General Article

Advances in Methods and Practices in Psychological Science
July-September 2025, Vol. 8, No. 3, pp. 1–36
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459251348431
www.psychologicalscience.org/AMPPS

Bestiary of Questionable Research Practices in Psychology

Tamás Nagy[1], Jane Hergert[2,3], Mahmoud M. Elsherif[4,5], Lukas Wallrich[6], Kathleen Schmidt[7], Tal Waltzer[8], Jason W. Payne[9,10], Biljana Gjoneska[11], Yashvin Seetahul[12], Y. Andre Wang[9], Daniel Scharfenberg[13], Gabriella Tyson[14], Yu-Fang Yang[15], Aleksandrina Skvortsova[16], Samuel Alarie[17,18], Katherine Graves[19], Lukas K. Sotola[20], David Moreau[21,22], and Eva Rubínová[23]

[1]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary; [2]Humanistische Hochschule Berlin, Berlin, Germany; [3]Department of Work and Organizational Psychology, FernUniversität in Hagen, Hagen, Germany; [4]Department of Psychology and Vision Sciences, University of Leicester, Leicester, England; [5]School of Psychology, University of Birmingham, Birmingham, England; [6]Birkbeck Business School, University of London, London, England; [7]Department of Psychology, Ashland University, Ashland, Ohio; [8]Department of Psychology, University of California San Diego, La Jolla, California; [9]Department of Psychology, University of Toronto, Toronto, Ontario, Canada; [10]Department of Psychology, St. Francis Xavier University, Antigonish, Nova Scotia, Canada; [11]Macedonian Academy of Sciences and Arts, Skopje, North Macedonia; [12]Department of Psychology, University of Innsbruck, Innsbruck, Tyrol, Austria; [13]Medical Psychology | Neuropsychology & Gender Studies, Center for Neuropsychological Diagnostics and Intervention (CeNDI), University Hospital Cologne and Faculty of Medicine, University of Cologne, Cologne, Germany; [14]Department of Psychology, King's College London, London, England; [15]Division of Experimental Psychology and Neuropsychology, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany; [16]Institute of Psychology; Health Medical and Neuropsychology, Leiden University, Leiden, Netherlands; [17]Department of Psychology, University of Montreal, Montreal, Quebec, Canada; [18]Department of Cardiology, Quebec Heart and Lung Institute Research Center - Université Laval, Quebec, Canada; [19]Department of Teacher Assessment and Preparation, University of Texas at Arlington, Columbia, Missouri; [20]Department of Psychology Pleasantville, Pace University, Ames, Iowa; [21]School of Psychology, University of Auckland, Auckland, New Zealand; [22]Centre for Brain Research, University of Auckland, Auckland, New Zealand; and [23]School of Psychology, University of Aberdeen, Aberdeen, Scotland

Abstract

Questionable research practices (QRPs) pose a significant threat to the quality of scientific research. However, historically, they remain ill-defined, and a comprehensive list of QRPs is lacking. In this article, we address this concern by defining, collecting, and categorizing QRPs using a community-consensus method. Collaborators of the study agreed on the following definition: QRPs are ways of producing, maintaining, sharing, analyzing, or interpreting data that are likely to produce misleading conclusions, typically in the interest of the researcher. QRPs are not normally considered to include research practices that are prohibited or proscribed in the researcher's field (e.g., fraud, research misconduct). Neither do they include random researcher error (e.g., accidental data loss). Drawing from both iterative discussions and existing literature, we collected, defined, and categorized 40 QRPs for quantitative research. We also considered attributes such as potential harms, detectability, clues, and preventive measures for each QRP. The results suggest that QRPs are pervasive

Corresponding Author:
Tamás Nagy, Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
Email: nagy.tamas@ppk.elte.hu

Creative Commons NonCommercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/), which permits noncommercial use, reproduction, and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

and versatile and have the potential to undermine all stages of the scientific enterprise. This work contributes to the maintenance of research integrity, transparency, and reliability by raising awareness for and improving the understanding of QRPs in quantitative psychological research.

Keywords
credibility crisis, expert consensus, metascience, open science, QRP, research integrity, research methods, open data

Received 5/20/24; Revision accepted 5/5/25

In recent decades, researchers have faced a reckoning with a crisis in replicability and credibility[1] of research, which has spurred efforts to improve psychological science (Eronen & Bringmann, 2021; Open Science Collaboration, 2015; Vazire, 2018; Yarkoni, 2020). Questionable research practices (QRPs) have been discussed as being one of the drivers of this credibility crisis. John et al. (2012) broadly defined QRPs as behaviors that "exploit . . . the gray area of acceptable [scientific] practice," which "can spuriously increase the likelihood of finding evidence in support of a hypothesis" (p. 524). These QRPs might be a direct or indirect result of researchers having too much flexibility when it comes to analytic strategies (often referred to as "the garden of forking paths"; Gelman & Loken, n.d.) and using this flexibility to their advantage in an undisclosed manner (often referred to as having too many "researcher degrees of freedom"; Simmons et al., 2011). Some QRPs are well known, and their corresponding terms are widely used in the scientific community. For example, "HARKing" (Kerr, 1998) refers to the practice of hypothesizing after the results are known, and "p-hacking" refers to repeatedly testing closely related hypotheses until the desired effect is statistically confirmed and reporting only the analysis strategy that "worked" (Rosenthal, 1979). Apart from these, many more QRPs exist, eroding the scientific enterprise.

Exploring the landscape of QRPs reveals a patchwork of studies showing that these practices are ubiquitous. In the United States, John et al. (2012) uncovered a breadth of QRPs among researchers, illuminating how common these practices are in academic settings. In Europe, Fiedler and Schwarz (2016) documented similar concerns in German research institutions, echoing the widespread nature of QRPs. The situation is reflected among the next generation of researchers as well; Brachem et al. (2022) reported on the attitudes and engagement with QRPs among psychology students in Germany, Austria, and Switzerland. Likewise, Gopalakrishna et al. (2022) identified QRPs in the research community in the Netherlands, thus painting a picture of a practice that transcends cultural and academic boundaries.

Previous Attempts to Define QRPs

Despite the considerable attention given to QRPs in the last decade, researchers have yet to converge on a clear definition of the concept. To our knowledge, the first definition and discussion of QRPs appeared in a panel report on responsible science (National Academy of Sciences [US] et al., 1992). This joint venture of the U.S. National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine defined QRPs as

> actions that violate traditional values of the research enterprise and that may be detrimental to the research process. Questionable research practices do not directly damage the integrity of the research process and thus do not meet the panel's criteria for inclusion in the definition of misconduct in science. (National Academy of Sciences [US] et al., 1992, pp. 5–6)

This definition did not include any reference to intentionality but clearly distinguished QRPs from research misconduct.

Years later, the widely influential article by Ioannidis (2005, "Why Most Published Research Findings Are False") mentioned systemic biases in research that lead to misleading results. He defined these biases as "a combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced" (Ioannidis, 2005, p. 697). Although he did not mention the term "QRP," he clearly referred to a similar phenomenon. The field of biomedicine has also been tackling QRPs—usually referred to as "selective reporting"—since the early 2000s (e.g., Al-Marzouki et al., 2005, 2008). A definition of scientific misconduct, which we believe better aligns with our definition of QRPs, was suggested in 2005 as "behaviour by a researcher, whether intentional or not, that falls short of good ethical and scientific standards, and in particular can arise in the context of clinical trials" (Al-Marzouki et al., 2005, p. 332). According to Banks, O'Boyle et al. (2016), QRPs represent "design, analytic, or reporting practices that have been questioned because

of the potential for the practice to be employed with the purpose of presenting biased evidence in favor of an assertion" (p. 7).

Lately, there have been more attempts to properly define QRPs. The Framework for Open and Reproducible Research Training (FORRT; Azevedo et al., 2019) defined QRPs very close to the first definition proposed by John et al. (2012) as "a range of activities that intentionally or unintentionally distort data in favor of a researcher's own hypotheses — or omissions in reporting such practices — including; selective inclusion of data, hypothesizing after the results are known (HARKing), and p-hacking" (Elsherif et al., 2025, para. 1). In contrast, Lakens (2022) defined QRPs with reference to the Netherlands Code of Conduct for Research Integrity as all practices that directly violate its requirements, which state the following: "Make sure that the choice of research methods, data analysis, assessment of results and consideration of possible explanations is not determined by non-scientific or non-scholarly (e.g., commercial or political) interests, arguments or preferences" (KNAW et al., 2018, p. 17). Therefore, Lakens explicitly categorized QRPs as forms of research misconduct, a stance that aligns with the definition proposed by Al-Marzouki et al. (2005). However, although the definition by Al-Marzouki et al. would place QRPs under the broader umbrella of "scientific misconduct," Lakens's perspective frames them as specific violations of the Netherlands Code of Conduct for Research Integrity.

