 0.5% (Benjamin et al., 2018), and in especially precarious cases even lower. In technical terms, this is called an *alpha level* or a *significance level*, and the calculated probability is called the *p-value*. p-values below a certain percentage (e.g., 5% in most social sciences[1]) are called *statistically significant*, or *significant at the 5% level*. Researchers would thus say that a coin is *not* fair if it lands on heads six times in a row (even at five times, which occurs in 3.125% of cases). In doing so, they accept that, if the coin actually is fair, they will draw a wrong conclusion in 5% of all cases.

On the other hand, it is entirely possible for a coin to be unfair, landing on heads 60% of the time and on tails 40% of the time, for example.

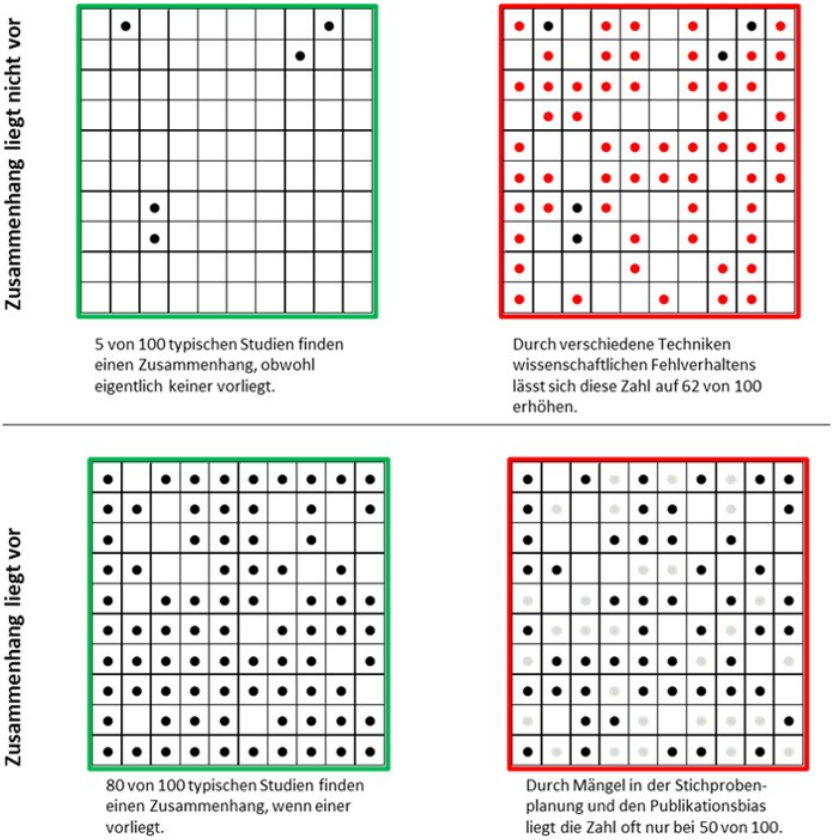

Figure 10: A single study does not yet lead to certain knowledge. Even if an investigated relationship does not actually exist, it can appear by chance in the data. And even if a relationship does actually exist, random fluctuations may prevent it from showing up in the data. Open science practices are meant to restore the state shown in the boxes on the left. From Röseler, L. (2021). Wissenschaftliches Fehlverhalten [Figure]. https://osf.io/uf7gz/. Licensed under CC BY-Attribution International 4.0.

1. In particle physics, statistical models are also used. There, instead of the 5% significance level (1.96 sigma), the 5-sigma criterion is applied, which corresponds to an alpha level of 0.0000573%.

Researchers' degrees of freedom

Running complete studies multiple times is very costly. Although it is a relatively reliable path to significant p-values, there are far more economical solutions. Most analyses are many times more complex than the coin-toss study described above. Let us consider the still very simple significance test for a correlation coefficient. The coefficient is a number between -1 and 1 and describes the type of relationship between two variables (e.g., income and life satisfaction). 0 means that there is no relationship; positive values mean that when one variable has high values, the other also has high values; and negative correlations mean that when one variable has high values, the other tends to have low values. Figure 11 shows various correlations.

Correlation

In statistical reports, *r* denotes a *correlation coefficient*, usually the *product-moment correlation* (also known as the Bravais-Pearson correlation). Correlation coefficients are standardized values for the relationship between two variables. These could be the intelligence and salary of several people, or the top speed and weight of several cars. Correlation coefficients always lie between -1 and 1. Values below 0 mean that the higher one variable is, the lower the other is (negative relationship). Values above 0 mean that the higher one variable is, the higher the other is too (positive relationship). 0 means that the two variables involved are independent of one another. The 98 in parentheses is the number of observations minus 2 and is referred to as degrees of freedom[2]. The correlation value of .420 (or 0.42) means that a positive relationship was observed. It is also

important to note that this only captures an overall positive or negative relationship (*linearity*). So if, for example, a U-shaped relationship is present (bottom right in the figure), this will not be reflected in the correlation.

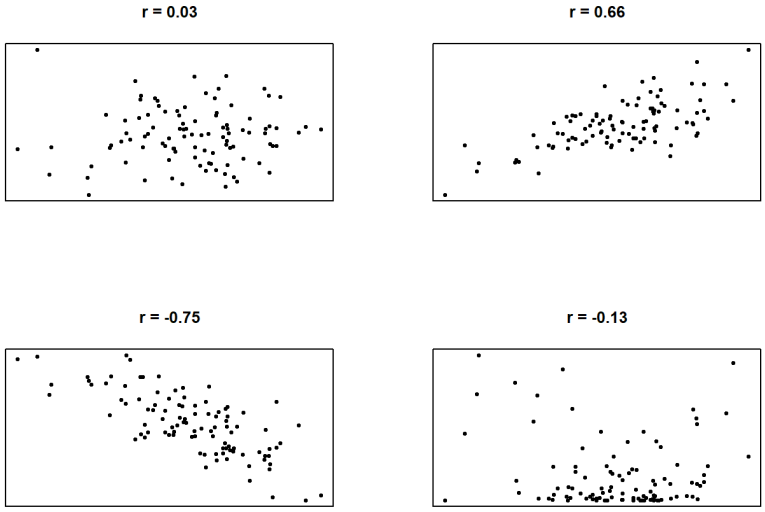

Figure 11: Various relationships between two variables and their correlation coefficients (simulated data).

Although this is a very simple test, it involves many decisions. Even after data collection, decisions must be made: Which of the surveyed people will be used for the test? Should any people be excluded, and if so, why (e.g., extreme values or implausible values)? How are the values of the variables computed? Which type of correlation should be used (e.g.,

2. The logic behind the term is comparable to a formula with several variables and the question: "How many variables' values do I need to know in order to calculate the rest?"

Bravais-Pearson, Kendall, or Spearman)? Is there an expectation about the direction of the correlation (directionality of the hypothesis)?

These questions correspond to degrees of freedom – that is, researchers have flexibility regarding which options they choose. None of the options is inherently superior to all the others, and each decision can be justified to some extent. The problem with this flexibility is that the results depend on it, and depending on the decisions made, the result can turn out to be a positive correlation, a negative correlation, or no correlation at all. The more complex the investigation and the statistical procedure, the greater the flexibility in data analysis. In itself, these degrees of freedom are not a bad thing; the problem only arises when just those results that are easy to publish or that fit researchers' beliefs are presented. This practice is called HARKing (hypothesizing after the results are known) and constitutes a form of circular reasoning. The hypothesis that was tested comes from the data, which of course confirm it. Various solutions allow for the reduction or complete elimination of degrees of freedom (e.g., preregistration). It is also possible to communicate the approach as exploratory, i.e., not planned or determined in advance.

In the data analysis process, the analogy of the "garden of forking paths" is used. In a simplified (!) example in Figure 12, we have 3x4x4x4 = 192 different results, which together cover the entire spectrum of possible conclusions – regardless of whether our hypothesis is correct or not.

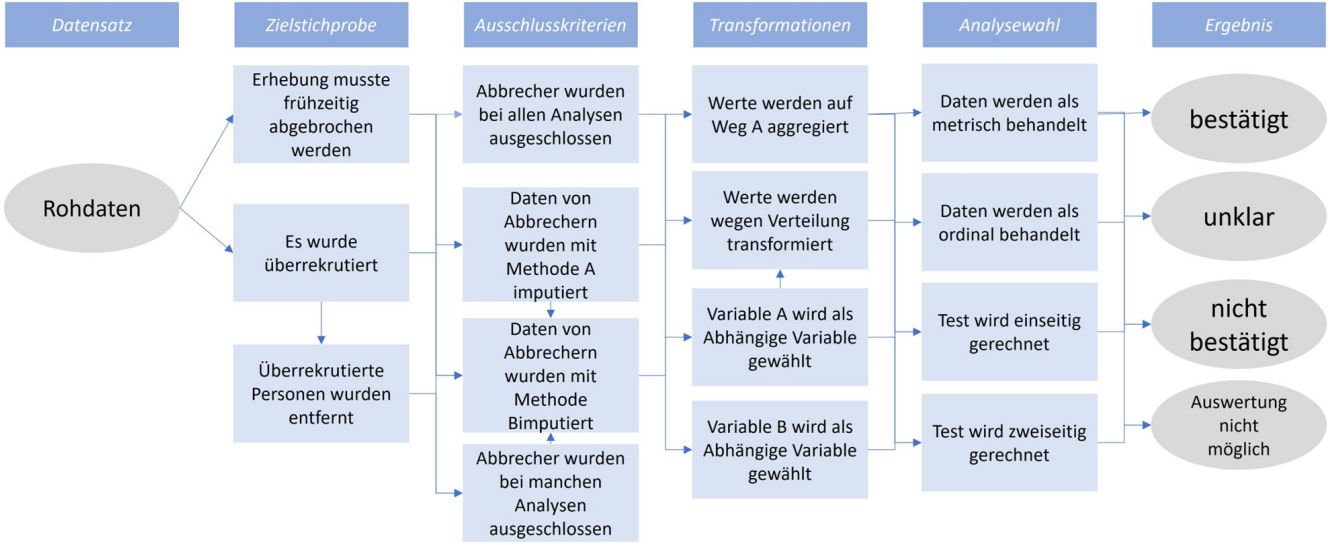

Figure 12: 192 different paths from a dataset to the (desired) result.

Demonstrations of the *garden of forking paths* exist for a wide variety of fields. The dependence of results on analytical choices has already been shown for evolutionary biology (Gould et al., 2023), social policy (Breznau et al., 2022), structural equation modeling (Sarstedt et al., 2024), and linguistic analysis (Coretta et al., 2023).

Typos

Data are often analyzed using advanced software, and the results then have to be laboriously transferred into the report. This is where typos quickly creep in. Nuijten et al. (2016) developed an algorithm that automatically detects reported significance tests, recalculates them, and flags inconsistencies. They found that, in major psychology journals between 1985 and 2013, roughly half of all articles contained at least one error. These "typos" were not entirely random; rather, erroneous values tended to favor positive findings. Such transcription errors also occur in meta-analyses (Lopez-Nicolas et al., 2022). And even citations are frequently erroneous: across various scientific disciplines, (Smith & Cumberledge, 2020) found that in 25% of all examined citations, the claims attributed to the cited work were not actually supported by the original articles.

P-hacking

The p-value in statistical tests indicates how probable the observed pattern is, given a previously assumed pattern. For a correlation, this usually means: how probable is it to observe a correlation of the magnitude found, if there is actually no relationship (i.e., $r = 0$) between the variables under investigation? Concretely, this could mean: how probable is it that, in

my dataset of 100 people, the correlation between intelligence and age is exactly $r(98) = .420$, if I actually assume that the two variables are unrelated?

The assumption of no relationship built into the significance test is called the *null hypothesis*. If the observed pattern is extremely improbable under the null hypothesis (often below 5%), this is referred to as a statistically significant relationship. It is important to note that significance here is to be understood purely in the statistical sense. The question of how meaningful a finding is for the world and for life cannot be answered with statistics within this framework. Because p-values are probabilities, they lie between 0 and 100%.

Among the QRPs (questionable research practices), p-hacking is another category that in turn encompasses several distinct techniques. P-hacking refers to researchers using their degrees of freedom to make the p-value "significant," i.e., to bring it below 5%. A frequently mistaken assumption about p-values is that high p-values indicate the *absence* of a relationship, or that p-values are only low when a relationship actually exists. In fact, p-values tend to be small when a relationship exists that can also be detected given the amount of data collected. If no relationship exists, p-values are uniformly distributed, meaning that all p-values occur equally often. Given the definition above, it follows directly that out of 100 studies conducted, roughly five will tend to show a significant relationship *even if none actually exists*. This fact enables a variety of p-hacking methods. Simmons et al. (2011) showed that the probability of obtaining a significant result, when there is actually no relationship in the data, can rise from 5% to roughly 60%. Figure 13 shows the distribution of p-values at various levels of statistical power (i.e., the probability of detecting a relationship of a given size when it actually exists).

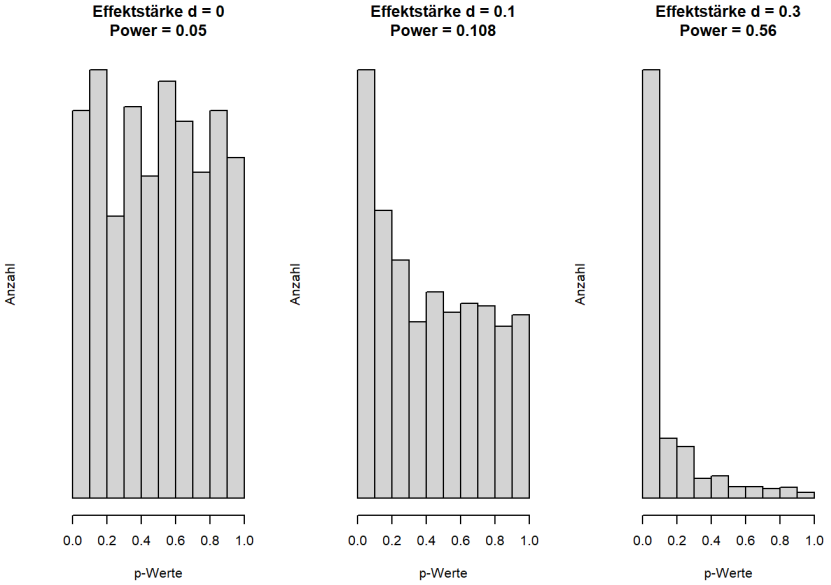

Figure 13: p-values are uniformly distributed when there is no actual difference or relationship, i.e., when the null hypothesis holds. The higher the statistical power, the more the distribution shifts into the range of statistical significance.

The chance of obtaining significant p-values even when the tested hypothesis is not actually true can be increased by "slicing" the sample (e.g., analyzing only women, only employed people, or only people older than 30), by collecting additional data ("optional stopping"), or by using several central variables (for example, measuring intelligence with three different tests and correlating each test individually with age). Even changing small parameters in the statistical tests (e.g., using a non-parametric Spearman correlation instead of the Bravais-Pearson correlation) increases the chances of a significant result (see Table 6). Some forms of *p-hacking* can be tried out here, for example: https://shinyapps.org/apps/p-Hacker/ (Schönbrodt, 2016). Wicherts, Veldkamp, Augusteijn, Bakker, Aert, et al. (2016) propose a checklist for avoiding p-hacking.

Table 6: Probability of obtaining a significant result through the application of various p-hacking techniques, following Simmons et al. (2011), Table 1. The proportion of significant results should correspond here to the fixed 5% alpha error rate.

Technique	**Proportion of significant results**
Multiple dependent variables with a correlation of r = .5 among them	9.5%
Collecting 10 additional observations per group	7.7%
Including an additional variable (e.g., sex) in the model	11.7%
Excluding (or retaining) one of three groups	12.6%
All techniques combined	60.7%

> **Faking data for dummies**
>
> Hussey and Hughes (2018), and building on this, Sarstedt and Adler (2023), self-mockingly proposed methods to make p-hacking even easier. On this website, users can generate random numbers within a desired range: https://mktg.shinyapps.io/extra-p_ointless/.

Selective reporting

When planning a study in the social sciences, the question of how a particular construct should be measured often arises. For intelligence, political opinion, life satisfaction, and many other variables, there is no single test but rather many measurement instruments, some of which are only weakly related to one another. At the same time, the theories being tested are usually vague and do not dictate which measure should be used to assess a construct. Theories are thus often agnostic with respect to measurement methods – or put differently, according to the theory it does not matter how the variable is measured. If a study then chooses different measurement methods for a construct, the theory would need to be confirmed equally across all tests. If that is not the case, the theory should be revised. Contrary to this recommendation, and in order to maximize the chances of publishing their results, researchers often report results *selectively*. Instead of all results, only the "best-fitting" or "most exciting" ones are reported. As has become clear above in the discussion of p-hacking and researchers' degrees of freedom, this leads to relationships being found that do not actually exist. If, for example, three different and independent measures are used to test a hypothesis, the probability of at least one significant result rises from 5% to 14%.

Selective reporting: out of a hundred correlations involving unrelated random numbers, on average five significant ones are to be expected – and this occurs *purely by chance*. None of them, taken together, are significant. In order to improve their chances of publication, and thereby their chances of a permanent position, researchers frequently report only the most exciting part of their results, thereby distorting the overall picture.

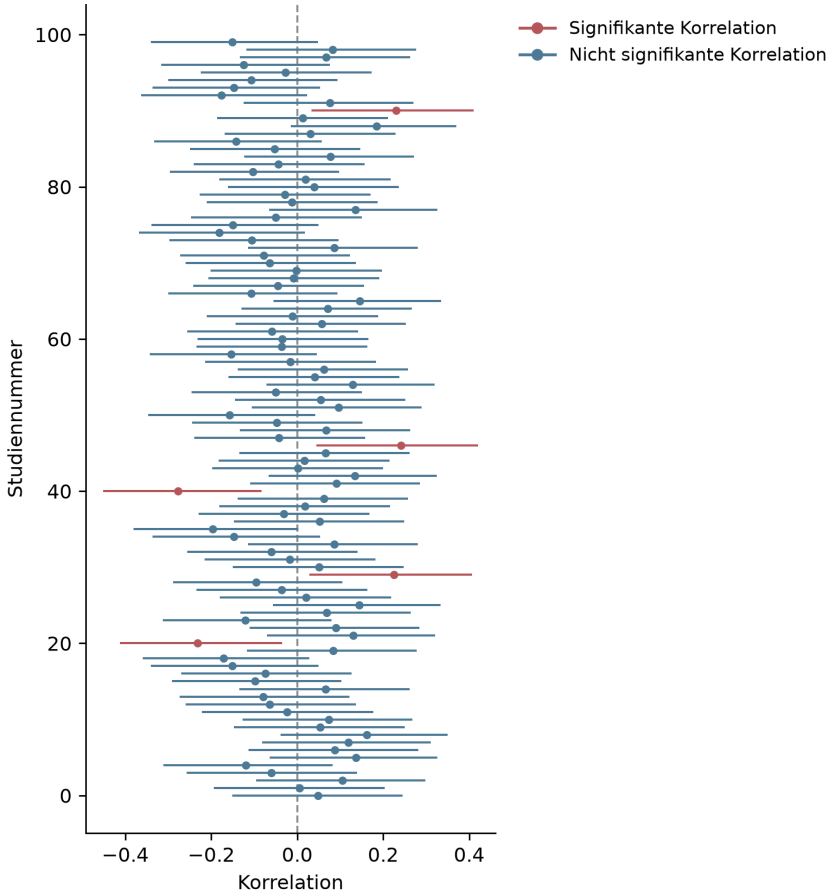

Figure 14: One hundred correlations, five of which are significant by chance.

Optional stopping

If a study's test is rerun after each new observation and the p-value is examined each time, there are two possible trajectories for the p-value: if a relationship actually exists between the variables being studied, the p-value will *converge* – that is, it will approach a particular value, namely 0. The probability of the observed result becomes ever lower as the sample grows larger. A coin landing only on heads is more unusual if it has done so 100 times than if it has done so three times. If no relationship exists, the p-value will not, as is often expected, tend toward 1, but rather *fail to converge*. It will instead behave chaotically, sometimes high and sometimes low – and will also turn out significant more often than expected. Researchers exploit this fact through *optional stopping*: they keep collecting data until their hypothesis is confirmed. Incidentally, this problem does not arise for correlations and other effect size measures. These converge, depending on their magnitude, from around 250 observations onward (Schönbrodt & Perugini, 2013).

Presenting calibrated models as planned models (*overfitting*)

Complex statistical models have many adjustable parameters. It is possible to make all the countless decisions before applying a model to data, but typically other calibrations are tried out, and one that was not originally planned turns out to fit better. This means, for example, that certain variables are included in a model in order to maximize its predictive power. Many models even involve various algorithms that decide, based on fixed rules, what the model should look like. A model is thus fitted

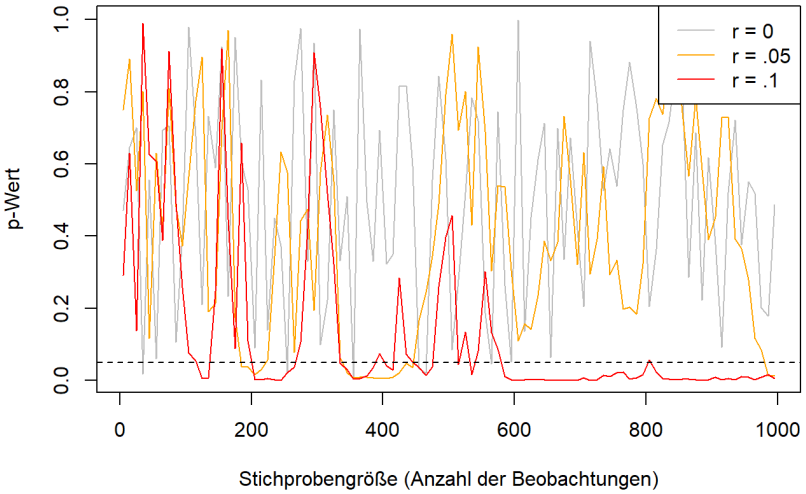

Figure 15: Convergence of p-values and effect sizes depending on effect size: effect sizes (here: correlation coefficients) converge with large samples; p-values only converge if the correlation is not 0.

to a data pattern. If this procedure is disclosed transparently, that is perfectly fine. Problems arise when the best model found is presented as though it had been the planned model. In social science research, the pattern present in the data almost always also contains noise – that is, fluctuations attributable to measurement imprecision or other unknown influences. These influences fluctuate by definition (in psychological test theory, for instance, this is referred to as *error*, an unsystematic fluctuation that averages out with repeated measurement). In future investigations, a model fitted to past data and the noise it contains will necessarily perform worse, because the noise in the new data is different. This is referred to as an "overfitted" model, or overfitting.

People's tendency to confirm themselves (*confirmation bias*)

A particular problem in scientific methods is confirmation bias. The phenomenon is not clearly defined in the scientific literature (Nickerson, 1998); here, I mean by it the tendency of people (or, in this context, researchers) to find the patterns they expect to find. Confirmation bias is itself based on scientific findings (Oswald & Grosjean, 2004) and has been applied by researchers to themselves (Mynatt et al., 1977; Yu et al., 2014). These considerations lead close to logical absurdities and paradoxes – self-mockingly, Nickerson (1998) himself notes the possibility that all findings on confirmation bias could themselves merely be products of the same bias, which would in turn confirm the existence of confirmation bias (p. 211). In practice, there is a risk that researchers do not uncover truths but instead twist everything so that their preconceptions are confirmed. Ludwik Fleck (1935/2015), in his sociology of science – which forms

the foundation for Thomas Kuhn's work on scientific revolutions (Kuhn, 1970/1996) – goes a few steps further still: he argues for a model of scientific progress in which the goal is not to get closer to the truth, but to understand problems, to the best of one's knowledge, against their social background. This does not mean that there is no truth, only that truth is not simply the correspondence of statements with facts. In place of this *correspondence theory of truth*, often held by researchers, Fleck offers a *consensus theory* of truth: the agreement of many people is what matters. Scientific facts are not "discovered" by individuals but created by a collective. Confirmation bias then manifests itself in the way findings that contradict the consensus are screened out, and current views are clung to for as long as possible. Although philosophical theories of truth go beyond the scope of this book, it should be noted that none of the three theories of truth (correspondence, consensus, and coherence) is tenable (Albert, 2010, *Münchhausen trilemma*).

Data fabrication

The practices discussed so far are often described as questionable. Some researchers consider this a euphemism, since as a researcher one ought to know enough to recognize that the techniques described above are *not* scientific and clearly do not serve the generation of knowledge. They clearly hinder scientific progress, endanger trust in science, and lead to enormously high costs. Unfortunately, many researchers today are still unaware of these problems. "That's just how we were taught, and how it's always been done," people say. That certain studies could not be replicated was, in some cases, already known to many people – they simply did not think it possible to record this in the scientific record. In any case, the

term “questionable research practices” suggests that researchers engaging in them are operating in a gray area. In my view, this is only the case because, if researchers lost their jobs for having engaged in p-hacking, not many researchers would be left.

The situation is different when it comes to fabricating and manipulating data. How often data manipulation or fabrication occurs is uncertain, and estimates are difficult to make. A meta-analysis of surveys on this topic estimated that between 0.86 and 4.45% of all researchers admitted to having manipulated data. 72% reported having engaged in questionable research practices (Fanelli, 2009). Stroebe et al. (2012) later compiled examples of data fabrication and recommended peer review and replication as fraud detectors. A more recent and extremely extensive study by Gopalakrishna, Wicherts, et al. (2021) reported that 8.3% of all respondents had manipulated or fabricated data and 51.3% had engaged in questionable research practices (Table 2), confirming the scale of the problem. Depending on the discipline, further problems arise, such as the reuse of previously published biomedical images, which was found in approximately 3.8% of all published articles (Bik et al., 2016). While it was long assumed that fraud occurs only in very rare cases, the real problem is above all that it is rarely uncovered. How often research is “fake” has hardly been studied, and estimates vary widely (Heathers, 2024). Cases of fraud that have come to light have resulted in the retraction of the respective scientific articles and often in consequences for the careers of those responsible. Retractionwatch.org maintains the world’s largest database of retracted articles (as of December 2023: 49,628 articles): http://retractiondatabase.org/.

It is a rather bleak fact that methods for fabricating data are, on the one hand, becoming ever easier (Naddaf, 2023), while researchers who expose

errors are occasionally sued. For example, the authors of Datacolada.org, who have already exposed problems on multiple occasions, were sued by Francesca Gino over one of their publications (https://datacolada.org/109), in response to which thousands of researchers raised funds to support the financial cost of the legal proceedings (https://www.gofundme.com/f/uhbka-support-data-coladas-legal-defense).

Further information

- This podcast discusses how, and whether, fraud in science can be stopped: https://freakonomics.com/podcast/can-academic-fraud-be-stopped/.
- Daniel Lakens describes what the replication crisis felt like from the perspective of a young researcher: http://daniellakens.blogspot.com/2020/11/why-i-care-about-replication-studies.html.
- Nagy et al. (2024) systematically catalog questionable research practices (QRPs) in their *Bestiary*.

Theories

Scholarship works with theories. What these look like exactly differs considerably between disciplines. While the natural sciences often work with mathematical models, that is, formulas that describe the relationship between variables explicitly and unambiguously and allow predictions, the social sciences often work with verbal theories in the style of "X and Y are positively related" or "the higher X, the higher Y," and the traditional humanities work, for example, with verbal explanations. Verbal theories have the advantage of tending to be easy to understand and broadly applicable, but the terms they use are often subject to individual, cultural, or temporal influences, and discussants risk talking past one another in scholarly discourse.

For formal theories, all variables involved are precisely defined, and such theories often have a strongly restricted scope of validity (e.g., many physical laws hold only under tightly controlled conditions, such as in a vacuum, at a specific temperature, and so on). A concern raised in the context of the replication crisis is that theories are not clear enough to predict when replications will succeed, and that this is one of the causes of low replication rates (Buzbas & Devezer, 2023a; P. Smaldino, 2019). A theory about the consequences of identifying with gender roles, for example, must account for changes in gender roles and their particularities across different countries. It is hardly surprising that one and the same experiment on this topic yields different results in the USA in 1980 than in Germany in 2020. What is problematic, however, is that – even though such qualifications seem sensible and necessary for many social-science theories – statements to this effect are rarely made.

Verbal theories are not inherently less scientific: within their respective fields, scientific theories always stand out from everyday explanations through their particularly high degree of systematicity (Hoyningen-Huene & Kincaid, 2023). However, fields that place value on predicting events cannot do without formal theories (Muthukrishna & Henrich, 2019). It should be emphasized that certain disciplines place *no value* on prediction (e.g., history, or fields that proceed primarily hermeneutically). Fields such as psychology, quantitative sociology, and parts of the humanities ("digital humanities") are currently moving closer to formal models – in social psychology, there was already a call to formalize theories once before, during a crisis in the 1960s (Lakens, 2023). Because theories lacking objectivity are rarely used by different researchers and, due to their flexible interpretation, are difficult to refute, an enormous quantity of useless theories has emerged there (C. J. Ferguson & Heene, 2012). Among these are also mutually contradictory theories: for instance, Banker et al. (2017) argued that "ego depletion," that is, the depletion of self-control resources, causes people to rely more on cues from other people (p. 2), whereas Francis et al. (2018) conjectured the opposite – that depletion prevents cues from being processed at all. Both provided data supporting their respective theories, yet a follow-up investigation found that both were probably wrong (Röseler, Schütz, et al., 2020).

Robinaugh et al. (2021) discuss examples of the conversion of verbal theories into formal ones. This process results in new, more specific predictions that can be derived. When a theory makes more precise predictions and the set of possible events that would contradict the theory grows, this represents an increase in *empirical content* (Glöckner & Betsch, 2011; Popper, 1959/2008).

Empirical Content and Strong Inference

Theories can differ in their empirical content. Concretely, this refers to how specific their predictions are. The more *possible* observations *would* refute a theory, the higher its empirical content.

Let us take the case where our theory allows us to make predictions about what kind of car will drive along a particular street at a particular time. The figure shows all possible cars. For simplicity, our example world contains only nine different cars, which differ in terms of the features *color* (green, black, blue), *rear spoiler* (with, without), and *wheel color* (gray, yellow).

