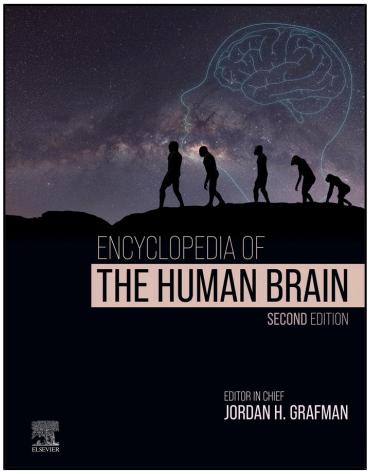

Hirschhorn, R., Schonberg, T., 2025. Replication. In: Grafman, J.H. (Ed.), Encyclopedia of the Human Brain, Second Edition, vol. 5, pp. 171–184. USA: Elsevier.
ISBN: 9780128204801

Replication

Rony Hirschhorn[a] **and Tom Schonberg**[a,b,c], [a] Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel; [b] School of Neurobiology, Biochemistry and Biophysics, Faculty of Life Science, Tel Aviv University, Tel Aviv, Israel; and [c] The Strauss Center for Computational Neuroimaging, Tel Aviv University, Tel Aviv, Israel

Key points

- Define replication and replicability
- Discuss the replication crisis and its effect on neuroscience research
- Discuss how a replication study's success is assessed
- Present causes for failed replications, and demonstrate how some of them can be mitigated
- Present novel initiatives and future directions in the field

Encyclopedia of the Human Brain, Second Edition, Second Edition, 2025, 171–184

Glossary

fMRI Functional Magnetic Resonance Imaging (fMRI) is a functional neuroimaging technique measuring changes associated with changes in blood flow in the brain and specifically blood-oxygenation levels. These changes are considered to be coupled with neural activity patterns. In fMRI studies, subjects usually perform a task while MR images are continuously acquired with the whole brain often acquired during 2 s. Then, the time-series data is analyzed, usually with relation to a task participants performed during the scan to produce activation maps. Those maps depict the probability that a certain region of the brain was activated during a task. There is also a common fMRI scanned called resting state, whereby participants do not perform a task

EEG Electroencephalography (EEG) is a neuroimaging technique for recording electrical activity on the scalp. Small metal discs (electrodes) capture the macroscopic activity of the surface layer of the brain. This electrical activity is a summation of postsynaptic potentials of large groups of neurons that fire continuously. In EEG studies, these signals are continuously recorded, creating EEG waveforms. Then, these waveforms are averaged with respect to an event of interest, to produce event-related potentials (ERPs): activities that are temporally related to an event in the experiment

MEG Magnetoencephalography (MEG) is a functional neuroimaging technique mapping neural activity by recording magnetic fields produced by electrical currents occurring naturally in the brain using sensitive magnetometers. Similar to EEG, MEG data provides a high temporal resolution that allows the assignment of neural activity to specific events. MEEG refers to both EEG and MEG

Sample size The number of observations collected in an experiment. Usually, the sample size is the number of participants who took part in the study

NHST Null Hypothesis Significance Testing (also referred to as the "frequentist approach") is a statistical inference method testing whether, given experimental data, the "null hypothesis" can be rejected. A research hypothesis usually assumes a relationship between two or more variables (e.g., height and weight). The null hypothesis is then formulated as the assumption that no such relationship exists. In the NHST framework, the hypotheses are tested while knowing the distribution of the test statistic under the null hypothesis. The test then asks whether the observed data is likely to be derived from the null distribution. If the observed data is extreme enough, then the null hypothesis can be rejected. The "extremeness" of the test statistic is reflected by the probability value ("p-value"), which measures how likely the test statistic is, under the assumption of the null hypothesis. When the p-value of a statistic is small (below some threshold, usually 0.05), it is considered unlikely, and the null hypothesis is rejected

Effect size In the NHST framework, an effect size is a standardized measure of the strength of the relationship between variables of interest, or the magnitude of the difference between groups

Statistical power In the NHST framework, the statistical power of a null-hypothesis test is the probability of rejecting the null hypothesis when the null hypothesis is indeed false. Meaning, the statistical power of a test is the probability to correctly reject the null hypothesis ("true positive"). The statistical power is complementary to the probability of a type-II error (the probability of not rejecting the null hypothesis when the null hypothesis is false; a "false negative"). Together they sum up to one.A study can be considered as lacking statistical power (or, "underpowered"). This means that the combination of the study design and the chosen statistical test have a low probability of detecting a true effect of a certain size of interest. Thus, in underpowered studies, the probability of making a type-II error is high. Furthermore, Button et al. (2013) have shown that the probability of making a type-I error (a "false positive") in underpowered studies increases as well. Meaning, in studies lacking statistical power, the probability of wrongfully rejecting the null hypothesis (when the null hypothesis is true) is also increased.

Power analysis An analysis method used to calculate the minimum sample size required to detect an effect of a given size. For a given level of statistical power (e.g., 80%), a statistical test (e.g., a student's t-test), and a p-value (e.g., 0.05), a power analysis provides the number of observations that the experimenter needs to collect

Replication The process of repeating a study with new observations

Replicability The ability of a study to produce similar results when it is executed on a new sample of participants

Reproducibility The ability of a study to reach the same results when performing the original study's analysis on the original data

Robustness The ability of a study to produce similar results upon reanalysis of the study's data using different methods

Generalizability The ability of a study to produce similar results when it is executed on a new sample of participants and data analysis methods are (somewhat) different from the ones originally used

Abstract

Most scientific studies are aimed to test different hypotheses, deepen our understanding of various phenomena, provide information about a specific process, or uncover the works of some mechanism. Eventually, the results of a research

experiment are aimed at equipping us with knowledge about ourselves and the world around us. Some of the factors that affect the results originate from the underlying research question, while others are related to researchers' specific choices when performing the study. To discover whether the results obtained in an experiment are reproducible, scientists should perform a replication study. Replication is the act of performing a study on a new sample to test whether the original scientific findings are recreated. In recent years, the importance of replications has been emphasized across fields, from social sciences to cancer biology. In this article, we review how replications are defined, executed, and evaluated. This highlights the complexities researchers face when performing and assessing the success of replication attempts. We discuss what affects a study's replicability and highlight progress made in brain research following the appreciation of replications' contribution to the field.

Introduction

Most scientific studies test a hypothesis the researchers have about the value of a single parameter, or the relationship between multiple parameters. For example, in neuroscience, one might examine the relationship between a negative emotional response to a stimulus and the neural activity in a specific area of the brain associated with it. Thus, researchers design and develop experiments to test their hypotheses, usually on a group of participants. Then, the researchers analyze the study's results using various statistical methods to evaluate whether the collected data can provide information about the hypothesis. Based on the analysis, the researchers reach a conclusion about the relationship they tested. However, this conclusion is derived from a single experiment, performed once on a limited group of participants. The question that comes up is, are the results from this study reliable?

Replication is the process of assessing a study's reliability by repeating it with a different sample of collected data. The replicability of an experiment is the ability to achieve similar results and reach similar conclusions when conducting it again with different participants. Intuitively, if the results are reliable, repeating the experiment's design and analyses on a new sample of participants should yield consistent results. Thus, replicability is a crucial feature of a scientific study, especially when examining the brain and behavior. Conclusions derived from neuroscience studies affect not only our understanding of the brain, but also have a great impact on daily life: from policymaking (e.g., COVID-19 recommendations: Van Bavel et al., 2020), to clinical diagnosis (e.g., an EEG-based index for assessing consciousness in unresponsive patients: Casali et al., 2013), and mental health treatments (e.g., video-based clinical treatments: Matheson et al., 2020). These tools rely on the premise that the research findings establishing their usefulness are replicable. However, until recently, replication has not been at the forefront of research as novelty of results was highlighted as the goal of many studies. In this article, we overview the replication crisis and review the progress made in neuroscience toward increasing replicability.

What is replication?