Some researchers have investigated questionable practices pertaining to specific parts of the research process or subdisciplines of the field. For example, Gerrits et al. (2019) used an expert-consensus approach to identify "questionable reporting practices" and arrived at the following definition: "to report, either intentionally or unintentionally, conclusions or messages that may lead to incorrect inferences and do not accurately reflect the objectives, the methodology or the results of the study." Likewise, Wigboldus and Dotsch (2016) stated that "what makes certain research practices questionable or fraudulent are not the practices in themselves. It is the way they are reported (or not reported) that makes them questionable or fraudulent" (p. 30). They even argued that no data-analysis practices in and of themselves can be considered to be QRPs because QRPs are only about transparency—a conclusion we disagree with. Lack of transparency is often but not always the defining feature of QRPs. Moreover, Flake and Fried (2020) introduced the concept of "questionable measurement practices," which they defined as "decisions researchers make that raise doubts about the validity of measure use in a study, and ultimately the study's final conclusions" (p. 458). And even the field of metascience appears vulnerable to QRPs; "questionable metascience practices" have been defined as "a research practice, assumption, or perspective that several commentators have identified as potentially problematic for metascience and/or the science reform movement" (Rubin, 2023).

Table 1 summarizes some of the attempts that have been made to define QRPs. All of these definitions touch on the idea that QRPs involve actions or decisions in the research process that can lead to bias, misrepresentation, or distortion of research findings. However, only some acknowledge that engaging in QRPs can be unintentional (Gerrits et al., 2019; Elsherif et al., 2025). Lakens (2022) saw QRPs as actions that breach a code of conduct. From that perspective, behaviors such as research misconduct and fraud are also subsumed under the broader definition of QRPs. Finally, only one definition of QRPs explicitly excludes research misconduct (National Academy of Sciences [US] et al., 1992).

We believe that several questions remain open that are not addressed by the reviewed definitions of QRPs. First, where do different QRPs stand on the continuum between unintentional (honest) research errors and intentional fraud or research misconduct? More specifically, can engaging in QRPs be unintentional, and should motivations therefore be excluded from a comprehensive QRP definition? Second, how wide or narrow should a definition of QRPs be? Should questionable measurement practices and other more specific practices be viewed as separate entities, or are they simply a subset of QRPs? Are different QRPs conceptually independent, or do QRPs constitute families of behaviors? Third, which stages of the research process are prone to QRPs?

Following previous definitions of QRPs, some researchers have examined their prevalence and attempted to quantify associated harms. In the next sections, we review the literature focused on the prevalence and negative impact of QRPs and then outline the aims of the current research.

Estimating the Prevalence of QRPs

Several studies in the field of psychology have endeavored to collect the most frequent QRPs. Many of these studies are based on the list of QRPs created by John et al. (2012), who reported self-admission rates between 1.7% and 66.5% for 10 different QRPs. The less severe and more justifiable a potential QRP was seen to be, the more often their sample of 2,155 American psychologists admitted having engaged in that practice at least once in their career.

Fiedler and Schwarz (2016) rephrased the 10 items John et al. (2012) used to make them less ambiguous, which resulted in considerably lower rates of self-admission among 1,138 German psychologists. In addition, the authors computed the estimated prevalence of

Table 1. Previous Definitions of Questionable Research Practices

Authors, year	Definition	Actions	Intention	Outcome	Misconduct included	Research phase
National Academy of Sciences [US] et al. (1992)	"Actions that violate traditional values of the research enterprise and that may be detrimental to the research process. … Questionable research practices do not directly damage the integrity of the research process and thus do not meet the panel's criteria for inclusion in the definition of misconduct in science." (pp. 5–6)	Yes	No	Yes	No	All
Al-Marzouki et al. (2005)	"Scientific misconduct[a] has been defined as behaviour by a researcher, whether intentional or not, that falls short of good ethical and scientific standards, and in particular can arise in the context of clinical trials." (p. 332)	Yes	Yes	No	Yes	All
John et al. (2012)	"Exploitation of the gray area of acceptable practice" (p. 524)	No	Yes	No	Yes[b]	All
Banks, O'Boyle, et al., (2016)	"Design, analytic, or reporting practices that have been questioned because of the potential for the practice to be employed with the purpose of presenting biased evidence in favor of an assertion." (p. 7)	Yes	Yes	Yes	Yes	Planning Data collection Data processing Data analysis Write-up
Agnoli et al. (2017)	"Questionable research practices (QRPs) are methodological and statistical practices that bias the scientific literature and undermine the credibility and reproducibility of research findings." (p. 1)	Yes	No	Yes	Yes	Planning Data collection Data processing Data analysis
Rabelo et al. (2020)	"They are strategies used during the research process that . . . can maximise the chances of finding apparent evidence to support an expected result or to produce an attractive, counterintuitive research conclusion that is more likely to be published by scientific journals." (p. 1)	Yes	Yes	Yes	Yes	All
Elsherif et al. (2025)	"A range of activities that intentionally or unintentionally distort data in favour of a researcher's own hypotheses – or omissions in reporting such practices – including; selective inclusion of data, hypothesising after the results are known (HARKing), and p-hacking." (para. 1)	Yes	Yes	Yes	Yes	All
Lakens (2022)	Questionable research practices generally describe practices that violate the requirement from the code of conduct to "Make sure that the choice of research methods, data analysis, assessment of results and consideration of possible explanations is not determined by non-scientific or non-scholarly (e.g. commercial or political) interests, arguments or preferences." (para. 2 in chap. 15.1)	Yes	Yes	No	Yes	Planning Data collection Data processing Data analysis Write-up

Note: We examined whether each definition a questionable research practice addresses researcher actions, discusses intentionality, considers the outcome of the action, and includes misconduct. The following research phases were considered: planning, data collection, data processing, data analysis, write-up, and publication.

[a]Although this definition is about scientific misconduct, we see this to be a better fit for questionable research practices as they are regarded today.

[b]Despite mentioning a "gray area," they explicitly included data falsification as a questionable research practice.

QRPs by multiplying the proportion of yes responders with the average repetition percentage, which resulted in one-fifth of the prevalence estimate than what John et al. reported. Note that Fiedler and Schwarz did not deny the existence and problematic nature of QRPs; however, they did not share the pessimistic view that QRPs are the norm by which psychologists in academia conduct their research.

Recently, researchers have also examined the prevalence of QRPs in disciplines outside of psychology, including ecology and evolution (Fraser et al., 2018; Gould et al., 2023), linguistics (Coretta et al., 2023), health care (Artino et al., 2019), and communication science (B. N. Bakker et al., 2021), confirming prevalence estimates similar to those in psychology. Ravn and Sørensen (2021) found substantial variability in what is considered a QRP by researchers across disciplines. They conducted 22 focus-group interviews with researchers ($N = 105$) from the humanities, medical sciences, technical sciences, social sciences, and natural sciences, which yielded 40 QRPs. Only one of those QRPs repeatedly emerged across all five disciplines, and 19 of them came from only one of the disciplines.

A meta-analysis on the prevalence of research misconduct and QRPs in a wide array of research disciplines, such as medicine, biomedicine, economics, and psychology ($k = 42$), reported an estimate of 12.5% (95% confidence interval = [12.4%, 19.2%]) of researchers having engaged in one or more QRPs over the published research up to 2020 (Y. Xie et al., 2021), a number not as concerning as 66.5% but substantial nevertheless. Banks, O'Boyle et al. (2016) conducted a comprehensive narrative overview of evidence on engagement in QRPs from a multimethodological approach. The authors noted that only six studies out of 64 reported little to no evidence of engaging in QRPs, whereas 58 found overwhelming evidence that researchers engage in QRPs.

To summarize, studies collectively suggest that up to 50% of researchers across scientific disciplines may engage in QRPs to some extent (e.g., Brachem et al., 2022; Fiedler & Schwarz, 2016; Gopalakrishna et al., 2022; John et al., 2012; Y. Xie et al., 2021), highlighting the urgency for a concerted effort to globally address these practices (Lakens, 2022). Moreover, QRPs may be even more widespread in nonacademic research (Bespalov et al., 2018), for instance, advertising research (Bergkvist, 2020), market research, or public opinion polls.

The Negative Impact of QRPs

The first discussion of QRPs stated that "they can erode confidence in the integrity of the research process, violate traditions associated with science, affect scientific conclusions, waste time and resources, and weaken the education of new scientists" (National Academy of Sciences [US] et al., 1992, p. 6). Testing some of these potential harms is challenging, yet recent efforts have been made to assess the impact of QRPs on scientific conclusions, specifically, false-discovery rates and replication success.

More than a decade ago, Simmons et al. (2011) first demonstrated with simulated data how common QRPs—such as "optional stopping" and "selectively reporting conditions"—can dangerously inflate the likelihood of obtaining statistically significant results for effects that do not exist. Kravitz and Mitrof (2023) demonstrated the individual effect of three common QRPs ("mixing pilot- and main-study data," optional stopping, "not publishing null findings") and their combined effect on false-discovery rate. Although not publishing null findings led to the most serious inflation of false-discovery rate (up to .50), all three QRPs substantially inflated false-discovery rate.

Likewise, in a recent study, Stefan and Schönbrodt (2023) simulated the impact of 12 QRPs on false-discovery rate for four sample sizes (30, 50, 100, 300). The QRPs in question included optional stopping, "alternating between hypothesis tests" (parametric vs. nonparametric), "scale redefinition," "outlier exclusion," and "favorable imputation," among others. The highest false-discovery rate varied between strategies from around 8% to 40%. Two trends emerged: The severity increased with increasing dissimilarity between tested data sets and an increasing number of tests conducted. Equally, one of the key findings of Banks, Rogelberg, et al. (2016) is that the extent to which engagement in QRPs is problematic likely varies by type and engagement frequency.