- The purple theory states: *The observed car has gray wheels*. Without a theory, all cars would be equally likely to us; the purple theory "forbids" the car from having yellow wheels. It rules out 3/9 of the cars.
- The red theory states: *The observed car is blue*. The probability of refuting it would be higher in our sample world, namely 6/9. Because the red theory is, so to speak, a riskier bet *a priori* – that is, without further prior knowledge – it has higher empirical content.
- The orange theory has the highest possible empirical content: *The observed car is green, has no rear spoiler, and has gray wheels*. It rules out all but one case (8/9).

The example with the nine possible car types is, of course, highly simplified. In certain fields, however, researchers occasionally manage to reduce the results of experiments to a few possible outcomes and thereby weigh theories against one another. Platt (1964)

calls this the *method of strong inference* and argues that fields proceeding this way experience rapid progress. Building on this, P. E. Smaldino (2017) calls for more theories or models and argues that researchers should always offer several explanations simultaneously. This can have the advantage that researchers do not commit to a single possibility and that theories are not treated as someone's property. As long as a theory can be clearly attributed to one person, there is a risk that criticism of the theory will be confused with criticism of the person.

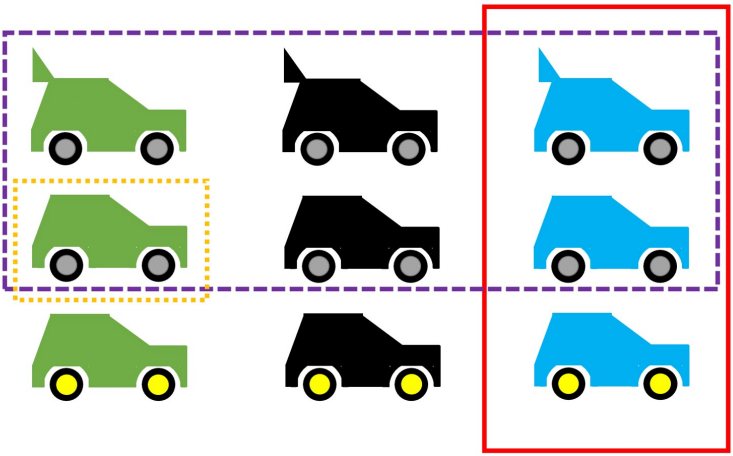

Figure 16: Visualization of theories with different empirical content.

Deduction and Induction

Methods are being reformed, and scientists discuss how science works, how it should proceed, and which methods are sensible or nonsensical. As becomes clear from the hermeneutic circle, one path to knowledge consists of combining a set of observations into a regularity or law (*induction*),

while another consists of deriving predictions about observations not yet made from a law or theory (*deduction*). This distinction is repeatedly neglected or obscured in scientific discourse. For example, a debate in consumer psychology revolved for years around which path was better, even though both paths are equally legitimate and complement one another (Calder et al., 1981). Something similar applies to conflicts between qualitative and quantitative approaches, which, formally considered, tend to proceed inductively or deductively, respectively (Borgstede & Scholz, 2021). In replication research, the inductive side has traditionally received more attention (Hüffmeier et al., 2016; Yamashita & Neiriz, 2024): every difference between a replication study and the original study is cited as a possible cause for the failure of the replication attempt, in order to preserve the trustworthiness of the original findings (R. F. Baumeister & Vohs, 2016). This overlooks the fact that minor differences between the original and replication study (e.g., the measures used, the average age of participants, the language of the instructions) are not captured by theories – and should therefore, according to the theories themselves, be irrelevant – and that a failed replication clearly reveals the limits of the theory, from which recommendations for modifying the theory can be derived (Cesario, 2014; Dijksterhuis, 2014). An overview of the approaches can be found in the following table.

Table 7: Characteristics of inductive and deductive approaches; taken and adapted from an unpublished manuscript by Röseler & Leder.

Facet	**Deductive Approach (Theory-driven)**	**Inductive Approach (Phenomenon-driven)**

Generalizability lies in…	the theory: it is maximally general a priori (e.g., it holds for all people until demonstrated otherwise).	the data: only diverse observations across different contexts allow the assumption that the phenomenon is universally valid.
Change in generalizability	Generality decreases with more observations.	Generality increases with more observations (provided they are confirmatory in nature).
Type of test	Predictions of the theory are primarily subjected to attempts at *refutation*.	Repeated observations *confirm* the original individual case.
Choice of study setting	Student samples from a single country or laboratory studies are unproblematic.	The context of the study should reflect the target conditions (e.g., when applying the findings in practice) as closely as possible.

Auxiliary Hypotheses

Replication failures can be explained via the following paths:

1. Type I error in the original study: The original finding was merely a chance finding or arose through scientific misconduct (see the chapter "Researchers' Degrees of Freedom").
2. Type II error in the replication study: The original study was correct; the replication study made an error (e.g., too small a sample, poor calibration of instruments, or scientific misconduct).

3. Boundary condition of the phenomenon: Both studies are trustworthy. The replication study *differs* in a way that matters for the theory (e.g., the replication study was conducted with people from a different country, and the theory only holds for people from the "original country").

Option 3 is constructive and accepts both individual findings as robust. This requires a theoretically relevant difference between the original and replication study, which, given the infinite number of possible important factors, applies in most cases (Smedslund, 2015). This path can then be used to modify the theory or to formulate an additional theory that must likewise be taken into account for the context of the study. Things become difficult when researchers conduct a replication to the best of their knowledge, it "fails" (that is, it does not demonstrate what it was meant to demonstrate), and other researchers criticize the replication for having done something "wrong." After Hagger et al. (2016), in consultation with Roy Baumeister, tested his ego depletion theory with a large-scale study, R. F. Baumeister & Vohs (2016) criticized that it should have been expected from the outset that the study would not work, and described the study as misguided. Vohs, who was a co-author of the critique, conducted another large-scale replication study a few years later. Although this time the researchers were able to follow their own advice, they again failed to find the expected effect (Vohs et al., 2021).

Further Information

- Ramminger (2023) discusses a philosophical perspective on the relationship between theory, measurement, and replication.

- Yarkoni (2019) argues that replication problems originate in the generalization of results to theories.
- In a talk, Fanelli discusses the complexity of research as a reason for replication failures and proposes a theory for measuring complexity (Fanelli, Tan, et al., 2022). A video of a talk is available online: https://www.youtube.com/watch?v=CEAV7420jBk.

Epistemic Problems

Epistemology is the theory of knowledge and a subfield of philosophy. It examines how science works, how knowledge can be produced, and what the difference is between knowledge and belief. Epistemological explanations for replication failures go beyond the problems arising from the scientific system or scientific methods, but they are also subject to different standards of testability. Whereas approaches from the philosophy of science sometimes allow for predictions that can be tested empirically (that is, through observation), the philosophical problems themselves can only be thought through and discussed.

Robustness and Historicity of Phenomena

Under what conditions is it unsurprising that replication attempts fail? One possibility is to assume that the phenomena under investigation are extremely sensitive or unstable. Discovering regularities in human behavior analogous to planetary motion may simply not be possible (Smedslund, 2015). The current assumption held by many social scientists is that regularities exist that do not apply to every single person without exception but hold "on average." In this regard, Lewin (1930) distinguishes between *Aristotelian* and *Galilean* lawfulness, with *Aristotelian* regularities being the ones predominantly studied in the social sciences. A regularity or natural law in the Aristotelian sense might be, for example, that men are taller than women.

An even more extreme theory holds that people are aware of the knowledge that exists about them and dynamically adjust their behavior accordingly,

meaning that the behavioral sciences are always historical, or time-bound: if it is discovered that people tend not to change their decisions even when doing so would improve their situation, this fact is brought to their attention through science, and they can then adjust their behavior. This, incidentally, is the status quo bias, whereby people prefer the current situation over an alternative one (Samuelson & Zeckhauser, 1988; Xiao et al., 2021). Another example is the gender pay gap, that is, salary differences between men and women independent of qualification. Ideally, once people become aware of the problem, they change the system so that the pay gap no longer exists. This perspective cannot be applied to all areas of the social sciences, however. For instance, knowing about one's own pain sensitivity, intelligence, or personality has no influence on that very trait.

How strongly phenomena can differ due to seemingly minor differences in study design has already been examined at the meta-scientific level. Landy et al. (2020) had several hypotheses tested by several different researchers, and factors that, according to the underlying theories, should have made no difference at all led to opposite results. Applied to replication research, this means that in certain research areas it may be entirely unclear under which conditions particular relationships can be observed. An even more extreme example is a study by Breznau et al. (2022). The researchers had 73 teams use the same data to test the hypothesis that "immigration reduces public support for social policy measures." The findings ranged from large negative to large positive associations and could not be explained by methodological choices. They recommend exercising modesty when researching complex human behavior.

Further Reading

- Fleck (1935/2015) proposes a theory of scientific progress in which knowledge is always subject to a so-called thought style that is optimal for the given social situation. In doing so, he draws on elements of evolutionary theory, sociology, and psychology.
- Shiffrin et al. (2018) discusses the apparent paradox between fallibility – that is, the fact that science is not always right – and scientific progress.

Solutions

Let us take a look at what has changed across the academic disciplines since 2012. Many of the developments discussed originate from psychology. The changes are categorized according to which part of the problem they primarily address: is the purpose of a new requirement to improve the system, the methodology, or the theories of research? Some solutions are linked to many problems simultaneously, while others are tailored to specific issues. The figure provides a rough breakdown.

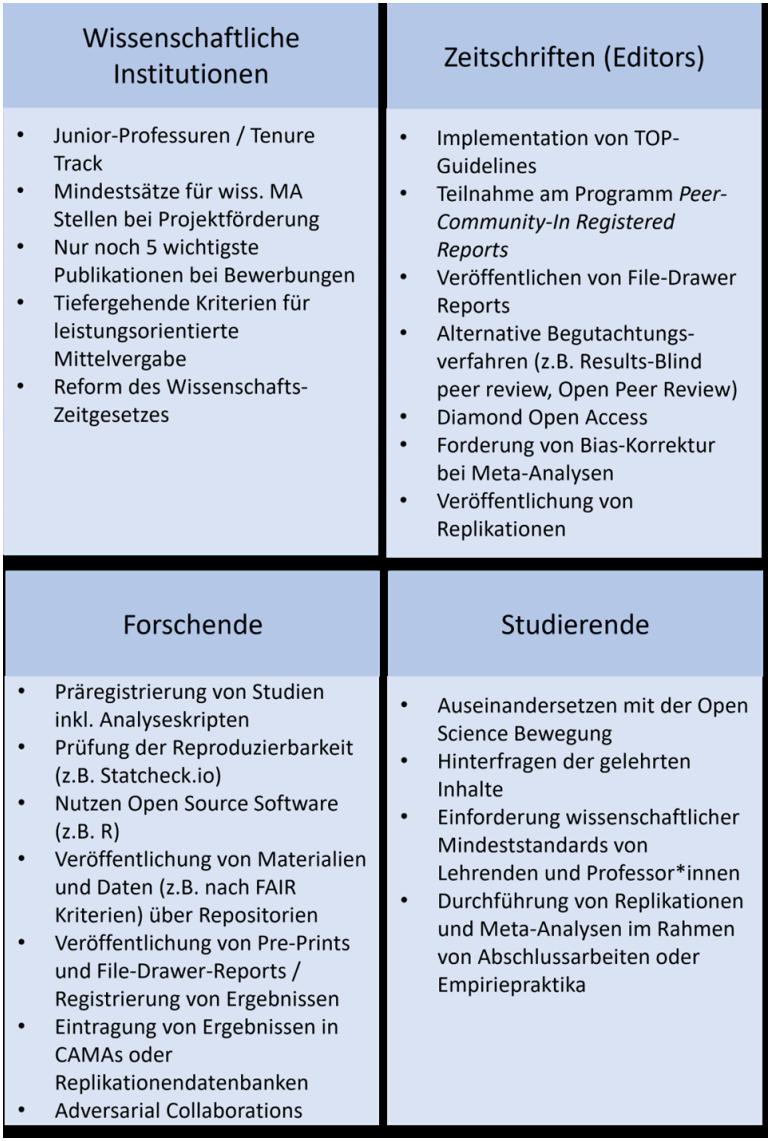

Figure 17: Proposed solutions and the problems in research they address.

The System

Approaches that aim to change the system hold the greatest potential, because in order to earn a living within the system, researchers have to play by its rules. And as long as publications are the currency, and papers with catchy titles and clear-cut results are judged to be of higher quality, researchers are motivated to search for catchy titles and clear-cut results rather than for the truth.

Overall, a positive development is visible (Korbmacher et al., 2023), and a shift in the incentive structure is being pursued. This shift can be understood as an alignment of the scientific system with the *Mertonian norms* (named after Robert Merton) (Merton, 1973): (1) Communalism: scientific knowledge should belong equally to all researchers, in order to foster collaboration. (2) Universalism: scientific merit is independent of the socio-political status and personal attributes of those involved. (3) Disinterestedness: scientific institutions act in the interest of science and not for personal gain. (4) Organized skepticism: scientific claims should be subjected to critical scrutiny before they are accepted.

In a blog post, Nosek recommends a framework of measures by which the desired changes should be made, one after another …

1. possible (e.g., through infrastructure such as online repositories where research materials can be uploaded publicly and free of charge),
2. easy (e.g., through low-barrier offerings, multilingual instructions),
3. normative (e.g., through scientific communities that jointly stand behind calls for improvement),
4. rewarded (e.g., through designated awards), and

5. required (e.g., through minimum standards demanded by journals or funders)

… in that order. How the various approaches look in practice for the different actors — politics, universities, or journals — is discussed below.

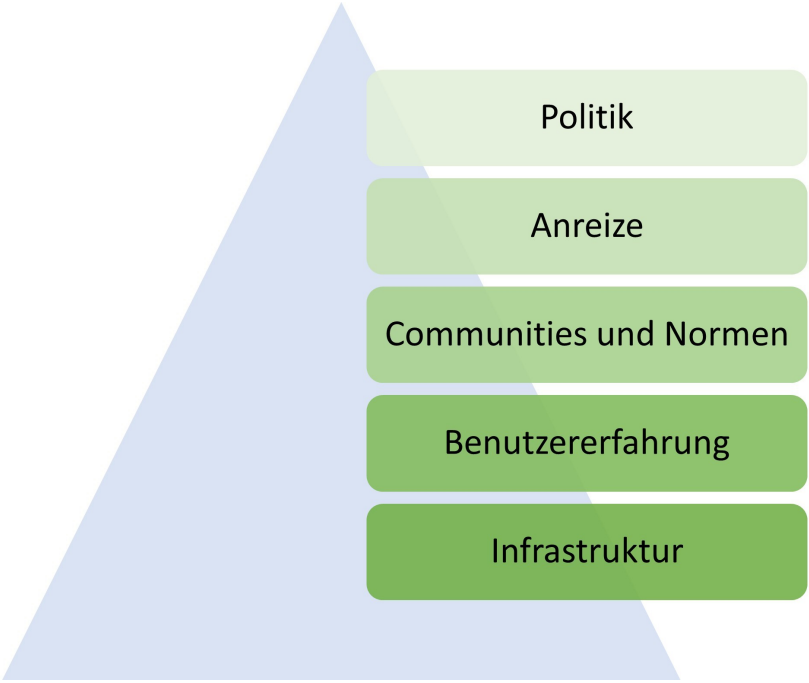

Figure 18: Culture change in science, according to Nosek (https://www.cos.io/blog/strategy-for-culture-change).

Politics

Internationally, political parties and associations clearly stand behind open science and open access. For example, UNESCO recommends universal access to scientific knowledge regardless of country of origin, gender role, political borders, ethnicity, or economic or technological

barriers (UNESCO, 2020, p. 3). Working groups for policy instruments, funding, and infrastructure have been established accordingly. The G7 advocates for scientific integrity, academic freedom, and open science. Open access as well as transparency of scientific procedures are likewise called for by the European Union. Infrastructure (e.g., the European Open Science Cloud) and various open science research projects are being funded specifically. In addition, the free publication and peer-review platform Open Research Europe is available for all research projects funded by the EU.

In Germany, the government of the 2021–2025 term had set out, in its coalition agreement, to "establish open access … as a common standard" (my translation from the German). Individual federal states such as North Rhine-Westphalia have, through alliances of their respective universities, developed open *access* strategies (DH.NRW | AG Openness, 2023) and are currently working on open *science* strategies. Other countries, such as Sweden, have already developed national guidelines on open science. Regarding the problem of abuse of power, the issue is acknowledged, for example, in a position paper by the state of North Rhine-Westphalia, but it is understood as an isolated case rather than a systemic problem (commentary on this).

Universities

Which universities are doing something? (Examples)

The topic of open science has already found resonance at many universities. While most German universities have (time-limited) funds for open access publications, some have gone further and established open science policies (e.g., FAU Erlangen), open science centers (e.g., LMU Open Science Center; Cologne Open Science Center; Münster Center for Open Science; Mannheim Open Science Office). In addition, the Leibniz Information Centre for Economics in Kiel and the Leibniz Institute for Psychology support replication research, for example with a replication journal (https://www.jcr-econ.org) or through a *Juniorprofessur* (a fixed-term professorship, roughly comparable to a tenure-track assistant professorship) in psychological metascience. At TU Dortmund, a needs assessment on research data management was conducted (Kletke et al., 2024). One of the largest centers for metascience in Europe has formed in Tilburg, the Netherlands. Germany's 15 largest universities (U15) have come out in favor of transparent and fair evaluation of research performance and have signed CoARA, and an alliance of many large scientific societies and communities (e.g., the DFG and the Fraunhofer Society, Germany's largest applied-research organisation) has likewise spoken out in favor of traceable research assessment. The Berlin universities have developed a joint open access declaration, the Berlin Declaration, and in France the University of Sorbonne is a pioneer: together with the University of Amsterdam and University College London, it signed a declaration on the publication of research data. Since 2024, the University of Sorbonne has also termi-

nated its contract with Clarivate for the use of the "Web of Science" research database and has since been working with the open source software "OpenAlex" (Priem et al., 2022a).
Using the Road2Openness tool (https://road2openness.de/tool/), institutions can conduct a self-assessment across various open science topic areas and receive a report.

Universities bear a particular responsibility for the long-term development of scholarship, since they employ researchers and make the selection decisions for the few permanent positions in academia. If, for decades, professorships are awarded on the basis of subjective, non-reproducible criteria that are secondary to good scholarship (e.g., number of publications in journals), this can have a negative effect on the development of scholarship. Addressing this problem, a research award from the Berlin Institute of Health (BIH), given annually for projects that promote scientific integrity, was awarded to a project developing objective and meaningful selection criteria for professorships (Gärtner et al., 2022; Schönbrodt, Gärtner, Frank, Gollwitzer, Ihle, Mischkowski, Phan, Schmitt, Scheel, & Schubert, 2022). To reduce the role of quantitative indicators, some universities and DFG grant applications already apply the "N-best" or, commonly, "5-best" rule (Frank, 2019): only the five best research articles may be listed in an application, and evaluation may be based only on these.

Within universities, libraries also play an educational role regarding research data management and publication culture (Schmidt et al., 2024). Through them, the research process can be supported with appropriate infrastructure (e.g., for storing and publishing research materials and results) (Quan, 2021), preventing a "dependency on a few commercial

providers" that "limits what is achievable in research in terms of working methods and questions" (Siems, 2024). Because of the high value placed on "freedom of research," universities have hardly any means of obligating their members to comply with open science strategies. This could put them at risk of becoming less attractive to researchers: if, for example, prominent journals can no longer be accessed through the university because contracts with closed-access journals have been terminated, this makes researchers' work harder. For instance, the law faculty of the University of Konstanz successfully challenged the university's attempt to make use of a statutory secondary publication right (*Zweitveröffentlichungsrecht*, § 38(4) of the German Copyright Act) compulsory — unlike in most countries, where self-archiving depends on the publisher contract, researchers in Germany hold a statutory right allowing them to also publish their own work themselves once it has appeared in a journal (e.g., via their own websites), but in Konstanz they could not be obligated to make use of it.

Another lever available to universities is the subsidizing of publication costs at journals. If, for example, the scientific quality of a journal is called into question, a university can stop this subsidy.

An often neglected role also falls to university teaching. Because of the freedom of research and teaching, and because degree programs are typically planned well in advance, integrating new topics such as open science is difficult. Researchers for whom the topic plays a role in their teaching proactively share their materials, jointly develop curricula, and network in large international initiatives such as the Framework for Open and Reproducible Research Training (FORRT.org) (C. R. Pennington & Pownall (2024), for example, have published concrete proposals for integrating open science into teaching).

Institutes and associations

Scientific fields live above all through communities — that is, through everyone doing research in that area. They organize themselves into societies (e.g., the German Psychological Society), interest groups, or similar communities. In Germany, the German Research Foundation (DFG) holds a special position, distributing several billion euros in research funding, provided jointly by the federal government and the sixteen federal states (*Bund und Länder*). As one of the most important national institutions, its stance on open science carries considerable weight (Deutsche Forschungsgemeinschaft, 2022a). Experience shows, however, that change tends not to originate from the DFG itself; rather, the DFG waits for impulses from the individual disciplines. In psychology, ZPID (the Leibniz Institute for Psychology, Germany's national psychology infrastructure centre) also promotes infrastructure through journals, preprint servers, and other means, and in economics, the ZBW (Leibniz Information Centre for Economics) manages important information and journals. Interdisciplinary organizations such as CERN and international actors such as the American Psychological Association also commit to openness and transparency.

Many new associations have also emerged as part of the open science reform. FORRT (Azevedo et al., 2019), an interdisciplinary initiative led largely by early-career researchers, advocates for embedding open science in teaching. The Society for the Improvement of Psychological Science (SIPS) is devoted to improving psychological research. So-called "grassroots" initiatives (that is, movements driven by young researchers) have emerged at numerous universities and joined together into networks such as the Network of Open Science

Initiatives (NOSI) and "Reproducibility Networks" such as the German network *GRN* (https://reproducibilitynetwork.de), the UK's *UKRN* (https://www.ukrn.org), and others. There are also alliances of professors aiming to change short-term contracts (e.g., the Network for Sustainable Science, https://netzwerk-nachhaltige-wissenschaft.de).

Journals

Scientific journals are seen as the stage of scholarly discourse and largely determine which elements of the research process become part of the "scientific record" and are thus deemed relevant. They are also responsible, as organizers of the peer-review process, for quality assurance in science. In response to the lack of quality, there are already detailed recommendations for how journals should be designed; new journals are being founded; tools for automated checking, such as the Problematic Paper Screener, are being developed; and entirely new review and publication models are being proposed and widely implemented. The Journal of Open Source Software, for example, is built on publicly viewable program code, and its infrastructure can easily be copied and adapted for other journals.

Recommendations

Editors who want to engage with open science practices for the journal they manage can now draw on a comprehensive guide (Silverstein et al., 2023). Proposals were collected via a discussion platform (Journal Editors Discussion Interface, JEDI), which explains what things like Registered Reports, open peer review, diversification, and open access actually are and how they can be implemented at a journal. Potential wor-

ries and fears are addressed and answered. The Committee on Publication Ethics (COPE) likewise engages in education and training to help editors, universities, and research institutions deal with problems in the publication system. It offers, for example, guidelines on the circumstances under which publications should be retracted or corrected, and on what ethical standards a review process should meet. Editors who manage journals for commercial publishers and want to switch to systems entirely in the hands of researchers can get help migrating from commercial to open, free systems through university libraries, and can manage journals, for instance, with the Open Journal System (see, e.g., the OJS network). The Directory of Open Access Journals (DOAJ) maintains a database of more than 20,000 open-access journals. Reviewers of research articles can access training materials and guides through the Reviewer Zero Initiative (https://www.reviewerzero.net, https://osf.io/e7z5k/wiki/Resources/).

Scholar-led publishing

An alternative to publishing research in traditional, commercial journals is offered by Diamond Open Access journals, which are run by researchers themselves. Diamond open access specifically means that peer-reviewed articles can be published and read free of charge (Armengou et al., 2024). Six practical guidelines for how to go about this are summarized in Wrzesinski (2023). This development goes hand in hand with the cancellation of contracts between university libraries and publishers, such as MIT's boycott of Elsevier journals, which saves roughly 2 million US dollars annually. A discussion paper by the Leopoldina, Germany's National Academy of Sciences, further proposes that the rights to journals should remain

> with researchers or scientific societies, and that journals should be financed in the long term through public funds.

Highlighting open science practices

It is known from psychology that motivation works just as well through reward as through punishment (Balliet et al., 2011). In road traffic, where until fairly recently there were only penalties for breaking the rules, this insight is reflected, for example, in speed displays that give drivers a cheerful smiley face for keeping to the speed limit. Scientific journals highlight compliance with recommendations (e.g., publicly available datasets) with badges. Since 2024, the journal *Psychological Science* has gone so far as to abolish badges again, because open science is now the standard for all articles there. The website topfactor.org lists journals and their compliance with various standards in a ranking. Finally, every reward system carries the problem that actors orient their behavior toward the rewards and, in doing so, try to take shortcuts (Klonsky, 2024).

Review systems

Science is characterized by systematicity (Hoyningen-Huene, 2013). Arguably the most systematic way to guarantee scientific quality assurance would be a study that compares different systems. While current research is attempting something similar (Soderberg et al., 2021), alternative review systems are currently being tried out. As a reminder: researchers write articles, which they submit to journals. There, a person (the editor) is responsible for ensuring that the article, provided it fits the journal, is sent out to reviewers.

To check whether judgments in the review process agree between reviewers, Etzel et al. (2024) surveyed various reviewers on classic criteria. They found that this was not the case and propose criteria that have a clear value and can be assessed reliably (e.g., whether data are publicly available).

Open peer review

Ever since researchers began engaging critically with the review system, there has also been discussion of what happens to the reviews themselves: traditionally, they remain confidential. The journal publishes the final article, and all previous versions are known only to the authors, editors, and reviewers. These reviewers, moreover, usually remain anonymous, meaning that if they did not put in much effort, it will probably never be noticed. Furthermore, researchers have no incentive to write reviews — at most they receive proof that they reviewed a manuscript for a journal. Some journals have since introduced *open peer review*. The name only half fits, because reviews are published only if the article is accepted by the journal. This carries the risk that serious problems — the kind normally required for a rejection — never come to light. Researchers can then submit the unrevised article to another journal, where the problems may go undetected. This practice leads to duplicated work and enormous costs (Aczel et al., 2021). Zoltan Kekecs remarked, during a discussion at a conference, that even casinos that compete with one another exchange lists of cheaters. Journal editors do not yet do this. A model in which all evaluations are published is offered by Unjournal (https://unjournal.pubpub.org).

Table 8: Degrees of openness of peer review reports and their consequences.

Reviews remain confidential	**Reviews are published upon publication**	**Reviews are published upon both publication and rejection**
No proof of quality control is possible.	Rejected articles can be submitted to other journals without revision, causing extra work.	Quality control is traceable and transparent.

More contentious than the publication of reviews is the anonymity of reviewers: the standard is that reviews are anonymous but can be signed if desired. This protects doctoral students who give a poor assessment of an article by a potential future boss. Even professors can risk, when giving a negative assessment, that the colleague in question will later review one of their grant applications and retaliate for the criticism. On the other hand, anonymity can result in criticism being directed at the person rather than being constructive.

Another way peer review can be open is that anyone can take part in it. This is possible, for example, at Meta-Psychology. The problem here, however, is the relatively low participation. At the journals Synlett and ASAPbio, groups are coordinated in a way that resembles the *journal clubs* already used to discuss exciting articles — just under the name "Pre-Print Review Club."

Review of preprints

What is a preprint?

A preprint is a manuscript *at* or *before* the stage of submission to a journal. It may not yet have been reviewed, or it may have been rejected by a journal after review. The article has been published on a website, is citable (e.g., assigned a DOI), and can be downloaded free of charge. In some cases it may make sense — out of concern about censorship or because of the long duration of the review process — *not* to subject a scholarly contribution to review *before publication* (e.g., for commentaries, position pieces, or public exchange). Typically, peer-reviewed contributions count for more in an academic career.

The purpose of preprints is manifold: they increase the availability of knowledge, allow faster publication (e.g., mathematical proofs must be laboriously checked over the course of up to several years), and prevent other researchers from beating someone to an innovative idea. In medicine and the social sciences, they were used extensively during the COVID-19 pandemic because of the faster exchange they enabled (Fraser et al., 2020). In epidemiology, it could not be shown that preprints are of better or worse quality (L. Nelson et al., 2022). When choosing a preprint server, researchers should make sure to use providers that are in scientific hands and run open source code (e.g., arxiv.org or osf.io/preprints), since servers run by commercial publishers (e.g., preprints.org, run by the controversial publisher *MDPI*) can be misused to track researchers' activity or to promote their own journals. Long-term availability is also relevant. Preprints

published, for example, via the Open Science Framework (osf.io) are guaranteed to remain available for at least the next 50 years.
The term preprint ("before printing") comes from the era when scientific journals appeared exclusively in print and through university libraries. "Printing" referred to the reproduction and printing carried out by the publisher. At that time, it also made sense for researchers to hand over reproduction rights to publishers, since they themselves would not have been able to print and distribute articles in that way. In the internet age, the reproduction role of journals has disappeared.

Preprints are reviewed through platforms and communities such as *PCI*, *f1000research*, or MetaROR (https://researchonresearch.org/project/metaror/). They are then submitted not to a journal but to the relevant organization, and reviewed there. This model is called *publish-review-curate (PRC)*: preprints are uploaded first, then reviewed, and finally grouped into thematic collections. Quality assurance is organized by researchers themselves and is independent of commercial publishers. It is recommended that students be actively involved in this process (Holford et al., 2024). The model at *Peer Community In (PCI)* is similar: researchers whose preprints receive a positive review through the process organized by PCI can publish their articles published, without further review, at one of the participating ("PCI-friendly") journals. The publish-review-curate model is thus split across preprint servers (publish), PCI (review), and traditional journals (curate). F1000research.com functions like a journal in which articles are directly accessible and change over time through peer review. Because non-reviewed articles are also published there, there are conflicts over whether such articles are indexed in databases used for research assessment. Similarly, the *Zeitschrift für digitale*

Geisteswissenschaften uses a peer-review traffic light: after submission, articles are immediately available there, and a traffic light indicates whether they are under review and, if reviewed, whether there are (still) serious problems or not.

Memory of reviews

Another way to permanently attach reviews to research articles is through platforms that allow quick commenting. On Pubpeer.com or hypothes.is, for example, any scientific contribution (including, e.g., data) can be commented on publicly and, if desired, anonymously. This is possible for preprints too, so that these comments remain permanently linked to them. Browser plugins then flag articles for which there are discussions on Pubpeer. Other platforms include alphaxiv.org, Disqus, and scirev.org.