Replication has many definitions, stemming from different scientific eras and disciplines (for a review of different definitions in psychological sciences, see Schmidt, 2009). Lykken (1968) recognized the importance of replicability in the social, clinical, and personality research domains. In psychology research, the effects in question often involve complex relationships with latent variables that are hard to separate from noise. Therefore, the effects that researchers aim to test for in such experiments might stem from factors other than the actual relationship of interest. Thus, a single experiment cannot be considered to provide enough evidence for the examined theory. In order to test a theory's predictions against the possibility that the theory is false (Popper, 1959), Lykken suggested that "demonstrating an empirical fact must involve a claim of confidence in the replicability of one's findings". Consequently, he defined several types of replications: A *literal replication* involves an exact repetition of the sampling methods, experimental conditions, measuring techniques and analysis methods used in the original study, done by the same research group. An *operational replication* is one where a different group duplicates the original study (repeating the original procedures from the original report, also known as the material realization: Radder, 1992). A *constructive replication* (also known as *conceptual replication*) attempts to replicate the original findings while deliberately using a different procedure. The conclusions of a specific study cannot be generalized without constructive replications, demonstrating that the construct (phenomenon) of interest can be derived even with new procedures and collected data. While constructive replications test a study's generalizability, an operational replication tests whether the original findings can be duplicated when the methods are the same.

At their core, different attempts to define "replication" all agree on the simple notion of assessing a study's reliability: repeating an experiment should lead to the same results (National Academies of Sciences Engineering and Medicine, 2019). Notably, replication and replicability are often confused with similar terms, which have different meanings and implications. First, reproducibility describes achieving the same results when the original analysis is performed on the original data without collecting new observations (Nosek et al., 2022). Second, robustness is achieved when the original dataset is analyzed differently, and the same conclusions are reached (Nosek et al., 2022). Finally, generalizability is the combination of robustness and replicability: when the original results are achieved with a different dataset (replicability) and using different analysis pipelines (robustness), they are generalizable (Nosek et al., 2022).

Thus, replication is the process of repeating a study with a newly collected sample, and replicability is the ability to achieve the same results as the original study (Goodman et al., 2016). A replication study aims to estimate the replicability, or reliability, of the original experiment.

The replication crisis

In 2005, a provocative paper in genetics research, titled "Why most published research findings are false", was published (Ioannidis, 2005). Addressing a trend of failure to replicate previous findings, Ioannidis made a conceptual statistical argument for why most published experiments reporting positive results are false (the statistical assumptions made by Ioannidis were criticized by Goodman and Greenland (2007)). By "false", he meant that the relationship found in the study (reported based on its statistical significance) does not truly exist. As such, previously published findings of significant relationships could be, in fact, "false positives" (also known as type-I errors or false alarms).

Based on this logic, Ioannidis concluded that the following conditions increase the likelihood that a study's result is false (and so, would fail to replicate): a small sample size (i.e., too few subjects or observations in the collected data), a small effect size (i.e., a weak relationship found between two or more variables), a large number of tested relationships in the scientific field (i.e., more exploratory than confirmatory experiments), flexibility in the methods of the scientific field (i.e., a lack of standardization in study designs, definitions, analysis steps, etc.), conflicts of interest (i.e., a financial or other interest that might bias the research), and incentives to make extreme claims (due to the competitive need to publish papers with novel results).

Yet the pitfalls Ioannidis deduced when making his statistical argument have been known for decades. For example, Tversky and Kahneman (1971) explained that a small sample size is underpowered, contributing to a failure to reject the null hypothesis assuming that the research hypothesis is true (known as a type-II error or a false negative result). They explicitly warned that this could lead to failure to replicate research findings and recommended calculating the power of an experiment before it was performed. Relatedly, though not directly addressing Ioannidis' point, Cohen (1962) years earlier, recommended increasing the sample size to increase the statistical power of experiments. Using power analysis, he demonstrated the large impact that statistical power has on finding a significant effect size. An effect of miniature magnitude, even when statistically significant, means that the strength of the tested relationship (or the difference between the groups of interest) is "trivially small" (Cohen, 1990). He also warned against taking a test's statistical significance as a proxy of effect size.

Despite these warnings, statistical significance has in fact become an essential requirement for findings to be published in scientific journals. This too has been recognized as a problem: Sterling (1959) described how journals restrict publications to statistically significant results while many insignificant results are never published. He then deduced that type-I errors (false positives), rather than actual effects, might drive a massive part of the literature in various scientific disciplines. Twenty years later, Rosenthal (1979) revisited the issue, which from that point on has been known as the "file drawer problem": when publications that do not reach significance are not published and are "left in the drawer". Thus, "positive results" (statistically significant ones) have become a desirable goal for researchers, as they were necessary for the results to be reported and published.

Despite the fact that scientists having recognized these issues long ago, the rate of underpowered studies with small effect sizes and a small number of subjects remained consistent (e.g., Smaldino and McElreath, 2016). In addition, replication studies have rarely been carried out (e.g., the replication rate in psychology was estimated to be only 1.07%, according to Makel et al., 2012). Moreover, scientific journals refrained from publishing such studies, or those reporting negative (i.e., insignificant) results (e.g., Aldhous, 2011). Thus, Ioannidis' work exposed a cross-disciplinary gap between theoretical knowledge and scientific practice: the causes for unreliable published works were known, but researchers barely addressed them and did not have any incentives to do otherwise.

The publication by Ioannidis and other works (as reviewed by Nelson et al., 2018) catalyzed many replication efforts across different fields (e.g., computation: Donoho et al., 2008; psychology: Open Science Collaboration, 2012; cancer biology: Errington et al., 2014), in what would be coined as "the replication crisis": one report after another, canonical results in various disciplines failed to replicate. First, only 36% effects in psychology experiments were replicated (Open Science Collaboration, 2015), then a major reduction in effect sizes in cancer biology (Errington et al., 2021), and 61% in behavioral economics (Camerer et al., 2016). The newly developed interest in replication raised the question: What makes replications fail?

Causes for failed replication

Intuitively, a replication attempt should be simple: mimic the steps taken in the original study, collect a new sample and test whether similar results are achieved (assuming enough details were provided about the original studies' methods). If so, then the confidence in the reliability of the phenomenon increases. If the replication attempt failed, one might conclude that the original findings could have been erroneous. However, this might not be accurate, as many factors affect a study's replicability. Below we provide an overview of these factors.

Deviation from the original study

A direct replication (Nosek and Errington, 2017; Zwaan et al., 2018) is a study that recreates all the "relevant" settings of the original experiment. While in Lykken's definition of an operational replication the entire study is duplicated, a direct replication only considers the settings that are relevant. Nevertheless, the decision of what is relevant depends on factors such as the researchers' priorities and the availability of the information about the original study (see "insufficient reporting" below). Thus, when a study's findings fail to replicate, it may as well be due to the replication experiment failing to fulfill an essential requirement of the original study. Though it may point out issues in the hypothesis' generalizability, the initial results' reliability is not truly assessed when deviating from the original study. For example, the facial feedback hypothesis in social psychology proposes that one's emotion

is affected by facial expression ("pen-in-mouth" effect; Strack et al., 1988). A large collaboration across 17 labs performed a direct replication and failed to reproduce the effect (Wagenmakers et al., 2016). However, the replication study's setup included a video camera that recorded participants. This was not included in the original setup, and might have interfered with participants' emotional behavior (Strack, 2016). A recent adversarial collaboration tested several types of manipulations across multiple designs and found that facial feedback (though not specifically the pen-in-mouth manipulation) can increase and even generate the feeling of happiness (Coles et al., 2022). Thus, the relevance of some settings to the effect in question might become evident only after a failed replication attempt.

Null-hypothesis significance testing (NHST)

Neuroscience and psychological statistical analyses are often based on null hypothesis significance testing. In this statistical inference process, a value of a sample statistic derived from the experimental data falls somewhere within a distribution of values, created based on the null hypothesis (according to which there is no significant relationship between the parameters of interest). Suppose the sample statistic is extreme enough (with the size of the extreme region pre-determined by the researcher): in this case, the null hypothesis is rejected, and the test result is deemed significant. Otherwise, the null hypothesis cannot be rejected. This popular inference process typically leads researchers to examine studies dichotomously: was the null hypothesis rejected? Often, rejection of the null hypothesis (due to a statistically significant sample statistic) is interpreted as evidence supporting the alternative hypothesis (e.g., "a relationship was found", or "the relationship was replicated").