In another recent Monte Carlo simulation study focusing on effect-size estimates instead of false-discovery rate, Anderson and Liu (2023) demonstrated the effect of two "cherry-picking" strategies, one in which the outcomes pertained to the same construct and another in which they pertained to different constructs. They found that these QRPs are biasing the magnitude and increasing the heterogeneity of effect-size estimates, particularly when population effect sizes and the correlations of outcomes are small. Overall, existing evidence suggests that QRPs can have detrimental effects on scientific conclusions, especially on false-discovery rate and effect-size estimates.

QRPs may distort the original results and create unrealistic benchmarks for replication (Freuli et al., 2023). Practices such as cherry-picking outcomes or questionable interim analyses may inflate effect sizes and lower p values in the original study, setting a replication standard that is difficult to achieve under unbiased conditions. For instance, inflated original effect sizes lead to a phenomenon known as "shrinkage" in replication

studies, in which the observed replication effect size is significantly smaller than the original. This discrepancy can cause the replication to fail the criterion of statistical significance even when the true effect exists (Freuli et al., 2023).

Increased false-discovery rate and biased effect-size estimates should result in reduced replicability; however, some researchers hold diverging, although less common, viewpoints. Ulrich and Miller (2020) argued that replicability might be low in research areas in which true effects are generally rare solely for statistical reasons and not just because certain QRPs run rampant. The study included a quantitative model of replication rate that took into account factors such as the level of significance, power, base rate of true effects, and influence of individual QRPs (i.e., "selective reporting" of significant studies, "failing to report all dependent measures," "data peeking," and selective outlier exclusion). This model allowed the estimation of the relative contribution of each of these factors to the replication rate. In line with other simulation studies (e.g., Witt, 2019), Ulrich and Miller showed that the modeled QRPs indeed inflated the false-discovery rate but also indicated that the relative contribution of QRPs may be small under certain circumstances. For example, when the base rates of true effects and true effect sizes are small, the net influence of QRPs on replicability seems to be small and can be compensated by increased statistical power. Unsurprisingly, the negative effect on replicability was more pronounced when QRPs were used more extensively.

To summarize, context and nuance are needed in the discussion of the potential negative effects of QRPs on individual empirical work and their threat to the overall quality of scientific outputs. In light of the work discussed here, we agree with the notion that QRPs pose a substantial threat to scientific conduct and quality. Although some studies have investigated important negative consequences of QRPs (e.g., inflated false-discovery rate, inflated effect size, and replication success), other harms (e.g., reduced confidence in science) might be more difficult to quantify. To be able to investigate the effects of QRPs, the field first needs a comprehensive list of QRPs and their associated negative effects.

Aims of This Study

In this study, we aimed to (a) reach a definition that focuses on researcher actions—rather than consequences—and distinguishes QRPs from similar but not identical concepts, research misconduct, and researcher errors and (b) list and categorize QRPs. This work provides a firm foundation for researchers to critically reflect and improve on their practice, for lecturers and students to teach and learn better practice, and for reviewers and editors to identify and assess the risk stemming from QRPs. This work can also provide a solid framework for future research investigating, for example, the prevalence of QRPs, the extent of associated harms, and the effectiveness of preventive measures.

Method

As a community of scientists, we used the expert-consensus procedure to allow for a collaborative effort in collecting and categorizing QRPs (referred to hereafter as "community consensus"). By bringing together individuals with diverse expertise, we aimed to create a comprehensive list of QRPs that could be used in future research. A hybrid hackathon and online group work provided opportunities for participants to collaborate regardless of location and time zone, creating a more inclusive and efficient data-collection and -analysis process.

Collaborators

At the onset, 37 collaborators from various branches of psychology joined the project in a hackathon event. The majority of initial participants hailed from the United States (n = 12), Canada (n = 6), Germany (n = 5), the United Kingdom (n = 5), and New Zealand (n = 2). In addition, one participant originated from Austria, Australia, Hungary, the Netherlands, North Macedonia, Poland, and Singapore. The collaborators represented a diverse range of academic positions, comprising one full professor, two associate professors, eight assistant professors, three research associates, nine postdoctoral fellows, 11 doctoral students and candidates, and three students (bachelor or master).

Nineteen collaborators continued to contribute off site and online until the manuscript's completion. We collected a number of metrics indicating the expertise of the collaborators and demographic information indicating the diversity of the collaborators. Selected metrics are reported below; for a complete summary, see the online supplemental materials at https://osf.io/f7uqh/.

At the time of submission, the collaborators included one associate professor, six assistant professors, three research professionals, seven postdocs, and two PhD students. The majority (n = 18) of collaborators were trained in psychology; one was trained in special education. The collaborators have conducted research in the following fields: metascience (n = 10); social psychology (n = 9); cognitive psychology (n = 8); experimental psychology (n = 7); quantitative psychology (n = 6); personality psychology (n = 5); psychometrics (n = 4); education, health, industrial/organizational psychology, and neuropsychology (n = 3 each); clinical, cultural, developmental, environmental, and positive psychology (n = 2

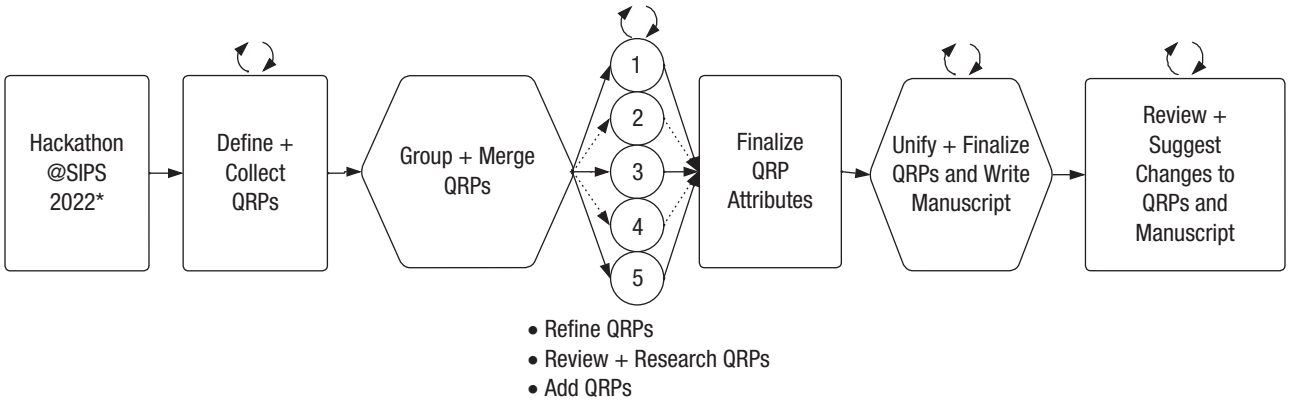

Fig. 1. Flowchart of the community-consensus process. Rectangles = group activities of all collaborators; circles = small-group work; hexagons = lead authors' work; dotted lines = asynchronous processes; solid lines = synchronous processes; circular arrows = iterative processes. The initial group at the hackathon was larger than the final number of collaborators. Asterisks indicate hybrid work, online and in person. The rest of the work took place exclusively online, both synchronously and asynchronously.

each); and affective psychology, evolutionary psychology, forensic psychology, methods in psychology, and psychophysiology (n = 1 each). The collaborators have conducted research using the following methods: cross-sectional studies, experimental, and survey and questionnaire methods (n = 17 each); correlational studies (n = 16); meta-analysis (n = 12); longitudinal and qualitative methods (n = 10 each); observational methods (n = 9); randomized controlled trials (n = 8); quasi-experimental methods and simulation and computational models (n = 6 each); psychophysiological methods (n = 5); neuroimaging techniques (n = 3); archival research (n = 2); and case studies (n = 1).

The collaborators had between 2 and 16 years of experience conducting research (including PhD-related research; M = 8.8 years, SD = 4.0, Mdn = 9) and between 2 and 12 years of experience conducting metascientific research (e.g., involvement in Many Labs, Many Analysts, the Psychological Science Accelerator, the FORRT, the Collaborative Replication and Education Project; M = 5.2 years, SD = 2.8, Mdn = 4). The number of publications of peer-reviewed articles ranged between two and 137 across collaborators (M = 25.7, SD = 33.6, Mdn = 14), and their h-index ranged between 2 and 42 (M = 10.5, SD = 10.4, Mdn = 6.5).

Procedure

Overview of the process. We initiated the process of collecting QRPs in a hackathon[2] at the annual meeting of the Society for Improving Psychological Science (SIPS) in 2022 in Victoria, British Columbia, Canada. One of the lead authors (T. Nagy) organized the hackathon aiming to find a definition of QRPs, collect as many QRPs as possible, provide key characteristics and examples of each QRP, and ultimately, create a bestiary of QRPs. Participants of the hackathon collaborated both in person and online. In follow-up online sessions, participants engaged in reviewing relevant literature and merged and grouped the QRPs based on concept similarity. This process started at the conference and continued online (see Fig. 1). Note that the QRP definition was created during the SIPS 2022 conference and reflects the consensus reached by all initial collaborators. Subsequently, during the revision process, we decided to remove any notions of intentionality from the definition.