Attention to the topic in existing journals

Requirements placed on scientific articles by journals underwent major changes starting in 2010. In the social sciences, numerous journals now follow the "Transparency and Openness Promotion" (TOP) guidelines and receive corresponding TOP factors (https://topfactor.org/summary, Grant et al., n.d.). These record the degree of openness and transparency required for research data and materials, and whether the journal in question publishes replications. New editors at existing journals have made major changes. For example, Hardwicke & Vazire (2023) announced, for the journal *Psychological Science*, a standard recalculation of all reported results starting in 2024, by working together with the Institute for Replication (https://i4replication.org). Some journals have published

special issues focused on replication studies or the reproducibility of results (Carriquiry et al., 2023).

Journals for "non-innovative" work

Because of the selection of exciting results, there is no platform for research that is not groundbreaking yet still highly relevant. Estimates suggest that up to 40% of all studies conducted remain unpublished within four years of completion (Ensinck & Lakens, 2023). Other researchers cannot learn from this work, and the resources that went into collecting and analyzing the data, the time of participants, and the lengthy preparation of the research all end up wasted. New journals and formats have emerged to address this problem. In economics, following calls to action (Zimmermann, 2015), a journal for replications and comments has been established. The journal Meta-Psychology offers a format for replications and one for "file-drawer reports." The latter is intended for studies that would otherwise end up in the file drawer because of unsurprising results or errors in execution, but that nonetheless contain important information. To capture the failure that is typical of the research process, the Journal of Trial and Error was also founded, and reports on reproducibility and replications can be published at ReScience C and ReScience X.

Publication models

An even more radical proposal than adapting existing journals is to replace the current system with an entirely new one. This constitutes a social dilemma, in which millions of researchers must suddenly behave differently and, in doing so, act against the rules of the scientific system as it

currently stands (Brembs et al., 2023). The dilemma was designed by commercial publishers, who profit from it. Brembs et al. (2023) have developed a precise proposal for building a decentralized system, modeled on social networks such as Mastodon, organized by researchers themselves, that values research products such as data or software just as much as traditional research reports. Platforms that already implement such a micro-publishing system include Research Equals and Octopus.ac (Hsing et al., n.d.). Opposing a system in which all products are reviewed but the review process does not itself guarantee quality, these platforms use traffic-light systems and public commenting to signal what was reviewed, whether it was reviewed at all, what criticisms arose, and how they were addressed.

Researchers

Regardless of nationality, field, or university, many researchers have rethought and adjusted their working practices in the course of open science. Hundreds have signed public declarations on research transparency and calls for open science practices in their role as reviewers. Individual researchers conduct replication studies as part of teaching (Boyce et al., 2023; Jekel et al., 2020; Korell et al., 2023), join forces worldwide to carry out projects that no individual could manage alone (e.g., the Psychological Science Accelerator, https://psysciacc.org), and develop collections of information (Open Scholarship Knowledge Base, https://oercommons.org/hubs/OSKB), guides, and glossaries (Parsons et al., 2022) to make open science easier to access. Communication among researchers now takes place independently of journals and faster than before, via social media such as Mastodon, Bluesky, and personal blogs.

Blogs, in turn, are searchable and networked through platforms such as *Rogue Scholar* (https://rogue-scholar.org). One demand that remains unmet is for researchers to reform the system through unionization (Rahal et al., 2023).

A particular form of resistance is journal boycotts. Here, researchers commit themselves, publicly, to no longer publishing their research in commercial journals and to no longer providing reviews for them. Taylor discusses potential collateral damage in a blog post: researchers who depend on publishing in high-ranking journals are disadvantaged by this.

Evaluation criteria

With the San Francisco Declaration on Research Assessment (DORA), many institutions and researchers began, in 2012, to publicly and clearly oppose evaluating research solely on the basis of citation metrics (e.g., the impact factor). By 2024, there were already more than 25,000 signatures from 165 countries. The declaration consists of a general recommendation followed by specific ones (for funding organizations, institutions, publishers, and so on). The general recommendation reads: "Do not use journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions." (https://sfdora.org/read/).

In psychology, evaluation criteria for researchers are being developed systematically, using methods from personality psychology and psychological assessment (Gärtner et al., 2022; Schönbrodt, Gärtner, Frank, Gollwitzer, Ihle, Mischkowski, Phan, Schmitt, Scheel, Schubert,

Steinberg, et al., 2022). One problem here is that researchers divide up the work within groups, and some people benefit from this more than others (e.g., because they specialize in methods and therefore appear less often as first author). An analogous problem occurs in football: if players were evaluated only on goals scored, defenders, midfielders, and even those responsible for the assist would be disadvantaged. Against this background, Tiokhin et al. (2023) propose a stepwise evaluation: first, the groups in which researchers work should be evaluated, and only then the individual members.

Evaluation criteria are also being further developed for individual research articles — above all in peer review. M. Elsherif et al. (2023) have developed a template for reviewing quantitative psychological studies and replications. As part of the "Peer Reviewer Openness Initiative" (PRO, https://www.opennessinitiative.org), researchers have publicly committed to giving a positive assessment of a research article only if they have access to all necessary materials and data (unless there are good ethical reasons why this is not possible).

A fundamental problem in developing evaluation criteria is that they often disadvantage anything outside the standard case (Hostler, 2024). Criteria impose a uniform yardstick across many different researchers and research disciplines. Economists who are evaluated by impact factor like to publish articles in journals that overlap with other disciplines, because citation counts are higher there. If methodological criteria are used for evaluation (e.g., whether a central study is preregistered), those who do qualitative research — for whom classical preregistration is not helpful at all — are disadvantaged.

Alternatives to the impact factor

While the San Francisco Declaration ("DORA") clearly condemns the journal impact factor, it leaves open which alternatives should be chosen. Building on this, the Coalition for Advancing Research Assessment (coara.eu) recommends the use of qualitative characteristics. Nosek et al. (2015) developed the Transparency and Openness Promotion Guidelines (TOP Guidelines), which evaluate journals on the basis of ten criteria and allow journals to be ranked (topfactor.org).

i TOP factor versus Hirsch index

To calculate the TOP factor for a journal, one must first record, for every facet, how open and transparent the journal is. For example, regarding the transparency of materials, one checks whether the journal's guidelines say anything about this at all (0 points), whether the journal merely requires a statement of whether materials are available (1 point), whether materials must be uploaded to a database (2 points), or whether, beyond that, someone will reproduce (i.e., recheck) the analyses (3 points). The points across all facets are then summed — so, strictly speaking, it is a TOP "sum" rather than a "factor" (just as the impact factor is actually a ratio). From a psychometric standpoint, it is questionable to sum across the different facets in this way, as though 3 points on one facet could straightforwardly compensate for a missing point on another.

The Hirsch index (or h-index) indicates that, of all publications, at least this many have been cited at least this many times. An index of 4 would mean that at least 4 articles have been cited at least 4 times each.

H-indices and TOP factors are publicly available for many journals. The data can easily be downloaded and related to one another (materials for this are available online: https://osf.io/utzfs). To be able to infer quality from citation counts, Peroni & Shotton (2012) developed a system for specifying what kind of citation is involved (Citation Typing Ontology, CiTO), which has already been implemented by journals (Willighagen, 2023). Whether this actually improves the assessment of research remains to be seen. Not entirely seriously, a Free Lunch Index has also been proposed for medicine, reflecting the sum of gifts received from industry (Scanff et al., 2023).

Infrastructure (open infrastructure)

In the pyramid of culture change, infrastructure forms the foundation. It enables various open science practices to be put into effect. For example, research data cannot simply be published if there is no place or website for it. Considerable progress has been made here, and sharing research materials, data, or publications has never been easier. Guidelines for research infrastructure (Bilder et al., 2020) specify how infrastructure should ideally be built: for example, it should be designed sustainably and financed for the long term, and used across disciplines, institutions, and locations. In Germany, the association "National Research Data Infrastructure (NFDI)" advocates for organizing data as a shared resource. Within discipline-specific consortia, structures are being created that make it possible to share research data. An interactive map of open infrastructure is available online (https://kumu.io/access2perspectives/open-science#disciplines/by-os-principle/open-infrastructure).

> **The cost of an "open book"**
>
> This book was written entirely using free software (GNU R, RStudio, Quarto) and is hosted free of charge on GitHub. A further step would be to use exclusively open source software — that is, programs whose code has been made public and can be used for one's own purposes. Currently, the availability of this book depends on GitHub remaining free. An additional publication through a library (digital and print, e.g., the University and State Library of Münster) would cost 100 euros.

Table 9: Examples of open science infrastructures.

Service	Purpose	Provider
Literature database	Managing and searching the literature	OpenAlex
Data repository	Publication of research data	re3data.org, osf.io, researchbox.org, zenodo.org
Preprint server	Publication of research articles	arXiv.org, Zenodo.org
Journal system (editorial manager)	Management of scientific journals (submission, review, publication, indexing)	Open Journal System
Post-publication peer review	Review and commenting on research after publication	Pubpeer.com

Service	Purpose	Provider
Identification of researchers, institutions, and research		Open Researcher and Contributor ID (ORCID.org), Research Organization Registry (ROR.org), Digital Object Identifier (DOI.org)

Open access publications

To increase access to scientific knowledge, more and more research is being published as "open access." Specifically, this can happen in many different ways: people can publish articles in non-commercial journals (for a collection of more than 20,000 such journals, see, e.g., https://doaj.org); some commercial journals make articles freely available after a certain period has elapsed; or authors can pay money so that the article appears openly accessible in a traditional journal. The costs for this can range from a few hundred euros up to 9,000 euros at "prestigious journals." The funds for this usually come from university budgets (e.g., open access funds, which are subsidized or covered by the German Research Foundation [DFG] using tax revenue) or from project funds — meaning that funds were requested, when applying for project money, specifically to cover the costs of open access publication. While open access originally explicitly did not refer to this pay-to-publish model, journals have since "re-commercialized" it (Morgan & Smaldino, 2024) — that is, they have made use of it for their own benefit. The various types are neatly listed in the open access strategy of the universities of North Rhine-Westphalia (DH.NRW | AG Openness, 2023). Besides open access at first publication, researchers usually retain the right, under "green open access," to upload

the articles they have written to their own website, their institution's website, or subject-specific repositories. Zumstein (2024) recommends, for example, that at subscription-based ("subscription model") journals, one should not pay extra for the open access option, since this is not sustainable, and should instead make the secondary publication openly available.

Table 10: Types of open access at first publication, following DH.NRW | AG Openness (2023).

Type	Rule
Diamond	All publications are immediately and freely available, and no publication fees are charged
Gold	All publications are immediately and freely available; authors pay a fee per publication (Article Processing Charges / Book Processing Charges)
Hybrid	The journal has a subscription model. Individual articles are made publicly available for a fee.
Moving wall	The journal has a subscription model. After a set period elapses (6–48 months), articles become freely accessible.
Promotional	Individual articles are made freely available to promote the journal.

How prestige can blind us to openness

In some places there are unwritten rules such as "if you publish in a high-ranking journal during your PhD, you'll get the top grade." This shifts ever more power to prestigious journals. People are congratulated extraordinarily for having published something in Psychological Bulletin or even Nature Human Behaviour. It plays no role that nobody without a university affiliation (that is, working or studying there) can read the article in Psychological Bulletin, or that

> publishing in Nature can cost up to 9,000 euros (usually paid from tax revenue).

Besides commercial and open access journals, there is another actor in the accessibility of scientific knowledge: through *shadow libraries* such as Sci-Hub or Anna's Archive, people can access articles that would otherwise sit behind a paywall ("guerrilla open access"). Sci-Hub (https://en.wikipedia.org/wiki/Sci-Hub), the best-known shadow library, contains nearly 70% of all 81.6 million scientific articles published up to 2018 (Himmelstein et al., 2018). Strecker (2019) has analyzed usage statistics from Germany. The distribution and downloading of such content is legally contested. Other ways to obtain research articles for free include 12ft.io, the hashtag #canihazpaper on social media, or personally writing to authors by email or through academic social networks (Researchgate.net, Academia.edu). Brembs offers an overview of methods for reading scientific articles for free in a blog post.

Choosing a journal

Aside from prestige or journal impact factors, researchers choosing where to publish their research can look for Diamond Open Access — that is, free public access (e.g., via https://doaj.org or https://freejournals.org/current-member-journals/) — and take the TOP factor into account (top-factor.org). Various relevant characteristics are compiled by oa.finder (https://finder.open-access.network). Researchers should also check with their libraries: at subscription-model (hybrid open access) journals, a market for paid open access publication has emerged. Journals and publishers publish extremely large numbers of articles without

rigorous review and earn money through open access fees. In such cases the journals are marketed as "open access journals," and the costs are called "Article Processing Charges (APCs)." A dashboard run by Bielefeld University breaks down which journals and publishers have received how much money from German universities (https://treemaps.openapc.net/apcdata/openapc/#publisher/). In 2022, for example, German universities spent 68 million euros on commercial open access publications. The economic cost of publishing research is estimated at roughly $400, or 380 euros, per article (Grossmann & Brembs, 2021). For the 31,517 articles they published open access that year, this would put the estimated maximum cost at just under 12 million euros — whereas, through Diamond Open Access journals, costs far below this would be possible. The publisher MDPI — the publisher receiving the most German APC payments, ranked first by article count, with nearly 9,000 German-authored articles, and revenues of more than 16 million euros from German institutions alone — is particularly controversial: researchers (https://predatoryjournals.org/news/f/is-mdpi-a-predatory-publisher) have shown that processing times (duration of review, revision, and publication) are unrealistically uniform, and that there are multiple special issues per day (a journal typically has 1–2 special issues per year). Beall published a controversial list on his website (https://beallslist.net) identifying journals as unscientific, and describes his experiences in an article (Beall, 2017). Tools for identifying disreputable journals and conferences are also available (https://thinkchecksubmit.org, https://thinkcheckattend.org).

Monitoring

How much research is published as open access, and how much this costs, can be tracked using various tools. *OpenAPC* lists open access costs by publisher, journal, and university (https://treemaps.openapc.net/apcdata/openapc/), and the Open Access Monitor breaks down how many publications in Germany are published under which open access models (https://open-access-monitor.de).

Currently (as of summer 2024), 54.4% of all indexed journals have no open model, and the bulk of all research is published through them. A further 19.2% of journals operate under a transformative agreement. These are intended to achieve a transition from a subscription model to an open access model, with all previously published articles also being made openly accessible. Probably the largest player here is the DEAL consortium (https://deal-konsortium.de/publizierende): German scientific organizations have joined forces to negotiate a nationwide contract with publishers that allows all German researchers to publish open access without additional cost.

> **i** Open textbooks
>
> While journal articles are primarily used for exchange among researchers, textbooks play the special role of forming the bridge between experts and interested parties (e.g., students). The relationship between textbook authors and publishers is less strained than that between researchers and journals — even though these are largely the same publishers. Two particular differences may account for this: 1. textbook authors earn money from sales, and 2. they can declare their own works relevant for exams. As a result, it is students,

> and not the authors themselves, who bear the costs. In the United Kingdom, pilot projects already exist that aim to shift teaching as fully as possible onto open textbooks (Farrow et al., 2020).

Mass resignations of editors

It is becoming increasingly common for a journal's entire editorial board to resign. The reason is usually that the publisher wants to raise publication costs or the number of published articles. A list of such resignations is maintained by Retractionwatch.org ("Editorial Mass Resignations," https://retractionwatch.com/the-retraction-watch-mass-resignations-list/). In these cases, a community of researchers has spent years working hard to build and promote a journal's reputation, only to be punished by having to pay even more money to exchange their research with one another.

In many such cases, the researchers go on to found a new journal, usually open access, following their resignation. While university libraries support the founding of new journals, and the process is technically unproblematic, there are legal and social hurdles: publishers such as Taylor & Francis often agree "non-compete clauses" with researchers, forbidding them from working for another journal for one or several years after their editorial position ends. A scientific community that forms the core of a journal must also stand unanimously behind such a change. In one case, the editorial board made clear that it would be completely socially unacceptable to continue supporting the old journal. Newly founded journals also do not yet have an impact factor, because no citable articles have yet appeared and none could yet be cited. The abandoned journals are referred to as "zombie journals." Publishers counteract mass resignations by fre-

quently rotating members of the editorial board, so that they find it harder to coordinate. In-person meetings are also made more difficult by the international nature of research.

> **Openness through building blocks**
>
> Openness in science can take many forms: one particular type of public access is currently spreading in biotechnology research. So that scientists, engineers, and the general public can gain access to designs, this field increasingly turns to interlocking building blocks such as those from the company LEGO® (Boulter et al., 2022). The patent on "the Lego brick" has already expired, so various companies or individuals with 3D printers can now produce their own building blocks. The files for 3D printers are freely available on the internet. Using this approach, researcher David Aguilar was able to design a prosthetic arm out of interlocking building blocks.

Preprints

As already explained in the chapter on reviewing preprints, preprints are always publicly and freely available. Researchers can assign various licenses — for example, prohibiting commercial use — though these are nearly always very permissive. Beyond the advantages of preprints already discussed, revisions to them are also many times faster: researchers can respond to criticism and correct errors at any time. In an article on the coronavirus, an error was noticed shortly after publication and corrected within two days. With a journal article, this process can drag on for years. Especially for serious problems that would normally lead to a retraction, publishers are comparatively slow. Authors can take

a preprint offline at any time — though it will likely still be findable via search engines.

Preprints can also help ensure that resources are used more efficiently: through faster communication between researchers, it becomes apparent more quickly when different groups are working on the same question. This can lead to collaborations forming, or to groups shifting their focus.

> **Unwarranted fear of idea theft**
>
> In some scientific disciplines, researchers are hesitant about preprints. They fear that someone will steal their idea and publish an article on it in a journal faster than they can. While this fear also exists under the traditional system — for example, through malicious reviewers or conference attendees — and can happen even with published journal articles, the advantage of preprints is that they are dated, making it clearly traceable, for everyone, what was published when.

Approaches against the selection of exciting results

The results of a study are the thing least under the control of the researcher conducting it (or at least, they should be — after all, what interests us is the truth, not a researcher's skill at making the data look as favorable as possible). It is all the more frustrating, then, that journals use the result as a criterion for publication. Among the many submissions, journals mostly select those articles that achieved exciting results, or that confirmed their initial hypothesis (*confirmation bias*). The following approaches solve this problem, for example, by excluding results from the review process.

Results-blind peer review

The simplest approach is simply to redact or omit the results section. Various journals offer this as an option. Since this is currently (2024) still the exception rather than the rule, most reviewers are well aware that those who choose the results-blind peer review option are usually those whose results are not "pretty enough" for the standard route.

Registered Report

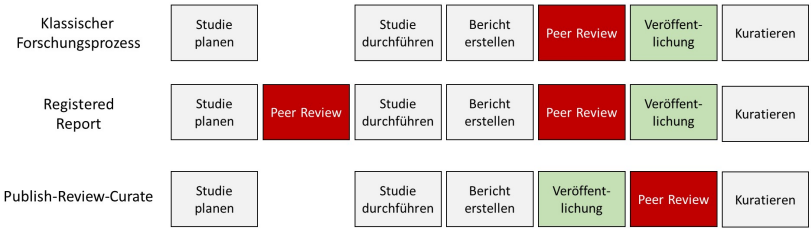

Figure 19: Classic research process compared to Registered Reports and the PRC model. With Registered Reports, an additional review takes place before the study is conducted. With PRC models, publication takes place before review.

A more radical approach than reviewing an article without its results section is reviewing the article before any results exist at all. This format is called a *Registered Report*. Here, the manuscript — containing the theory to be tested, the methodology, and the analysis plan — is submitted to the journal before any data have been collected. If this "half-finished" article is accepted (*in-principle acceptance*), the data are then collected, analyzed as planned, and a further round of review follows. Here it is stipulated that the authors may not change anything in the parts already written, and that reviewers may not, after the fact, criticize the already-approved

approach. The only remaining question is whether the plan was followed and whether the conclusions are grounded in the planned analyses. This is meant to prevent articles from being rejected because the results are not exciting enough, innovative enough, or in line with expectations. Initial investigations can already show that this improves research quality compared to the traditional approach (Soderberg et al., 2021). An overview of journals offering this format is available online (https://www.cos.io/initiatives/registered-reports → Participating Journals, Chambers, 2018; Chambers & Tzavella, 2022). This format also makes clear that research suffers from massive publication bias (i.e., mostly studies that confirmed their hypotheses get published, while studies that did not are rarely published): Scheel et al. (2021) showed that the proportion of results consistent with expectations in Registered Reports is 44%, markedly below the 96% typically found in psychology.

Preprint-based models

With preprints — that is, not-yet-reviewed manuscripts — being published ever more often, new avenues for review are opening up, such as the publish-review-curate model (see the section "Review of preprints"). So-called overlay journals (elife) select among preprints those they send to reviewers to obtain their opinions. Provided the preprint's authors agree, they then receive reviews, and their article is ultimately published in the journal.

Depending on the field, varying numbers of journals have agreements with PCI initiatives to publish accepted articles without their own peer review. More participating journals — especially well-known ones — make PCIs more attractive to researchers. Journals interested in

a PCI but not wanting to give up review entirely can list themselves as "PCI-interested journals" (as opposed to "PCI-friendly journals"). Participating journals save work this way and stay relevant, as their function shifts toward collecting and disseminating thematically relevant research. In one case, a journal that had a "PCI-friendly" agreement with PCI-RR sent a manuscript out for another round of peer review despite a *recommendation*, and was immediately removed by the PCI partners. Researchers who want to organize review processes for PCIs — analogous to editors at traditional journals — can become *recommenders*, which requires a minimum level of knowledge and completing a training (https://rr.peercommunityin.org/about/recommenders).

Table 11: Summary of the various review models and how each handles research results.

Review procedure	**Handling of results**
Traditional, PCI, PRC	Results are visible and can factor into the assessment
Results-blind peer review	Results exist but are withheld from reviewers
Registered Report	Results do not yet exist
Peer Community In Registered Report (PCI-RR)	Results do not yet exist

i Anecdotes: my worst experiences with peer review

Peer review is brutal. What follows is **my personal experience** with the process. Early in my career, I had to learn that in many cases it is a lottery how a manuscript's evaluation turns out. Renowned scientists told me that they had articles which they had resubmitted to journals, again and again, for years, and which were eventually accepted without them having changed much. Beyond the acceptance of an

article for publication, the reasons for rejection matter too: reviewers frequently do not read articles carefully, and criticisms are not constructive. Here are my worst experiences.

Submission to Collabra: This concerned an article that sat between disciplines. It wasn't just about expectations, or product reviews, or the method of downloading data directly from the internet. This "odd-one-out article" had already been rejected by three journals. At none of them was it even sent out to reviewers, because it never fit the journal. The journal *Collabra*, where we finally submitted it, was only a few years old at the time and had a broad scope. After submitting in May 2020, we waited more than six months for a review. Waiting longer is initially a good sign: it suggested the article had actually been sent out to reviewers. In this case, however, we were bitterly disappointed: in November 2020, upon inquiry, we learned that 14 reviewers had been approached, and the editor ultimately decided to reject the manuscript based on a single review. The reason given was that the method was unsuitable for the research question because the data were incidental. In psychology and economics, the problem with "found data" has been known for decades, and we had already discussed it extensively in the article.

Submission to the Journal of Experimental Social Psychology: We had conducted a replication of a study published there — the finding did not replicate. My thinking was to give the journal that had originally published the non-robust finding the chance to correct itself. The reviews were fair and positive; there were a few points to discuss, but it was clear to us that these were misunderstandings, not substantive issues. Not so for the editor: he explained that the finding was not novel enough and that it was already clear it would

not replicate. I explained to him that nobody had yet attempted to replicate the finding, and that we ourselves, before analyzing the results, had even taken a vote on what outcome we expected: it was almost perfectly split, 50-50. Moreover, another article had since been published with the opposite result. We made clear that we could easily address all the issues raised and would welcome the chance to revise. The editor would have had hardly any additional work beyond sending the article back out to reviewers afterward. In response to our email, we received only a brief reply: *Hi. I know my decision is disappointing, but I'm going to stick with my decision on this one.* At this point I found myself at a crossroads: why would a researcher not be willing to discuss the reasons for a scientific decision? We decided to contact a different editor directly. Shortly afterward, we received an invitation to revise. The article was eventually published in its revised form.

Submission to the European Journal of Personality: The central claim of this article was that different measures of a supposed trait did not correlate with one another. The finding called into question whether this was even a coherent trait at all. The reasons given for rejection by the two reviewers and the editor made it clear: nobody had actually read the article. One reviewer noted that something was wrong with the values, since, according to one of the tables, they did not correlate with each other. That was exactly our point. We showed that this was not due to our data but occurred the same way in other datasets. Had he read the table's caption, the paragraph before it, or the one after it, this would have been clear. He hadn't…

Reviews of grant applications at the German Research Foundation (DFG): Successfully securing DFG funding is an important

step in psychology on the way to a permanent position, even though everyone knows it's a lottery. For two rejected applications, the wait for the review was around nine months in both cases, and the criteria applied showed just how absurd the process can be. For example, one review evaluated the institution where I had been working at the time of submission. That I no longer worked there by the time of the decision — which is entirely normal given how common short-term contracts are — was not taken into account by the reviewer. In another case, our application was rejected because another research group had planned a similar study. That this other group was, at the same time, in discussions with us about letting us take over that part of the work entirely, played no role. Review processes at Germany's most important funding body, incidentally, are so anonymous that I am not permitted to publish the reviews. Outsiders therefore cannot see the extent to which quality assurance actually takes place. When applications are rejected because the reviewer misunderstood something, the decision cannot be appealed either; instead — if the specific program still allows it — a new application must be submitted. These are just brief excerpts from dozens of submissions and rejections. Beyond this, almost every researcher can recount baseless, personal insults. In my experience, anonymous versus open peer review is like comparing anonymous comment sections with non-anonymous ones: under anonymity, personal insults and falsehoods dominate the dialogue.

Replication research

There is frequent talk of a replication crisis — that is, a crisis of insufficient replicability. Unsurprisingly, this has affected the role of replications in the social sciences and beyond. Numerous avenues have been pursued to raise the standing of replication studies and thereby their frequency in research. In doing so, the sciences must now make up for something they should have addressed from the outset: specifying what standards of replicability should apply and how these should be tested. By replicability, I mean that a hypothesis can also be confirmed using data other than that of the original study. This is about a minimal degree of generalizability, not primarily about a deeper understanding of theories — even though the latter is sometimes still criticized as missing, despite nobody having claimed that replications should solve the theory problem as well (Feest, 2019).

The imprecision of replication studies

An as-yet unresolved problem is the imprecision of both original *and* replication studies. Ting & Greenland (2024) criticize that the imprecision of replication studies is often disregarded. What they overlook, however, is that replication studies almost always have a far more precise study design and adhere to higher methodological standards than any preceding studies. Embellishing results in the sense of p-hacking (Simmons et al., 2011) is in principle also possible in replications (Protzko, 2018), though harder to do given the higher standards. While researchers do appropriately take replication findings into account in their judgments (McDiarmid et al., 2021), there

> is so far no research on how susceptible replications are to data fabrication compared to original studies.

What *should* replicate?

Against the backdrop of robustness, it becomes clear when a successful replication should be expected and when it should not. If a hypothesis is formulated in an original study as universally valid (e.g., shortly after birth, male babies weigh more on average than female babies), it should also be demonstrable repeatedly. This contrasts with cases where a hypothesis is not formulated as universally valid (e.g., where something is tied to a specific context, as in qualitative research). There are entire disciplines for which replicability is irrelevant — for example, in archaeology, an excavated object is torn from its context, thereby "destroying" the original. This process is not repeatable, and nobody expects it to be. It is also possible to replicate artifacts (that is, findings that arise only because of the methods used) if their origin is based on some general regularity (Devezer et al., 2021). For example, it can happen that a statistical model reliably "triggers," identifying patterns in findings, even when these do not arise from the actual explanation but only because the model is poorly calibrated or its assumptions were violated. For a long time, for instance, it appeared that left-handed people die younger than right-handed people. This finding could be reproduced for a while across various German samples, and explanations were already being proposed for why this might be. With today's data, the replication is no longer possible, because it turned out to be an artifact: until a few decades ago, all children were trained to write and cut with their right hand. Eventually, this practice of re-training stopped. To show up as a left-handed person in death statistics in the years

that followed, one would have had to die quite young — since there were hardly any older people who wrote with their left hand, precisely because of the earlier re-training. As a result, left-handed deceased people were, on average, a full 10 years younger than right-handed deceased people. Replicability is therefore not sufficient for validity or truth, but it is necessary in order to underscore the validity of a hypothesis or to defend its claim to truth.

Fletcher (2021) lists the conditions under which something can be replicated:

1. There are no errors in the data analysis.
2. The finding is not attributable to statistical imprecision.
3. The finding does not depend on neglected background factors.
4. There was no fraud or other research misconduct.
5. The finding generalizes to a population larger than the sample of the original study.
6. The hypothesis remains valid even when tested in a completely different way.

> Necessary and sufficient conditions
>
> Replicability is a necessary but not a sufficient condition for generalizability — what does that mean? The terms “necessary” and “sufficient” may be familiar to many people from school mathematics or introductory logic. Their precise meaning is especially relevant in logic:

- Necessary means "it can't happen without this." Having cocoa powder is necessary for me to be able to make hot cocoa. That is, I cannot make hot cocoa if I don't have cocoa powder. At the same time, it is not sufficient: just because I have cocoa powder doesn't mean I can make hot cocoa. I might still be missing milk, water, a cup, or a microwave.
- Sufficient means: if it is present, that alone is enough, and no further conditions need to be met. If I hear a plane in the sky, that is sufficient for there to be a plane flying overhead. But it is also possible for a plane to fly overhead without my hearing it (for example, because it is flying very high, the ambient noise is very loud, or it is a quiet glider).

Unlike in some examples, the condition need not always occur before the event. So this is not about a causal relationship in which one thing *leads to* the other. A particular feature of conditions is that if A is necessary for B, then B is sufficient for A. That I have made hot cocoa is therefore sufficient for me to have cocoa powder. And that a plane is flying overhead is necessary for me to be able to hear it.

Who replicates?