However, rejection of the null hypothesis does not necessarily provide support for the alternative hypothesis. Criticisms of this inference process have been raised for decades; as Rozeboom (1960) put it, "a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected". NHST's common use and binary interpretation ignore the hypothesis testing as a degree of belief (or disbelief) in the hypothesis, with confidence in either position strengthening based on experimental evidence. Instead, the conventional use of the NHST framework wrongfully reduces the inference process into a simple binary test. A small difference in the test statistic can turn a negative result into a positive one (i.e., rejection of the null), with the turning point being completely arbitrary (usually, a p-value of 0.05; Rozeboom, 1960), while ignoring the full context of the NHST framework and its shortcomings (Szucs and Ioannidis, 2017). Colquhoun (2014) used simulated t-tests to demonstrate that using NHST with a criterion of p-value being 0.05 leads to a 30% false alarm rate (notably, on its own, a single test with p-value of 0.05 will have a 5% false alarm rate, assuming all assumptions are met). Thus, the reliance of common neuroscientific analysis methods on NHST contributes to replication failures.

An additional issue with the frequentist approach is that when multiple comparisons are made, the chances of receiving a significant effect increases. This issue is highly concerning in neuroimaging data analysis. In brain imaging, hypothesis tests are done simultaneously over many data points, to examine which of them are significant. For example, in fMRI, neural data are recorded and transformed into a series of three-dimensional brain images. Each such image is segmented into voxels (three-dimensional pixels), such that the image is divided into hundreds of thousands of voxels. Hundreds of such images are acquired (and segmented into voxels) during an experiment. Then, when researchers aim to test which areas of the brain are activated during the task, they perform a statistical test for each to examine whether the activity is significant. Thus, multiple hypothesis tests are performed, each of them with a criterion of 5% type-I error (i.e., a p-value of 0.05). As Lindquist and Mejia demonstrate (2015), for every 100,000 tests performed with a p-value threshold of 0.05, 5000 type-I errors will be observed. A similar issue arises in EEG studies, where hundreds of scalp electrodes record thousands of times per second. A point-by-point analysis of the electrical activity in these electrodes over time results in a similar multiple comparisons issue. Notably, the multiple comparisons issue is not unique to imaging; when a study performs multiple statistical tests to compare groups or recognize a relationship, the probability of finding a significant effect due to sampling error increases. Thus, whenever multiple statistical tests are performed simultaneously, the multiple comparisons issue arises. In the next section ("Increasing Replicability"), we will discuss how the multiple comparisons issue is addressed and corrected.

Questionable research practices (QRPs)

Despite the critique of the NHST framework, neuroscience and psychological studies still widely adopt this framework with a threshold of $p \leq 0.05$ for rejecting the null hypothesis. The 5% false-positive rate threshold (when correcting for multiple comparisons) is expected to minimize type-I errors. Yet, even with this strict limit, questionable research practices (QRPs; National Academy of Sciences, 1992) make it very easy to report a statistically significant result that is false. These behaviors derive from the researchers' degrees of freedom when performing an experiment and are characterized by a trial-and-error behavior where different analyses are conducted until a significant result is reached (also termed "fishing expeditions" or "p-hacking", Gelman and Loken, 2013). Notably, the underlying assumption is that, unlike deliberate misconduct behaviors, QRPs are performed without dishonest intention. These practices might stem from old habits, reliance on old literature, biases that researchers have due to being incentivized based on the publication of statistically significant results, lack of knowledge, or ill-defined research hypotheses.

Choice of sample size

Two problematic practices with respect to sample size can lead to false results. First, when not defining the sample size of one's experiment in advance, a researcher can perform an interim analysis of the data and see if the result is significant. When testing the hypothesis on the collected sample, the researcher can decide to continue collecting data, until a significant result is obtained. As Simmons et al. (2011) have shown in a simulation study, the practice of stopping data collection when significance is reached

can easily lead to inflated false-positive rates, especially in small sample sizes. Second, under the NHST framework, small sample sizes contribute to underpowered study designs (Simmons et al., 2011), which increases the probability that any significant effect is a type-I error (Button et al., 2013).

Hypothesizing after the results are known (HARKing; Kerr, 1998)

A confirmatory scientific study is driven by hypotheses, which are operationalized and tested in an experiment. Once the data is collected, the a priori hypotheses according to which the study was designed can be tested. HARKing is a sub-cluster of QRPs in which hypotheses are tested, reported, or omitted after the results of the study are already known. Murphy and Aguinis (2019) recognized four types of HARKing: hypothesis proliferation, THARKing (transparent HARKing), cherry-picking, and question trolling.

- Hypothesis proliferation is a case in which the researcher tests hypotheses that were not originally planned. The data itself might expose interesting trends and even inspire the generation of additional research questions. However, as not originally included in the research plan, these post-hoc hypotheses should be reported as such. Presenting a newly observed result as if it confirms a pre-existing hypothesis yields type-I errors, because "testing" a hypothesis on the data that generated the hypothesis to begin with ("double-dipping") invalidates the statistics (Wagenmakers et al., 2012). Thus, reporting post-hoc hypotheses as if they were a priori hypotheses is a form of HARKing, leading to a failure to replicate the original result.
- THARKing occurs when researchers form novel hypotheses in the discussion section of the paper, based on the study's results - for example, presenting a hypothesis that was contradicted by the data as a "competing hypothesis". This can lead to overfitting of the data and impedes a theory's potential for predictive precision (Hitchcock and Sober, 2004). A novel hypothesis based on data that were collected with the goal of testing a different hypothesis, was never truly tested. Thus, a replication attempt of such a hypothesis may fail to demonstrate the same trend.
- Cherry-picking is the act of reporting only the measures that were found to have a significant effect, when in fact, other dependent measures were collected as well. Retroactively dropping experimental tests which did not yield significant results increases studies' false-positive rate (Simmons et al., 2011). Cherry-picking is also related to the multiple comparisons problem: when reporting only a subset of multiple statistical significance tests that were made without correcting for multiple comparisons, the probability of type-I errors increases.
- Question trolling is another form of exploratory analysis, in which researchers sift through existing data without a clear hypothesis to find results that might be of interest. This is a data-driven approach but reported as confirming a hypothesis. Thus, it distorts the scientific process and biases the scientific literature as the problem of "double-dipping" once again increases the probability of a type-I error rate. This exploratory method can be acceptable, as long as it is reported and interpreted as such. When "massaging" the data to find an interesting result and formulate a hypothesis, it can no longer be used to test the hypothesis it just helped to formulate.

Thus, a replication of a finding that was achieved by HARKing can fail to reach similar conclusions as the original study, as those conclusions were tailored to the originally collected data.

Data analysis choices

If not pre-defining how exactly each hypothesis will be tested, researchers might "fish" for statistical tests that yield the desired results ("p-value fishing", Neuroskeptic, 2012; or "data massaging", Wagenmakers et al., 2012). One might think that executing many different tests until one yields a significant result is a sign of a thorough procedure of "truth-seeking". However, reporting an exploratory study as a confirmatory test of a hypothesis is a QRP. Presenting an exploratory result as a confirmatory test using a selected method leads to type-I errors (Wagenmakers et al., 2012). Thus, the result of an exploratory analysis (presented as confirming a hypothesis) can fail to replicate.

Outlier exclusion

If not pre-defining which observations need to be excluded from the study, researchers have the freedom to create an exclusion threshold based on their collected observations. This allows researchers to discard data that, once viewing the results, might seem to "ruin" the expected result. This creates a lack of standardization in the field and makes findings hard to interpret considering the different criteria between studies. In addition, post-hoc decisions regarding outlier exclusion affect the significance of the result, which leads to an increase in false-positive findings (John et al., 2012), especially when researchers exclude data based on its impact on analysis results. Without a clear justification for excluding samples ahead of time, replication attempts can fail to find the original effect (that might as well have been a type-I error).