After the hackathon, we formed five online groups of three to five collaborators to further refine the list of QRPs. The lead authors (T. Nagy and J. Hergert) assigned each group seven to 10 QRPs, and the members worked individually and together to define the attributes of each QRP, including its name, aliases, research phase, potential harms, preventive measures, detectability, clues, examples, and references. Describing the QRPs was an iterative process: Each member provided suggestions, and all members could provide feedback and contribute to discussions. The result was moved to the next phase only when a group consensus was reached.

After reviewing and researching the assigned QRPs, the groups reconvened to pool their findings and reach a consensus on each QRP and its description. This was also an iterative process during which all members provided suggestions and feedback until a consensus was reached. T. Nagy and J. Hergert then reviewed the list, unified the language, and finalized the QRP attributes. Finally, all authors had the opportunity to suggest final changes to the QRP list.

Description and classification of QRPs. For each QRP, we first formulated a descriptive name that would capture the actions and behaviors involved, collected alternative names, generated a definition, and provided

examples. Next, to clarify that researchers may engage in different QRPs at different stages of the research process, we determined the research phase for each QRP: planning, data collection, data processing, data analysis, write-up, and publication. "Planning" encompassed all activities before data collection begins. "Data collection" referred to the period from the start to the end of obtaining data. "Data processing" involved operations on the data before hypothesis testing. "Data analysis" covered statistical modeling and inference, during which hypotheses and research questions are evaluated. The "write-up" phase included drafting the manuscript until submission, and "publication" encompassed all aspects of the publication process. The research stage thus provided one way of categorizing the QRPs.

Another categorization emerged during the iterative collection process because we recognized that some QRPs belong to the same family of research behaviors. We grouped these conceptually related QRPs under QRP umbrella terms.

Subsequently, we considered the consequences of each QRP in terms of potential harms to research outcomes or credibility. After describing harms, we collected possible preventive measures (methods and best practices) that researchers can take to avoid engaging in QRPs or indicate their lack of engagement with QRPs. Note that we focused on the researchers' perspective rather than other stakeholders because considering preventive measures or remedies from the perspective of editors, publishers, reviewers, universities, and funding agencies would vary significantly. However, given that science constitutes a community of practice that has to be largely built on trust and researcher independence, we believe that researcher behaviors are an important starting point.

Finally, we assessed whether it is possible to detect the presence of each QRP. Assuming a scenario in which (a) an individual who possesses knowledge and skills of research methods and data analysis that are typically acquired in PhD programs (Ord et al., 2016) and (b) is somewhat knowledgeable of the research topic has read the publication, we classified detectability as either yes (detectable), no (not at all detectable), or maybe (detectable only with extra effort and/or extra material, e.g., preregistration or data). To complement detectability, we compiled clues that could indicate the presence of each QRP, noting that the mere presence of a single clue may not necessarily serve as evidence of a QRP but that the presence of multiple clues could justify further investigation. We additionally provided references to further resources that discuss each QRP.

Results

Data and code for reproducing the tables and summary statistics are available at https://osf.io/f7uqh/.

The definition of QRPs

We arrived at a definition that was accepted by all collaborators:

> Questionable research practices (QRPs) are ways of producing, maintaining, sharing, analyzing, or interpreting data that are likely to produce misleading conclusions, typically in the interest of the researcher. QRPs are not normally considered to include research practices that are prohibited or proscribed in the researcher's field (e.g., fraud, research misconduct). Neither do they include random researcher error (e.g., data loss).

The definition provides a clear understanding of the types of research practices that are considered questionable and can lead to misleading conclusions. First, the definition specifies a wide variety of actions that may be associated with QRPs, such as research activities, including data collection and analysis, data presentation and interpretation, and data sharing. Second, the definition highlights the outcome of using QRPs because they are likely to produce misleading conclusions, which are typically in the interest of the researcher. We recognize that researchers often employ QRPs to support a hypothesis, but we also consider research in which QRPs are used to collect evidence for the absence of an effect. Third, the definition excludes fraud and random errors as QRPs. We considered fraud, for example, the fabrication of data, clearly unacceptable and therefore not questionable. We also acknowledge that errors may happen in the research process for many reasons (Kovacs et al., 2021).

While working on this manuscript, we had several discussions about the role of researchers' intentions in QRPs. Our view is that QRPs lie on a continuum between (but not including) fraud and random error. QRPs may be strongly motivated or simply based on a lack of knowledge or social pressure, and it is important not to assume that all engagement in QRPs is intentional. Ultimately, we decided to stay clear of judgments of intentionality and instead focus on providing a detailed description of the behaviors and their attributes, indicators, and preventive measures.

List of QRPs

A total of 40 QRPs were collected; five QRPs pertained to the planning phase, four QRPs were dedicated to data collection, seven QRPs focused on data processing, five QRPs centered around data analysis, 14 QRPs involved the write-up process, and five QRPs concerned publication. Table 2 shows the identified QRPs by research phase, and Table 3 shows all QRPs and attributes.

Table 2. List of Identified Questionable Research Practices by Research Phase

Planning	Data collection	Data processing	Data analysis	Write-up	Publication
Choosing biased manipulations	Placing undue influence on participants	Discretizing continuous variables	Choosing a poor model specification	Citing unreliable research	Creating multiple publications from the same study
Choosing biased measurements	Mixing pilot- and main-study data	Excluding data points	Choosing unjustified *p* value adjustment	Hypothesizing after the results are known (HARKing)	Declaring false authorship
Choosing overlapping measures to find significant results	Optional stopping	Missing data hacking	Neglecting assumptions for statistical models	Incorrect reporting of test statistics	Not linking the prersegistration to the published study
Preregistering after the results are known (PARKing)	Selective sampling	Modifying measurements	Selecting a favorable random-number-generator seed	Making unsupported conclusions	Not making data accessible
Performing inappropriate power analysis		Redefining group-membership rules	Using ad hoc covariates	Not disclosing deviations from preregistration	Publishing studies selectively
		Using ad hoc exclusion criteria for participants		Omitting important details of the scientific process	
		Variable transformation fishing		Selective citing	
				Selective reporting of hypotheses	
				Selective reporting of indicator variables	
				Selective reporting of outcomes	
				Selective test reporting	
				Using irrelevant references	
				Using unjustified references	

QRP attributes

QRP umbrella terms. Certain QRPs have gained recognition as an extensive range of behaviors, including p-hacking and "citation engineering." To enhance precision in delineating these practices, we aimed to disaggregate these overarching QRPs into more specific ones while retaining the commonly acknowledged terms. Consequently, we introduced the notion of QRP umbrella terms. These terms collect several QRPs that share common characteristics or features (see Table 4).

"P-hacking" was the most common QRP umbrella term, consisting of 13 QRPs, referring to manipulating statistical analyses and procedures to obtain significant *p* values at the expense of scientific integrity. We categorized seven QRPs into the "cherry-picking" umbrella term, meaning to selectively choose and present data or information to support a specific hypothesis or conclusion while ignoring or omitting other relevant data. Six QRPs were categorized under "citation engineering," involving the manipulation of citations and references to bolster the credibility of a study or a claim or to inflate publication statistics. Four QRPs were classified as "sample curation," which refers to manipulating or selectively including/excluding participants to achieve desired results. Three QRPs were classified as "influencing