Despite strong efforts, replication studies are still relatively rare. Estimates range from 5% down to under 0.1%, depending on the research discipline. By their own account, most researchers have carried out a replication at some point, but were unable to confirm the original findings (M. Baker, 2016). Much of the data in the following table comes from a tweet by Gilad Feldman.

Table 12: Overview of how much replication is done, and where.

Discipline	Source	Finding
Economics	Ankel-Peters et al. (2023)	Depending on the journal, the share of replications and reproductions ranges between 0 and 11%.
Economics	Mueller-Langer et al. (2019)	0.1% of published studies are replications.
Experimental linguistics	Kobrock & Roettger (2023)	Fewer than 0.001% of studies include replications.
Psychology (1900 – ca. 2012)	Makel et al. (2012)	About 1.07% of all articles since 1900 include replications, and the share has recently increased.
Psychology (2014 – 2017)	Hardwicke et al. (2022)	Replications are reported in 3–8% of studies.
Psychology (2010 – 2021, high-impact journals)	Clarke et al. (2023)	0.2% (169 of 84,834) of articles included direct replications.
Ecology and evolution	Kelly (2019)	Fewer than 1% of studies are described as replications.
Second language research	Marsden et al. (2018)	0.25% of studies are replications.
Education	Makel & Plucker (2014)	0.13% of articles included replications.
Special education	Makel et al. (2016)	0.5% of articles included replication studies.
Criminology	McNeeley & Warner (2015)	Just over 2% of articles include replication studies.
Behavioral ecology	Kelly (2006)	Replications are rarely conducted.

The standing of replication studies

Replications of Bem's infamous study on precognition ("feeling the future") were rejected without review ("desk rejected") by the journal that had published the original. The editor explained that the journal does not publish replication studies, no matter the outcome, since it did not want to become the journal of Bem replications (Lakens, 2023). Among the 3,185 journals with a TOP factor, 341 journals (10.7%) stated in 2024 that they accept replications. Replications are defined as confronting current theoretical expectations with current data, and their important role in the scientific process is being increasingly recognized (Nosek & Errington, 2020). Much of this development has been driven by social psychology; other fields have changed little, or only slowly (Torka et al., 2023), and even there, discussions of replication findings can still be harsh, with replications sometimes misrepresented and unjustly criticized (Chandrashekar & Feldman, 2025).

In economics, the Institute for Replication (I4R, https://i4replication.org/reports.html) has formed, organizing so-called Replication Games in which studies are reproduced and replicated. Building on an earlier call to action (Zimmermann, 2015), the Journal of Comments and Replications in Economics (JCRe) was founded. From 2024, I4R is focusing on conducting replications for specific journals. Ironically, it does not do this for open access journals but, for example, for Nature (Brodeur, Dreber, et al., 2024). The Journal of Applied Econometrics also publishes replications of studies from numerous economics journals.

Types of replication projects

While the majority of replication studies resemble classic studies, specialized models have also emerged. For example, researchers can use StudySwap to find groups willing to replicate their results before publication (Lakens, 2023, p. 11). In Registered Replication Reports, large teams at multiple sites focus on a previously agreed-upon study and carry it out simultaneously everywhere. One of the most fruitful approaches is the use of theses for replications. As part of their studies, all students must complete a thesis (e.g., a bachelor's or master's thesis). Through replications, they can learn to reproduce an important study while simultaneously testing the robustness of the original findings. While Quintana (2021) proposed using theses in this way, it is already established practice in comparative political science (Korell et al., 2023), in psychology (Boyce et al., 2023; Feldman, 2021; Jekel et al., 2020; Wagge et al., 2019), and in economics (https://home.uni-leipzig.de/lerep/). Replication series also began in 2024 at the Berlin Social Science Center (WZB) (https://www.wzb.eu/de/forschung/markt-und-entscheidung/verhalten-auf-maerkten/labsquare) and at the Sports Science Replication Center (https://ssreplicationcentre.com).

Table 13: Types of replication.

Type	Explanation	Example
Registered Replication Report	Researchers at many different sites carry out the same replication study (Simons et al., 2014).	Cheung et al. (2016)

Type	Explanation	Example
Internal replication	A research group reports, within a research article, its *own* replication of one of its studies.	Ongchoco et al. (2023)
Close replication	An original study is repeated as closely as possible by other researchers.	Xiao et al. (2021)
Conceptual replication	Other researchers test the same hypothesis again, but deliberately use different methods.	Sobkow et al. (2021)

Who publishes replications?

Srivastava (2012) proposed a "Pottery Barn" rule for scientific journals. This refers to the rule common in shops selling fragile goods: "you break it, you buy it." Applied to replications, this would mean that a journal must also publish all replications of a study it has published. In practice, no journal does this — possibly out of fear of reducing their impact factor or of being accused of inadequate quality assurance. A weaker version of this idea is proposed by Hüffmeier & Kühner (2024) with replication markets: journals would advertise central publications for replication; researchers could then apply, with a research proposal (a Registered Report), to replicate these central studies; and the journal would ultimately select a proposal and fund it with money for data collection. Aside from the replication journals JCRe (https://www.jcr-econ.org) and ReScience X (rescience.org/x), there are only a few journals that publish replication

studies. In psychology, these include, for example, Meta-Psychology and the Journal of Trial and Error.

To increase the findability of replications nonetheless, replication *databases* have emerged. In these, researchers collect replication studies and present them in a user-friendly way. For economics, Jan Höffler did foundational work with the Replication Wiki (ReplicationWiki) (Höffler, 2017). In psychology, LeBel introduced curatescience.org. That project died out fairly quickly; the idea was later picked up by FORRT with "Replications and Reversals" (forrt.org/reversals/) and, in 2023, merged with the "Replication Database" into the "FORRT Replication Database." With meta-analytic functions and the ability to check bibliographies for replication studies, the FORRT Replication Database is currently the most comprehensive project of its kind (forrt.org/replication-hub). Most of these projects are carried out by large communities, with participants painstakingly contributing data by hand. Automated methods are in development, but are still immature (Ruiter, 2023).

What makes a good replication?

Replication methods are being developed across many fields. Building on work first done in psychology (Brandt et al., 2014; Hüffmeier et al., 2016), there are now guidelines for replications in quantitative sociology (Freese & Peterson, 2017), the humanities (Schöch, 2023), and marketing (Urminsky & Dietvorst, 2024). Methods for sample size planning (Bourazas et al., 2024; Pawel et al., 2023; Simonsohn, 2015), for study selection (Howard & Maxwell, 2023; Isager et al., 2024), for evaluation (S. K. Yeung & Feldman, 2023), and for communication (Janz

& Freese, 2021) of replication studies have all been developed. Below, recommendations are structured in the form of a replication guide.

A short guide to conducting replication studies

1. **Choosing a study**

   The study to be replicated should be relevant and open to doubt. Relevance can be reflected in high citation counts or in the extent to which subsequent research builds on the finding. If many replications already exist, and it is already clear whether the finding is replicable or not, a further replication study is not very informative. Because of the replication crisis, doubtfulness is usually present (e.g., through p-values close to the 5% threshold). Meta-analytic anomalies (Adler et al., 2023) can also be used to check for doubtfulness.

   When selecting a study, it is also advisable to take into account any existing discussions (comments in journals or on Pubpeer.com), if there are any. For theses, feasibility must also be considered: a longitudinal study lasting 10 years is not very feasible. There should also be established replication standards for the relevant field. For brain-scanner data, for instance, this is not yet the case, since hundreds or thousands of correlations are compared there and the original values are often unavailable.

2. **Conducting the replication**

   It is advisable to take stock of all publicly available materials. Where possible, original results should be reproduced and checked. Even without the data, values can be checked

for correctness using, for example, statcheck.io (Nuijten et al., 2016). For more recent studies, the original authors may be able to provide data and materials.

A particular challenge in research is planning the sample size (statistical power). While this problem is actually least severe for replication studies, since there is already a finding to orient around, that finding is usually too imprecise to use directly for the calculations. The small-telescopes approach (Simonsohn, 2015) helps here, and equivalence tests may need to be used as well (Lakens, 2017).

Replicators rarely get around having to adapt the original method: materials are outdated, need to be translated into another language, or need to be adapted to a particular sample of people. Similar studies, which can be found via replication databases, can help here (Röseler, Kaiser, et al., 2024). Extensions that increase the informational value of the study are also often worthwhile.

3. **Analysis**

   As with the methods, it is often useful to reproduce the original analysis (confirmatory) and then run further tests (exploratory). Comparing the results — including with regard to methodological differences — then allows for a comprehensive picture.

4. **Discussion**

   The conditions under which the replication is interpreted as successful should be specified beforehand in a preregistration. Proposals for such criteria are made by Brandt et al. (2014),

LeBel et al. (2019), and Anderson et al. (2017). Deviations from the preregistration and from the original study should be discussed extensively. Whether a comment from the original study's authors is helpful is debated, even though such comments are required for publication at some journals.

5. **Report**
   Comprehensive reports by Feldman and colleagues (Ziano et al., 2021) provide good guidance for one's own manuscripts. For short reports, a standardized form is also available, which can be used to enter findings into the replication database. Brandt et al. (2014) also recommend registering the results, which enables their further use (Röseler et al., 2022). Finally, publishing the preprint is advisable. For the social sciences and medicine, an interdisciplinary replication journal is currently in preparation, to which the study can be submitted for publication.

Modeling replicability

Replicating all studies in a field is currently unrealistic and would require enormous resources. Various methods are therefore being developed to predict, based on the properties of a study, what replication outcome should be expected. In the repliCATS project, part of the SCORE project on measuring scientific quality, researchers' judgments, reproduction attempts, and replication attempts are combined. Similar projects use complex statistical models (e.g., large language models) or meta-analytic methods such as p-curve or z-curve to predict replication rates. Such

models usually require very large numbers of observations, and predictions are only meaningful at the level of hundreds of studies. For example, Boyce et al. (2023) and the Hagen Cumulative Science project (Jekel et al., 2020) computed moderation analyses — that is, they examined which study properties affect the replication outcome (e.g., switching from an in-lab study to an online study). Comprehensive databases such as the FORRT Replication Database (Röseler, Kaiser, et al., 2024) should enable more precise models to be built in the future. Current results should be treated as preliminary and interpreted with caution, since they are based on non-random samples, study properties were not varied systematically and randomly, and replicating such meta-analytic findings is itself difficult.

Publishing all results

To solve the long-known file-drawer problem (Rosenthal, 1979; Sterling, 1959), intensive efforts are underway to develop statistical models that can "correct out" this bias. None of the current models works for all data (Carter et al., 2019), which means researchers currently have no way around publishing all their results.

In medicine, there is the special case that studies on humans must be publicly registered before they are conducted. A publicly viewable database therefore exists, through which it can be traced who conducted a study and when. At the same time, registration numbers must be provided when publishing articles. Combining both pieces of data lets us see how many studies are published within a certain period after registration — broken down even by individual people and institutions (Quest Dashboard; https://quest-cttd.bihealth.org/). The Open Trials

project (https://opentrials.net/about/) is also working to link information on registrations and research articles across various databases.

In other fields there is no obligation to register studies before they are conducted, so it remains completely unclear which and how many studies "end up in the file drawer." The journal Meta-Psychology offers an article format for psychological file-drawer reports, in which researchers can publish all studies on a given topic, including failed or flawed studies. The Journal of Negative Results (https://www.jnr-eeb.org/index.php/jnr/about) and the Journal of Articles in Support of the Null Hypothesis (https://www.jasnh.com) specialize in results that do not match expectations. As part of the *PsychFileDrawer* project, an online database of unpublished psychological studies existed for a while, and "All Results Journals" published results from the natural sciences (http://www.arjournals.com). Of the latter, however, only the biology journal is still active, while physics and chemistry have not published anything for a long time.

Topic-specific databases have become established, especially in psychology, that bring together all studies on a given topic, allowing studies to be searched and, sometimes, statistically summarized. Such summaries are often meta-analyses (studies of studies, or analyses of many studies' results), and thanks to these databases, they are dynamic: studies can be filtered, the databases grow over time, and users can determine the analyses themselves. Because of the central role of statistical results, these databases hold great promise for cumulative science — that is, they make it easier for researchers to build on one another's work. Besides their topics and functions, the databases also differ in how they collect results. Some come from individual researchers, while others arose through crowdsourcing, meaning that a group provides the database

infrastructure and other researchers send in their data. This has the advantage that the work is shared, and contributors make their data more visible through publication in the database. The table below lists a number of topic-specific databases.

Table 14: Selected meta-analysis and replication databases.

Name	Link	Topics
MetaLab	https://langcog.github.io/metalab	Language development, cognitive development
metaBUS	metabus.org	Social sciences
SOLES	https://camarades.shinyapps.io/AD-SOLES/	Animal models of Alzheimer's disease
OpAQ	t1p.de/openanchoring	Anchoring effects
FReD	https://forrt-replications.shinyapps.io/fred_explorer/	Replication studies
Power Posing	https://metaanalyses.shinyapps.io/bodypositions/	Effects of body posture
Metadataset	metadataset.com	Agriculture
METAPSY	https://www.metapsy.org	Psychotherapy

Tools are also increasingly being developed to help researchers build such databases. Tools such as *MetaUI*, *Dynameta*, or *Breathing Life into Meta-Analyses* make it easier to create a website on which data can be interactively analyzed and downloaded [metaUI, https://github.com/lukaswallrich/metaui; Dynameta, https://github.com/gls21/Dynameta; *Breathing Life into Meta-Analyses, Allbritton et al. (2024)*]. The project PsychOpen CAMA (https://cama.psychopen.eu) provides further materials.

Dealing with errors

Research errors can have massive consequences. For example, the Wakefield scandal — in which parents were paid to make false statements about their children's development after vaccination — led to vaccine skepticism. This means that parents, out of concern over the (false) consequences, do not have their children vaccinated, and the children can then fall ill and die, even though the original concern was false and has long since been repeatedly disproven (DeStefano & Shimabukuro, 2019). In many scientific fields, there is a tense error culture: as soon as someone points out an error, it is taken as a personal attack, and critics are insulted (R. F. Baumeister & Vohs, 2016). Because of strict hierarchies, it can be fatal for researchers without a permanent position to criticize professors, since the latter review their articles and decide on their employment contracts. In the Wakefield case, it took many years — even after the error was clear — before the articles were formally retracted by the journal (Eggertson, 2010). And, in the end, researchers do not even notice when articles are retracted, unless they actively look for it.

Various approaches aim to make error culture more open and to solve these problems. Ideally, everyone should understand that errors happen again and again, and that everyone benefits when errors are corrected. In scientific practice, unfortunately, it is not always *everyone* who benefits; rather, the person who made the error may have gained a Nobel Prize or a professorship thanks to it. To spur discussion on this, psychologists from Switzerland launched a bounty program (https://error.reviews), in which people can apply and then earn money by finding errors made by others, or earn money in the role of reviewer if they made no errors, or only minor ones. Platforms like Pubpeer.com also allow anonymous commenting

on research. Of particular interest there is research by prominent figures, such as Nobel laureate Thomas Südhof, whose articles have sparked discussions with as many as more than 50 comments. Comments on Pubpeer have also led to numerous retractions of published articles. Retractions are collected in the Retraction Database (retractiondatabase.org). Using a browser plugin, researchers can highlight articles for which there are Pubpeer discussions. To make retractions more visible, a study is currently underway (as of summer 2024) in which people who cite retracted articles in preprints receive an email notification (RetractoBot, https://www.retracted.net).

When a retraction occurs, a short text is published explaining why the article was retracted. Such texts are hardly standardized and are often opaque. The Committee on Publication Ethics (https://publicationethics.org) has developed recommendations on the conditions under which articles should be retracted, and Ivory & Elson (2024) recommend standardized texts to lay out the reasons for a retraction. A worrying development is the alteration of scientific articles by publishers without any notification. Aquarius and colleagues document such covert corrections as "stealth corrections" (Aquarius et al., 2025).

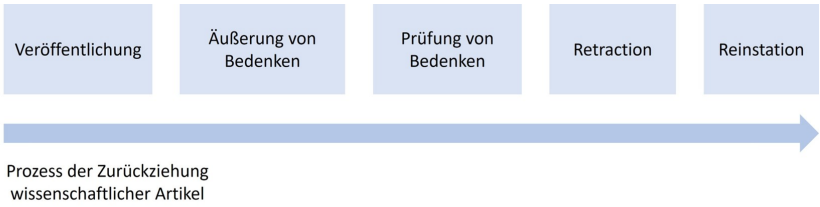

Figure 20: The retraction process for articles: after publication, editors can raise concerns based on reports, and then investigate them. If the concerns are confirmed, the article is retracted. In rare cases, it turns out there was no error after all, and the retraction can itself be withdrawn (reinstatement).

A few scientists check large numbers of articles. The most famous among them is Elisabeth Bik. She has contributed to the retraction of nearly 600 articles and the correction of nearly 500 articles (https://www.buzzfeednews.com/article/stephaniemlee/elisabeth-bik-didier-raoult-hydroxychloroquine-study). She examines figures in biology research articles. It sometimes happens that so-called Western blots appear twice, even though they are supposed to represent different figures. This can happen when converting articles into a journal's format, and when researchers create figures, whether by accident or with malicious intent.

Experience report: from error to correction

In 2021, I wrote a research article about a dataset that had been created through the collaboration of 99 researchers (Röseler et al., 2021). In 2022, the article was published in the Journal of Open Psychology Data. However, I had made a mistake. Below I describe how it was identified and fixed:

1. October 2022: During the revision of the article, data from additional researchers were added. Up to that point, 64 people had been involved in writing the article; the list then grew to 74 people. I sent the revised manuscript, with a table of the 74 people, to the journal. In the journal's system, authors could *additionally* be entered into designated fields. These fields therefore listed the 64 people, while the article itself listed 74. The submitted article was then accepted and converted by the journal into its publication format. The table was converted as well, but in the process, someone specifically deleted the

ten people who had been alphabetically inserted into the table, cross-checking against the people entered into the system at the time of the first submission. I received the finished article and had one week to give feedback on it. I did not notice the missing people.

2. In 2023, one of the missing authors contacted me: his doctoral student wanted to list the citation on her CV, but could not find herself among the contributors. It turned out that an entire team of six people was missing. I wrote to the journal's editor, apologized for the error, and asked whether a correction was still possible. It was not possible to amend the published version, but a formal correction was possible.

3. I decided to go ahead with the correction. In the intervening months, the journal had changed its submission management software, and I had to re-enter all 74 people for the new submission. Including all the checks, this took roughly a full working day — but the system also required the nationality of everyone involved, which I did not know. About half of them were students, and many others had changed institutions, so it was no longer possible for me to find out everyone's nationality. I flagged the problem and waited for an exception to be made during the correction process. At the same time, my six-month parental leave began, including a move and a change of job. After returning, I contacted one of the editors again. Seven further months of silence followed — despite repeated follow-ups, I received no reply.

4. In January 2024, I posted a Pubpeer comment. I had lost hope that the process would ever come to an end, but wanted to draw attention to the error. It included the correct list of contributors, which the journal also had on file.

5. In July 2024, I wrote to several editors again. This time I received a reply. The journal suddenly had the nationalities of the people involved, apparently entered by hand by someone at the time of the original publication. I was now able to re-enter all authors, this time with nationality. In doing so, I noticed that, besides the six missing people, four more were also missing who had never contacted me. I informed everyone affected about what had happened and apologized again. The journal then very quickly prepared the formatted article. Again, I had one week to suggest corrections. Together with colleagues and research assistants, we noted about 10 further errors in names and institutions and asked that the corrected version be sent back to us again, so that we would not have to spend years chasing a correction a second time.

6. In August 2024, the copyediting was completed and the correction went to publication (Röseler, Weber, et al., 2024).

Open teaching / Open Educational Resources

The term *Open Educational Resources* refers to any materials that can be used for teaching. This can include books — like this one — research articles, presentation slides, lecture notes (e.g., on the topic of microeco-

nomics), recorded lectures (e.g., on the topic of philosophy of science), or explainer videos. Open, here, usually means that they can be downloaded from the internet free of charge. Sharing these resources goes to the heart of open science: nobody should be excluded, whether by lack of local availability, high prices, or language. Open science advocates often upload their materials to portals (e.g., https://www.oerbw.de, https://www.twillo.de/oer/web/, https://portal.hoou.de, https://oercommons.org). Platforms such as Zenodo (https://zenodo.org) or the Open Science Framework (osf.io) allow for free long-term archiving — that is, guaranteed availability of at least 20 or 50 years, respectively.

Among the various actors in the field of Open Educational Resources, FORRT deserves special mention: in its *Educational Nexus*, many different courses and materials on open science specifically are made available (https://forrt.org/syllabus/). Online seminars are run, and a multilingual glossary and a knowledge base are maintained (Parsons et al., 2022). Other notable knowledge bases include the ROSiE Knowledge Hub and the Open Economics Guide.

Further information

- Henderson & Chambers (2022) describes ten simple rules for writing a Registered Report.
- Lakens et al. (2024) discuss the benefits of Registered Reports and preregistration.
- An explainer video on PCIs is available online (https://www.youtube.com/watch?v=4PZhpnc8wwo).

- Royal Netherlands Academy of Arts and Sciences (2018) reports on what replication looks like across various disciplines.
- Using a wiki and a webinar, students can learn about replications through the ReplicationWiki.
- Johanna Gereke and Anne-Sophie Waag discussed open science in university teaching in a talk.
- Levendis (2018), in his textbook on time series analysis, uses exclusively reproductions of analyses from actual publications.
- The OER World Map offers a worldwide overview of Open Educational Resources: https://oerworldmap.org.
- The German Research Foundation (DFG) describes guidelines for safeguarding good scientific practice in a code of conduct (Deutsche Forschungsgemeinschaft, 2022b).
- A website enabling the opening and analysis of research data is currently being developed in the Netherlands: https://opendataexplorer.org/.
- Bernhard Mittermaier discusses, in an article, why transformative agreements are a dead end (Mittermaier, 2025).

Methods

In science there is no *one method* (Feyerabend, 1975/2002); rather, methods are developed for problems, problems are solved, and methods are refined or abandoned. Methods are thus like tools, and not everything can be assembled with a single screwdriver. Open science reforms bring countless methodological innovations, improvements, and proposals that researchers often find overwhelming: advancing research projects, teaching seminars and lectures, raising external funding, and now also open science? While many problems are attributed to inadequate methods training (Lakens, 2021b), which researchers then have to make up for on their own time, some methods actually make the work easier. For doctoral students in biological psychology, for example, the guide ARIADNE was developed (https://igor-biodgps.github.io/ARIADNE/graph/graph.html). This chapter offers an overview of methodological developments and debates in the social sciences and in disciplines that primarily work with statistical methods.

Meta-Analyses

Meta-analyses are studies of studies. Researchers typically extract results from already published studies, write to other researchers in a field to ask about unpublished studies, and use statistical methods to analyze the similarities and differences between the results. Fletcher (2022) argues that only meta-analyses can demonstrate the generalizability of (statistical) phenomena. In an ideal world, everyone could carry on as before and meta-analytic models would simply correct for the problems. Given the

devastating extent of publication bias, however, this is currently not possible. How meta-analyses can nevertheless be informative is discussed in general terms by Carlsson et al. (2024), and I list specific problems and solutions below.

Assessing Publication Bias and P-Hacking

For meta-analyses, the rule is: garbage in, garbage out. Anyone who combines many poorly conducted studies in a meta-analysis ends up with a poor summary. This is, for instance, what happened when Hagger et al. (2010) found a substantial effect for a model of willpower in their meta-analysis, yet subsequent large-scale replication attempts and analyses all failed to find an effect of comparable size (Dang et al., 2020; Friese et al., 2018; Hagger et al., 2016; Vohs et al., 2021). What can still be done — and should be done in every meta-analysis — is an assessment of data quality, for example the extent of publication bias. There are methods that check whether unpublished studies exist, and methods that correct for potentially missing studies. Some of these only work with more than 200 studies, while others can already be applied to a dozen studies.

Funnel Plot

One of the oldest methods is the funnel plot (Light & Pillemer, 1984). It plots the precision and effect size of individual studies in a single diagram. Ideally, the points should form the shape of a funnel: the more precise a study is (for example, due to a large sample), the closer the association it measures should lie to the true mean. Less precise studies deviate unsystematically, sometimes above and sometimes below. Because

non-significant results are rarely published, funnel plots almost never actually show a funnel shape — instead, the non-significant results are simply missing.

The figure below shows a funnel plot for a study on the relationship between math anxiety and math performance. The pattern is nearly symmetrical, indicating only weak publication bias. The small associations at the bottom edge are skewed slightly to the right, and the precise effects at the top are not all the same size but vary considerably.

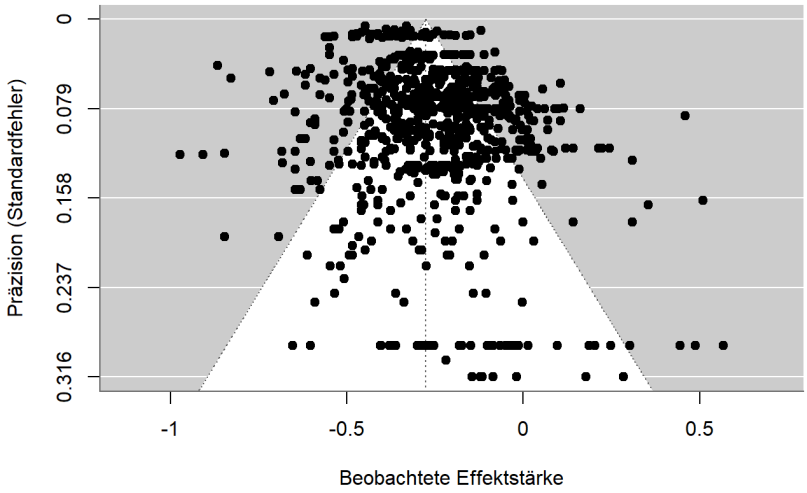

Figure 21: Funnel plot for the relationship between math anxiety and math performance.

P-Curve

In p-hacking, data are analyzed in multiple ways and only the analysis yielding a low, and therefore significant, p-value is reported. Results thus

become significant not because the hypotheses are correct, but because the data were analyzed until they became significant. In the chapter on p-hacking we saw how p-values are distributed depending on whether the hypothesis is correct or not. The p-curve (Simonsohn et al., 2014b, 2014a; Simonsohn et al., 2015) makes use of exactly this fact. The p-values from a set of studies are plotted in a diagram and their distribution is examined. If there is no p-hacking, the values are either uniformly distributed (all p-values occur equally often) or cluster near 0 (smaller p-values occur more often). P-hacking, however, causes the values to cluster near the 5% threshold, since the data need not be "hacked" any further than that. The method gained particular prominence when Simmons & Simonsohn (2017) applied it to one of the most famous phenomena in psychology, power posing, and found that p-hacking had likely occurred there. Critics later pointed out other ways a suspicious p-curve can arise even without p-hacking, and the method is now rarely used (Erdfelder & Heck, 2019).

Z-Curve

Instead of using p-values, meta-analytic findings can also be converted into so-called z-values. These are normally distributed, and additional algorithms can be used to estimate, based on observed effects, how many further effects there ought to be. This method, called the z-curve (Bartoš & Schimmack, 2022), can thus correct for the file-drawer effect and for p-hacking. Its output also includes an estimate of what the replication rate would be if all the analyzed studies were conducted again. According to recent studies, these estimates work quite well, although they require a large amount of data (Röseler, 2023; Sotola, 2023; Sotola & Credé, 2022).

Sensitivity Analysis

An approach that works not only for meta-analyses but for almost all statistical analyses is the so-called sensitivity or robustness analysis. Different analytical pathways are run through and the extent to which they affect the results is examined. In meta-analyses, for example, many possible procedures can be computed at once. Such a "shotgun" approach was coined by Kepes et al. (2017) and has since been adopted in other studies (Körner et al., 2022). Carter et al. (2019) provide an overview of various procedures and the conditions under which they are suitable for particular data.

Forensic Meta-Science

Various fields have developed methods for identifying implausible data patterns. Such techniques check, for instance, whether reported values are even *possible*: if 10 people each have a value of either 0 or 1, the mean of those 10 values cannot be 0.15, but only 0, 0.1, 0.2, 0.3, … 1.0. A collection of guides for checking such problems is COSIG (Collection of Open Science Integrity Guides) (Richardson, 2025).

Rules of Thumb for Assessing Individual Articles

A meta-analysis is laborious and can take several years. Even researchers who lack the funds for student assistants to help code and check hundreds or thousands of studies have little chance of conducting a proper analysis on their own. The following rules of thumb — and they are not meant to be more than that — offer shortcuts for assessing scientific quality.

Many Significant Studies

Since a crisis of confidence in social psychology in the 1960s (Lakens, 2023), many journals have required multiple studies per research article. As a result, resources have been invested in several smaller studies rather than one solid study. Most of these "multi-study papers" then report exclusively significant results across as many as 10 studies. While many studies with uniformly significant results may look impressive at first glance, closer inspection raises suspicion: individual studies typically have an 80-95% probability of yielding a significant result in the central analysis. This probability (statistical power) decreases when several studies are run in sequence. It is comparable to a marksman who hits a glass bottle with a rifle 99% of the time. The probability that he hits a single bottle with one shot is thus 99%. The probability that he hits 50 out of 50 bottles with 50 shots is lower, namely 99%^50 (to the power of fifty) = 60.5%. Scientific studies undergo a similar "power deflation." The probability of conducting 4 significant studies, each with 80% power, is 40.96%. Actually publishing exactly such a set of studies is then extremely unlikely (Lakens & Etz, 2017).

Effect Sizes "Just Barely Significant"

Following the logic of the p-curve, it is unlikely that p-values fall between 1% and 5%. Due to p-hacking, however, this happens frequently. A p-value close to 5% also comes with a confidence interval for the effect size that is close to 0, e.g., (Jané et al., 2024). Suppose someone conducts two studies on a topic and both have p-values near 5% with roughly equal numbers of participants; the question then arises why the sample size was

not increased for the later study — given a result that was only just significant, it is clear that the researcher was "lucky," since statistical power was not particularly high. When planning the sample for the next study, one should therefore build on the first study and adjust the plan accordingly, e.g., (Lakens, 2021a).