Errors

Errors can occur at any stage of the experimental process: from the study design to the experiment's execution, analysis pipelines, and statistical testing and interpretation. For example, given the NHST framework, a neuroscience meta-analysis showed that about 50% of the examined studies did not perform the correct statistical analysis required to actually test a difference between two groups (by performing subgroup analysis without interaction tests, Nieuwenhuis et al., 2011). Another famous example is the dead salmon experiment, performed by Bennett et al. (2009) to demonstrate the importance of correcting for multiple comparisons. In this fMRI study, the group found significant activations in the brain of a dead salmon. Thus, incorrect analysis of the data can lead to false-

positive findings. Although the need for correcting for multiple comparisons was known in the field, this well-known study promoted it even further. Thus, failed replication attempts can reveal statistical analysis biases and mistakes. For example, a study demonstrating the effects of perceptual load on the human visual cortex was retracted after a graduate student failed to replicate it and managed to attribute the original finding to unintentionally biased analysis methods (de Haas et al., 2014).

Errors can also occur during the implementation of the study. Experiments are heavily based on computer software: from the software used to run the study, to the code used to process and analyze the data, researchers rely on a mixture of software products to execute their experiments and extract results. Inevitably, these programs might have errors in them, which can lead both to erroneous conclusions that cannot be replicated, and to failed replications where a mistake hindered the inference process. For example, Eklund et al. (2016) revealed that even prominent software packages used for fMRI data analysis contained flaws that lead to an increase in false-positive rates. When software packages that researchers use contain flaws, it is not surprising that an in-house software that single researchers develop has errors. Indeed, coding errors do happen, and unfortunately, when found only after publication, they might lead to retraction of the published work (Dolk et al., 2019).

File drawer effect

The file drawer effect still exists, as high-impact journals continue to publish positive findings (results that are considered statistically significant under the NHST framework) rather than null results or replication attempts. Researchers' performance and skills are often assessed based on the volume of their published studies. As such, scientists are pressured to increase their number of publications, which in turn decreases their quality. In addition, as null results are often not accepted for publication, researchers hardly ever try to publish them (Smaldino and McElreath, 2016). This means that the published literature represents a skewed, partial sample out of all studies conducted in the field. Thus, a replication of a finding might fail due to it reflecting a false-positive result being an outlier out of an unknown amount of failed, unpublished experiments.

Insufficient reporting

Despite having standardized formats of publications in different journals, no uniform standard exists for the level of detail one must provide in the manuscript itself. Even when journals define basic reporting requirements, they are rarely enforced: for example, Clayson et al. (2019) showed that in EEG studies, published papers fulfilled only 63% of their corresponding journals' reporting guidelines. And so, it is not surprising that the vast majority of studies do not contain the sufficient amount of detail required to replicate them (Errington et al., 2021). Thus, a replication attempt might as well fail because the authors of the original manuscript failed to provide a transparent account of their experiment.

Deliberate misconduct: fabrication, falsification, and plagiarism (FFP)

FFP are assumed to be the worst research practices that scientists agree should be avoided (Steneck, 2006). Plagiarism does not directly affect the replication of a study; despite breaking the trust of other colleagues, as long as the reported results are real, the study can be assumed to be reliable. On the other hand, fabrication and falsification of a study entail presenting false results, which of course explains a failed replication attempt of the original "study". This behavior is, of course, unscientific and biases the knowledge aggregated in the literature.

The original result was a type I error

An experiment with a flawless methodology that is reported with full transparency can still fail to replicate, simply because of an actual type-I error. This might stem from using a small sample size, resulting in low statistical power. Effects found in underpowered studies are likely to represent a type-I error (Button et al., 2013). Alternatively, type-I errors can simply reflect new directions or novel methodologies to be explored, especially in less mature research fields. The odds that a study will replicate increase when the effect in question has already been thoroughly researched (Nosek et al., 2022).

The replicated result is a type II error

In a complimentary manner, a failed replication attempt can reflect a false negative result, also known as a type II error. Of course, if the replication study itself suffers from small sample sizes or QRPs, it will fail to achieve the previously found results.

Evaluating replication results

What defines a successful replication? Intuitively, after repeating the experiment and the analyses, a replication succeeds if it yields similar results and allows us to reach similar conclusions. Clearly, we do not expect the results to be identical. If so, in what aspect should the replication results be similar to the original study for the replication to be considered successful?

Overall, there are two types of approaches for estimating the success of a replication study: a frequentist type, and a Bayesian type. **Under the NHST framework**, some might choose to define the replicability of a reported study with respect to the significance of the finding in the replication attempt, so that a significant effect is considered a successful replication (e.g., Soto, 2019). Yet, failing to reject the null hypothesis in a replication study is not equivalent to accepting the null hypothesis. In addition, the approach of comparing the p-values of the original and replication studies is problematic because a difference in significance between the original and replication p-values does not necessarily indicate that the difference between the studies is significant (Nieuwenhuis et al., 2011). **Bayesian methods** allow interpreting the evidence as indicating either a failed or a successful

replication. In addition, with Bayesian methods, replication results can be interpreted as supporting neither the null hypothesis nor the alternative, providing insight regarding the information that can be derived from the replication experiment. Still, Bayesian methods suffer from caveats as well. For example, one can use a default Bayes hypothesis test, which asks whether the original effect is observed in the replication study data (Dienes, 2011). In this framework, one hypothesis is that the original effect was spurious, and the alternative hypothesis is that the replication effect is compatible with the one found in the original experiment. The replication Bayes factor quantifies the change in evidence provided by the replication experiment, given that the evidence provided by the original study is already available. Yet, when the original study's p-value supports the alternative hypothesis while the replication attempt's Bayesian analysis supports the null hypothesis, it is unclear whether the original finding indeed failed to replicate (Verhagen and Wagenmakers, 2014). An alternative approach (Verhagen and Wagenmakers, 2014) is to calculate the posterior distribution based on the original study's data and use it as a prior for the replication study's assessment. This framework compares the null hypothesis (zero effect size) against the hypothesis that the effect derives from the posterior probability distribution of the original finding. However, generating the posterior distribution from the original study's data is not a trivial procedure, and it is hard to implement with statistical designs that include ANOVAs (rather than simpler comparisons). Thus, Ly et al. (2019) suggest calculating the replication Bayes factor by dividing the complete Bayes factor (based on evidence from both the original and replication study) by the Bayes factor of the original study.

Effect size

Researchers can assess the replication attempt by comparing the replication's effect size with the originally reported one, such that a weaker replicated effect size means the replication has failed (Hagger et al., 2016). This might be problematic in behavioral neuroscience domains in which effect sizes are generally small, making it difficult to set an effect size below which the hypothesis cannot be accepted. In addition, an unfeasibly large sample size might be required to replicate small effect sizes such that the replication has enough statistical power. For example, the "small telescopes approach", suggested by Simonsohn (2015), measures the detectability of the finding in question: instead of using the original effect size as the benchmark, a small effect size in this context is defined as one that would give the original study 33% power (given the sample size of the original study). Then, the replication effect size is tested against the small effect size, such that if the replication effect is smaller than the small effect size, the replication has failed, and the studied effect is not large enough to have been detectable with the original sample size (the "original telescope" is "too small"). Notably, if a replication attempt fails to fail the small telescope test (i.e., the replication's effect size is *not* smaller than the predefined small effect), it cannot be interpreted as a successful attempt. This is an example of a caveat of the NHST framework when comparing studies, that can be solved by taking a Bayesian approach.

Confidence intervals

Are sometimes mistakenly thought to be an approach to test replicability. However, as it is misguided for multiple reasons, it should be refrained from. The first downside of this approach is that underpowered studies have wide CIs, so the CI comparison can succeed simply because the original study lacked statistical power. As such, it is hard to test the replicability of the original finding using this method, as for a replication attempt to fail, the original study needs to be adequately powered. Second, the common interpretation of CIs (as indicating anything about the true parameter lying within the boundaries) is incorrect (Hoekstra et al., 2014): The correct interpretation of a 95% CI is that if we were to repeat the experiment, 95% of the CIs we would receive would have contained the true mean.