Table 3. Questionable Research Practices and Their Attributes

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Planning									
Choosing biased manipulations	—	Selecting an unjustified manipulation to reach a misleading outcome	• Influencing participants	Selecting subclinical drug doses to suggest no effect or only deliberately high doses to suggest an adverse effect Using images or videos to elicit emotions, but the stimuli are not eliciting emotions	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Compromised generalizability	• Run pilot studies to investigate if the manipulation can elicit an effect. • Use manipulation-check questions. • Use standard stimuli or dosages.	Maybe	• Lack of manipulation check • Using dosages outside of recommended values • Using stimuli that can elicit extreme responses • Using stimuli that were not previously tested and/or have no proven effect	• Joe and Leif (2020) • Marchetti and Schellens (2007)
Choosing overlapping measures to find significant results	Leveraging the jangle fallacy	Exploiting the conceptual similarity between measures, presenting them as distinct constructs, to get significant results	None	Researcher is using similar items in two seemingly different constructs and concludes that the constructs have a high correlation. Researcher finds a positive relationship between depression and suicidality, but the positive relation is in part due to the depression scale already containing items about suicide. A researcher estimates a positive relation between Construct A and Construct B in a meta-analysis, but the meta-analysis includes studies that have used the same task to operationalize either construct.	• Inflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability	• Be transparent about similar items. • Consider whether the measures used in a study distinctly operationalize different constructs. • Account for covariance that is due to similar items.	Yes	• Items or tasks are similar for the correlated constructs	• Flake and Fried (2020) • Hodson (2021) • Wang and Eastwick (2020)
Performing inappropriate power analysis	—	Selecting inappropriate parameters or methods for power analysis and/or misinterpreting/misusing power-analysis results	None	A researcher chooses a small sample size to get null results in a study about the harmful effects of smoking. A researcher uses default parameters for the power analysis to show that power analysis had been performed; however, the analysis is uninformative.	• Inflated confidence in the research • Inflated Type II error • Reduced replicability	• Be transparent about how power analysis was conducted. • Use meaningful parameters that are based on field standards, literature, or common sense.	Yes	• Details of power analysis are not reported or justified. • Parameters of the power analysis are generic and do not fit the study. • Relatively low sample size • The results of power analysis are misinterpreted.	• Heckman et al. (2022) • Kovacs et al. (2022) • Lakens (2022)
Preregistering after results are known (PARKing)	—	Preregistering a hypothesis or analysis after knowing the outcome of the analysis	• Misusing open-science practices	Researchers realize that their manuscript will not be accepted without a preregistration, so they create one post hoc and link it to their study.	• Inflated confidence in the research	• Be transparent about when the preregistration was done. • Disclose the familiarity (if any) of the researcher with the data. • Preregister before analyzing the data.	Yes	• Data collection occurred previous to the preregistration date. • Date of preregistration is unrealistically close to the first submission without the manuscript being a registered report.	• Yamada (2018)
Choosing biased measurements	—	Selecting an instrument that is biased or invalid to support a desired narrative	• Influencing participants	Researcher uses loaded, leading, or suggestive questions, e.g., "Does the lack of respect schoolchildren have for their teachers, in your opinion, influence everyday teaching methods in schools?" Researcher assesses a psychological construct using an ad hoc questionnaire with no proven validity. Researcher uses a scale that does not capture the intended construct properly to support a desired narrative.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability	• Frame questions in a neutral way. • Get external opinions on questionnaires/study materials or conduct a registered report. • Use questionnaires that are unbiased and psychometrically sound.	Yes	• Ad hoc questionnaires are used instead of validated instruments. • Measurement items use suggestive or biased language. • Measurement is based on single-item questions. • No discussion of the psychometric properties of the instruments	• Flake and Fried (2020)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Data collection									
Mixing pilot- and main-study data	Double dipping, retaining pilot data	Including data from a pilot study if the results support the hypothesis	• Sample curation	Researcher conducts a pilot study to check the protocol and analyzes pilot data. The pilot data and main study data are aggregated in the data analysis if pilot-data results are in line with expectations.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability	• Do not include pilot data in the analyzed data set. • Report pilot data separately.	Maybe	• If data are shared, timestamps may fall into two distinct periods. • Sample sizes reported throughout the article might not match.	• Kravitz and Mitroff (2020) • Kriegeskorte et al. (2010)
Optional stopping	Peeking, data peeking	Monitoring hypothesis tests during data collection and stopping when statistical inference is favorable without controlling for sequential testing	• Sample curation	Researcher is collecting responses and tests the hypothesis after every participant. When significance is reached, the researcher stops collecting data.	• Inflated Type I error • Reduced replicability	• Preregister stopping rules and adjustments for Type I error inflation. • Preregister the estimated sample size.	Maybe	• Absence of preregistration • Low sample size • The *p* values are just below the significance threshold (usually .05). • Relatively large effect size compared with other studies in the field • Vague or absent reason for sample size	• de Heide et al. (2021) • Lakens (2022) • Schönbrodt and Perugini (2013) • Schönbrodt et al. (2017) • Wicherts et al. (2016)
Selective sampling	Biased sampling	Collecting a sample in a way that biases the findings	• Sample curation	Researcher tests the likeability of chocolate on a group of children only to find that everyone loves it. Using uncomparable groups: The researcher tests if men are more aggressive than women. For comparison, women from a university are compared with men from a prison. Picking a subsample of a panel data set to find the desired results	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Compromised generalizability • Reduced replicability if sampling bias is not disclosed	• Consider using statistical control for confounding variables. • Make sure that the sample represents the population. • Preregister the sampling process. • Report the sampling process transparently. • Use a sampling method that does not bias the results. • Use comparable (or matched) groups.	Yes	• Compared groups are from different populations. • Convenience sample • Unclear rationale for sample selection	• DeepChecks (n.d.) • Marks (1947)
Placing undue influence on participants	—	Affecting participants to make them give responses that support the desired narrative	• Influencing participants	Researchers tell the participants that they believes the treatment will work. Researchers use a briefing document that is trying to influence participants' attitudes about the topic that is assessed in the study. During data collection, the same organization that runs the study also runs a marketing campaign to influence public opinion on the same topic.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability	• Avoid exposing participants to any cues that might influence their responses. • Blind the experimenter as possible. • Researchers interacting with participants should remain neutral and follow scripts during testing. • Use automated research procedures (e.g., research-presentation software) instead of human research assistants whenever possible.	No	• Absence of blinding • Absence of procedure scripts • Suggestive questions in the survey	• McCambridge et al. (2012)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Data processing									
Excluding data points	—	Exclusion of data points or outliers without proper justification and transparent reporting	• *P*-hacking	Removing individual reaction-time trials based on post hoc criteria Trying different outlier cutoff criteria until an effect is statistically significant	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Compromised generalizability • Reduced replicability • Reduced reproducibility	• Perform blinded data analysis. • Perform sensitivity analysis. • Preregister the study. • Report post hoc changes in exclusion criteria.	Maybe	• Absence of open data • Absence of preregistration • Unexplained discrepancy between the recruited and analyzed sample sizes and degrees of freedom	• M. Bakker and Wicherts (2014a, 2014b) • Osborne and Overbay (2004)s
Missing data hacking	Favorable imputation	Choosing the strategy to handle missing data based on the impact on the results	• *P*-hacking	A researcher tries three ways of handling missing data, for example, listwise deletion, multiple imputation, and inverse probability weighting. The expected results appear only with inverse probability weighting. The researcher reports only this strategy in the article and leaves out results with listwise deletion and multiple imputation. Can also be within a single method, specifically multiple imputation, because it uses one or more variables to replace missing data and because the choice of these variables is up to the researcher but can also be statistically based.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced reproducibility • Reduced replicability	• Perform blinded data analysis. • Perform sensitivity analysis. • Preregister missing-data approach.	Maybe	• Absence of open data • Lack of rationale or references supporting the missing-data approach • No mention of the missing-data approach or the missing data at all • Unexplained discrepancy between the recruited and analyzed sample sizes	• Enders (2010) • Woods et al. (2024)
Using ad hoc exclusion criteria for participants	—	Exclusion of participants without proper justification transparent reporting	• *P*-hacking • Sample curation	Researcher finds that a correlation between two variables is not significant. After removing two participants—who should be included—the association becomes significant. Then, the researcher comes up with post hoc exclusion criteria for those participants. Researchers do not find an expected association between perceived stress and personality. When looking at only the top 25% of perceived-stress scores, the association is there. Researchers go on to report the top 25% scores as their population of interest and do not disclose that they looked at the rest of the sample population.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Compromised generalizability • Reduced replicability • Reduced reproducibility	• Perform blinded data analysis. • Perform sensitivity analysis. • Preregister the study. • Report post hoc changes in exclusion criteria.	Maybe	• Absence of open data • Absence of preregistration • Sample too narrow for recruitment methods • Unexplained discrepancy between the recruited and analyzed sample sizes	• Lang et al. (2019) • Nüesch et al. (2009)