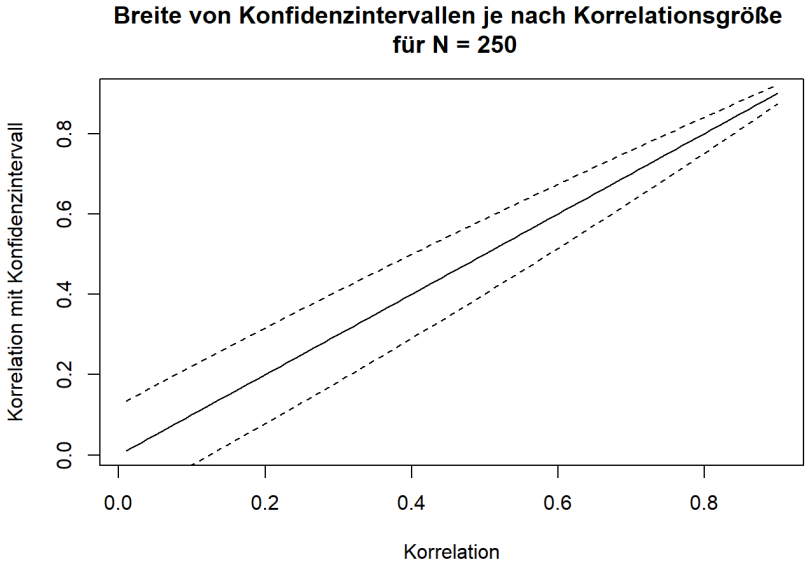

Figure 22: Width of confidence intervals as a function of correlation size for N = 250.

Checking for Reproducibility

If a finding is tested again using the same data and, ideally, the same program or analysis code, to check whether the same numbers come out — not merely whether the hypothesis is confirmed again — this is called a reproduction of the results. Unlike a replication, no new data are collected. That results are reproducible ought to be the absolute minimum standard

for scientific reports, yet it is far from being met. Reproduction studies are still rare. The average success rate across fields (economics, education, biomedicine, health sciences, geosciences) is around 50% (C. Chang & Li, 2022; Cobey et al., 2023; Koukouraki & Kray, 2023), and even the success rate for simple analyses using programs that explicitly output reproducibility protocols is extremely low (Thibault et al., 2024).

! Terminological Confusion

While the term *replication* in economics refers both to testing an existing study with new data and to re-testing with the same data, psychology uses the term *reproduction* for the latter. In biological reproduction research, the term "reproducibility" is used instead to avoid ambiguity. In yet other cases, such as Open Science Collaboration (2015), replications (new data) are referred to as "reproducibility," while repeated tests with the same data are called "computational reproducibility." Finally, in some areas the boundaries blur — for example, when a replication of the PISA study findings uses partly the same and partly new data, or when the data are computer-generated and the same program can use a pseudo-random number generator to produce different data with the same underlying structure.

For a few years now, the journal Meta-Psychology has been one of the first in psychology to conduct reproducibility checks for all published articles. These are carried out by researchers voluntarily or as part of their work for the journal. While this practice has already been called for at other journals (Lindsay, 2023), it is still the exception. Reproduction checks from all sorts of disciplines can be published at Rescience (http://rescience.github.io). Independent of journals, the Codecheck

community offers to check the code of any research article to verify that it runs and that the results can be recomputed (Nüst & Eglen, 2021). Articles with verified code can then reference the CODECHECK report, which is uploaded as a public article on Zenodo.org.

For 2024, the Institute for Replication announced that it would reproduce studies from the journal Nature Human Behaviour ("Promoting Reproduction and Replication at Scale," 2024). Nature Human Behaviour is one of the most prestigious journals in research on human behavior — though prestige should not be equated with scientific quality. It is managed by the Springer publishing group and charges the highest article processing fee, around €9,000 per article. The strategic decision to focus on articles from that journal has the advantage that the people conducting the reproducibility checks may be able to publish them there, and that the reproductions receive considerable attention. Given the quality standards such journals claim for themselves, and the fact that free journals like Meta-Psychology can carry out the procedure without external help from the Institute for Replication, we again see the familiar pattern of publishers exploiting their prestige to extract free, profit-generating labor from the scientific community. In the end, it is once again not the journal itself that contributes to scientific quality assurance, but the Institute for Replication.

A shortcut for checking correctness, used by many journals, is the program statcheck. It automatically detects classic statistical tests and checks, based on the reported values, whether they are internally consistent. Hartgerink (2016) checked results from over 50,000 articles with the program and had the articles commented on via PubPeer. Because the algorithm — as openly acknowledged in the comments — occasionally flags values as erroneous incorrectly, and because the authors of the arti-

cles were not warned about the comments beforehand, the DGPs (German Psychological Society) condemned the practice (German-language statement). The responses from the Statcheck group and from Chris Hartgerink are no longer available.

> **Reproduction at the Push of a Button**
>
> *Push-button replications* refer to results that can be recomputed by any researcher with little effort — at the push of a button, as it were. While social science journals increasingly require that data and analysis code be published in a form that allows results to be recomputed, the journal Image Processing Online (IPOL, https://www.ipol.im) embodies the ideal of this approach: for every article published there, a demo is available in which, after selecting an image, the algorithm published in the article is run live.

Large-Scale Reproduction Projects

Various research disciplines have launched large-scale projects to estimate reproducibility for entire disciplines. A pioneer in this field was Höffler's ReplicationWiki (https://replication.uni-goettingen.de). Subsequent projects such as the Replication Network (https://replicationnetwork.com) drew heavily on the data compiled there. For economics, Brodeur, Mikola, et al. (2024) reported a reproducibility rate of 70%, and in management science 55% (Fišar et al., 2024). The Institute for Replication (I4R) also overlaps with the Social Science Reproduction Platform of the Berkeley Initiative for Transparency in the Social Sciences (BITSS; https://www.bitss.org/resources/social-science-reproduction-platform). While I4R published a database of all results in 2024, the BITSS platform

has already been available for some time. Alongside these predominantly economics- and political-science-oriented initiatives, the interdisciplinary Multi100 project investigates the robustness of 100 results from various fields (https://osf.io/q5h2c). A completed project originating in psychology was a university's offer to check researchers' work for reproducibility (D. Baker et al., 2023). In the geosciences, CODECHECK (https://codecheck.org.uk) offers a community in which researchers certify the one-time reproducibility of code. A journal that publishes reproducibility checks is Rescience C (https://rescience.github.io).

Open Code

Publicly available data and code are necessary for reproduction and robustness checks. Journals face a trade-off here between making submission harder — and thus making themselves less attractive — by imposing higher requirements, and promoting scientific quality assurance. A similar problem exists among the operators of panels in which large surveys or performance tests are regularly conducted, such as the international PISA study or the German Socio-Economic Panel (SOEP), a long-running household panel run by DIW Berlin. In published analyses of SOEP data, code is shared in only 20% of cases (German-language source: https://www.wifa.uni-leipzig.de/fileadmin/Fakultät_Wifa/Institut_für_Theoretische_Volkswirtschaftslehre/Professur_Makroökonomik/Economics_Research_Seminar/ERS-Paper_Marcus.pdf).

Robustness Analyses

Similar to sensitivity or robustness analyses in meta-analyses, individual studies can also explore further paths in the "garden of forking paths." As a reminder: the path from data to results is long and involves many different decisions. To show that the result does not actually depend on these decisions, one can demonstrate what the results look like if different decisions had been made. The most extreme form of such robustness analyses is the *multiverse analysis* (see, e.g., Mazei et al., 2025 for recommendations). Here, an attempt is made to make all possible decisions simultaneously. The resulting set of results is then analyzed or presented in some form (e.g., averaged) (Jacobsen et al., 2024). Another option is the *multi-analyst study*. This concerns the dependence of results on the decisions of different researchers, with many people analyzing the data independently of one another. In the end, it is checked how strongly the results agree across researchers.

The figure below shows the results of different analysis methods applied to a fixed dataset (fictional data). Different types of correlations, different sample sizes, and different hypotheses were used. The result changes slightly each time, so that the value ranges between 0.20 and 0.35, but the positive (and significant) correlation is preserved throughout.

Reproducible Manuscripts

In the online version of this book, it is possible to display the code behind a figure before the figure itself. With this code, the figure can easily be reconstructed or reproduced. Research articles, too, can be written in this way. Text, programming code, and results in the form of numbers, tables,

Figure 23: Results of different analysis methods for a fixed, fictional dataset.

and diagrams are all written within a single program, sparing researchers the need to copy or retype numbers. Recomputing results is also made much easier. These programming environments are based on the so-called Markdown language, and countless other programming languages can be embedded within such documents.

While researchers often lack the expertise or time to make their manuscripts reproducible, pilot projects already exist (D. Baker et al., 2023), as do journals that support researchers in doing so (Carlsson et al., 2017). In combination with multiverse analyses, it is also possible, for research articles published online, to write the text so that it reacts interactively to alternative analytical decisions. Readers can thus make decisions within the manuscript and directly see how the results change.

Statistics

Probably every research discipline that uses statistics has been affected by the replication crisis. Very much in the spirit of "post hoc ergo propter hoc" (after this, therefore because of this), this fact is often interpreted to mean that the use of statistical methods is the cause of the replication problems. While counterarguments hold that the methods are simply being used *incorrectly* (Lakens, 2021c), some researchers also propose changes or alternatives. A group of 72 psychologists, for instance, called for lowering the significance threshold for new findings from 5% to 0.5% (Benjamin et al., 2018), making p-hacking harder. Others propose banning null hypothesis significance testing (NHST) altogether and using other methods instead: Wagenmakers (2007) advocates for Bayesian statistics, and the journal Basic and Applied Social Psychology has banned the use of significance tests entirely — a move that may actually have increased the problem of false-positive findings (Fricker Jr et al., 2019).

Open Data and Open Materials

Because of frequent fixed-term contracts and repeated moves between universities, but also because of discontinued doctorates or people leaving academia due to the Wissenschaftszeitvertragsgesetz (Germany's fixed-term academic contract law), projects must often be handed over to other researchers. If the research materials and data are not documented and prepared properly, time and effort are lost in the process. In extreme cases, animals were raised and operated on in a laboratory, and the investigation cannot be continued. Open data and materials are meant to prevent this, to enable collaborative work, and to make errors correctable.

In more extreme cases, researchers attempt to publish articles based on fabricated data. Only where openly accessible data exist can such fraud be detected (Carlisle, 2021).

Numerous studies have already shown that data are frequently not shared even upon request, and that this has not changed in the wake of the replication crisis (Vanpaemel et al., 2015). Moreover, whether data are shared says nothing about whether they contain errors (Claesen et al., 2023). More and more journals require the publication of data (e.g., https://topfactor.org/journals?factor=Data+Transparency), funding bodies require data management plans, tools for automated data preparation are under development (https://leibniz-psychology.org/das-institut/drittmittelprojekte/datawiz-ii), and numerous research data repositories — websites where data can be uploaded and found — have emerged.

The most important prerequisites for researchers to be able to share data are the consent of participants (where applicable), anonymization where necessary (e.g., for data on health or political views), and holding the rights to the data. Participant consent is routinely obtained before a study begins; anonymization takes place either during data collection or afterward (e.g., for qualitative data using AMNESIA); and the rights are usually only lacking when the data were collected on behalf of a company. The most difficult part is probably anonymization. Campbell et al. (2023), for example, report how they anonymized accounts from survivors of sexual assault using a multi-stage process in which names, dates, locations, trauma histories, and other sensitive information were redacted.

Where Are Research Data Uploaded?

Depending on the field and institution, research data are archived in different places. Universities often have their own services, but field-specific repositories are usually better for discoverability: psychologists frequently use the Open Science Framework (OSF), PsychArchives from the Leibniz Institute for Psychology, or Researchbox.org. For the social sciences, the Leibniz Institute for the Social Sciences offers various resources via GESIS. In chemistry, LISTER is software under development that semi-automatically describes data based on electronic lab notebooks. Re3data provides an overview of research data repositories (e.g., by field).

Table 15: Repositories for research data.

Field	Repository
Interdisciplinary, mainly psychology	osf.io
Social sciences	data.gesis.org
Political science and social sciences	icpsr.umich.edu
Life sciences	Pangaea.de
Arts and humanities	de.Dariah.eu
Linguistics	Clarin.eu
Biology	gfbio.org
Materials science	nomad-lab.eu
Qualitative data	qdr.syr.edu
Interdisciplinary	openbis.ch
Interdisciplinary	about.coscine.de
Interdisciplinary	frdr-dfdr.ca/repo

⚠ Where Are Research Data Published?

Repositories allow research data to be published with a mouse click. If additional features such as interactive analyses or peer review are desired, researchers use dedicated tools and specialized journals. Psychological datasets, for example, can be published in the Journal of Open Psychology Data, the R package PsyMetaData (Rodriguez & Williams, 2022) contains data from psychological meta-analyses, and the journal Inggrid publishes data from engineering research. On dedicated websites, researchers can access chat log data voluntarily shared and anonymized by individuals at MOCODA, download data on human cooperation at CODA, or analyze estimation judgments at OpAQ.

Criteria for Preparing Research Data

Merely uploading research data to a website is not enough to make research more transparent. Typically, the data are linked in the research article in which they were used, and a codebook is provided that explains what the various values mean.

The table below shows an excerpt from a fictional dataset with three variables. Real datasets usually contain many more variables (e.g., the Abitur grade broken down by subject, demographic data such as age and gender, date of the survey, and sometimes variables with cryptic names like “V1_Z01” or “V0815”). Here, the three variables (i.e., columns) are “ID,” a sequential number identifying each participant; “IQ,” the measured IQ score from a particular intelligence test; and “Abitur grade,” the grade from participants’ German school-leaving exam, which participants

reported themselves. (German school-leaving grades run from 1.0, the best, to 6.0, the worst — the inverse of a US-style GPA, where a higher number is better.) The codebook contains this information.

Table 16: Example of a dataset.

ID	IQ	Abitur grade
1	103	2.6
2	86	2.4
3	112	1.8

FAIR and CARE

The FAIR principles for research data were developed on an interdisciplinary basis. They recommend that data be archived so as to be findable (**F**), accessible (**A**), interoperable (**I**) across different computer systems, and reusable (**R**). De Waard (2016) structures these requirements as a pyramid, with storage as the foundation, sharing above that, and quality assurance at the top. According to an EU report, the annual cost of data not complying with the FAIR principles amounts to €10.2 billion. Discipline-specific templates for making shared data FAIR are currently being developed by the Center for Open Science (www.cos.io/blog/cedar-embeddable-editor).

The CARE principles go a step further. They were designed for data collection involving Indigenous peoples, and originate from the Indigenous data sovereignty movement (e.g., the Global Indigenous Data Alliance). Building on FAIR, they call for collective benefit (**C**) of the data — for example, usability by society. Local examples from Germany include flood hazard maps (German-language) and the city of Münster's "Cool

City Map" (German-language), which lets residents mark places that stay cool on hot days. The people represented in the data must be given authority to control (**A**) how the data are used — that is, they should have a say in how the data appear. To preserve their self-determination, data should also be shared responsibly (**R**), and their rights and well-being should be central to the research (**E**thics). When non-scientists actively participate in data collection or preparation, this is referred to as citizen science. For example, people can send their collected love letters to the German-language Liebesbriefarchiv (Love Letter Archive), which allows the German language, social conventions, and cultural change to be studied more comprehensively than would be possible using only a handful of famous scholars, or people can operate space telescopes over the internet from their own home computer (http://www.aim-muenster.de). This does not mean that laypeople conduct research on the basis of which they overturn "classic" research findings (Levy, 2022), but rather that they gain insight into the research process and can contribute to knowledge generation under guidance.

i Requesting Research Data from Governments

When governments commission companies to answer research questions, citizens can request the underlying data. In the United Kingdom, for example, there is the platform WhatDoTheyKnow; in Germany, the equivalent is FragDenStaat. In one study, Maier et al. (2024) used such data to check whether recommendations regarding open science practices had been taken into account.

Concerns About Open Data

Data Police

The discourse around open science is at times highly charged: researchers who ask for data, or use data to identify errors, are labeled data parasites, data police, or even data terrorists. The claim that "those who have nothing to hide have nothing to fear" has a dystopian undertone, and in politically charged times, and in the face of plagiarism hunters, researchers are understandably reluctant to make themselves fully transparent. In this context, however, it should be kept in mind that science is not a solitary hobby but a profession with social responsibility. Anyone hiding something here should rightly be suspected of not doing proper science. Fittingly, the outside wall of the University and State Library in Münster bears large red letters reading "Gehorche keinem" — "obey no one." Form your own opinion instead — go to the researchers' work and data and see the truth for yourself. When researchers do not share their data and publish articles behind paywalls, this constitutes an unnecessary obstacle to independent opinion-forming.

Data Theft

Sharing data often conflicts with the goal of maintaining a competitive advantage over other researchers. This means that researchers withhold their data, publish as many articles as possible based on it, and only share the data once there is "nothing left to extract" from it. The concern is that other researchers might be quicker to publish articles based on the data. In practice, however, the data end up not being published at all. It is important to understand, regarding this concern, that *sharing* data does

not mean *giving away* data. When sharing, researchers can attach a license that, for example, specifies how the data must be cited. If other researchers fail to comply, they risk their careers. Furthermore, it is easier to prove who originally collected the data if that person published it early on.

Citation Counts

Open data is sometimes promoted with the argument that it leads to more citations. This motivates some scientists more than good scientific practice does, but is probably not actually true (Colavizza et al., 2020).

> **i** What Does Long-Term Archiving Mean?
>
> Funding bodies sometimes require long-term archiving. Depending on the context, this means that data must be stored and retrievable for 20 or 50 years. A somewhat extreme variant of long-term archiving was carried out with data from Github.io: all data that had been uploaded there as of February 2, 2020 were saved and transported to a disused coal mine in Norway. There they are meant to remain retrievable for up to 1,000 years.

Increasing Replicability

The approaches discussed so far, such as meta-analyses or reproducibility checks, can often be applied to existing projects. The following proposals, however, are more difficult to apply retroactively — they are mainly suited to new research and aim to increase the replicability of newly published studies. This includes stricter methodological standards or replications

carried out before publication ("internal replications"), as is standard practice, for instance, in genetic epidemiology. A positive side effect here is that research groups join forces for replication studies and exchange data (Royal Netherlands Academy of Arts and Sciences, 2018). As discussed in the book's conclusion, it is difficult to make a general, cross-disciplinary statement about whether these proposals actually affect replicability, given how rare replication studies still are. Even in social psychology, replications are currently still the least-implemented measure among all open science recommendations (Glöckner et al., 2024). Either way, the value of measures that generally increase the transparency of research is clear with respect to replication difficulties and p-hacking.

Transparent Reports

Aczel et al. (2020) designed a *Transparency Checklist* that researchers can work through point by point to check whether their research report is transparent. In the online app (https://www.shinyapps.org/apps/TransparencyChecklist/), a report can then be generated and attached to the research article. The checklist is divided into the topics of preregistration, methods, results, and data/code/materials. Regarding results, for example, it asks whether the number of observations was reported for all groups. The checklist is currently available in about 30 languages. The checklist by Aczel et al. (2020) is, however, primarily suited to quantitative studies. For qualitative and mixed-methods studies, Symonds & Tang (2024) have developed an assessment scheme.

A further and shorter variant is the 21-word solution. Here, a proposed statement (Simmons et al., 2012) is included in the report, assuring readers that no studies (or results) were withheld. It is far

> less rigorous and comprehensive than the transparency checklist, but it lowers the barrier to engaging with the transparency of research reports.
> For animal studies, the ARRIVE guidelines were developed. Building on these, the LAG-R standards aim to ensure clear reporting of the genetic make-up of the animals studied (Teboul et al., 2024).

Preregistration

Research is either confirmatory (i.e., researchers have thought through in advance which data they will collect, which analyses they will run, and what the possible results might be) or exploratory (i.e., researchers try to approach a topic without preconceptions, generating questions for future, possibly confirmatory, research). Whenever research is confirmatory, it should be *preregistered*. This means that all the thinking researchers should have done in advance should be written down before data collection begins. Preregistrations are meant to prevent researchers from ending up in a *garden of forking paths* — or rather, they help fix the path in advance, so that it is no longer possible afterward to change the path in order to produce the desired results.

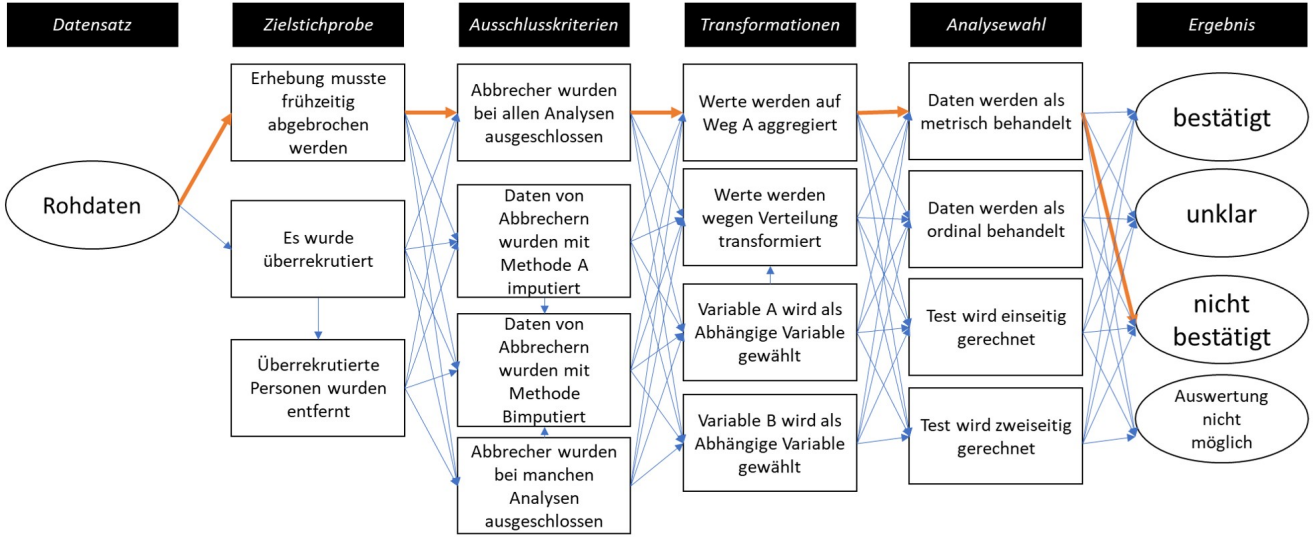

Figure 24: Preregistration fixes the path through the garden of forking paths in advance.

A good preregistration is characterized by researchers stripping themselves of all degrees of freedom (Wicherts, Veldkamp, Augusteijn, Bakker, van Aert, et al., 2016) and specifying in advance all the decisions they will make on the way from data to results. Preregistrations should also be well structured, so that other people can easily verify what was determined beforehand (Simmons et al., 2020). To record all necessary decisions (e.g., how to handle missing values or the planned number of participants) in a structured way, preregistration templates exist. These exist for all kinds of areas, such as social-psychological experiments (van 't Veer & Giner-Sorolla, 2016) or replication studies (Brandt et al., 2014). A collection of more than 20 templates along with their intended uses is available here. Researchers with little experience in a given area are supported by such templates. For instance, preregistration templates for meta-analyses include overviews of literature databases and already incorporate gold standards for transparent reporting (PRISMA, David Moher et al., n.d.). Moreover, the preregistration materials can later be reused for the manuscript itself.

The method of preregistration developed within psychology, drawing on the *registration* of studies involving human participants in medicine and the registration of meta-analyses. Like some other models, it transfers to many other fields without major adaptation — for example, to paleontology (Drage & Wong Hearing, 2023).

Preregistration Should Always Include an Analysis Plan and Analysis Code

Regarding the spread of preregistration, there is an ongoing debate about whether it should be introduced in as simple and resource-efficient a form

as possible, or whether extensive and carefully prepared preregistrations should be required. For instance, the widely used template from aspredicted.org consists of only 11 questions and does not allow analysis code or other files to be attached. Since the purpose of preregistration is to prevent p-hacking, compromises on the scope of preregistrations make little sense.

Analysis code is based on a particular program and corresponding programming language and carries out the preparation of the data (e.g., computing scores, excluding observations) as well as the analyses themselves. Traditionally, it is written after the data have been collected. This can render studies worthless, because it may only become apparent after data collection that the planned analyses are not feasible, and because studies may then be analyzed in whatever way produces the desired results (p-hacking). Akker et al. (2023) were unable to demonstrate, for psychology research articles, that preregistration protects against p-hacking, but they did not distinguish between preregistrations with and without an analysis plan. For Registered Reports, which typically require a fully worked-out analysis plan, Scheel et al. (2021) were able to show that the proportion of hypothesis-confirming results drops sharply. Using thousands of statistical tests from economics research articles, Brodeur, Cook, et al. (2024) showed that preregistration only protects against p-hacking when an analysis plan is present. An analysis *plan* differs from analysis *code* in that the latter is unambiguous. A plan might state, for example, "we will compute the correlation between variable X and variable Y and test whether it is significant." This does not specify which significance level will be used, whether a one-tailed or two-tailed test will be performed, which type of correlation will be computed (Pearson, Spearman, Kendall), how missing values will be handled, and so on. The probability of obtaining a signifi-

cant result when the true correlation is 0 then rises from the nominal significance level of, say, 5% to 10%. This cannot happen with programming code (e.g., "cor.test(x, y)"), since the default settings used when something is left unspecified are clearly defined there. Reviewers of research should therefore pay close attention to exactly what is included in a preregistration (Thibault et al., 2023).

Deviations from Preregistrations

"The more deliberately people proceed, the more effectively chance strikes them" (my translation), as Friedrich Dürrenmatt puts it in point 8 of his *21 Points to* The Physicists (Dürrenmatt, 2012). The same holds for preregistrations: researchers cannot anticipate every eventuality. It must therefore be expected that, due to programming errors, reasoning errors, or previously unconsidered arguments, researchers will need to proceed differently than specified in the preregistration. Deviations from preregistrations are very common (Heirene et al., 2021) — for instance, the actual sample size differs from the target sample size in more than 50% of cases. Once again, transparency is what distinguishes the right way to handle this: every deviation should be clearly communicated, and it should be discussed why the deviation occurred and how it affects the results (e.g., by checking whether results change when the target sample rather than the full, larger-than-planned sample is used) (Heirene et al., 2021; see also Lakens, 2024). Currently, reviewers do not check preregistrations or deviations from them (Syed, 2023; TARG Meta-Research Group and Collaborators et al., 2021). Tools that do this automatically are, however, under development (e.g., RegCheck, https://regcheck.app).

Where Are Studies Preregistered?

A broad infrastructure already exists for preregistering studies. Researchers can, for example, preregister studies via the Open Science Framework (osf.io) and also upload data, materials, analyses, and manuscripts there, as well as later publish preprints. Projects and preregistrations can also remain private or anonymized, and preregistrations can be placed under a multi-year embargo — a period during which the preregistration is not yet public and can only be accessed via a special link. All public preregistrations are automatically indexed by Google Scholar and are thereafter findable, citable, and also visible in the OSF search engine. Within OSF, either open templates or specific existing templates, such as the one by Brandt et al. (2014), can be used for preregistration (further templates will be incorporated in the future).

Another preregistration provider is aspredicted.org, though no files can be attached there, and its 11 questions are primarily suited to classic psychological studies. Medical studies are registered — usually without the details that are important for genuine preregistration — via https://clinicaltrials.gov, and meta-analyses via PROSPERO (http://www.crd.york.ac.uk/PROSPERO/).

Preregistration Checklist

Before conducting the study

- The methodology and analysis plan of the study have been described completely and in a structured way (ideally using a template)
- The analysis script has been written and tested using test data or random numbers, and it works
- The preregistration has been time-stamped and archived before the study (e.g., via OSF.io)

After conducting the study

- A link to the preregistration is included in the manuscript (e.g., in the methods section or in a paragraph on the transparency and openness of the study, ideally paired with links to data and materials)
- All deviations from the preregistration are listed and justified

What If the Desired Results Do Not Come Out?

A question researchers often ask me when discussing preregistration once again demonstrates the discrepancy between *research that is good for science* and *research that is good for one's career*. A preregistration prevents results from being dressed up (p-hacking). This increases the risk of obtaining results that are hard to publish, for example because they are not groundbreaking or not as expected. While a preregistration thus dis-

tinguishes researchers who are genuinely interested in good science, it can currently — in various research disciplines — have negative career consequences for not embellishing one's data.

Planning Sample Size

Another method, recommended mainly for confirmatory research using statistical methods, is planning the sample size. This planning can be based on available resources or other constraints, but in the social sciences and medicine it is often done via power analyses. Statistical power is the probability of detecting an association of a given size if such an association actually exists. The underlying logic is that an association should *not* go undetected simply because the study was not adequately designed to detect it.

Power analyses have been known for a long time, and it is regularly criticized, at least within psychology, that they are conducted too rarely (Cohen, 1988, 1992). While they are still often absent or insufficiently integrated into students' methods training, there are now extensive guides (Lakens, 2021a), software tools (Champely, 2020; Faul et al., 2007; Faul et al., 2009; Zhang & Mai, 2022), and video tutorials. Methods have meanwhile also been developed for alternative significance tests [equivalence tests; Lakens (2017); Lakens et al. (2018)] and for replications (Simonsohn, 2015). A particular type of sample planning is based on Bayesian statistics and is recommended especially in ethically sensitive situations such as animal behavior research: Richter (2024) proposes computing the so-called Bayes factor — which, unlike p-values from significance tests, also converges when there is no true association — after every observation or trial, and concluding the study once a certain

value has been exceeded. It is important to note that this approach is explicitly suited only to Bayesian analyses.

Effect Sizes Instead of P-Values

It is now clear that statistical methods are frequently misunderstood and misused (Gigerenzer, 2004; Lakens, 2021b; Perezgonzalez, 2014). The discussion is currently shifting from p-values (Uygun Tunç et al., 2023) toward effect sizes and the question of what role different values play. For example, Vohs et al. (2021) were able to demonstrate that the controversial effect of diminishing self-control in the *ego depletion paradigm* does exist (i.e., it is not equal to 0) — but in their study it was so small that a successful demonstration alone would require over 6,700 participants, more than any experiment ever conducted on the topic, and almost twice the 3,524 participants used by Vohs et al. (2021). It remains controversial whether researchers should focus on easily demonstrable phenomena, so-called "low-hanging fruit" (R. Baumeister, 2020), whether there is a threshold below which phenomena are practically undetectable and thus of no interest (Primbs et al., 2023), and what practical relevance such small effects have (Anvari et al., 2021).