A meta-analytic approach for assessing replicability

Examine replication attempts of the same study across labs, such that heterogeneous results imply the original finding is unreliable (e.g., Hagger et al., 2016). If the heterogeneity is small, this indicates agreement across the studies and can support the original conclusion (such as in Klein et al., 2018). However, once again, studies with small sample sizes may lead to more heterogeneous results, making it difficult to deem a replication as a failure.

Notably, there is no convergence on a standardized metric to evaluate replication results. Some researchers claim that there should be no single standardized method to test the replication results (Anderson and Maxwell, 2016), as different criteria test different aspects of the replication compared with the original results. Thus, the assessment of a replication attempt can benefit from using multiple complementary metrics (Marsman et al., 2017).

Increasing replicability

Despite the lack of convergence in measuring replicability, the replication crisis drew attention to causes for replication failures stemming from substandard research conventions. Awareness of these tendencies has facilitated attempts to improve research practices, hoping to increase replicability. Below we provide an overview of possible solutions (summarized in **Table 1**). None of them guarantees that a finding will replicate; yet, by improving scientific practices, researchers can produce research of high quality that contributes to scientific knowledge. Then, thorough replication attempts could be executed, to test whether the original findings are reliable.

Table 1 Factors contributing to replication failure and suggested improved scientific practices to mitigate them. NHST: Null Hypothesis Significance Testing; QRPs: Questionable Research Practices; FFPs: Fabrication, Falsification, and Plagiarism.

Contributing factor for replication failure		*Improving scientific practices*
NHST	Multiple comparisons	- Correct for multiple comparisons (Bonferroni, FWE, FDR)
	Significance interpretations	- Consider effect sizes and confidence intervals, interpret with caution - Bayesian framework
QRPs	Sample size	- Determine (and justify) sample size based on power calculations - Preregister the size and do not add observations during data collection
	HARKing: hypothesis proliferation THARKing cherry-picking Question trolling	- Preregister all the study's hypotheses - Plan the experimental design including all conditions and dependent measures - Preregister the study design - Report exploratory analyses as such
	Data analysis choices	- Define what methods will be used to test each hypothesis - Preregister these choices
	Outlier exclusion	- Decide on outlier exclusion criterion in advance - Preregister the criterion
Errors	Hypothesis errors	- Choose the appropriate statistical analyses for the hypothesis, make sure to know their limitations and what corrections need to be performed - Preregister all the methods and pipelines
	Implementation errors	- Don't reinvent the wheel; always prefer existing tools - Test your processing pipelines as thoroughly as you can - Be transparent: share all the processing and analyses codes
File Drawer effect		- Registered reports
Insufficient reporting		- Be transparent: share all the materials used in the experiment: experimental software, stimuli, software - Provide a detailed, comprehensive report of all the settings and methods used in the study.
FFP		- Do your best

Deviation from the original study

Variations stemming from a lack of knowledge are due to insufficient reporting of the original research. As detailed below, they can be solved by increasing transparency and sharing the study materials (for example, by using and sharing checklists; Nichols et al., 2017). Deviations stemming from subjective differences in what researchers define as "relevant" for the scientific question can be solved in one of two ways: one option is to ensure that a direct replication recreates as much of the original study as possible, without intentionally leaving out parameters and procedures that seem unimportant. The other option is to perform a conceptual replication instead of a direct replication. The goal of a conceptual replication is different from a direct one. Conceptual replications test a hypothesis using different methods, trying to test whether the underlying theory is correct (Nosek and Lakens, 2014). In this framework, the question is whether a theory is true and can be generalized above and beyond the methods used to test it (Goodman et al., 2016; Nosek and Errington, 2017).

Null-hypothesis significance testing (NHST)

Methods for controlling type-I errors in the NHST framework have been used to try to refrain from false-positive results that would fail to replicate. Corrections for multiple comparisons such as Bonferroni (1936), False Discovery Rate (FDR; Benjamini and Hochberg, 1995), and Familywise Error (FWE; Hochberg and Tamhane, 1987) have been implemented, especially in neuroimaging studies where multiple comparisons are abundant. Yet, even when no multiple comparisons are made, researchers have argued that the traditional frequentist approach with a p-value threshold of 0.05 is likely to generate type-I errors. Suggestions such as changing the criterion to 0.005 and even 0.001 have been made (e.g., Johnson, 2013), as well as calls for caution when interpreting results of significance testing (Colquhoun, 2014; Lakens et al., 2018).

Instead, researchers of a to-be-replicated study can choose to replace the NHST framework altogether and adopt Bayesian statistics when testing predictions. In Bayesian approaches, a prior probability (derived from a theory or previously collected data) and a likelihood function (derived from a statistical model of the current observed data) are used to create a posterior probability—thus, updating the belief in a certain hypothesis. In addition to their ability to provide support for the null hypothesis (as the null is a competing model to the research hypothesis), Bayesian methods can also be useful in studies with small sample sizes. The Bayesian framework can be used instead of the frequentist one in many common statistical neuroscience analyses. For example, Rouder et al. (2009) suggested using a Bayesian alternative for the t-test, to be able to collect evidence in support of the null hypothesis. In addition, Penny et al. (2005) developed methods to use Bayesian analysis in fMRI experiments. Thus, the NHST approach can be replaced with a Bayesian approach, to collect evidence for both supporting and opposing hypotheses, and allow explicit updating the belief in a theory in light of new evidence. Notably, for some analyses the transition is not trivial, and entails computational and implementation challenges.

Preregistration

Minimizing the prevalence of QRPs requires researchers to carefully think about their study design choices, analysis plans, and predicted results, and document them. Preregistration (Wagenmakers et al., 2012) allows researchers to document their planned experiments in advance, to make sure the executed study is indeed confirmatory (rather than exploratory). Notably, many of the issues stemming from QRPs can be partially mitigated by preregistering the relevant parameters, with registries such as OSF (Foster and Deardorff, 2017) and AsPredicted (Bedics, 2018; see "Relevant Websites"). However, despite being necessary, preregistrations alone are not sufficient, and further actions are needed:

Well-powered study

A well-powered study requires an adequate number of data points; either collecting data from many subjects or collecting many observations (e.g., trials) in case the number of participants is small (Smith and Little, 2018). In addition, to prevent waste of resources (by performing the experiment on more subjects than needed for a well-powered experiment, or on too few subjects to allow drawing conclusions), it is crucial to define the desired number of observations based on power calculations (e.g., Brysbaert, 2019) rather than on arbitrary or historical practices (thus it is advised to use tools such as G*Power, and Neuropower; see "Relevant Websites").

Hypothesizing after the results are known (HARKing)

The NHST framework is designed for testing predictions, not for hypothesis generation. And so, exploratory studies cannot be considered as strong evidence supporting a hypothesis. For a study to be confirmatory, researchers must explicitly preregister all the hypotheses, and all the dependent measures and study conditions. Preregistration of all the study's hypotheses and dependent measures creates a distinction between planned and post-hoc results, allowing transparency regarding initial and postdictive hypotheses (Nosek et al., 2018). Preregistration reduces cherry-picking, as all parameters are listed in advance, and so need to be reported once the results are analyzed. Any novel hypothesis that was not originally listed and is based on observing the data is thus treated as exploratory, mitigating both hypothesis proliferation and THARKing. Thus, a detailed and thorough preregistration, that outlines all the hypotheses and their predictions, is a first step in reducing HARKing.

Data analysis choices

For a study to be confirmatory, researchers should preregister all the analyses and statistical tests ahead of time. As the reader will see below ("Insufficient Reporting"), once again, "the devil is in the details", and registering the planned tests alone still leaves many degrees of freedom to the researchers (both the original authors and the replicators). Yet, without an explicit, preregistered analysis plan, a study is not confirmatory, and replication cannot be carried out without sufficient information.