Discretizing continuous variables	Dichotimizing variables, median split	Taking a continuous variable and making it categorical without proper justification and transparent reporting	• *P*-hacking	Researcher does not find an association between depression and continuous age variables and recodes age into young and old categories. After that, age groups show a significant association with depression. An independent samples t test is reported instead of a correlation.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability • Compromised generalizability	• Perform blinded data analysis. • Perform sensitivity analysis. • Preregister the study. • Use original measurement levels.	Yes	• Absence of open data • Scale or response options in materials or methods do not match how they are reported in the results. • Test statistics do not match expected data-analysis strategy.	• Cohen (1983) • DeCoster et al. (2011) • MacCallum et al. (2002)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Modifying measurements	—	Changing the properties of a measure/ measurement to produce favorable results without proper justification and/or transparent reporting	• *P*-hacking	Researcher uses only a portion of the items from a longer scale. Researcher combines items from different scales into a single measure. Researcher chooses which EEG electrodes to aggregate based on the results.	• Reduced replicability • Reduced reproducibility • Reduced validity of the measure • Inflated or deflated reliability of the measure • Inflated Type I or Type II error • Inflated or deflated effect-size estimates	• Describe and justify any modifications on measurements. • Perform blinded data analysis. • Perform sensitivity analysis. • Publish study materials. • Preregister the study. • Use conventional measurements/ measures.	Maybe	• Absence of open data • Absence of open study materials • Discrepancy between the reported version of measurement/ measure and original or conventional measure	• Flake and Fried (2020)
Redefining group-membership rules	—	Post hoc (re)definition of grouping criteria without proper justification and transparent reporting	• *P*-hacking	Collapsing the multicategorical variable of sexual orientation into heterosexual and nonheterosexual Trying different age ranges in cross-sectional age comparisons to maximize group differences	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability • Reduced reproducibility • Compromised generalizability	• Report post hoc changes in grouping rules and report results using original grouping rules as well. • Preregister the study. • Perform blinded data analysis. • Perform sensitivity analysis.	Maybe	• Absence of open data • Absence of preregistration • Oversimplified sample description • Response options in materials/methods different than reported groups	• Wicherts et al. (2016).
Variable transformation fishing	—	Selecting variable transformations that produce favorable results without proper justification and/ or transparent reporting	• *P*-hacking	Researcher runs a statistical test using several different transformations (e.g., changing levels of measurement, log-transformations, rescaling of the outcome) and reports only the one that produces a significant result.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability • Reduced reproducibility	• Describe and justify any variable transformations. • Perform blinded data analysis. • Perform sensitivity analysis. • Preregister conditional transformations.	Maybe	• Absence of open data • Reported values are outside of regular range. • Transformation is applied without justification. • Using transformations that are unconventional for the measure	• Lee (2020)
Data analysis									
Choosing a poor model specification	Overfitting or underfitting models, bias-variance trade-off	Creating too complex models on too small data sets, causing the model to learn the noise and random fluctuations instead of generalizable patterns Alternatively, creating too simple models that do not adequately fit the data	• *P*-hacking	Overfitting: The researcher fits a regression model with 25 predictors on a sample of 100 participants. Underfitting: The researcher uses linear regression to investigate a nonlinear association.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error	• Perform blinded data analysis. • Perform sensitivity analysis. • Use a theoretically justified model in confirmatory studies. • Underfitting: Visualize data and the model. • Use methods that prevent overfitting in exploratory research (e.g., use separate train and test data sets, use cross-validation resampling methods, use regularization or other feature-selection methods).	Yes	• Improper prediction selection (e.g., no regularization) • No mention of holdout (or test) data set or cross-validation • Overfitting: very high R^2 value (close to 1) • The number of included predictors in the model is large. • The number of observations is low. • Underfitting: Data visualization shows high model bias.	• Babyak (2004)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Choosing unjustified *p*-value adjustment	Not adjusting or overadjusting p values	Not adjusting or overadjusting *p* values when running multiple tests	• *P*-hacking	A researcher decides (a) whether to adjust for multiple tests (e.g., in an ANOVA), (b) which adjustment method to use, and (c) which (i.e., how many) comparisons to include depending on results obtained. Researcher uses Bonferroni correction when correlating several variables to prove that an association does not exist.	• Inflated Type I or Type II error	• Perform blinded data analysis. • Perform sensitivity analysis. • Preregister *p*-value adjustment plans. • Preregister planned contrasts.	Yes	• Multiple tests are made that would require *p*-value adjustment. • The *p*-value adjustment is not mentioned. • Using *p*-value correction that is too strict (e.g., Bonferroni) without proper justification	• Bender and Lange, S. (2001) • Cramer et al. (2016) • Midway et al. (2020)
Neglecting assumptions for statistical models	—	Using statistical models even though requirements are not met	• *P*-hacking	Analyzing data using parametric tests, such as *t* tests, but the data require a nonparametric test Analyzing dependent data using a statistical model that does not account for dependency	• Inflated Type I or Type II error • Reduced replicability • Reduced reproducibility	• Perform and report necessary assumption checks.	Yes	• Evidence of assumption breaches (e.g., nonnormality, nonindependent data, largely different standard deviations by group) • Not reporting assumption checks	• Knief and Forstmeier (2021) • Meteyard and Davies (2020)
Using ad hoc covariates	Selectively including control variables	Addition or removal of covariates to influence the estimates or significance for the effect of interest	• *P*-hacking	A researcher opportunistically decides which background variables (e.g., age, gender) to control for without a causal theory or a preregistration. A researcher decides whether to control for a baseline value in an experimental design depending on the results of statistical tests. Researcher avoids the inclusion (or measurement) of theoretically justified moderators (e.g., severity of a condition or socioeconomic status) to be able to imply greater generalisability.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Compromised generalizability • Reduced replicability • Reduced reproducibility	• Perform blinded data analysis. • Preregister complete models/ analytical plan. • Report robustness checks.	Maybe	• Different covariates in different analysis steps are used. • There is a lack of justification for the selection of the covariates.	• Becker et al. (2016) • Stefan and Schönbrodt (2022) • VanderWeele (2019) • Wysocki et al. (2022)
Selecting a favorable random-number-generator seed	Resampling lottery	Trying different random seeds until getting a favorable result, potentially in combination with small number of replications	• *P*-hacking	A researcher keeps on bootstrapping a CI (e.g., for a mediation indirect effect) with different seeds until the 95% CI just excludes 0.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability • Reduced reproducibility	• Report significance only if the results are robust across random seeds. • Use a large number of replications (e.g., bootstrap samples). • Perform blinded data analysis.	Maybe	• The *p* values are just below the significance threshold (usually .05).	• Götz et al. (2021) • Stack Exchange (2018)
Write-up									
Selective test reporting	—	Repeatedly testing a hypothesis in different ways until the desired result is found and then selectively reporting the findings that support the desired conclusion	• *P*-hacking • Cherry-picking	Researchers analyze the data using (a) multiple statistical methods (multiple t tests, ANOVAs, different random structures in LMEMs) and/or (b) multiple data-eligibility specifications. Based on the results, they choose to present only one analysis that gives a significant result.	• Inflated confidence in the research • Inflated Type I or Type II error • Reduced reproducibility • Reduced replicability	• Perform blinded data analysis. • Perform specificity-curve analysis. • Preregister data-processing (e.g., missing data approach) and statistical-analysis strategy. • Report all performed hypothesis tests.	Maybe	• Absence of preregistration • Arbitrary data-processing steps and/or statistical methods • The *p* values are just below the significance threshold (usually .05).	• Chuard et al. (2019) • Head et al. (2015) • Olejnik and Algina (1987) • Simmons et al. (2011) • Wagenmakers (2007)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Hypothesizing after the results are known (HARKing)	Texas sharpshooter fallacy; post hoc, ergo propter hoc	Presenting a hypothesis that is based on observed results (post hoc or a posteriori) as if it were presumed before obtaining results (a priori)	None	Researcher claims to have predicted an unexpected result. Researchers have no hypotheses originally and form hypotheses after exploring the data and presenting the hypotheses as if they had those from the beginning. Researchers have a hypothesis (e.g., a mediation hypothesis) and tests it, and if the results do not confirm the hypothesis but rather indicate an alternative pattern (e.g., a moderation), the researchers claim that this is what they hypothesized all along. Post hoc directional hypotheses: The researcher presents a hypothesis as if it were unidirectional (i.e., Group A's mean is larger than Group B's, or a correlation will be positive), although the original hypothesis was bidirectional. This change will make the hypothesis test significant.	• Inflated confidence in the research • Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Reduced replicability	• Clearly separate exploratory and confirmatory findings. • Form hypotheses before analyzing the data. • Perform blinded data analysis. • Preregister confirmatory hypotheses. • Use robust exploratory research practices (e.g., holdout data set, cross-validation, multiverse analysis, blinded data analysis).	Maybe	• Absence of preregistration • Unexplained and unconventional choices in the methods and results sections	• Andrade (2021) • Brookes et al. (2001) • Kerr (1998) • Leung (2011) • Weston et al. (2019)
Making unsupported conclusions	—	Interpreting research findings or their implications in a way that is not backed by evidence	None	Researcher concludes that a treatment is effective for groups and contexts that were not considered in the study. Researcher concludes that a treatment worked; however, the treatment effect did not differ from the effect of the control condition (or no control condition was used). Researcher implies causality based on a research design that does not allow causal inference.	NA	• Make it clear that the evidence is limited to certain contexts. • Make sure that every interpretation is properly supported by evidence. • Use conditional statements for cases in which evidence is weak or the researcher uses extrapolation.	Yes	• Causal claims are made without the methodology or analysis allowing causal inference. • Results are generalized to contexts outside of the study's scope. • The chosen methodology and statistical analysis do not allow the researcher to answer the hypothesis. • The statistical results do not match the conclusions.	• Simons et al. (2017) • Yarkoni (2020)