Qualitative Research

In qualitative research, there is often no clear research question fixed in advance; instead, it is developed over the course of the research process. Alongside anonymization (Campbell et al., 2023), preregistration therefore also plays a special role. Nevertheless, there are typical procedures

— for example, decisions about how to code responses — that can be pre-registered using dedicated templates.

Open Source Software

Research without programs for literature search, reference management, data analysis, and academic writing is unthinkable in most disciplines. Even the systems used by academic journals to structure the submission process (submission, review, revision, publication) are based on software. This relevant area within software is called "research software engineering," and the challenge is that researchers, despite having had little prior contact with writing software (e.g., analysis code), must acquire expertise in this area. This ranges from general recommendations on the "idioms" of programming languages (e.g., the Zen of Python) all the way to collaborative work using systems like GitHub or GitLab.

Open source means that a program's source code is readable and can be freely reused. This makes it possible to trace exactly what the program does. While, for example, the statistical software IBM SPSS has no such visible code, in GNU R every step of a statistical calculation can be traced. Errors that can jeopardize entire branches of science (Soergel, 2014) are easier to detect in open source programs. In cryptography there is talk of Kerckhoffs's principle, which states that even encryption methods are more secure when only the key — not the encryption method itself — is kept secret. For example, software for analyzing fMRI data (a type of "brain scan") was not properly validated until 15 years after its introduction, and its quality turned out to be inadequate (Eklund et al., 2016). Furthermore, proprietary software — software "owned" by a company or an individual — carries the risk of *lock-in*: people come to rely on the

program and align their workflows closely with it. At some point they become dependent on it, for example being able to manage their data only with that one program, or facing enormous costs if they were to switch (Brembs et al., 2023). A license that identifies software as open source is the GNU license. Widely used proprietary software outside of research includes Microsoft Windows and the Office suite, Adobe Acrobat for reading PDF files, or Zoom for video calls. While the German government of the 2021-2025 legislative period set itself the goal of putting software projects out to tender as open source projects, it failed to follow through on this.

In short: for research in which traceability plays a primary role, open source software should be used wherever possible. For some methods this is currently not possible; there, it is the task of university libraries, as custodians of knowledge (Quan, 2021), and of professional societies, as the disciplines' points of contact, to provide or commission the necessary infrastructure.

> **i** A "Copy" of the Open Science Framework
>
> One of the most valuable tools in the social sciences is the Open Science Framework (osf.io). It gives researchers the ability to preregister studies and publish data, research reports, and materials. The entire codebase is available online and was copied for another project (GakuNin RDM) and adapted to its needs. Licensing terms clearly permit the code to be copied and reused. This openness of knowledge makes it possible to create similar offerings without having to start from scratch.

Table 17: Commercial and non-commercial software.

Infrastructure	Commercial Software	Non-Commercial Software
Literature database	SCOPUS	OpenAlex (Priem et al., 2022b)
Reference management	Citavi, Mendeley	Zotero (Puckett, 2011)
Data collection	Unipark, Millisecond Inquisit	PsychoPy (Peirce et al., 2022)
Statistical data analysis	IBM SPSS, Stata	GNU R (R Core Team, 2018), PSPP (Yagnik, 2014), JASP (Love et al., 2019)
Review and publication	Editorial Manager	Open Journal System (Willinsky, 2005)

Slow Science

Many open science developments can be summarized under the heading of "slow science." The traditional mass production of low-quality research articles stands in contrast to a more mindful engagement and thorough scrutiny. Hyman (2024), for example, criticizes the fact that many of the proposed solutions put researchers at a disadvantage in the competition to publish. He proposes instead to solve these problems using artificial intelligence. Publishers such as Elsevier are already using AI for peer review, even though models like ChatGPT 4.0 are not actually suited to this task (Thelwall, 2024).

Big Team Science

Large-scale research projects such as ManyBabies (Byers-Heinlein et al., 2020), replication projects (Open Science Collaboration, 2015), initiatives like FORRT (Azevedo et al., 2019), and software development efforts such as GNU R (R Core Team, 2018) have shown what large groups of researchers are capable of, and that many current problems cannot be solved by individual researchers alone. In physics, the record for the largest number of authors on a single study is held by a paper from CERN on the Higgs boson. About 15 of the 33 pages consist solely of the names of the people involved, with additional pages for their institutions (Aad et al., 2015). In this context, the concept of "authors" is shifting toward that of "contributors" (A. Holcombe, 2019).

To clearly state who did what, researchers can use the Contributor Roles Taxonomy (CRediT, A. O. Holcombe, 2019), which consists of standardized role descriptions. Apps can then be used to generate simple tables or lists showing the contributions of everyone involved (A. O. Holcombe et al., 2020).

Further Reading

- Kepes et al. (2023) have developed a beginner's guide for assessing publication bias.
- For preclinical studies, CAMARADES Berlin maintains a wiki on systematic reviews: https://www.camarades.de.

- Harrer et al. (2021) have written a freely available book on conducting meta-analyses: https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/.
- Adler et al. (2023) have compiled various tools for the heuristic assessment of the trustworthiness of scientific findings. At https://mktg.shinyapps.io/Toolbox_App/, researchers can enter data and run the various procedures.
- Wilson et al. (2017) compile recommendations that all researchers should follow to ensure reproducibility.
- Giehl et al. (2024) discuss the technical hurdles involved in sharing human neuroimaging data.
- The recommendations of the German Psychological Society (DGPs) on handling research data (Schönbrodt et al., 2017) are available online (in German): https://www.uni-muenster.de/imperia/md/content/fb7/ethikkommission/dgps_datenmanagement_deu_9.11.16.pdf.
- https://www.datalad.org is a free program that automatically documents how knowledge is derived from data.
- Typical myths about open data are discussed in a blog post by Elina Takola: https://www.sortee.org/blog/2024/04/12/2024_open_data_myths/.
- With the FAIR-Aware quiz (https://fairaware.dans.knaw.nl), researchers can test their knowledge about sharing research data.
- Open educational materials on FAIR data are provided by Jürgen Schneider and by the University Library of Mannheim.

- Examples of citizen science and ways to get involved can be found at mitforschen.org and Wissenschaft-im-Dialog.de (both German-language).
- An overview and translations of the Contributor Roles Taxonomy are available online: CRediT translation project on GitHub.
- A guide to open source software (in German) is provided by the German IT industry association Bitkom: https://www.bitkom.org/Bitkom/Publikationen/Open-Source-Leitfaden-Praxisempfehlungen-fuer-Open-Source-Software.
- On the YouTube channel "Veritasium," open source software is discussed in the context of a security vulnerability: https://www.youtube.com/watch?v=aoag03mSuXQ.

Theories

In science, theories often serve the function of a signpost. They make it possible to estimate what the results of a study will look like before it is conducted. If the results are therefore not as expected—for example, when a replication fails—there is a suspicion that the theory might be wrong or incomplete. In psychology, where theories are frequently formulated in everyday language rather than formalized (i.e., as mathematical formulas), the value of formal variants has long been debated (Meehl, 2004; Platt, 1964; Wilcox & Rousselet, 2024), and replication failures are attributed to theories being too imprecise, ambiguous, and poorly developed. The prevailing motto is that theories are like toothbrushes, and one should never use someone else's. This problem has existed for decades without much having changed (Lakens, 2023; Muthukrishna & Henrich, 2019). In the meantime, however, there are efforts to solve the problem as well (Buzbas & Devezer, 2023b; Sion & Earp, 2024), for example by highlighting the advantages of established formal theories (Robinaugh et al., 2021), developing tools for formalizing theories (https://www.theorymaps.org), or through the efforts of individual researchers. The value of replication studies is seen, within the context of the "theory crisis," in the fact that only repeated investigation can establish the conditions under which a particular phenomenon can be demonstrated (Buzbas & Devezer, 2023b).

Specification and Formalization

The problem of a theory expressed in everyday language can be illustrated as follows: Wagenmakers et al. (2016) attempted to replicate a study

on the facial feedback hypothesis. The hypothesis states that people find things funnier when they hold a pen in their mouth in a way that engages the muscles also used when grinning, compared to holding the pen differently. To check whether participants held their pens correctly in their mouths, they were filmed in the replication study. This was not the case in the original study. The replication studies failed, and the authors of the original study cited the difference in camera presence as the reason the replication failed (Strack, 2016). Whether and how the presence of a camera in the experimental setup is relevant was not part of the theory or hypothesis. Because the theory is formulated in ordinary language, an infinite number of such "post hoc" explanations can be introduced into the discussion, explanations that relativize any replication failure. Such misunderstandings are rarer with a formal theory.

The fact that theories are hard to understand and not very informative results in most theories never being falsified, yet also never being investigated further ("zombie theories," C. J. Ferguson & Heene, 2012). Something similar applies to psychological scales, most of which are used only once (Elson et al., 2023). Flexible theories also make p-hacking easier: since the theory contains no specification for how the data should be processed, whichever variant confirms the theory can be chosen (Van den Akker et al., 2022).

Adversarial Collaborations

One approach intended to prevent uninformative studies from being conducted and to keep discussions from focusing on differing interpretations of theories is adversarial collaboration (*adversarial collaborations*, Cleeremans, 2022). In this approach, proponents of two incompatible

positions work together, for example to design a study intended to settle the conflict (Cowan et al., 2020; Mellers et al., 2001).

Further Information

- Trafimow & Earp (2016) describe how replications are possible even without a theory.

The World

Having discussed systemic, methodological, and theoretical reasons for the replication crisis, the discussion ends with a general, ontological point: the expectation that findings must be repeatedly reproducible rests on the assumption, or theory, that the world – or at least the particular regularity being studied – is stable. Given fluctuating patterns in the world or in human behavior, replication failures are unsurprising. For psychology, for instance, Smedslund (2015) has criticized this assumption as false and has cited this *historicity* as one of the reasons why psychology cannot be a natural science. Gergen (1973) made a similar argument. Both perspectives may have their origin in the foundational debates of psychology: at the beginning of the 20th century, as psychology gradually distanced itself from philosophy and worked increasingly empirically (that is, through observation rather than mere reasoning), Dilthey & Riedel (1970) divided the natural sciences and the humanities on the grounds that the natural sciences are not concerned with re-experiencing and understanding, but with abstract, general (*nomothetic*) laws. The humanities, by contrast, examine processes within a concrete and unique (*idiographic*) context, together with all their particular features. In the long run, it was the natural-scientific approach that became established, one that had already yielded fruitful insights 50 years earlier (Fechner, 1860).

Chance also plays a special role in generalizing from individual observations to universally valid regularities. What is distinctive about statistical methods, for example, is that uncertainty is expressed as a number and thereby *quantified*. The important question in the context of replication problems is whether differences between findings arise by chance or

because of some actual difference. These two kinds of chance are referred to as ontological and epistemic chance. Ontological chance means that there simply is no reason, whereas with epistemic chance we merely do not know the reason. The outcome of a coin flip – that is, whether the coin lands on heads or tails – is, for example, a case of epistemic chance: if we knew all the relevant parameters (the height of the hand, the force of the throw, which side faces up when tossed, and so on), we could predict with certainty how the coin would land. Incidentally, fair coins mostly land on the same side they started on (Bartoš et al., 2023).

Conclusion and Outlook

Researchers work under considerable pressure, and the academic system rewards innovation and clear-cut results more than it rewards care and openness. In the social sciences, more and more researchers are joining forces to address these problems. Emerging from the trust crisis, large parts of psychology already find themselves in a renaissance. Discussions are being held every day, at the personal, institutional, and political level, about how science should be conducted. It has become clear that, unlike some economic markets, science does not have an "invisible hand" that brings everything into balance – instead, it is researchers themselves who actively work toward the self-correction of science. Before an intervention to improve research can take place, however, an inventory is first needed (where are improvements necessary?), followed by observation. This book offers an inventory of problems and proposed solutions. It is now the role of *meta-science*, that is, science about science, to examine which approaches work, and how well.

Excellent Research Is Open Research

In some areas, Open Science has developed from something positive that was not missed when absent into something taken for granted. In the discourse, this is occasionally referred to as the Open Science Kano model. While this does not hold for all facets of Open Science (for example, not for citizen science), one thing is clear: enabling a (rapid) growth of knowledge requires that knowledge be verifiable. This is only possible with open research reports (Open Access), open data (Open Data), and verifiable calculations (Open Code, reproducibility).

What Has Improved?

I hope the chapters on solutions have made clear that there is no problem for which at least one proposed solution does not already exist, and that these solutions are already being implemented in diverse ways (Korbmacher et al., 2023). For example, guidelines have been developed for replication research, and replications are increasingly seen not as personal attacks but as a useful tool (Nosek et al., 2022). Stricter requirements for reporting statistical tests have had a marked effect on their traceability (Giofrè et al., 2023), preregistration has reduced p-hacking (Brodeur, Cook, et al., 2024), and in some cases extremely high replication rates have been achieved using modern methods (Protzko et al., 2023; see also the critique by Bak-Coleman & Devezer, 2023). The proportion of researchers in psychology using these approaches rose sharply between 2010 and 2020, from 49% to 87% (J. Ferguson et al., 2023).

What Remains to Be Done?

In many disciplines, active engagement with Open Science is still pending. While social and personality psychology play a pioneering role among the social sciences, engagement in consumer psychology, or beyond psychology in medicine, education science, or sociology, only began a few years ago. And even within psychology, substantial gaps remain. For example, only 34 of 88 journals published replication studies, which accounted for just 0.2% of all published studies overall (Clarke et al., 2023; Feldman, 2024).

In general: to move these discussions beyond mere intuition, a meta-scientific evaluation of the new methods is indispensable (Dudda et al., 2024; Phaf, 2024; Soderberg et al., 2021). This brings us back to the *spectrum of possible reactions to replication failures*: should a theory be discarded immediately after a failed replication, or would that be too radical, with a certain persistence on the part of researchers instead helping to establish strong theories (Phaf, 2024)? Interdisciplinary collaboration would also be valuable: for instance, it would save a great deal of effort if terminology were developed across disciplines rather than separately (and with much redundancy) for psychology (Hüffmeier et al., 2016), the humanities (Schöch, 2023), and marketing (Urminsky & Dietvorst, 2024). In some areas it remains entirely unclear what replication success even looks like, making it difficult to judge whether two studies examining the same question arrive at the same result. Even where this is clear, the success rate depends heavily on how replication success is measured (https://forrt.org/FReD/articles/success_criteria.html).

Open Science is occasionally equated with *Open Access*. While Open Science is much more than that, the debate over Open Access is one of the oldest and, to some extent, independent of issues of research integrity. The social dilemma of how researchers can become independent of commercial publishers is still far from resolved. Whether transformative agreements between universities and publishers (for example, Germany's nationwide DEAL agreements, negotiated collectively by German research organisations) will succeed remains uncertain.

Open Washing

Even research that addresses the topic of Open Science sometimes fails to meet the very criteria it calls for. For example, an article by Open Science advocates (Protzko et al., 2023) has drawn criticism because it deviated from its preregistration without transparently reporting this.

What Is Happening Right Now?

The ideas and findings discussed here are a current snapshot of a dynamic discourse. More and more universities are founding Open Science centers and engaging with the associated topics. Germany's national research funder, the DFG, funds a Priority Programme (*Schwerpunktprogramm*) on meta-science and replicability, coordinated from LMU Munich (Ludwig Maximilian University of Munich). The American Center for Open Science is working on numerous projects, and within the European FORRT initiative, projects and teams are being organized. It is now only a matter of time before all researchers become aware of the problems and proposed solutions and take part in improving scholarship.

War and Open Science

A topic still relatively little addressed at present is that of war. Many research institutions are subject to embargoes that require collaborations with researchers from certain countries, such as China or Russia, to be reported or even avoided altogether. Depending on one's epistemological standpoint, different developments are conceivable:

- Positivism: Assuming that there is *one single truth* (Wittgenstein, 1922/2004) that everyone merely needs to uncover, opposing nations must be seen as competitors with whom knowledge should not be shared. Science would therefore need to close itself off toward them. How this could be achieved with respect to online databases is primarily a technical question.

- Relationist: In the sense of Fleck (1935/2015), truth is socially anchored, and different societies can, in theory, operate with different truths. Researchers from opposing nations might, as a result, have no interest at all in each other's viewpoints, or might dismiss them as politically tainted.

- Nationalist: From this perspective, one could argue that a society that funds science should be the primary beneficiary of it.

- Altruistic/idealistic: Drawing on the Mertonian norms (Merton, 1968), science can be understood as a commons belonging to all people, to which no one has privileged access (the communism of scientific knowledge) and from which all people can benefit. Under this view, science does not stop at political borders. From this perspective, Open Science could be understood as a diplomatic strategy for scientific exchange despite political differences.

Which perspective will ultimately prevail remains uncertain. In the context of the war in Israel and Gaza and an open letter from

> researchers on the subject, the German Federal Ministry of Education and Research compiled, in June 2024, lists of university faculty with particular political views and examined the possibility of cutting their research funding. This funding affair is perceived across the academic community – irrespective of its political background – as a threat to academic freedom, and is more compatible with nationalist and positivist perspectives.

There is a tension in the communication between researchers and society: among themselves, researchers must communicate that reforms and improvements in science are necessary – which is most easily achieved by pointing to problems. In dialogue with society, however, pointing to these problems can undermine trust in science (Penders, 2024). It is therefore worth emphasizing that science is not finished (indeed, strictly speaking, it is *never* finished), but this does not mean it is any less trustworthy than rumors, conspiracy theories, individual experiences, or religion. Rather, a self-critical stance is a central component of science and a sign that its self-correction mechanisms are working.

Further Information

- Retractionwatch reports on current news about errors in research: https://retractionwatch.com.
- On her YouTube channel, Mai Thi Nguyen-Kim, a German science communicator, discusses problems such as publication bias and self-correction mechanisms (in German): https://www.youtube.com/watch?v=DHyRaUeHcGY.

- Sabine Hossenfelder addresses, in a YouTube video, the question of how climate scientists (should) communicate their findings: https://www.youtube.com/watch?v=gMOjD_Lt8qY.
- The German network of Open Science initiatives connects researchers for talks and workshops on Open Science: OSF wiki of the network.
- An introduction to Open Science for students is offered by C. Pennington (2023), *Matters of Significance* (2024), and Chambers (2017).
- A glossary of Open Science terms is available online via FORRT: https://forrt.org/glossary/english.
- Open Science is not yet the norm; Kohrs et al. (2023) summarize how everyone can contribute to it.

References

Aad, G., Abbott, B., Abdallah, J., Abdinov, O., Aben, R., Abolins, M., AbouZeid, O., Abramowicz, H., Abreu, H., & Abreu, R. (2015). Combined measurement of the higgs boson mass in pp collisions at s= 7 and 8 TeV with the ATLAS and CMS experiments. *Physical Review Letters*, *114*(19), 191803.

Abduh, A. J. (2023). *A critical analysis of the world's top 2% most influential scientists: Examining the limitations and biases of highly cited researchers lists*. https://doi.org/10.22541/au.167435298.80209125/v1

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, *6*, 1–8. https://doi.org/10.1186/s41073-021-00118-2

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, D. S., Morey, C. C., Munafò, M., Newell, B. R., … Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*(1), 4–6. https://doi.org/10.1038/s41562-019-0772-6

Adler, S. J., Röseler, L., & Schöniger, M. K. (2023). A toolbox to evaluate the trustworthiness of published findings. *Journal of Business Research*, *167*, 114189. https://doi.org/10.1016/j.jbusres.2023.114189

Akker, O. R. van den, Assen, M. A. van, Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2023). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*, 1–10. https://doi.org/10.31222/osf.io/fhdbs

Albert, H. (2010). *Traktat über kritische vernunft* (Nachdr. d. 5., verb. und erw. Aufl., Vol. 1609). Mohr Siebeck.

Allbritton, D., Gómez, P., Angele, B., Vasilev, M., & Perea, M. (2024). Breathing life into meta-analytic methods. *J. Cogn.*, *7*(1), 61. https://doi.org/10.5334/joc.389

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., … Zwaan, R. A. (2014). Registered replication report: Schooler and engstler-schooler (1990). *Perspectives on Psychological Science : A Journal of*

the Association for Psychological Science, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

Ankel-Peters, J., Fiala, N., & Neubauer, F. (2023). Do economists replicate? *Journal of Economic Behavior & Organization*, *212*, 219–232. https://doi.org/10.1016/j.jebo.2023.05.009

Anvari, F., Kievit, R., Lakens, D., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2021). *Evaluating the practical relevance of observed effect sizes in psychological research*. https://doi.org/10.31234/osf.io/g3vtr

Aquarius, R., Schoeters, F., Wise, N., Glynn, A., & Cabanac, G. (2025). The existence of stealth corrections in scientific literature—a threat to scientific integrity. *Learn. Publ.*, *38*(2). https://doi.org/10.1002/leap.1660

Armengou, C., Bargheer, M., Gingold, A., Holsinger, S., Laakso, M., Mitchell, D., Mounier, P., Pölönen, J., Rooryck, J., Ševkušić, M., Souyioultzoglou, I., & Varachkina, H. (2024). *Operational diamond OA criteria for journals*. Zenodo. https://doi.org/10.5281/zenodo.12721408

Azevedo, F., Parsons, S., Micheli, L., Strand, J. F., Rinke, E. M., Guay, S., Elsherif, M. M., Quinn, K. A., Wagge, J. R., Steltenpohl, C. N., Kalandadze, T., Vasilev, M. R., Oliveira, C. M., Aczel, B., Miranda, J. F., Baker, B. J., Galang, C. M. O., Pennington, C. R., Marques, T., … FORRT. (2019). *Introducing a framework for open and reproducible research training (FORRT)*. https://doi.org/10.2218/eorc.2022.6968

Bak-Coleman, J., & Devezer, B. (2023). *Causal claims about scientific rigor require rigorous causal evidence.* https://doi.org/10.31234/osf.io/5u3kj

Baker, D., Berg, M., Hansford, K., Quinn, B., Segala, F. G., & English, E. (2023). *ReproduceMe: Lessons from a pilot project on computational reproducibility*. https://doi.org/10.31234/osf.io/k8d4u

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594. https://doi.org/10.1037/a0023489

Banker, S., Ainsworth, S. E., Baumeister, R. F., Ariely, D., & Vohs, K. D. (2017). The sticky anchor hypothesis: Ego depletion increases susceptibility to situational cues. *Journal of Behavioral Decision Making*, *87*(1), 23. https://doi.org/10.1002/bdm.2022

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Bartoš, F., Sarafoglou, A., Godmann, H. R., Sahrani, A., Leunk, D. K., Gui, P. Y., Voss, D., Ullah, K., Zoubek, M. J., Nippold, F., Aust, F., Vieira, F. F., Islam, C.-G., Zoubek, A. J., Shabani, S., Petter, J., Roos, I. B., Finnemann, A., Lob, A. B., … Wagenmakers, E.-J. (2023). *Fair coins tend to land on the same side they started: Evidence from 350,757 flips*. https://doi.org/10.1080/01621459.2025.2516210

Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.2720

Baumeister, R. (2020). *Do effect sizes in psychology laboratory experiments mean anything in reality?* https://doi.org/10.31234/osf.io/mpw4t

Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(4), 574–575. https://doi.org/10.1177/1745691616652878

Beall, J. (2017). What I learned from predatory publishers. *Biochem. Med. (Zagreb)*, *27*(2), 273–278. https://doi.org/10.11613/bm.2017.029

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200. https://doi.org/10.1037/h0024835

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. de, Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bik, E. M., Casadevall, A., & Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, *7*(3). https://doi.org/10.1128/mbio.00809-16

Bilder, G., Lin, J., & Neylon, C. (2020). *The principles of open scholarly infrastructure*. The Principles of Open Scholarly Infrastructure.

Bohorquez, N. G., Weerasuriya, S., Brain, D., Senanayake, S., Kularatna, S., & Barnett, A. (2024). *Health and medical researchers are willing to trade their results for journal prestige: Results from a discrete choice experiment*. https://doi.org/10.31219/osf.io/uwt3b

Bol, T. (2023). Gender inequality in cum laude distinctions for PhD students. *Sci. Rep.*, *13*(1), 20267. https://doi.org/10.31235/osf.io/s5b6j

Borgstede, M., & Scholz, M. (2021). Quantitative and qualitative approaches to generalization and replication-a representationalist view. *Frontiers in Psychology*, *12*, 605191. https://doi.org/10.3389/fpsyg.2021.605191

Boulter, E., Colombelli, J., Henriques, R., & Féral, C. C. (2022). The LEGO® brick road to open science and biotechnology. *Trends Biotechnol.*, *40*(9), 1073–1087. https://doi.org/10.1016/j.tibtech.2022.02.003

Bourazas, K., Consonni, G., & Deldossi, L. (2024). Bayesian sample size determination for detecting heterogeneity in multi-site replication studies. *TEST*, 1–20. https://doi.org/10.1007/s11749-023-00916-4

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G. H., Cornelissen, G., Døssing, F. S., Espín, A. M., Evans, A. M., Ferreira-Santos, F., Fiedler, S., Flegr, J., Ghaffari, M., Glöckner, A., Goeschl, T., Guo, L., Hauser, O. P., … Wollbrant, C. E. (2017). Registered replication report: Rand, greene, and nowak (2012). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Boyce, V., Mathur, M., & Frank, M. C. (2023). Eleven years of student replication projects provide evidence on the correlates of replicability in psychology. *Royal Society Open Science*, *10*(11), 231240. https://doi.org/10.1098/rsos.231240

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

Brembs, B. (2018). Prestigious science journals struggle to reach even average reliability. *Frontiers in Human Neuroscience*, *12*, 37. https://doi.org/10.3389/fnhum.2018.00037

Brembs, B., Huneman, P., Schönbrodt, F., Nilsonne, G., Susi, T., Siems, R., Perakakis, P., Trachana, V., Ma, L., & Rodriguez-Cuadrado, S. (2023). Replacing academic journals. *Royal Society Open Science*, *10*(7). https://doi.org/10.1098/rsos.230206

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(44), e2203150119. https://doi.org/10.1073/pnas.2203150119

Brockman, J. (2022). *Adversarial collaboration: An EDGE lecture by daniel kahneman*. https://www.edge.org/adversarial-collaboration-daniel-kahneman

Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2024). Do preregistration and preanalysis plans reduce p-hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement. *Journal of Political Economy Microeconomics*, *2*(3), 527–561. https://doi.org/10.1086/730455

Brodeur, A., Dreber, A., Hoces de la Guardia, F., & Miguel, E. (2024). Reproduction and replication at scale. *Nature Human Behaviour*, *8*(1), 2–3. https://doi.org/10.1038/s41562-023-01807-2

Brodeur, A., Mikola, D., & Cook, N. (2024). *Mass reproducibility and replicability: A new hope*. https://doi.org/10.2139/ssrn.4790780

Brohmer, H., Hofer, G., Bauch, S. A., Beitner, J., Berkessel, J., Corcoran, K., Garcia, D., Gruber, F. M., Giuliani, F., Jauk, E., Krammer, G., Malkoc, S., Metzler, H., Mües, H. M., Otto, K., Rahal, R.-M., Salwender, M., Stahlberg, D., Sczesny, S., … Athenstaedt, U. (2024). *Effects of the generic masculine and its alternatives in germanophone countries - a multi-lab replication and extension of stahlberg, sczesny, and braun, 2001*. https://doi.org/10.31219/osf.io/q56rk

Buttliere, B. (2024). Was this registered report pilot tested? Examination of vaidis, sleegers, van leeuwen, DeMarree, ... & priolo, d. (2024). In *PsyArXiv*. https://doi.org/10.31234/osf.io/c6r8x

Buzbas, E. O., & Devezer, B. (2023b). Tension between theory and practice of replication. *J. Trial Error*, *4*(1), 73–81. https://doi.org/10.36850/mr9

Buzbas, E. O., & Devezer, B. (2023a). Tension between theory and practice of replication. *Journal of Trial & Error*, *4*(1).