Outlier exclusion

To avoid confirmation bias, in which an extreme observation seems like it should be dropped to "make sense" of the analysis, a clear definition of what makes an observation an outlier should also be decided upon in advance. Notably, preregistration alone does not solve the standardization problem: without a clear justification for the choice of exclusion thresholds, researchers are free to set their different criteria based on their own preferences. However, a pre-defined exclusion criterion greatly contributes to the credibility of a replication attempt of the original study.

Errors

Errors are inevitable. Researchers should work hard attempting to avoid errors but know that, nevertheless, they will occur. Working with this assumption in mind, researchers should create systems of checks and balances that help them to validate their processes:

Hypothesis errors

Preregistration of the study's sample size, hypothesis, experimental conditions, and dependent variables does not prevent researchers from making mistakes such as choosing an unfitting statistical test or not correcting for multiple comparisons.

Researchers should consult their community to ensure that their hypotheses' operationalization is correct and that they are using the right tools to test them. Mistakes can still occur, as no scientific Oracle exists. However, awareness of the possibility of making such mistakes can help researchers look for potential errors in their own design and drive scientists to plan and execute proper solutions.

Implementation errors

Software-based errors are also prevalent and, in a way, unpreventable (Sharma, 2016). However, large software packages have higher chances of finding (and then fixing) implementation errors than ad-hoc software programmed by a single researcher, especially if they are open-source. Open-source software is one that is available for the public to access and modify. When the code is accessible to everyone, it increases the probability of characterizing and solving software-based errors (Linus' law: "given enough eyeballs, all bugs are shallow"). Thus, relying on existing solutions is always better than reinventing the wheel. Another aspect that is key to airing out mistakes and executing a proper replication is transparency: sharing the codes for the entire processing and analysis pipeline. By the mere act of opening the codes for general inspection, the community can help in locating and correcting errors, increasing efficiency, and performing better replications based on public materials (Trisovic et al., 2022).

File drawer effect

A solution to the file-drawer effect depends on a paradigm shift in the way scientific journals operate. Below we describe actions taken by journals in that direction.

Publishing null results

Some journals announced that they encourage publications of null results (e.g., Nature, 2020). Publications of null findings might reveal unfavorable scientific directions (as opposed to promising ones) or other methodological constraints researchers and funding agencies should be aware of.

Registered reports

Many journals have adopted the registered report model (Chambers, 2012; Chambers et al., 2014; Nosek and Lakens, 2014). According to this model, scientists submit their study proposals (including the hypothesis, methods, and analysis plans) to be peer-reviewed and accepted by the journal. Crucially, this step occurs before the data is collected and analyzed. Thus, the decision to accept the study for publication occurs regardless of whether the data supports the hypothesis. Once the study is concluded, the modified manuscript, with deviations from the planned protocol clearly marked, is submitted to the journal for publication. By having to accept or reject publication when the results are still unknown, registered reports emphasize the scientific process rather than the outcome. Similar to preregistrations, registered reports also aid in mitigating HARKing and THARKing.

Post-publication peer review

Alternatively, some journals embraced a post-publication peer review approach (Markie, 2015). In this model, a study is published in an online, open-access platform where the reviewers' comments (and any revisions made following) are also published (Chambers and Tzavella, 2021).

Insufficient reporting

Two important guiding principles can help mitigate issues stemming from insufficient reporting: transparency, and detail.

Transparency

Preregistration is a powerful tool that allows documentation of the planned study design and analyses, but it does not specify the level of detail the report needs to include. That is why a clear distinction between what was preregistered vs. exploratory must be made clear. Despite the need to thoroughly register as much as possible in advance, reporting each aspect and step of the study is impossible during the preregistration phase (as not everything is known in advance). In addition, a comprehensive report of the entire study's parameters and settings might be unfeasible at the publication stage due to the limitations of the publishing journal. This is why transparency—open materials (data, study paradigm) and codes (pre-processing and analyses pipelines)—is crucial. Openness and transparency are the foundation of increasing experiments' reliability. Together, preregistration, detailed reporting, and open resources facilitate a proper replication attempt of the original study.

Detail

There have been attempts to create a standard basic level of detail, required for reporting a study. The Organization of Human Brain Mapping (OHBM) erected Committees on Best Practices in Data Analysis and Sharing (COBIDAS) to create standard reporting practices. They have developed best practices in data analysis and result sharing for neuroimaging studies (fMRI: Nichols et al., 2017; MEEG: Pernet et al., 2018). Even those are sometimes considered not comprehensive or clear enough (Pernet et al., 2020), and recommendations continue to proliferate as different imaging technologies continue to develop. Yet, the fundamental principle is identical: transparent, comprehensive, detailed reporting (together with sharing of all the study's materials, methods, pipelines, and settings) is crucial for a proper replication to occur.

Deliberate misconduct: fabrication, falsification, and plagiarism (FFP)

Do your best

The competitive atmosphere of the scientific community might encourage some researchers to look for loopholes, allowing them to increase their publication volume in mischievous ways. Science is done to gain knowledge about the world; neuroscience is a field that aspires to improve our understanding of creatures' minds and behaviors. It is important to remember that despite the importance of publications for a researcher's career, publishing results of scientific inquiries is, first and foremost, for the greater good: helping increase the existing body of knowledge regarding a particular phenomenon. Scholars should bear this in mind when conducting research and avoid FFPs altogether.

Challenges and future directions

Detailed, thorough reporting has proven to be crucial for a study's replicability and the ability to assess its contribution to the field. Yet, journals do not enforce reporting and data sharing guidelines (Clayson et al., 2019). Thus, researchers are still able to publish their work with limited reports that lack crucial information, and without sharing all the study's resources (e.g., materials, pipelines, codes). Insufficient reporting and lack of sharing hinder replication studies, making it impossible to execute a proper replication attempt. Moreover, when a replication attempt fails, the interpretation of the failure is unclear; were the original study's findings erroneous? Or did the replication study fail to capture a real effect? Thus, the definition of what makes a failed replication attempt and the conclusions that can be derived from it are unclear. Notably, a successful replication attempt is also subject to interpretation, and might be affected by the method used to define success. As reviewed in the article, replication attempts vary in their type, extent, and the metrics used to evaluate their outcome.

In addition, even with clear reporting guidelines and standards, there remain many degrees of freedom in processing and analyzing experimental data. For example, as previously mentioned, the COBIDAS standards are extensive, but not comprehensive enough to capture all the parameters that differ between different research groups. The impact of researchers' degrees of freedom can be immense: For example, in fMRI data analysis, there are multiple ways to implement motion correction, spatial smoothing, normalization, and high-pass filtering (see Box 3 in Poldrack et al., 2017). Furthermore, different approaches are used across research groups even in implementing corrections for multiple comparisons. Thus, at each step of the process, a researcher makes a data analysis choice that can influence the results. Demonstrating this effect, Botvinik-Nezer et al. (2020) have recently shown that when 70 research groups were given the same dataset and set to test the same hypotheses, their conclusions differed considerably. Thus, detailing the analyses pipelines is necessary to its fullest extent.

These issues have been emphasized in recent years, as more studies examine their ramifications in neuroscience. New initiatives such as multi-lab collaborations are on the rise (e.g., fMRIflows: Notter et al., 2021; EEGManyLabs: Pavlov et al., 2021). Registered reports are becoming more common (Goffin et al., 2019). The hostility that failed replications encounter was suggested to be resolved by making clear precommitments regarding replication success (Nosek and Errington, 2020). Awareness of the importance of increasing replicability and open science has also reached scientific funding agencies, with new ideas and initiatives being highlighted (e.g., Smaldino et al., 2019). Thus, although much more work still needs to be done, neuroscience studies are progressing toward improved replicability via community accepted research practices, increased transparency, and data sharing.