Incorrect reporting of test statistics	—	Not using statistical-test-reporting conventions to obscure exact results and assume that they are above or below threshold values	None	Researcher is "rounding off" a p value in an article (e.g., reporting that a p value of .054 is less or equal to .05). Researcher reports p values only and conceals test statistics. Researcher reports a correlation without disclosing the degrees of freedom, number of observations, or CI, so it seems like the effect is large (e.g., $r = .70$, $n = 15$, 95% CI = [.01, .90]). Researcher does a model comparison and reports only fit statistics that are in favor of the preferred model.	• Inflated Type I or Type II error • Reduced reproducibility	• Adhere to reporting conventions (e.g., APA). • Publish data. • Publish processing and analysis code. • Use literate programming (e.g., RMarkdown, quarto, jupyter). • Work with software that supports you in producing and checking your write-up (e.g., papaja, statcheck).	Yes	• Absence of open code • Absence of open data • Anomalies in reported statistics (e.g., test statistics are incompatible with p values). • Fit statistics are missing without proper explanation. • Statistics are not reported according to conventions (e.g., three digits for p values, reporting of df).	• M. Bakker and Wicherts (2011) • Jackson et al. (2009) • Nuijten et al. (2016, 2017)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Omitting important details of the scientific process	Incomplete methods or results sections	Not reporting important details of the methodology and statistical analysis	• Cherry-picking	Researcher omits sample characteristics, such as the sample was recruited on MTurk or participants received compensation for participation. Researcher reports correlations without specifying the type (e.g., Spearman). Researcher does not share study materials on request or does not report exact questionnaire items. Researcher fails to mention pilot studies that were conducted to arrive at the final design.	• Compromised generalizability • Reduced replicability • Reduced reproducibility	• Preregister the study or written research plan before conducting the study. • Report every important detail of the scientific process. • Use a lab log during data collection to keep track of changes in the scientific process.	Maybe	• Absence of open study materials • Details that are usually shared are missing. • Replication is not possible from published method.	• Gernsbacher (2018) • National Academies of Sciences, Engineering, and Medicine et al. (2019) • Wagenmakers et al. (2021)
Selective reporting of hypotheses	Cherry-picking hypotheses, chrysalis effect, fishing expedition	Reporting hypothesis test only if it fits the researcher's expectation	• Cherry-picking	Researcher formulates five hypotheses, of which only three are supported by the data, and only these three get reported in the final research report (chrysalis effect). Fishing expedition: Researchers surveys college students about the outfit they are wearing and their scores on several tests, which allows for many possible analyses (examining different colors, types of clothing, tests, score cutoffs, etc.). They end up reporting only a subset of findings to claim college students perform significantly better on tests when they are wearing green. See also modifying measurement, selective reporting of indicator variables, and selective reporting of outcomes.	• Inflated confidence in the research • Inflated or deflated effect-size estimates • Inflated Type I or Type II error	• If some hypotheses get left out because of the scope of the write-up, be transparent about it. • Preregister the study. • Report all hypotheses in the write-up regardless of whether they were confirmed.	Maybe	• Number of hypotheses in preregistration (or dissertation) exceeds the number in the publication.	• Andrade (2021) • O'Boyle et al. (2017) • Simmons et al. (2011)
Visualizing data in a misleading way	—	Choosing suboptimal visualizations or altering figure properties to exaggerate or diminish effects	None	Researcher truncates the *y*-axis so it is not starting at zero and/or does not add error bars. This makes differences seem larger and more significant than they are in reality. Researcher uses arbitrary categories to present interval data on a map. Researcher displays a pie chart with percentage numbers falling below or exceeding 100.	NA	• Follow best practices on how to visualize data.	Yes	• Chartjunk (e.g., three-dimensional elements, ornaments) is present on the plot. • In a plot, *y*-axis is starting at an arbitrary point. • Only summary statistics are shown, without individual data points. • Scale or response options in text do not match how they are presented in a plot. • Statistical uncertainty (e.g., error bars) is not shown on plots. • Visualization does not match the reported results in the text.	• Nguyen et al. (2021) • Weissgerber et al. (2019) • Wilke (2019)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Citing unreliable research	—	Citing an unreliable publication to support the study's narrative	• Citation engineering	Researcher cites a publication that presents low-level evidence to support a claim with no reference to the study's limitations or other studies. The researcher cites a retracted article.	• Increasing the credibility of low-evidence research • Inflated credibility of statements	• Never cite a retracted study without clearly indicating its retraction. • Cite only publications that properly support their claims.	Yes	• The cited publication provides no or low-quality evidence to its claims. • The cited publication is retracted or otherwise discredited, or its claims are refuted.	• American Psychological Association (2023) • Balshem et al. (2011) • Letrud and Hernes (2019)
Selective reporting of indicator variables	Cherry-picking indicator variables, cherry-picking conditions/ groups	Reporting only the indicator variables (or predictors, features, independent variables) that are used in analyses that produce expected results	• Cherry-picking	Researcher reports indicators that are associated with the outcome rather than including all measured indicator variables in the results section. Researcher drops one or more conditions or groups/merges two or more groups into one/splits a group into more groups than were initially planned depending on statistical results.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Compromised generalizability • Reduced replicability	• Perform blinded data analysis. • Preregister the study. • Report all indicators.	Maybe	• Indicators are reported in supplemental material but not mentioned in main text. • Measures get reported in the methods section but not in the results section. • Number of preregistered indicators exceeds the number of indicators in publication. • Reported mean time of participation does not match the number of reported measures.	• Gernsbacher (2018) • Simmons et al. (2011)
Selective reporting of outcomes	Cherry-picking outcomes	Reporting only the outcomes (or dependent variables) that are used in analyses that produce expected results	• Cherry-picking	Researcher uses several scales to measure the same construct but reports only the one that produces expected results. Researcher tests effectiveness of a new intervention for depression by measuring its effects on anxiety, sleep quality, and stress and reports only the outcome that shows the desired effect.	• Inflated or deflated effect-size estimates • Inflated Type I or Type II error • Compromised generalizability • Reduced replicability	• Perform blinded data analysis. • Preregister the study. • Report all outcome.	Maybe	• Measures get reported in the methods section but not in the results section. • Number of preregistered outcomes exceeds number of outcomes in publication. • Outcomes are reported in Supplemental material but not mentioned in main text. • Reported mean time of participation does not match number of reported measures.	• Gernsbacher (2018) • Pigott et al. (2013) • Simmons et al. (2011)
Using unjustified references	—	Selectively citing works by specific researchers or journals to inflate citation metrics or boost a journal's impact factor	• Citation engineering	Researchers selectively cites their own publications for boosted citation metrics. Citation networks: Researcher cites a colleague's unrelated work to get cited in a similar way.	• Inflated credibility of publications • Inflated credibility of journals	• Cite only relevant studies. • Provide comprehensive coverage of related scholarly literature.	Yes	• Publications are cited without relevance to the claims. • Specific authors or journals are cited disproportionately frequently.	• Fong and Wilhite (2017) • Mehregan (2022)
Not disclosing deviations from preregistration	—	Deviating from the preregistration without transparency and proper justification in the publication	• Misusing open-science practices	Researcher preregisters collecting data from a nonstudent sample but ends up including students and does not disclose this deviation. Researcher preregisters a data analysis using linear regression but uses robust regression instead without reporting the discrepancy. See also first example in selective reporting of hypotheses.	• Inflated confidence in the research • Reduced replicability	• Avoid vagueness in preregistration. • Disclose and justify every divergence from the preregistration. • Use methods that will provide robust results even when preregistration is not specific at points (e.g., blind data analysis, cross-validation).	Yes	• Link to the preregistration in the manuscript does not work or leads to a page that cannot be accessed. • Preregistration and published study differ on important aspects.	• Adam (2019) • Claesen et al. (2021) • Lakens (2024) • Nosek et al. (2018)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Selective citing	Cherry-picking citations	Avoiding to mention studies that do not support the hypothesis of the research or even those that do support the hypotheses to make the study appear more novel	• Citation engineering • Cherry-picking	Researchers overly cites empirical work that supports their hypotheses and withholds citing work that did not find the effect at all or even the opposite. Researcher omits other null findings to maximize the perceived value of a null finding.	• Inflated credibility of statements • Inflated confidence in the research	• Provide comprehensive coverage of related scholarly literature	Yes	• Cited studies point in only one direction. • Important studies and experts are missing from references. • Systematic reviews and meta-analyses are not cited.	• Duyx et al. (2017) • Gøtzsche (2022)
Using irrelevant references	—	Using citations that are not connected to the claims to increase the credibility of a statement	• Citation engineering	Researcher supports a statement with three citations, and two of them are unrelated to the statement.	• Inflated credibility of statements	• Cite only relevant studies.	Yes	• Publications are cited without relevance to the claims..	• Penders (2018) • Teixeira da Silva and Vuong (2021)
Publication									
Not linking the preregistration to the published study	—	Creating a preregistration but not associating it with the published study	None	Researchers preregister a study, and after conducting the research, they do not mention the in the manuscript because of too many diversions.	NA	• Always link the preregistration to the manuscript and report discrepancies.	Maybe	• A preregistration that fits the study is findable.	• Wicherts et al. (2016)
Creating multiple publications from the same study	Publication overlap, salami slicing	Breaking up of research findings from the same data set into several publications without proper justification and the disclosing of related manuscripts	• Citation engineering	Researcher conducts a study measuring several outcomes (or predictors) and publishes results in several articles, with each article focusing on just one outcome (or predictor), while not disclosing the other articles. A study on cross-cultural differences with 20 participating labs from 20 countries results in 10 publications in which each article reports comparisons of two countries.	• Biased effect-size estimates in meta-analyses (because of nonindependence of results) • Inflated Type I or Type II error that is due to unknown family-wise error rate	• Preregister the publication strategy. • Publish study results in one single publication or disclose all related articles.	Maybe	• Absence of open data • Description of the sample is the same over several studies by the same researcher or lab. • Several articles exist with similar outcomes or predictors based on the same data set by the same researcher or lab. • The methods suggest a large study, but the scope of the article is narrow.	• Broad (1981) • Hilgard et al. (2019) • Kaiser et al. (2021) • J. S. Xie and Ali (2023)
Declaring false authorship	—	Attribution and arrangement of authorship that does not correspond to the authors' contributions to influence the publishing process and increase the credibility of the study	• Citation engineering	Honorary authorship: Researcher adds a coauthor who did not contribute to the manuscript. Ghost authorship: The researcher excludes a coauthor who significantly contributed to the project. Controversial researcher writes a manuscript and publishes it under a pseudonym so it seems that more than one person shares the same view.	Lack of deserved credit for contributing authors Inflated confidence in the research based on the reputation of authors who were included in (or excluded from) the author list	Explicitly declare contributions to the project (e.g., CRediT taxonomy). Include everyone who made a significant contribution to the project. Include authors only who contributed to the project.	No	None	• Fong and Wilhite (2017) • Holcombe (2019) • Wislar et al. (2011)