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., & Liu, L. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 349. https://doi.org/10.31234/osf.io/dmhk2

Byrne, J. A., & Christopher, J. (2020). Digital magic, or the dark arts of the 21st century-how can journals and peer reviewers detect manuscripts and publications from paper mills? *FEBS Letters*, *594*(4), 583–589. https://doi.org/10.1002/1873-3468.13747

C. Chang, A., & Li, P. (2022). Is economics research replicable? Sixty published papers from thirteen journals say "often not." *Crit. Fin. Rev.*, *11*(1), 185–206. https://doi.org/10.1561/104.00000053

Cabanac, G., & Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, *72*(12), 1461–1476. https://doi.org/10.1002/asi.24495

Cabanac, G., Labbé, C., & Magazinov, A. (2021). *Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals*. https://doi.org/10.48550/arXiv.2107.06751

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, *8*(2), 197. https://doi.org/10.1086/208856

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science (New York, N.Y.)*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, R., Javorka, M., Engleton, J., Fishwick, K., Gregory, K., & Goodman-Williams, R. (2023). Open-science guidance for qualitative research: An empirically validated approach for de-identifying sensitive narrative data. *Advances in Methods and Prac-*

tices in Psychological Science, *6*(4), Article 25152459231205832. https://doi.org/10.1177/25152459231205832

Carlisle, J. (2021). False individual patient data and zombie randomised controlled trials submitted to anaesthesia. *Anaesthesia*, *76*(4), 472–479. https://doi.org/10.1111/anae.15263

Carlsson, R., Batinović, L., Hyltse, N., Kalmendal, A., Nordström, T., & Topor, M. (2024). *A beginner's guide to open and reproducible systematic reviews in psychology*. https://doi.org/10.31234/osf.io/59vsw

Carlsson, R., Danielsson, H., Heene, M., Innes-Ker, Å., Lakens, D., Schimmack, U., Schönbrodt, F. D., van Asssen, M., & Weinstein, Y. (2017). Inaugural editorial of meta-psychology. *Meta-Psychology*, *1*, a1001. https://doi.org/10.15626/MP2017.1001

Carriquiry, A. L., Daniels, M. J., & Reid, N. (2023). Editorial: Special issue on reproducibility and replicability. *Stat. Sci.*, *38*(4). https://doi.org/10.1214/23-sts909

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Chambers, C. D. (2017). *The seven deadly sins of psychology*. Princeton University Press. https://doi.org/10.2307/j.ctvc779w5

Chambers, C. D. (2018). *Registered reports guidelines JESP 2018*.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Champely, S. (2020). *Pwr: Basic functions for power analysis. R package version 1.3-0*. https://CRAN.R-project.org/package=pwr

Chandrashekar, S. P., Adelina, N., Zeng, S., Chiu, Y. Y. E., Leung, G. Y. S., Henne, P., Cheng, B. L., & Feldman, G. (2023). Defaults versus framing: Revisiting default effect and framing effect with replications and extensions of johnson and goldstein (2003) and johnson, bellman, and lohse (2002). *Meta-Psychology*, *7*. https://doi.org/10.31234/osf.io/nymh2

Chandrashekar, S. P., & Feldman, G. (2025). *On the process and value of direct close replications: Reply to shafir and cheek (2024) commentary on chandrashekar et al.(2021)*. https://doi.org/10.1017/jdm.2025.10011

Charlton, A. (2022). *Replications of marketing studies*. https://openmkt.org/research/replications-of-marketing-studies/

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutog˘lu, B., Bahník, Š., Bowen, J. D., Bredow, C. A., Bromberg, C., Caprariello, P. A., Carcedo, R. J., Carson, K. J., Cobb, R. J., Collins, N. L., Corretti, C. A., DiDonato, T. E., Ellithorpe, C., Fernández-Rouco, N., Fuglestad, P. T., … Yong, J. C. (2016). Registered replication report: Study 1 from finkel, rusbult, kumashiro, & hannon (2002). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(5), 750–764. https://doi.org/10.1177/1745691616664694

Choi, H. H. (2023). In defense of the resampling account of replication. *Journal of Theoretical and Philosophical Psychology*, *43*(4), 249–251. https://doi.org/10.1037/teo0000224

Claesen, A., Vanpaemel, W., Maerten, A.-S., Verliefde, T., Tuerlinckx, F., & Heyman, T. (2023). Data sharing upon request and statistical consistency errors in psychology: A replication of wicherts, bakker and molenaar (2011). *Plos One*, *18*(4), e0284243. https://doi.org/10.1371/journal.pone.0284243

Clarke, B., Lee, P. Y., Schiavone, S. R., Rhemtulla, M., & Vazire, S. (2023). The prevalence of direct replication articles in top-ranking psychology journals. In *PsyArXiv*. https://doi.org/10.31234/osf.io/sa6rc

Cleeremans, A. (2022). Theory as adversarial collaboration. *Nature Human Behaviour*, *6*(4), 485–486. https://doi.org/10.1038/s41562-021-01285-4

Cobey, K. D., Fehlmann, C. A., Christ Franco, M., Ayala, A. P., Sikora, L., Rice, D. B., Xu, C., Ioannidis, J. P., Lalu, M. M., & Ménard, A. (2023). Epidemiological characteristics and prevalence rates of research reproducibility across disciplines: A scoping review of articles published in 2018-2019. *Elife*, *12*, e78518. https://doi.org/10.7554/elife.78518

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(3), 409–410. https://doi.org/10.1037//0033-2909.112.3.409

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PloS One*, *15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Cole, N. L., Kormann, E., Klebel, T., Apartis, S., & Ross-Hellauer, T. (2024). The societal impact of open science: A scoping review. *Royal Society Open Science*, *11*(6), 240286. https://doi.org/10.31235/osf.io/tqrwg

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L. G., Willis, M. L., Foroni, F., Reggev, N., Mokady, A., Forscher, P. S., Hunter, J. F., Kaminski, G., Yüvrük, E., Kapucu, A., Nagy, T., Hajdu, N., Tejada, J., Freitag, R. M. K., … Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the many smiles collaboration. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01458-9

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., Arantes, P., Athanasopoulou, A., Baese-Berk, M. M., Bailey, G., Sangma, C. B. A., Beier, E. J., Benavides, G. M., Benker, N., BensonMeyer, E. P., … Roettger, T. B. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231162567

Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., Naveh-Benjamin, M., Barrouillet, P., Camos, V., & Logie, R. H. (2020). How do scientific views change? Notes from an extended adversarial collaboration. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *15*(4), 1011–1025. https://doi.org/10.1177/1745691620906415

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science*, *284*(5420), 1670–1672. https://doi.org/10.1126/science.284.5420.1670

Cronbach, L. J., Snow, R. E., & Wiley, D. E. (1991). *Improving inquiry in social science: A volume in honor of lee j. cronbach*. Lawrence Erlbaum Associates.

Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., Buczny, J., Chen, Z., Cristofaro, V. de, Vries, L. de, Dewitte, S., Giacomantonio, M., Gong, R., Homan, M., Imhoff, R., Ismail, I., Jia, L., Kubiak, T., Lange, F., … Zinkernagel, A. (2020). A multilab replication of the ego depletion effect. *Social Psychological and Personality Science*, 194855061988770. https://doi.org/10.1177/1948550619887702

David Moher, Larissa Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, & Lesley A Stewart. (n.d.). *Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015 statement*. https://doi.org/10.1186/2046-4053-4-1

De Waard, A. (2016). Research data management at elsevier: Supporting networks of data and workflows. *Information Services & Use*, *36*(1-2), 49–55. https://doi.org/10.3233/isu-160805

DeStefano, F., & Shimabukuro, T. T. (2019). The MMR vaccine and autism. *Annual Review of Virology*, *6*(1), 585–600. https://doi.org/10.1146/annurev-virology-092818-015515

Deutsche Forschungsgemeinschaft. (2022a). *Open science als teil der wissenschaftskultur. Positionierung der deutschen forschungsgemeinschaft*. Zenodo. https://doi.org/10.5281/zenodo.7193838

Deutsche Forschungsgemeinschaft. (2022b). *Guidelines for Safeguarding Good Research Practice. Code of Conduct*. Deutsche Forschungsgemeinschaft. https://doi.org/10.5281/zenodo.6472827

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), 200805. https://doi.org/10.1098/rsos.200805

DH.NRW | AG Openness. (2023). *Open-Access-Strategie der hochschulen des landes NRW (version verabschiedete fassung vom 9. August 2023)*. Zenodo. https://doi.org/10.5281/zenodo.8322048

Dijksterhuis, A. (2014). Welcome back theory! *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(1), 72–75. https://doi.org/10.1177/1745691613513472

Dilthey, W., & Riedel, M. (1970). *Der aufbau der geschichtlichen welt in den geisteswissenschaften*. Suhrkamp Frankfurt am Main.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS One*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

Drage, H., & Wong Hearing, T. (2023). Diamond open access with preregistration: A new publishing model for palaeontology. In *EarthArXiv*. https://doi.org/10.31223/x5kh30

Dudda, L., Kormann, E., Kozula, M., DeVito, N. J., Klebel, T., Dewi, A. P. M., Spijker, R., Stegeman, I., Van den Eynden, V., Ross-Hellauer, T., & Leeflang, M. (2024).

Open science interventions to improve reproducibility and replicability of research: A scoping review preprint. In *BITSS*. https://doi.org/10.31222/osf.io/a8rmu

Dürrenmatt, F. (2012). *Die physiker: Eine komödie in zwei akten*. Diogenes Verlag AG.

Easley, R. W., & Madden, C. S. (2013). Replication revisited: Introduction to the special section on replication in business research. *J. Bus. Res.*, *66*(9), 1375–1376. https://doi.org/10.1016/j.jbusres.2012.05.001

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016). Registered replication report: Hart & albarracín (2011). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(1), 158–171. https://doi.org/10.1177/1745691615605826

Eggertson, L. (2010). Lancet retracts 12-year-old article linking autism to MMR vaccines. *Canadian Medical Association Journal (CMAJ)*, *182*. https://doi.org/10.1503/cmaj.109-3179

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Elsherif, M. M., Middleton, S. L., Phan, J. M., Azevedo, F., Iley, B. J., Grose-Hodge, M., Tyler, S. L., Kapp, S. K., Gourdon-Kanhukamwe, A., Grafton-Clarke, D., Yeung, S. K., Shaw, J. J., Hartmann, H., & Dokovova, M. (2022). *Bridging neurodiversity and open scholarship: How shared values can guide best practices for research integrity, social justice, and principled education*. https://doi.org/10.31222/osf.io/k7a9p

Elsherif, M., Feldman, G., & Yeung, S. K. (2023). *Quantitative manuscript peer review template*. OSF.

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren't toothbrushes. *Communications Psychology*, *1*(1). https://doi.org/10.1038/s44271-023-00026-9

Ensinck, E. N. F., & Lakens, D. (2023). An inception cohort study quantifying how many registered studies are published. In *PsyArXiv*. https://doi.org/10.31234/osf.io/5hkjz

Erdfelder, E., & Heck, D. W. (2019). Detecting evidential value and p-hacking with the p-curve tool. *Zeitschrift für Psychologie*, *227*(4), 249–260. https://doi.org/10.1027/2151-2604/a000383

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLIFE*, *10*. https://doi.org/10.7554/eLife.71601

Etzel, F., Seyffert-Müller, A., Schönbrodt, F. D., Kreuzer, L., Gärtner, A., Knischewski, P., & Leising, D. (2024). *Inter-rater reliability in assessing the methodological quality of research papers in psychology*. https://doi.org/10.31234/osf.io/4w7rb

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fanelli, D., Schleicher, M., Fang, F. C., Casadevall, A., & Bik, E. M. (2022). Do individual and institutional predictors of misconduct vary by country? Results of a matched-control analysis of problematic image duplications. *PLoS One*, *17*(3), e0255334. https://doi.org/10.1371/journal.pone.0255334

Fanelli, D., Tan, P. B., Amaral, O. B., & Neves, K. (2022). *A metric of knowledge as information compression reflects reproducibility predictions in biomedical experiments*. https://doi.org/10.31222/osf.io/5r36g

Farrow, R., Pitt, R., & Weller, M. (2020). Open textbooks as an innovation route for open science pedagogy. *Educ. Inf.*, *36*(3), 227–245. https://doi.org/10.3233/efi-190260

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fecher, B., & Friesike, S. (2014). *Open science: One term, five schools of thought*. Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2

Fechner, G. T. (1860). *Elemente der psychophysik*. Breitkopf und Härtel.

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Feldman, G. (2021). *Replications and extensions of classic findings in judgment and decision making*. https://doi.org/10.17605/OSF.IO/5Z4A8

Feldman, G. (2024). *The value of replications goes beyond replicability and is tied to the value of the research it replicates: Commentary on isager et al. (2024)*. OSF. https://doi.org/10.15626/mp.2024.4326

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Ferguson, J., Littman, R., Christensen, G., Paluck, E. L., Swanson, N., Wang, Z., Miguel, E., Birke, D., & Pezzuto, J.-H. (2023). Survey of open science practices and attitudes in the social sciences. *Nat. Commun.*, *14*(1), 5401. https://doi.org/10.1038/s41467-023-41111-1

Feyerabend, P. K. (1975/2002). *Against method* (Reprinted der 3. ed. 1993). Verso.

Fiorentino, C., & Montana Hoyos, C. (2014). The emerging discipline of biomimicry as a design paradigm shift. *Int. J. Des. Objects*, *8*(1), 1–15.

Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A. I., & Management Science Reproducibility Collaboration. (2024). Reproducibility in management science. *Management Science*, *70*(3), 1343–1356. https://doi.org/10.2139/ssrn.4620006

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Robumeta: Robust variance meta-regression*. https://doi.org/10.32614/cran.package.robumeta

Fleck, L. (1935/2015). *Entstehung und entwicklung einer wissenschaftlichen tatsache [formation and development of a scientific fact]: Einführung in die lehre vom denkstil und denkkollektiv [introduction to thinking style and thinking collective]* (10. Auflage, Vol. 312). Suhrkamp.

Fletcher, S. C. (2021). The role of replication in psychological science. *European Journal for Philosophy of Science*, *11*(1), 23. https://doi.org/10.1007/s13194-020-00329-2

Fletcher, S. C. (2022). Replication is for meta-analysis. *Philosophy of Science*, *89*(5), 960–969. https://doi.org/10.1017/psa.2022.38

Forster, N., & Lund, D. W. (2018). Identifying and dealing with functional psychopathic behavior in higher education. *Glob. Bus. Organ. Excel.*, *38*(1), 22–31. https://doi.org/10.1002/joe.21897

Francis, Z., Milyavskaya, M., Lin, H., & Inzlicht, M. (2018). Development of a within-subject, repeated-measures ego-depletion paradigm. *Social Psychology*, *49*(5), 271–286. https://doi.org/10.1027/1864-9335/a000348

Frank, M. C. (2019). N-best evaluation for academic hiring and promotion. In *PsyArXiv*. https://doi.org/10.31234/osf.io/tzaqh

Franzen, D. L., Carlisle, B. G., Salholz-Hillel, M., Riedel, N., & Strech, D. (2023). Institutional dashboards on clinical trial transparency for univer-

sity medical centers: A case study. *PLoS Medicine*, *20*(3), e1004175. https://doi.org/10.1371/journal.pmed.1004175

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2020). Preprinting the COVID-19 pandemic. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/2020.05.22.111294

Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, *43*(1), 147–165. https://doi.org/10.31235/osf.io/5bck9

Fricker Jr, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in basic and applied social psychology after their p-value ban. *The American Statistician*, *73*(sup1), 374–384. https://doi.org/10.1080/00031305.2018.1537892

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2018). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*. https://doi.org/10.1177/1088868318762183

Gärtner, A., Leising, D., & Schönbrodt, F. D. (2022). Responsible research assessment II: A specific proposal for hiring and promotion in psychology. In *PsyArXiv*. https://doi.org/10.31234/osf.io/5yexm_v2

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309.

Giehl, K., Mutsaerts, H.-J., Aarts, K., Barkhof, F., Caspers, S., Chetelat, G., Colin, M.-E., Düzel, E., Frisoni, G. B., & Ikram, M. A. (2024). Sharing brain imaging data in the open science era: How and why? *The Lancet Digital Health*, *6*(7), e526–e535.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Giofrè, D., Boedker, I., Cumming, G., Rivella, C., & Tressoldi, P. (2023). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices: Five years later. *Behav. Res. Methods*, *55*(7), 3845–3854.

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*(8), 711–721. https://doi.org/10.1017/s1930297500004149

Glöckner, A., Gollwitzer, M., Hahn, L., Lange, J., Sassenberg, K., & Unkelbach, C. (2024). Quality, replicability, and transparency in research in social psychology: Implementation of recommendations in germany. *Social Psychology*, *55*(3), 134–147. https://doi.org/10.1027/1864-9335/a000548

Gopalakrishna, G., Riet, G. ter, Cruyff, M. J. L. F., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. (2021). *Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the netherlands*. https://doi.org/10.31222/osf.io/vk9yt

Gopalakrishna, G., Wicherts, J. M., Vink, G., Stoop, I., van den Akker, O., Riet, G. ter, & Bouter, L. (2021). *Prevalence of responsible research practices and their potential explanatory factors: A survey among academic researchers in the netherlands*. https://doi.org/10.31222/osf.io/xsn94

Gould, E., Fraser, H., Parker, T., Nakagawa, S., Griffith, S., Vesk, P., Fidler, F., Abbey-Lee, R., Abbott, J., Aguirre, L., Alcaraz, C., Altschul, D., Arekar, K., Atkins, J., Atkinson, J., Barrett, M., Bell, K., Bello, S., Berauer, B., … Tompkins, E. (2023). *Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology*. https://doi.org/10.32942/X2GG62

Grant, S., Corker, K. S., Mellor, D. T., Stewart, S. L., Cashin, A., Lagisz, M., Mayo-Wilson, E., Moher, D., Umpierre, D., & Barbour, V. (n.d.). *TOP 2025: An update to the transparency and openness promotion guidelines*. https://doi.org/10.52843/cassyni.h54qxk

Greenwald, A. G. (1976). An editorial. *Journal of Personality and Social Psychology*, *33*(1), 1–7. https://doi.org/10.1037/h0078635

Grossmann, A., & Brembs, B. (2021). Current market rates for scholarly publishing services. *F1000Res.*, *10*, 20. https://doi.org/10.12688/f1000research.27468.2

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., Ridder, D. T. D. de, Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*(4), 495–525. https://doi.org/10.1037/a0019486

Hanson, M. A., Barreiro, P. G., Crosetto, P., & Brockington, D. (2023). *The strain on scientific publishing*. https://doi.org/10.48550/arXiv.2309.15884

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hardwicke, T. E., & Vazire, S. (2023). Transparency is now the default at psychological science. *Psychol. Sci.*, 9567976231221573. https://doi.org/10.1177/09567976231221573

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2021). *Doing meta-analysis with r: A hands-on guide*. Chapman; Hall/CRC.

Hartgerink, C. H. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results. *Data*, *1*(3), 14. https://doi.org/10.3390/data1030014

Haustein, S., Schares, E., Alperin, J. P., Hare, M., Butler, L.-A., & Schönfelder, N. (2024). *Estimating global article processing charges paid to six publishers for open access between 2019 and 2023*. https://arxiv.org/abs/2407.16551

Heathers, J. (2024). *How much science is fake?* https://doi.org/10.17605/OSF.IO/5RF2M

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*(4), 299–332. https://doi.org/10.3102/10769986021004299

Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. M. (2021). *Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison*. https://doi.org/10.31234/osf.io/nj4es

Henderson, E. L., & Chambers, C. D. (2022). Ten simple rules for writing a registered report. *PLoS Comput. Biol.*, *18*(10), e1010571. https://doi.org/10.1371/journal.pcbi.1010571

Hicks, D., Wouters, P., Waltman, L., Rijcke, S. de, & Rafols, I. (2015). Bibliometrics: The leiden manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Himmelstein, D. S., Romero, A. R., Levernier, J. G., Munro, T. A., McLaughlin, S. R., Greshake Tzovaras, B., & Greene, C. S. (2018). Sci-Hub provides access to nearly all scholarly literature. *Elife*, *7*. https://doi.org/10.7287/peerj.preprints.3100v2

Hoebel, M., Durglishvili, A., Reinold, J., & Leising, D. (2022). Sexual harassment and coercion in german academia: A large-scale survey study. *Sexual Offending: Theory, Research, and Prevention*, *17*. https://doi.org/10.5964/sotrap.9349

Höffler, J. H. (2017). ReplicationWiki: Improving transparency in social sciences research. *D-Lib Magazine*, *23*(3), 1. https://doi.org/10.1045/march2017-hoeffler

Holcombe, A. (2019). Farewell authors, hello contributors. *Nature*, *571*(7763), 147–148. https://doi.org/10.1038/d41586-019-02084-8

Holcombe, A. O. (2019). Contributorship, not authorship: Use CRediT to indicate who did what. *Publications*, *7*(3), 48. https://doi.org/10.3390/publications7030048

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLoS One*, *15*(12), e0244611. https://doi.org/10.1371/journal.pone.0244611

Holford, D., McLean, J., Holcombe, A. O., Puebla, I., & Kempe, V. (2024). Engaging undergraduate students in preprint peer review. *Active Learning in Higher Education*, 14697874241264495. https://doi.org/10.35542/osf.io/gzymh

Hostler, T. (2024). Research assessment using a narrow definition of "research quality" is an act of gatekeeping: A comment on gärtner et al. (2022). *Meta-Psychology*, *8*. https://doi.org/10.31234/osf.io/2zhwe

Howard, G. S., & Maxwell, S. E. (2023). ORMA: A strategy to reduce psychology's replication problems. *New Ideas in Psychology*, *68*, 100991. https://doi.org/10.1016/j.newideapsych.2022.100991

Hoyningen-Huene, P. (2013). *Systematicity: The nature of science*. Oxford Univ. Press.

Hoyningen-Huene, P., & Kincaid, H. (2023). What makes economics special: Orientational paradigms. *J. Econ. Methodol.*, 1–15. https://doi.org/10.1080/1350178x.2023.2192231

Hsing, P.-Y., Tukanova, M., Freeman, A., Munafo, M., & Thompson, J. (n.d.). *A snapshot of the academic research culture in 2023 and how it might be improved*. https://doi.org/10.5281/zenodo.8165703

Hüffmeier, J., & Kühner, C. (2024). Replication marketplaces would help science to become more self-correcting. *R. Soc. Open Sci.*, *11*(10), 240850. https://doi.org/10.1098/rsos.240850

Hüffmeier, J., Mazei, J., & Schultze, T. (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

Hume, D. (1748/2011). *Eine untersuchung über den menschlichen verstand [an enquiry concerning human understanding]* (Vol. 5489). Reclam.

Hussey, I., & Hughes, S. (2018). *Hidden invalidity among fifteen commonly used measures in social and personality psychology*. https://doi.org/10.31234/osf.io/7rbfp

Hyman, M. (2024). Freeing social and medical scientists from the replication crisis. *Available at SSRN 4898637*.

Isager, P. M., Lakens, D., Leeuwen, T. van, & Veer, A. E. van't. (2024). Exploring a formal approach to selecting studies for replication: A feasibility study in social neuroscience. *Cortex*, *171*, 330–346. https://doi.org/10.1016/j.cortex.2023.10.012

Ivory, J. D., & Elson, M. (2024). A tale of three retractions: A call for standardized categorization and criteria in retraction statements. *Current Psychology*, *43*(17), 16023–16029. https://doi.org/10.31234/osf.io/hy45v

Jacobsen, N. S. J., Kristanto, D., Welp, S., Inceler, Y. C., & Debener, S. (2024). Preprocessing choices for P3 analyses with mobile EEG: A systematic literature review and interactive exploration. *Psychophysiology*, *62*(1). https://doi.org/10.1111/psyp.14743

Jané, M. B., Xiao, Q., Yeung, S. K., Ben-Shachar, M. S., Caldwell, A. R., Cousineau, D., Dunleavy, D. J., Elsherif, M., Johnson, B. T., & Moreau, D. (2024). *Guide to effect sizes and confidence intervals*. https://doi.org/10.17605/OSF.IO/D8C4G

Janz, N., & Freese, J. (2021). Replicate others as you would like to be replicated yourself. *PS Polit. Sci. Polit.*, *54*(2), 305–308. https://doi.org/10.1017/s1049096520000943

Jaremka, L. M., Ackerman, J. M., Gawronski, B., Rule, N. O., Sweeny, K., Tropp, L. R., Metz, M. A., Molina, L., Ryan, W. S., & Vick, S. B. (2020). Common academic experiences no one talks about: Repeated rejection, impostor syndrome, and burnout. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *15*(3), 519–543. https://doi.org/10.1177/1745691619898848

Jekel, M., Fiedler, S., Allstadt Torras, R., Mischkowski, D., Dorrough, A. R., & Glöckner, A. (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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Johnson, E. J., & Goldstein, D. (2003). Medicine. Do defaults save lives? *Science*, *302*(5649), 1338–1339.

Kafkafi, N., Agassi, J., Chesler, E. J., Crabbe, J. C., Crusio, W. E., Eilam, D., Gerlai, R., Golani, I., Gomez-Marin, A., Heller, R., Iraqi, F., Jaljuli, I., Karp, N. A., Morgan, H., Nicholson, G., Pfaff, D. W., Richter, S. H., Stark, P. B., Stiedl, O., … Benjamini, Y. (2018). Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neurosci. Biobehav. Rev.*, *87*, 218–232. https://doi.org/10.1016/j.neubiorev.2018.01.003

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263. https://doi.org/10.2307/1914185

Kakinohana, R. K., Pilati, R., & Klein, R. A. (2022). Does anchoring vary across cultures? Expanding the many labs analysis. *European Journal of Social Psychology*. https://doi.org/10.1002/ejsp.2924

Kelly, C. D. (2006). Replicating empirical research in behavioral ecology: How and why it should be done but rarely ever is. *Q. Rev. Biol.*, *81*(3), 221–236. https://doi.org/10.1086/506236

Kelly, C. D. (2019). Rate and success of study replication in ecology and evolution. *PeerJ*, *7*, e7654. https://doi.org/10.7717/peerj.7654

Kepes, S., Bushman, B. J., & Anderson, C. A. (2017). Violent video game effects remain a societal concern: Reply to hilgard, engelhardt, and rouder (2017). *Psychological Bulletin*, *143*(7), 775–782. https://doi.org/10.1037/bul0000112

Kepes, S., Keener, S. K., McDaniel, M. A., & Hartman, N. S. (2022). Questionable research practices among researchers in the most research–productive management programs. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2623

Kepes, S., Wang, W., & Cortina, J. M. (2023). Assessing publication bias: A 7-step user's guide with best-practice recommendations. *Journal of Business and Psychology*, *38*(5), 957–982. https://doi.org/10.1007/s10869-022-09840-0

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J., Babalola, M. T., Bahník, Š., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Bocian, K., Brandt, M., Busching, R., Cai, H., Cambier, F., …

Nosek, B. A. (2018). *Many labs 2: Investigating variation in replicability across sample and setting*. https://doi.org/10.31234/osf.io/9654g

Kletke, O., Larres, I., Höhner, K., Kleina, W., Stapels, J., Zey, B., & Kasties, N. (2024). Bestands-und bedarfserhebung im forschungsdatenmanagement: Ergebnisse des projektes FDM-TUDO. *O-Bib. Das Offene Bibliotheksjournal/Herausgeber VDB*, *11*(2), 1–18.

Klonsky, E. D. (2024). Campbell's law explains the replication crisis: Pre-registration badges are history repeating. *Assessment*, 10731911241253430. https://doi.org/10.31234/osf.io/7czpr

Kobrock, K., & Roettger, T. (2023). Assessing the replication landscape in experimental linguistics. *Glossa Psycholinguistics*, *2*(1), 1–28. https://doi.org/10.31234/osf.io/fzngs

Kohrs, F. E., Auer, S., Bannach-Brown, A., Fiedler, S., Haven, T. L., Heise, V., Holman, C., Azevedo, F., Bernard, R., Bleier, A., Bössel, N., Cahill, B. P., Castro, L. J., Ehrenhofer, A., Eichel, K., Frank, M., Frick, C., Friese, M., Gärtner, A., … Weissgerber, T. L. (2023). Eleven strategies for making reproducible research and open science training the norm at research institutions. *Elife*, *12*. https://doi.org/10.31219/osf.io/kcvra

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, J. Ø.-S., Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., … Evans, T. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, *1*(1). https://doi.org/10.1038/s44271-023-00003-2

Korell, D., Reinecke, N., & Lott, L. (2023). Student-led replication studies in comparative politics: New findings by fresh eyes? *Zeitschrift für Vergleichende Politikwissenschaft*, *17*(3), 261–273. https://doi.org/10.1007/s12286-023-00578-4

Körner, R., Röseler, L., Schütz, A., & Bushman, B. J. (2022). Dominance and prestige: Meta-analytic review of experimentally induced body position effects on behavioral, self-report, and physiological dependent variables. *Psychological Bulletin*, *148*(1-2), 67–85. https://doi.org/10.1037/bul0000356

Koukouraki, E., & Kray, C. (2023). *Map reproducibility in geoscientific publications: An exploratory study*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kuhn, T. S. (1970/1996). *The structure of scientific revolutions* (3rd ed.). Univ. of Chicago Press. https://doi.org/10.7208/chicago/9780226458106.001.0001

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D. (2021a). *Sample size justification*. https://doi.org/10.31234/osf.io/9d3yf

Lakens, D. (2021b). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.31234/osf.io/shm8v

Lakens, D. (2021c). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2023). *Concerns about replicability, theorizing, applicability, generalizability, and methodology across two crises in social psychology*. https://doi.org/10.31234/osf.io/dtvs7

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, *10*(1). https://doi.org/10.31234/osf.io/ha29k

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, *8*(8), 875–881. https://doi.org/10.1177/1948550617693058

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and registered reports. *Evidence-Based Toxicology*, *2*(1), 2376046. https://doi.org/10.31234/osf.io/dqap7

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.843

Leonelli, S. (2023). *Philosophy of open science*. Cambridge University Press. https://doi.org/10.1017/9781009416368

Levendis, J. D. (2018). *Time series econometrics*. Springer. https://doi.org/10.1007/978-3-319-98282-3

Levy, N. (2022). Do your own research! *Synthese*, *200*(5), 356. https://doi.org/10.1007/s11229-022-03793-w

Lewin, K. (1930). Der übergang von der aristotelischen zur galileischen denkweise in biologie und psychologie. *Erkenntnis*, *1*(1), 421–466. https://doi.org/10.1007/BF00208633

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

Lindsay, D. S. (2023). A plea to psychology professional societies that publish journals: Assess computational reproducibility. *Meta-Psychology*, *7*. https://doi.org/10.15626/mp.2023.4020

Liu, C., Wang, L., Qi, R., Wang, W., Jia, S., Shang, D., Shao, Y., Yu, M., Zhu, X., Yan, S., Chang, Q., & Zhao, Y. (2019). Prevalence and associated factors of depression and anxiety among doctoral students: The mediating effect of mentoring relationships on the association between research self-efficacy and depression/anxiety. *Psychology Research and Behavior Management*, *12*, 195–208. https://doi.org/10.2147/PRBM.S195131

Lopez-Nicolas, R., Lakens, D., López-López, J. A., Aparicio, M. R., Sandoval-Lentisco, A., López-Ibáñez, C., Blázquez-Rincón, D., & Sánchez-Meca, J. (2022). Analytical reproducibility and data reusability of published meta-analyses on clinical psychological interventions. In *PsyArXiv*.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M., & Epskamp, S. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*, 1–17. https://doi.org/10.18637/jss.v088.i02

Mac Giolla, E., Karlsson, S., Neequaye, D. A., & Bergquist, M. (2022). Evaluating the replicability of social priming studies. In *PsyArXiv*. https://doi.org/10.31234/osf.io/dwg9v

Maier, M., Bartoš, F., Raihani, N., Shanks, D. R., Stanley, T., Wagenmakers, E.-J., & Harris, A. J. (2024). Exploring open science practices in behavioural public policy research. *Royal Society Open Science*, *11*(2), 231486.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty. *Educ. Res.*, *43*(6), 304–316. https://doi.org/10.3102/0013189x14545513

Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research: Necessary but far too rare. *Remedial and Special Education*, *37*(4), 205–212.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspect. Psychol. Sci.*, *7*(6), 537–542.