Summary

Following the replication crisis, the importance of replication became apparent across disciplines. In neuroscience, adopting practices that increase research quality (and, therefore, might also increase replicability) has been on the rise. Although not consistently enforced by scientific journals, improved scientific practices are becoming more routine. Studies are being preregistered, and their materials are shared with the public. New formats, such as registered reports, both facilitate transparency in research practices and allow for null results to be published. Guidelines aiming to improve research practices are becoming more popular among scientists. Direct replication attempts have been published, and their conclusions raised important discussions about the way replicability should be acknowledged in the scientific literature. And indeed, many questions about assessing replication and implementing standard research and report practices remain. However, the attention the field allocates to the importance of replicability has encouraged scientists to invest in better research practices. Will these practices improve replicability? Time will tell.

Acknowledgments

We would like to thank Prof. Liad Mudrik, Prof. Russell Poldrack, and Dr. Rotem Botvinik-Nezer for reviewing early drafts of this chapter.

References

Aldhous, P., 2011. Journal Rejects Studies Contradicting Precognition. New Scientist.

Anderson, S.F., Maxwell, S.E., 2016. There's more than one way to conduct a replication study: beyond statistical significance. Psychol. Methods 21 (1), 1–12. https://doi.org/10.1037/met0000051.

Bedics, J.D., 2018. AsPredicted. Available at. https://osf.io/gnxav.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57 (1), 289–300. https://doi.org/10.2307/2346101.

Bennett, C.M., Miller, M.B., Wolford, G.L., 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. NeuroImage 47 (Suppl. 1), S125.

Bonferroni, C., 1936. Teoria statistica delle classi e calcolo delle probabilita, vol. 8. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, pp. 3–62. Available at. https://cir.nii.ac.jp/crid/1570009749360424576.

Botvinik-Nezer, R., et al., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582 (7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9.

Brysbaert, M., 2019. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. J. Cognit. 2 (1). https://doi.org/10.5334/joc.72.

Button, K.S., et al., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14 (5), 365–376. https://doi.org/10.1038/nrn3475.

Camerer, C.F., et al., 2016. Evaluating replicability of laboratory experiments in economics. Science 351 (6280), 1433–1436. https://doi.org/10.1126/science.aaf0918.

Casali, A.G., et al., 2013. A theoretically based index of consciousness independent of sensory processing and behavior. Sci. Transl. Med. 5 (198), 198ra105. https://doi.org/10.1126/scitranslmed.3006294.

Chambers, C.D., 2012. Changing the Culture of Scientific Publishing From Within, NeuroChambers. Available at. https://neurochambers.blogspot.com/2012/10/changing-culture-of-scientific.html.

Chambers, C.D., et al., 2014. Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. AIMS Neurosci. 1 (1), 4–17. https://doi.org/10.3934/Neuroscience.2014.1.4.

Chambers, C.D., Tzavella, L., 2021. The past, present and future of Registered Reports. Nat. Human Behav. 6 (1), 29–42. https://doi.org/10.1038/s41562-021-01193-7.

Clayson, P.E., et al., 2019. Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: barriers to reproducibility and replicability. Psychophysiology 56 (11). https://doi.org/10.1111/psyp.13437.

Cohen, J., 1962. The statistical power of abnormal-social psychological research: a review. Soc. Psychol. 65 (3), 145–153. https://doi.org/10.1037/h0045186.

Cohen, J., 1990. Things I have learned (so far). Am. Psychol. 45 (12), 1304–1312. https://doi.org/10.1037/10109-028.

Coles, N.A., et al., 2022. A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. Nat. Human Behav. https://doi.org/10.1038/s41562-022-01458-9.

Colquhoun, D., 2014. An investigation of the false discovery rate and the misinterpretation of p -values. R. Soc. Open Sci. 1 (3), 140216. https://doi.org/10.1098/rsos.140216.

Dienes, Z., 2011. Bayesian versus orthodox statistics: which side are you on? Perspect. Psychol. Sci. 6 (3), 274–290. https://doi.org/10.1177/1745691611406920.

Dolk, T., et al., 2019. Retraction notice to "Auditory (dis-) fluency triggers sequential processing adjustments" [ACTPSY 191 (2018) 69-75]. Acta Psychol. 198, 102886. https://doi.org/10.1016/j.actpsy.2019.102886.

Donoho, D.L., et al., 2008. Reproducible research in computational harmonic analysis. Comput. Sci. Eng. 11 (1), 8–18. https://doi.org/10.1109/MCSE.2009.15.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. U. S. A. 113 (28), 7900–7905. https://doi.org/10.1073/pnas.1602413113.

Errington, T.M., et al., 2014. An open investigation of the reproducibility of cancer biology research. eLife 3, e04333. https://doi.org/10.7554/eLife.04333.

Errington, T.M., et al., 2021. Investigating the replicability of preclinical cancer biology. eLife 10, e71601. https://doi.org/10.7554/eLife.71601.

Foster, E.D., Deardorff, A., 2017. Open science framework (OSF). J. Med. Libr. Assoc. 105 (2). https://doi.org/10.5195/jmla.2017.88.

Gelman, A., Loken, E., 2013. The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There is No "Fishing Expedition" or "P-hacking" and the Research Hypothesis Was Posited Ahead of Time. Department of Statistics, Columbia University, p. 348.

Goffin, C., et al., 2019. Does writing handedness affect neural representation of symbolic number? An fMRI adaptation study. Cortex 121, 27–43. https://doi.org/10.1016/j.cortex.2019.07.017.

Goodman, S., Greenland, S., 2007. Assessing the Unreliability of the Medical Literature: A Response to "Why Most Published Research Findings are False". Dept. of Biostatistics Working Papers. Johns Hopkins University. Available at. https://biostats.bepress.com/jhubiostat/paper135.

Goodman, S.N., Fanelli, D., Ioannidis, J.P.A., 2016. What does research reproducibility mean? Sci. Transl. Med. 8 (341), 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027.

de Haas, B., et al., 2014. RETRACTED: perceptual load affects spatial tuning of neuronal populations in human early visual cortex. Curr. Biol. 30 (23), 4814. https://doi.org/10.1016/j.cub.2013.11.061.

Hagger, M.S., et al., 2016. A multilab preregistered replication of the ego-depletion effect. Perspect. Psychol. Sci. 11 (4), 546–573. https://doi.org/10.1177/1745691616652873.

Hitchcock, C., Sober, E., 2004. Prediction versus accommodation and the risk of overfitting. Br. J. Philos. Sci. 55 (1), 1–34. Available at. https://www.jstor.org/stable/3541832.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. John Wiley & Sons, Inc.

Hoekstra, R., et al., 2014. Robust misinterpretation of confidence intervals. Psychon. Bull. Rev. 21 (5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3.

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

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

Johnson, V.E., 2013. Revised standards for statistical evidence. Proc. Natl. Acad. Sci. U. S. A. 110 (48), 19313–19317. https://doi.org/10.1073/pnas.1313476110.

Kerr, N.L., 1998. HARKing: hypothesizing after the results are known. Pers. Soc. Psychol. Rev. 2 (3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4.

Klein, R.A., et al., 2018. Many labs 2: investigating variation in replicability across samples and settings. Adv. Methods Pract. Psychol. Sci. 1 (4), 443–490. https://doi.org/10.1177/2515245918810225.

Lakens, D., et al., 2018. Justify your alpha. Nat. Human Behav. 2 (3), 168–171. https://doi.org/10.1038/s41562-018-0311-x.

Lindquist, M.A., Mejia, A., 2015. Zen and the art of multiple comparisons. Psychosom. Med. 77 (2), 114–125. https://doi.org/10.1097/PSY.0000000000000148.

Ly, A., et al., 2019. Replication Bayes factors from evidence updating. Behav. Res. Methods 51 (6), 2498–2508. https://doi.org/10.3758/s13428-018-1092-x.

Lykken, D.T., 1968. Statistical significance in psychological research. Psychol. Bull. 70 (3, Pt.1), 151–159. https://doi.org/10.1037/h0026141.

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. https://doi.org/10.1177/1745691612460688.

Markie, M., 2015. Post-publication peer review, in all its guises, is here to stay. Insights UKSG J. 28 (2), 107–110. https://doi.org/10.1629/uksg.245.

Marsman, M., et al., 2017. A Bayesian bird's eye view of "Replications of important results in social psychology". R. Soc. Open Sci. 4 (1), 160426. https://doi.org/10.1098/rsos.160426.