Marefat, F., Hassanzadeh, M., & Hamidi, F. (2024). Incorporating replication in higher education: Supervisors' perspectives and institutional pressures. *Account. Res.*, 1–22. https://doi.org/10.1080/08989621.2024.2412054

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*, 523. https://doi.org/10.3389/fpsyg.2017.00523

Matters of significance. (2024). UCL Press. https://doi.org/10.14324/111.9781800086500

Mazei, J., Rudolph, C. W., Zacher, H., & Hüffmeier, J. (2025). Do not put all of your eggs in one basket: Multiverse analysis in applied psychology. *Journal of Applied Psychology*, *110*(11), 1511–1537. https://doi.org/10.1037/apl0001291

McDiarmid, A. D., Tullett, A. M., Whitt, C. M., Vazire, S., Smaldino, P. E., & Stephens, J. E. (2021). Psychologists update their beliefs about effect sizes after replication studies. *Nature Human Behaviour*, *5*(12), 1663–1673. https://doi.org/10.1038/s41562-021-01220-7

McNeeley, S., & Warner, J. J. (2015). Replication in criminology: A necessary practice. *Eur. J. Criminol.*, *12*(5), 581–597. https://doi.org/10.1177/1477370815578197

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2022). Variation and covariation in large-scale replication projects: An evaluation of replicability. *J. Am. Stat. Assoc.*, *117*(540), 1605–1621. https://doi.org/10.1080/01621459.2022.2054816

Mede, N. G., Schäfer, M. S., Ziegler, R., & Weißkopf, M. (2020). The "replication crisis" in the public eye: Germans' awareness and perceptions of the (ir)reproducibility of scientific research. *Public Understanding of Science (Bristol, England)*, 963662520954370. https://doi.org/10.1177/0963662520954370

Meehl, P. E. (2004). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Applied and Preventive Psychology*, *11*(1), 1. https://doi.org/10.1016/j.appsy.2004.02.001

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275. https://doi.org/10.1111/1467-9280.00350

Merton, R. K. (1968). The matthew effect in science. *Science (New York, N.Y.)*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Mittermaier, B. (2025). *Transformationsverträge sind eine sackgasse*. o-bib. Das offene Bibliotheksjournal / Herausgeber VDB.

Moore, A. D. (2000). Owning genetic information and gene enhancement techniques: Why privacy and property rights may undermine social control of the human genome. *Bioethics*, *14*(2), 97–119. https://doi.org/10.1111/1467-8519.00184

Morgan, T. J. H., & Smaldino, P. E. (2024). *Author-paid publication fees corrupt science and should be abandoned*. https://doi.org/10.31219/osf.io/3ez9v_v1

Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication studies in economics—how many and which papers are chosen for replication, and why? *Research Policy*, *48*(1), 62–83. https://doi.org/10.1016/j.respol.2018.07.019

Muhmenthaler, M. C., Dubravac, M., & Meier, B. (2022). The future failed: No evidence for precognition in a large scale replication attempt of bem (2011). *Psychology of Consciousness: Theory, Research, and Practice*. https://doi.org/10.1037/cns0000342

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *349*(Suppl 1), aac4716. https://doi.org/10.1038/s41562-018-0522-1

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, *29*(1), 85–95. https://doi.org/10.1080/00335557743000053

Naddaf, M. (2023). ChatGPT generates fake data set to support scientific hypothesis. *Nature*, *623*(7989), 895–896. https://doi.org/10.1038/d41586-023-03635-w

Nagy, T., Hergert, J., Elsherif, M. M., Wallrich, L., Schmidt, K., Waltzer, T., Payne, J. W., Gjoneska, B., Seetahul, Y., & Wang, Y. A. (2024). *Bestiary of questionable research practices in psychology*.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nelson, L., Ye, H., Schwenn, A., Lee, S., Arabi, S., & Hutchins, B. I. (2022). Robustness of evidence reported in preprints during peer review. *Lancet Glob. Health*, *10*(11), e1684–e1687. https://doi.org/10.1016/s2214-109x(22)00368-0

Neoh, M. J. Y., Carollo, A., Lee, A., & Esposito, G. (2023). Fifty years of research on questionable research practises in science: Quantitative analysis of co-citation patterns. *Royal Society Open Science*, *10*(10), 230677. https://doi.org/10.1098/rsos.230677

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

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

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

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

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.*, *73*(1), 719–748. https://doi.org/10.31234/osf.io/ksfvq

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

Nuijten, M. B., van Assen, M. A. L. M., Hartgerink, C. H. J., Epskamp, S., & Wicherts, J. M. (2017). *The validity of the tool "statcheck" in discovering statistical reporting inconsistencies*. https://doi.org/10.31234/osf.io/tcxaj

Nüst, D., & Eglen, S. J. (2021). CODECHECK: An open science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. *F1000Res.*, *10*, 253. https://doi.org/10.12688/f1000research.51738.2

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., Alshaif, N., Andringa, R., Aveyard, M., Babincak, P., Balatekin, N., Baldwin, S. A., Banik, G., Baskin, E., Bell, R., Białobrzeska, O., Birt, A. R., Boot, W. R., Braithwaite, S. R., … Zrubka, M. (2018). Registered replication report: Dijksterhuis and van knippenberg (1998). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *13*(2), 268–294. https://doi.org/10.1177/1745691618755704

O'Grady, C. (2021). Honesty study was based on fabricated data. *Science*, *373*(65-58), 950–951. https://doi.org/10.1126/science.373.6558.950

Ongchoco, J. D. K., Walter-Terrill, R., & Scholl, B. J. (2023). Visual event boundaries restrict anchoring effects in decision-making. *Proceedings of the National Academy of Sciences*, *120*(44), e2303883120. https://doi.org/10.1073/pnas.2303883120

Open Science Collaboration. (2015). Psychology: Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oswald, & Grosjean. (2004). *Oswald, m. E., & grosjean, s. (2004). Confirmation bias. In r. F. Pohl (ed.). Cognitive illusions. A handbook on fallacies and biases in thinking, judgement and memory. Hove and n.y. Psychology press*. Unpublished. https://doi.org/10.13140/2.1.2068.0641

Oviedo-García, M. (2024). The review mills, not just (self-) plagiarism in review reports, but a step further. *Scientometrics*, 1–9. https://doi.org/10.1007/s11192-024-05125-w

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Pawel, S., Consonni, G., & Held, L. (2023). Bayesian approaches to designing replication studies. *Psychol. Methods*. https://doi.org/10.1037/met0000604

Peirce, J., Hirst, R., & MacAskill, M. (2022). *Building experiments in PsychoPy*. Sage. https://doi.org/10.4135/9781036231378

Penders, B. (2024). Scandal in scientific reform: The breaking and remaking of science. *Journal of Responsible Innovation*, *11*(1), 2371172. https://doi.org/10.1080/23299460.2024.2371172

Pennington, C. (2023). *A student's guide to open science: Using the replication crisis to reform psychology*. McGraw-Hill Education (UK). https://doi.org/10.31234/osf.io/2tqep

Pennington, C. R., & Pownall, M. (2024). What have we learned from the replication crisis? Integrating open research into social psychology teaching. In *PsyArXiv*. https://doi.org/10.31234/osf.io/prx8w

Perezgonzalez, J. D. (2014). A reconceptualization of significance testing. *Theory & Psychology*, *24*(6), 852–859. https://doi.org/10.1177/0959354314546157

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semant.*, *17*, 33–43. https://doi.org/10.2139/ssrn.3198992

Phaf, R. H. (2024). Positive deviance underlies successful science: Normative methodologies risk throwing out the baby with the bathwater. *Rev. Gen. Psychol.* https://doi.org/10.1177/10892680241235120

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353.

Plesser, H. E. (2018). Reproducibility vs. Replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, *11*. https://doi.org/10.3389/fninf.2017.00076

Popper, K. R. (1959/2008). *The logic of scientific discovery* (Repr. 2008). Routledge Classics; Routledge. https://doi.org/10.4324/9780203994627

Priem, J., Piwowar, H., & Orr, R. (2022a). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*.

Priem, J., Piwowar, H., & Orr, R. (2022b). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv Preprint arXiv:2205.01833*.

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2023). Are small effects the indispensable foundation for a cumulative psychological science? A reply to götz et al.(2022). *Perspectives on Psychological Science*, *18*(2), 508–512. https://doi.org/10.31234/osf.io/6s8bj

Promoting reproduction and replication at scale. (2024). *Nat. Hum. Behav.*, *8*(1), 1. https://doi.org/10.1038/s41562-024-01818-7

Protzko, J. (2018). *Null-hacking, a lurking problem*. https://doi.org/10.31234/osf.io/9y3mp

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. W. (2023). High replicability of newly discovered social-behavioural findings is achievable. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-023-01749-9

Puckett, J. (2011). *Zotero: A guide for librarians, researchers, and educators*. Assoc of Cllge & Rsrch Libr.

Quan, J. (2021). *Toward reproducibility: Academic libraries and open science* (pp. 49–66). Cambridge University Press. https://doi.org/10.29085/9781783304615.004

Quintana, D. S. (2021). Replication studies for undergraduate theses to improve science and education. *Nat. Hum. Behav.*, *5*(9), 1117–1118. https://doi.org/10.1038/s41562-021-01192-8

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rahal, R.-M., Fiedler, S., Adetula, A., Berntsson, R. P.-A., Dirnagl, U., Feld, G. B., Fiebach, C. J., Himi, S. A., Horner, A. J., Lonsdorf, T. B., Schönbrodt, F., Silan, M. A. A., Wenzler, M., & Azevedo, F. (2023). Quality research needs good working conditions. *Nature Human Behaviour*, *7*(2), 164–167. https://doi.org/10.1038/s41562-022-01508-2

Ramminger, J. J. (2023). Vermessen? Zur möglichkeit philosophischer beiträge für den diskurs der quantitativen psychologie. *Cultura & Psyché*, *4*(2), 215–224. https://doi.org/10.1007/s43638-023-00081-3

Redden, J. P., & Hoch, S. J. (2009). The presence of variety reduces perceived quantity. *Journal of Consumer Research*, *36*(3), 406–417. https://doi.org/10.1086/598971

Reed, G., Hendlin, Y., Desikan, A., MacKinney, T., Berman, E., & Goldman, G. T. (2021). The disinformation playbook: How industry manipulates the science-policy process—and how to restore scientific integrity. *Journal of Public Health Policy*, *42*(4), 622. https://doi.org/10.1057/s41271-021-00318-6

Richardson, R. (2025). *The collection of open science integrity guides (COSIG): Expanding participation in post-publication peer review*. https://doi.org/10.5281/zenodo.15564777

Richter, S. H. (2024). Challenging current scientific practice: How a shift in research methodology could reduce animal use. *Lab Animal*, *53*(1), 9–12. https://doi.org/10.1038/s41684-023-01308-9

Riedel, N., Wieschowski, S., Bruckner, T., Holst, M. R., Kahrass, H., Nury, E., Meerpohl, J. J., Salholz-Hillel, M., & Strech, D. (2022). Results dissemination from completed clinical trials conducted at german university medical centers remained delayed and incomplete. The 2014 -2017 cohort. *Journal of Clinical Epidemiology*, *144*, 1–7. https://doi.org/10.1016/j.jclinepi.2021.12.012

Rife, S., Lambert, Q., Calin-Jageman, R., Matus, A., Banik, G., Barberia, I., Beaudry, J., Bernauer, H., Calvillo, D., & Chopik, W. (2024). Registered replication report: Study 3 from trafimow and hughes (2012). *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/esu9z_v2

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Robinson, E. (2011). Not feeling the future: A failed replication of retroactive facilitation of memory recall. *Journal of the Society for Psychical Research*, *75*(904).

Rodriguez, J. E., & Williams, D. R. (2022). *Psymetadata: An r package containing open datasets from meta-analyses in psychology*. https://doi.org/10.31234/osf.io/myxsj

Roe, C. A., Grierson, S., & Lomas, A. (2012). *Feeling the future: Two independent replication attempts*.

Röseler, L. (2023). *Predicting replication rates with z-curve: A brief exploratory validation study using the replication database*. https://doi.org/10.31222/osf.io/ewb2t

Röseler, L., Felser, G., Asberger, J., & Schütz, A. (2020). *No effect of variety on perceived quantity: Evidence from six studies*. https://doi.org/10.31234/osf.io/v643q

Röseler, L., Gendlina, T., Krapp, J., Labusch, N., & Schütz, A. (2022). *Successes and failures of replications: A meta-analysis of independent replication studies based on the OSF registries*. https://doi.org/10.31222/osf.io/8psw2

Röseler, L., Kaiser, L., Doetsch, C. A., Klett, N., Seida, C., Schütz, A., Aczel, B., Adelina, N., Agostini, V., Alarie, S., Albayarak-Aydemir, N., AlDoh, A., Al-Hoorie, A. H., Azevedo, F., Baker, B. J., Barth, C. L., Beitner, J., Brick, C., Brohmer, H., … Zhang, Y. (2024). *The replication database: Documenting the replicability of psychological science*. https://doi.org/10.31222/osf.io/me2ub

Röseler, L., & Schütz, A. (2022). Open science. In A. Schütz, M. Brand, S. Steins-Loeber, M. Baumann, J. Born, V. Brandstätter, C.-C. Carbon, P. M. Gollwitzer, M. Hallschmid, S. Lautenbacher, L. Laux, B. Marcus, K. Moser, K. I. Paul, H. Plessner, F. Renkewitz, K.-H. Renner, K. Rentzsch, K. Rothermund, … S. Steins-Löber (Eds.), *Psychologie* (pp. 187–198). Kohlhammer.

Röseler, L., Schütz, A., Baumeister, R. F., & Starker, U. (2020). Does ego depletion reduce judgment adjustment for both internally and externally generated anchors? *Journal of Experimental Social Psychology*, *87*, 103942. https://doi.org/10.1016/j.jesp.2019.103942

Röseler, L., Weber, L., Helgerth, K., Stich, E., Günther, M., Tegethoff, P., Wagner, F. S., Ambrus, E., Antunovic, M., & Barrera-Lemarchand, F. (2024). Correction: The open anchoring quest dataset: Anchored estimates from 96 studies on anchoring effects. *Journal of Open Psychology Data*, *12*(1).

Röseler, L., Weber, L., & Schütz, A. (2021). *OpAQ: Open anchoring quest*. https://osf.io/ygnvb

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Royal Netherlands Academy of Arts and Sciences. (2018). *Replication studies—improving reproducibility in the empirical sciences*.

Rubin, M. (2023). The replication crisis is less of a "crisis" in the lakatosian approach than it is in the popperian and naïve methodological falsificationism approaches. In *BITSS*.

Ruiter, B. de. (2023). Automatically finding and categorizing replication studies. *arXiv e-Prints*, arXiv–2311.

Sabel, B. A., Knaack, E., Gigerenzer, G., & Bilc, M. (2023). *Fake publications in biomedical science: Red-flagging method indicates mass production*. https://doi.org/10.1101/2023.05.06.23289563

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. https://doi.org/10.1007/BF00055564

Sarachik, M. P. (2009). Plastic fantastic: How the biggest fraud in physics shook the scientific world. *Physics Today*, *62*(10), 57–57.

Sarstedt, M., & Adler, S. J. (2023). An advanced method to streamline p-hacking. *Journal of Business Research*, *163*, 113942. https://doi.org/10.31234/osf.io/5ynfw

Sarstedt, M., Adler, S. J., Ringle, C. M., Cho, G., Diamantopoulos, A., Hwang, H., & Liengaard, B. D. (2024). Same model, same data, but different outcomes: Evalu-

ating the impact of method choices in structural equation modeling. *J. Prod. Innov. Manage.* https://doi.org/10.1111/jpim.12738

Satinsky, E. N., Kimura, T., Kiang, M. V., Abebe, R., Cunningham, S., Lee, H., Lin, X., Liu, C. H., Rudan, I., Sen, S., Tomlinson, M., Yaver, M., & Tsai, A. C. (2021). Systematic review and meta-analysis of depression, anxiety, and suicidal ideation among ph.d. students. *Scientific Reports*, *11*(1), 14370. https://doi.org/10.1038/s41598-021-93687-7

Scanff, A., Mauhe, N., Taburet, M., Savourat, P.-E., Clément, T., Bastian, B., Cristea, I., Braillon, A., Carayol, N., & Naudet, F. (2023). The "free lunches" index for assessing academics: A not entirely serious proposal. *Scientometrics*. https://doi.org/10.1007/s11192-023-04862-8

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467. https://doi.org/10.31234/osf.io/p6e9c

Schimmack, U. (2019). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, 1745691619863798. https://doi.org/10.1177/1745691619863798

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne*. https://doi.org/10.1037/cap0000246

Schmidt, B., Chiarelli, A., Loffreda, L., & Sondervan, J. (2024). Emerging roles and responsibilities of libraries in support of reproducible research. *LIBER Q.*, *33*(1), 1–21. https://doi.org/10.53377/lq.14947

Schneider, J. W., Allum, N., Andersen, J. P., Petersen, M. B., Madsen, E. B., Mejlgaard, N., & Zachariae, R. (2024). Is something rotten in the state of denmark? Cross-national evidence for widespread involvement but not systematic use of questionable research practices across all fields of research. *PLoS One*, *19*(8), e0304342. https://doi.org/10.31222/osf.io/r6j3z

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*(2-3), 373–403. https://doi.org/10.1007/s42803-023-00073-y

Schönbrodt, F. (2016). *P-hacker: Train your p-hacking skills!* http://shinyapps.org/apps/p-hacker/.

Schönbrodt, F., Baumert, A., Glöckner, A., Back, M., Arslan, R. C., Voracek, M., Horstmann, K. T., Rohrer, J. M., Mueller, E., Stahl, C., Fiebach, C., Miller, R., Schultheiss, O. C., Brohmer, H., Renkewitz, F., Janssen, L., Stefan, A. M., Heycke, T., Schmidt, N.-D., … Riedl, L. (2022). *Netzwerk der Open-Science-Initiativen (NOSI)*. Open Science Framework.

Schönbrodt, F., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., Phan, L. V., Schmitt, M., Scheel, A. M., & Schubert, A.-L. (2022). *Responsible research assessment i: Implementing DORA for hiring and promotion in psychology*.

Schönbrodt, F., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., Phan, L. V., Schmitt, M., Scheel, A. M., Schubert, A.-L., Steinberg, U., & Leising, D. (2022). *Responsible research assessment i: Implementing DORA for hiring and promotion in psychology*. https://doi.org/10.23668/psycharchives.8162

Schönbrodt, F., Gollwitzer, M., & Abele-Brehm, A. (2017). Der umgang mit forschungsdaten im fach psychologie: Konkretisierung der DFG-leitlinien. *Psychologische Rundschau*. https://doi.org/10.1026/0033-3042/a000341

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21). https://doi.org/10.1126/sciadv.abd1705

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, *115*(11), 2632–2639. https://doi.org/10.1073/pnas.1711786114

Siems, R. (2024). Subprime impact crisis. Bibliotheken, politik und digitale souveränität. *BIBL. Forsch. Prax.*, *0*(0). https://doi.org/10.1515/bfp-2024-0008

Sikorski, M., & Andreoletti, M. (2023). Epistemic functions of replicability in experimental sciences: Defending the orthodox view. *Found. Sci.* https://doi.org/10.1007/s10699-023-09901-4

Silveira, L. da, Calixto Ribeiro, N., Melero, R., Mora-Campos, A., Piraquive-Piraquive, D. F., Uribe Tirado, A., Machado Borges Sena, P., Polanco Cortés, J., Santillán-Aldana, J., Couto Corrêa da Silva, F., Ferreira Araújo, R., Enciso Betancourt, A. M., & Fachin, J. (2023). Taxonomia da ciência aberta: Revisada e ampliada. *Encontros Bibli Rev. Eletrônica Bibliotecon. Ciênc. Inf.*, *28*. https://doi.org/10.5007/1518-2924.2023.e91712

Silverstein, P., Elman, C., Montoya, A. K., McGillivray, B., Pennington, C. R., Harrison, C. H., Steltenpohl, C. N., Röer, J. P., Corker, K. S., Charron, L. M., Elsherif, M. M., Na, A., Hayes-Harb, R., Grinschgl, S., Neal, T. M. S., Evans, T. R., Karhulahti, V.-M.,

Krenzer, W. L. D., Belaus, A., … Syed, M. (2023). *A guide for social science journal editors on easing into open science* [Unpublished manuscript].

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2160588

Simmons, J. P., Nelson, L., & Simonsohn, U. (2020). Pre–registration is a game changer. But, like random assignment, it is neither necessary nor sufficient for credible science. *Journal of Consumer Psychology*. https://doi.org/10.1002/jcpy.1207

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*, *28*(5), 687–693. https://doi.org/10.1177/0956797616658563

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: A key to the file-drawer. *Journal of Experimental Psychology. General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to ulrich and miller (2015). *Journal of Experimental Psychology. General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Singh Chawla, D. (2024). The citation black market: Schemes selling fake references alarm scientists. *Nature*. https://doi.org/10.1038/d41586-024-01672-7

Sion, T. G. ap, & Earp, B. D. (2024). *Replication and theory development in the social and psychological sciences*.

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher (Ed.), *Computational social psychology* (pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14

Smedslund, J. (2015). Why psychology cannot be an empirical science. *Integrative Psychological & Behavioral Science*. https://doi.org/10.1007/s12124-015-9339-x

Smith, N., Jr, & Cumberledge, A. (2020). Quotation errors in general science journals. *Proc. Math. Phys. Eng. Sci.*, *476*(2242), 20200538. https://doi.org/10.1098/rspa.2020.0538

Sobkow, A., Surowski, M., Olszewska, A., Antoniewska, N., Bartkiewicz, U., Brzeska, A., Brzozowska, A., Oliwia, Bukowska, K., Choja, J., Choma, K., Chorbotowicz, P., Filimoniak, M., Filip, Ł., Gambuś, P., Gierlik, W., Gonczar, T., Goryczka, K., Góra, M., … Traczyk, J. (2021). *Conceptual replication study of fifteen JDM effects: Insights from the polish sample*. https://doi.org/10.31234/osf.io/5fc6z

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01142-4

Soergel, D. A. (2014). Rampant software errors may undermine scientific results. *F1000Research*, *3*. https://doi.org/10.12688/f1000research.5930.2

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

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Sotola, L. K. (2023). How can i study from below, that which is above? *Meta-Psychology*, *7*. https://doi.org/10.15626/MP.2022.3299

Sotola, L. K., & Credé, M. (2022). On the predicted replicability of two decades of experimental research on system justification: A z–curve analysis. *European Journal of Social Psychology*, *52*(5-6), 895–909. https://doi.org/10.1002/ejsp.2858

Srivastava, S. (2012). A pottery barn rule for scientific journals. *The Hardest Science*, *27*.

Stavrova, O., Kleinberg, B., Evans, A. M., & Ivanovic, M. (2024). Scientific publications that use promotional language receive more citations and social media mentions. In *PsyArXiv*. https://doi.org/10.31234/osf.io/da9xn

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

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(6), 929–930. https://doi.org/10.1177/1745691616674460

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768–777. https://doi.org/10.1037/0022-3514.54.5.768

Strecker, D. (2019). *Nutzung der schattenbibliothek sci-hub in deutschland*.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *7*(6), 670–688. https://doi.org/10.1177/1745691612460687

Sultan, M., Tump, A. N., Ehmann, N., Lorenz-Spreen, P., Hertwig, R., Gollwitzer, A., & Kurvers, R. (2024). Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. In *PsyArXiv*. https://doi.org/10.1073/pnas.2409329121

Syed, M. (2023). Some data indicating that editors and reviewers do not check preregistrations during the review process. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/nh7qw

Symonds, J. E., & Tang, X. (2024). *Quality appraisal checklist for quantitative, qualitative, and mixed methods studies*. https://doi.org/10.31234/osf.io/djmcq_v2

TARG Meta-Research Group and Collaborators, Thibault, R. T., Clark, R., Pedder, H., Akker, O. van den, Westwood, S., Thompson, J., & Munafo, M. (2021). Estimating the prevalence of discrepancies between study registrations and publications: A systematic review and meta-analyses. *MedRxiv*, 2021–2007.

Teboul, L., Amos-Landgraf, J., Benavides, F. J., Birling, M.-C., Brown, S. D., Bryda, E., Bunton-Stasyshyn, R., Chin, H.-J., Crispo, M., & Delerue, F. (2024). Improving laboratory animal genetic reporting: LAG-r guidelines. *Nature Communications*, *15*(1), 5574.

The Turing Way Community, & Scriberia. (2024). *Illustrations from the turing way: Shared under CC-BY 4.0 for reuse [figure]*. Zenodo. https://doi.org/10.5281/zenodo.13882307

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*. https://doi.org/10.2478/jdis-2024-0013

Thériault, F., R. (2023). *The missing majority in behavioral science dashboard*. https://remi-theriault.com/dashboards/missing_majority

Thibault, R. T., Pennington, C. R., & Munafò, M. R. (2023). Reflections on preregistration: Core criteria, badges, complementary workflows. *Journal of Trial & Error*, *2*(1), 10–36850. https://doi.org/10.31234/osf.io/w6tj4

Thibault, R. T., Zavalis, E. A., Malicki, M., & Pedder, H. (2024). An evaluation of reproducibility and errors in published sample size calculations performed using g* power. *medRxiv*, 2024–2007. https://doi.org/10.1101/2024.07.15.24310458

Tierney, W., Hardy III, J. H., Ebersole, C. R., Leavitt, K., Viganola, D., Clemente, E. G., Gordon, M., Dreber, A., Johannesson, M., & Pfeiffer, T. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes*, *161*, 291–309.

Ting, C., & Greenland, S. (2024). Forcing a deterministic frame on probabilistic phenomena: A communication blind spot in media coverage of the "replication crisis." *Sci. Commun.* https://doi.org/10.1177/10755470241239947

Tiokhin, L., Panchanathan, K., Smaldino, P. E., & Lakens, D. (2023). Shifting the level of selection in science. *Perspect. Psychol. Sci.*, 17456916231182568. https://doi.org/10.31222/osf.io/juwck

Torka, A.-K., Mazei, J., & Hüffmeier, J. (2023). Are replications mainstream now? A comparison of support for replications expressed in the policies of social psychology journals in 2015 and 2022. *Social Psychological Bulletin*, *18*, 1–22. https://doi.org/10.32872/spb.9695

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on klein. *Theory Psychol.*, *26*(4), 540–548. https://doi.org/10.1177/0959354316637136

UNESCO. (2020). *First draft of the UNESCO recommendation on open science*.

Urminsky, O., & Dietvorst, B. J. (2024). Taking the full measure: Integrating replication into research practice to assess generalizability. *Journal of Consumer Research*, *51*(1), 157–168. https://doi.org/10.1093/jcr/ucae007

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology*, *33*(3), 403–423. https://doi.org/10.31234/osf.io/af9by

Vaidis, D. C., Sleegers, W. W., Van Leeuwen, F., DeMarree, K. G., Sætrevik, B., Ross, R. M., Schmidt, K., Protzko, J., Morvinski, C., & Ghasemi, O. (2024). A multilab

replication of the induced-compliance paradigm of cognitive dissonance. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231213375.

van Aert, R. C. M., & van Assen, M. A. L. M. (2018). *P-uniform**. https://doi.org/10.31222/osf.io/zqjr9

Van den Akker, O., Assen, M. A. L. M. van, Enting, M., Jonge, M. de, Ong, H. H., Rüffer, F. F., Schoenmakers, M., Stoevenbelt, A. H., Wicherts, J. M., & Bakker, M. (2022). Selective hypothesis reporting in psychology: Comparing preregistrations and corresponding publications. In *BITSS*.

van Noorden, R. (2023). How big is science's fake-paper problem? *Nature*, *623*(7987), 466–467. https://doi.org/10.1038/d41586-023-03464-x

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

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

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., & Bunyi, A. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, *32*(10), 1566–1581. https://doi.org/10.31234/osf.io/e497p

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered replication report: Strack, martin, & stepper (1988). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing research with undergraduate students via replication

work: The collaborative replications and education project. *Frontiers in Psychology*, *10*, 247. https://doi.org/10.31234/osf.io/kbgzt

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Front. Psychol.*, *7*, 1832. https://doi.org/10.31219/osf.io/umq8d

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilcox, R. R., & Rousselet, G. A. (2024). *More reasons why replication is a difficult issue*. https://doi.org/10.31219/osf.io/9amhe

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, *322*(5901), 606–607. https://doi.org/10.1126/science.1162548

Willighagen, E. (2023). Two years of explicit CiTO annotations. *J. Cheminform.*, *15*(1), 14. https://doi.org/10.1186/s13321-023-00683-2

Willinsky, J. (2005). Open journal systems: An example of open source software for journal management and publishing. *Library Hi Tech*, *23*(4), 504–519.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology*, *13*(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510

Wittgenstein, L. (1922/2004). *Logisch-philosophische abhandlung: Tractatus logico-philosophicus* (1st ed., Vol. 12). Suhrkamp.

Wittgenstein, L. (1968). *Philosophical investigations*. Basil Blackwell.

Wrzesinski, M. (2023). *Wissenschaftsgeleitetes publizieren. Sechs handreichungen mit praxistipps und perspektiven*. Alexander von Humboldt Institut für Internet und Gesellschaft. https://doi.org/10.5281/zenodo.8169418

Xiao, Q., Lam, C. S., Piara, M., & Feldman, G. (2021). Revisiting status quo bias. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2020.2470

Yagnik, J. (2014). *PSPP a free and open source tool for data analysis*.

Yamashita, T., & Neiriz, R. (2024). Why replicate? Systematic review of calls for replication in language teaching. *Research Methods in Applied Linguistics*, *3*(1), 100091. https://doi.org/10.1016/j.rmal.2023.100091

Yarkoni, T. (2019). *The generalizability crisis*. https://doi.org/10.31234/osf.io/jqw35

Yeung, A. W. K. (2017). Do neuroscience journals accept replications? A survey of literature. *Frontiers in Human Neuroscience*, *11*, 468. https://doi.org/10.3389/fnhum.2017.00468

Yeung, S. K., & Feldman, G. (2023). *RRR assessment manuscript template*. OSF.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282. https://doi.org/10.3758/s13423-013-0495-z

Zhang, Z., & Mai, Y. (2022). *WebPower: Basic and advanced statistical power analysis (r package version 0.7)*. https://CRAN.R-project.org/package=WebPower

Ziano, I., Mok, P. Y., & Feldman, G. (2021). Replication and extension of alicke (1985) better-than-average effect for desirable and controllable traits. *Social Psychological and Personality Science*, *12*(6), 1005–1017. https://doi.org/10.1177/1948550620948973

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome Biology*, *17*, 1–3. https://doi.org/10.1186/s13059-016-1044-7

Zimmermann, C. (2015). On the need for a replication journal. *Wp*, *2015*(016). https://doi.org/10.2139/ssrn.2647280

Zumstein, P. (2024). *Open access und ein blick auf das wissenschaftliche publikationswesen*. PsychArchives. https://doi.org/10.23668/psycharchives.14438