Matheson, B.E., Bohon, C., Lock, J., 2020. Family-based treatment via videoconference: clinical recommendations for treatment providers during COVID-19 and beyond. Int. J. Eat. Disord. 53 (7), 1142–1154. https://doi.org/10.1002/eat.23326.

Murphy, K.R., Aguinis, H., 2019. HARKing: how badly can cherry-picking and question trolling produce bias in published results? J. Bus. Psychol. 34 (1), 1–17. https://doi.org/10.1007/s10869-017-9524-7.
National Academies of Sciences Engineering and Medicine, 2019. Reproducibility and Replicability in Science. National Academies Press (US).
National Academy of Sciences, 1992. Responsible Science, vol. I. National Academies Press, Washington, D.C. https://doi.org/10.17226/1864.
Nature, 2020. In praise of replication studies and null results. Nature 578 (7796), 489–490. https://doi.org/10.1038/d41586-020-00530-6.
Nelson, L.D., Simmons, J., Simonsohn, U., 2018. Psychology's renaissance. Annu. Rev. Psychol. 69 (1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836.
Neuroskeptic, 2012. The nine circles of scientific Hell. Perspect. Psychol. Sci. 7 (6), 643–644. https://doi.org/10.1177/1745691612459519.
Nichols, T.E., et al., 2017. Best practices in data analysis and sharing in neuroimaging using MRI. Nat. Neurosci. 20 (3), 299–303. https://doi.org/10.1038/nn.4500.
Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.-J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. Nat. Neurosci. 14 (9), 1105–1107. https://doi.org/10.1038/nn.2886.
Nosek, B.A., et al., 2018. The preregistration revolution. Proc. Natl. Acad. Sci. U. S. A. 115 (11), 2600–2606. https://doi.org/10.1073/pnas.1708274114.
Nosek, B.A., et al., 2022. Replicability, robustness, and reproducibility in psychological science. Annu. Rev. Psychol. 73, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157.
Nosek, B.A., Errington, T.M., 2017. Reproducibility in cancer biology: making sense of replications. eLife 6, e23383. https://doi.org/10.7554/eLife.23383.
Nosek, B.A., Errington, T.M., 2020. The best time to argue about what a replication means? Before you do it. Nature 583 (7817), 518–520. https://doi.org/10.1038/d41586-020-02142-6.
Nosek, B.A., Lakens, D., 2014. Registered reports. Soc. Psychol. 45 (3), 137–141. https://doi.org/10.1027/1864-9335/a000192.
Notter, M.P., et al., 2021. fMRIflows: a consortium of fully automatic univariate and multivariate fMRI processing pipelines. bioRxiv. https://doi.org/10.1101/2021.03.23.436650.
Open Science Collaboration, 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. Perspect. Psychol. Sci. 7 (6), 657–660. https://doi.org/10.1177/1745691612462588.
Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349. https://doi.org/10.1126/science.aac4716.
Pavlov, Y.G., et al., 2021. #EEGManyLabs: investigating the replicability of influential EEG experiments. Cortex 144, 213–229. https://doi.org/10.1016/j.cortex.2021.03.013.
Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24 (2), 350–362. https://doi.org/10.1016/j.neuroimage.2004.08.034.
Pernet, C., et al., 2018. Best Practices in Data Analysis and Sharing in Neuroimaging Using MEEG. https://doi.org/10.31219/osf.io/a8dhx.
Pernet, C., et al., 2020. Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research. Nat. Neurosci. 23 (12), 1473–1483. https://doi.org/10.1038/s41593-020-00709-0.
Poldrack, R.A., et al., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18 (2), 115–126. https://doi.org/10.1038/nrn.2016.167.
Popper, K., 1959. The Logic of Scientific Discovery. Basic Books, Oxford, England. https://doi.org/10.4324/9780203994627.
Radder, H., 1992. Experimental reproducibility and the experimenters' regress. PSA Proc. Bienn. Meet. Philos. Sci. Assoc. 1992 (1), 63–73. https://doi.org/10.1086/psaprocbienmeetp.1992.1.192744.
Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86 (3), 638. https://doi.org/10.1037/0033-2909.86.3.638.
Rouder, J.N., et al., 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon. Bull. Rev. 16 (2), 225–237. https://doi.org/10.3758/PBR.16.2.225.
Rozeboom, W.W., 1960. The fallacy of the null-hypothesis significance test. Psychol. Bull. 57 (5), 416–428. https://doi.org/10.1037/h0042040.
Schmidt, S., 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Rev. Gen. Psychol. 13 (2), 90–100. https://doi.org/10.1037/a0015108.
Sharma, V., 2016. Why Software Will Never Be Bugless. Available at. yourstory.com. https://yourstory.com/2016/06/software-bugless.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366. https://doi.org/10.1177/0956797611417632.
Simonsohn, U., 2015. Small telescopes: detectability and the evaluation of replication results. Psychol. Sci. 26 (5), 559–569. https://doi.org/10.1177/0956797614567341.
Smaldino, P.E., McElreath, R., 2016. The natural selection of bad science. R. Soc. Open Sci. 3 (9), 160384. https://doi.org/10.1098/rsos.160384.
Smaldino, P.E., Turner, M.A., Contreras Kallens, P.A., 2019. Open science and modified funding lotteries can impede the natural selection of bad science. R. Soc. Open Sci. 6 (7), 190194. https://doi.org/10.1098/rsos.190194.
Smith, P.L., Little, D.R., 2018. Small is beautiful: in defense of the small-N design. Psychon. Bull. Rev. 25 (6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8.
Soto, C.J., 2019. How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. Psychol. Sci. 30 (5), 711–727. https://doi.org/10.1177/0956797619831612.
Steneck, N.H., 2006. Fostering integrity in research: definitions, current knowledge, and future directions. Sci. Eng. Ethics 12 (1), 53–74. https://doi.org/10.1007/PL00022268.
Sterling, T.D., 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. J. Am. Stat. Assoc. 54 (285), 30–34. https://doi.org/10.1080/01621459.1959.10501497.
Strack, F., 2016. Reflection on the smiling registered replication report. Perspect. Psychol. Sci. 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. J. Pers. Soc. Psychol. 54 (5), 768–777. https://doi.org/10.1037/0022-3514.54.5.768.
Szucs, D., Ioannidis, J.P.A., 2017. When null hypothesis significance testing is unsuitable for research: a reassessment. Front. Hum. Neurosci. 11, 390. https://doi.org/10.3389/fnhum.2017.00390.
Trisovic, A., et al., 2022. A large-scale study on research code quality and execution. Sci. Data 9 (1), 60. https://doi.org/10.1038/s41597-022-01143-6.
Tversky, A., Kahneman, D., 1971. Belief in the law of small numbers. Psychol. Bull. 76 (2), 105–110. https://doi.org/10.1037/h0031322.
Van Bavel, J.J., et al., 2020. Using social and behavioural science to support COVID-19 pandemic response. Nat. Human Behav. 4 (5), 460–471. https://doi.org/10.1038/s41562-020-0884-z.
Verhagen, J., Wagenmakers, E.-J., 2014. "Bayesian tests to quantify the result of a replication attempt": correction to Verhagen and Wagenmakers (2014). J. Exp. Psychol. Gen. 143 (6), 2073. https://doi.org/10.1037/a0038326.
Wagenmakers, E.-J., et al., 2012. An agenda for purely confirmatory research. Perspect. Psychol. Sci. 7 (6), 632–638. https://doi.org/10.1177/1745691612463078.
Wagenmakers, E.-J., et al., 2016. Registered replication report: Strack, Martin, & stepper (1988). Perspect. Psychol. Sci. 11 (6), 917–928. https://doi.org/10.1177/1745691616674458.
Zwaan, R.A., et al., 2018. Making replication mainstream. Behav. Brain Sci. 41, e120. https://doi.org/10.1017/S0140525X17001972.

Relevant websites

AsPredicted. https://aspredicted.org/.
NeuroPower. http://neuropowertools.org/.
OSF. https://osf.io/.
G*Power. https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.