(continued)

Table 3. *(continued)*

QRP	Alias(es) and related concepts	Definition	QRP umbrella term(s)	Example(s)	Potential harms	Preventive measures	Detectability	Clues	Sources
Not making data accessible	—	The data sets and/or codebooks are not made accessible to the public and/or peer reviewers without justifiable cause.	• Misusing open-science practices	Researcher does not provide a publicly accessible repository link to the data set. Data repository link is accessible, but the data are not comprehensible (e.g., lacks cleaning and organization, codebook and instructions), hence it is difficult or impossible to use to reproduce findings.	• Restricted potential for secondary data analysis • Reduced reproducibility	• Data should be shared based on the FAIR (findable, accessible, interoperable, and reusable) principles and legislative context of the researcher. • If confidential and personal information makes participants identifiable, apply masking and anonymization and then share data. • Use synthetic data sets when original data cannot be shared.	Yes	• Data are not shared according to FAIR principles. • No information in the publication on the availability of the data	• Boué et al. (2018) • Ellis and Leek (2018) • Quintana (2020) • Wilkinson et al. (2016)
Publishing studies selectively	File-drawer problem	Choosing which study to publish or share based on whether the findings fit expectations	• Cherry-picking	Researchers run a study and find out the results do not support their hypothesis (e.g., no significant findings). Thus, the researchers do not try to publish or share the study publicly. Researchers run several studies and publish only those that support the hypothesis in a multistudy article.	• Inflated confidence in a multistudy article • Publication bias	• Preregister only on platforms that will eventually publish all preregistrations. • Publish all studies, even when the findings do not support hypotheses.	No	• Publication bias can be estimated in meta-analysis.	• Rosenthal (1979) • Simonsohn et al. (2014)

Note: QRP = questionable research practice; CI = confidence interval; ANOVA = analysis of variance; APA = American Psychological Association; NA = not applicable; LMEMs = Linear Mixed-Effects Models.

Table 4. QRP Umbrella Terms

QRP umbrella term	Description	QRPs
Cherry-picking	Selectively choosing and presenting data or information to support a specific hypothesis or conclusion while ignoring or omitting other relevant data	Omitting important details of the scientific process, publishing studies selectively, selective citing, selective reporting of hypotheses, selective reporting of indicator variables, selective reporting of outcomes, selective test reporting
Citation engineering	Manipulating citations and references to bolster the credibility of a study or a claim or inflate publication statistics	Citing unreliable research, creating multiple publications from the same study, declaring false authorship, selective citing, using irrelevant references, using unjustified references
Influencing participants	Practices that introduce bias or manipulate participants in a way that skews study results	Choosing biased manipulations, choosing biased measurements, placing undue influence on participants
Misusing open-science practices	Violating principles of transparency and openness in research	Not disclosing deviations from preregistration, not making data accessible, preregistering after the results are known (PARKing)
P-hacking	Manipulating statistical analyses and procedures to obtain significant *p* values, often at the expense of scientific integrity	Choosing a poor model specification, choosing unjustified *p*-value adjustment, discretizing continuous variables, excluding data points, missing-data hacking, modifying measurements, neglecting assumptions for statistical models, redefining group-membership rules, selecting a favorable random-number-generator seed, selective test reporting, using ad hoc covariates, using ad hoc exclusion criteria for participants, variable transformation fishing
Sample curation	Manipulating or selectively including/excluding participants to achieve desired results	Optional stopping, mixing pilot- and main-study data, selective sampling, using ad hoc exclusion criteria for participants
None	—	Choosing overlapping measures to find significant results, hypothesizing after the results are known (HARKing), incorrect reporting of test statistics, making unsupported conclusions, not linking the preregistration to the published study, performing inappropriate power analysis, visualizing data in a misleading way

Note: QRPs can belong to more than one umbrella term. QRP = questionable research practice.

participants," including practices that introduce bias or manipulate participants in a way that skews study results. In addition, there were three QRPs classified as "misusing open-science practices," containing practices that violate principles of transparency and openness in research. Seven QRPs did not belong to any specific umbrella term.

Potential harms. Although all QRPs result in misleading conclusions by definition, some QRPs lead to more specific forms of harm. Likewise, it is worth considering that QRPs may have direct and indirect harms. For example, reduced replicability can be a direct harm, whereas the loss of resources (e.g., time, money) because of attempting to replicate an unreliable study would be an indirect harm. In this study, we focused on specific and direct harms (see Table 5).

A range of negative consequences can affect the transparency, efficiency, and credibility of scientific research. Among the most prevalent potential harms, we identified "biased statistical error rates" (i.e., inflated or deflated Type I or Type II error) in 27 cases (see Table 5). This means that several QRPs may result in false-positive or false-negative conclusions. Twenty-three QRPs were associated with reduced replicability. Reduced replicability undermines the reliability and trustworthiness of research outcomes. In addition, 21 QRPs can cause "biased (i.e., inflated or deflated) effect-size estimates," and 13 QRPs were identified to "reduce reproducibility" of the research, which prevents other researchers from verifying findings. Twelve QRPs may directly "inflate credibility" by overstating the quality of key aspects of the scientific process. Such overstatements may affect how particular claims, referenced studies, and scholarly

Table 5. Potential Harms Associated With Each QRP

QRP	Biased effect-size estimate	Biased statistical error rate	Compromised generalizability	Inflated credibility	Reduced replicability	Reduced reproducibility	Other specific harm
Planning							
Choosing biased manipulations	X	X	X				
Choosing biased measurements	X	X			X		
Choosing overlapping measures to find significant results	X	X			X		
Preregistering after results are known (PARKing)				X			
Performing inappropriate power analysis		X		X	X		
Data collection							
Placing undue influence on participants	X	X			X		
Mixing pilot- and main-study data	X	X			X		
Optional stopping		X			X		
Selective sampling	X	X	X		X		
Data processing							
Discretizing continuous variables	X	X	X		X		
Excluding data points	X	X	X		X	X	
Missing-data hacking	X	X			X	X	
Modifying measurements	X	X			X	X	X
Redefining group-membership rules	X	X	X		X	X	
Using ad hoc exclusion criteria for participants	X	X	X		X	X	
Variable transformation fishing	X	X			X	X	
Data analysis							
Choosing a poor model specification	X	X					
Choosing unjustified *p*-value adjustment		X					
Neglecting assumptions for statistical models		X			X	X	
Selecting a favorable random-number-generator seed	X	X			X	X	
Using ad hoc covariates	X	X	X		X	X	

(continued)

Table 5. *(continued)*

QRP	Biased effect-size estimate	Biased statistical error rate	Compromised generalizability	Inflated credibility	Reduced replicability	Reduced reproducibility	Other specific harm
Write-up							
Citing unreliable research				X			
Hypothesizing after the results are known (HARKing)	X	X		X	X		
Incorrect reporting of test statistics		X				X	
Making unsupported conclusions							
Not disclosing deviations from preregistration				X	X		
Omitting important details of the scientific process			X		X	X	
Selective citing				X			
Selective reporting of hypotheses	X	X		X			
Selective reporting of indicator variables	X	X	X		X		
Selective reporting of outcomes	X	X	X		X		
Selective test reporting		X		X	X	X	
Using irrelevant references				X			
Using unjustified references				X			
Visualizing data in a misleading way							
Publication							
Creating multiple publications from the same study		X		X			
Declaring false authorship				X			X
Not linking the preregistration to the published study							
Not making data accessible						X	X
Publishing studies selectively				X			X

Note: Biased effect-size estimate: Several QRPs can inflate or deflate observed effect sizes, leading to over- or underestimation of the true effect size. Biased statistical error rate: QRPs can directly or indirectly influence the probability of Type I (false positive) or Type II (false negative) errors, distorting statistical inference and leading to incorrect conclusions about the presence or absence of effects. Inflated credibility: QRPs can cause an impression that the study is more robust or reliable than it is in reality. Compromised generalizability: QRPs can lead to claiming representativeness in which representativeness is limited, undermining the external validity of the research and the broader applicability of the findings. Reduced replicability: QRPs can decrease the likelihood that independent researchers achieve similar results when repeating a study with new data even if the methodology is identical. Reduced reproducibility: QRPs can impair the ability to recreate the analytical process using the original data set, which makes it difficult to verify the results of the original study. Other harms include those that appeared fewer than three times (see text). For three QRPs (making unsupported conclusions, visualizing data in a misleading way, and not linking the preregistration to the study), no unique harms beyond the generic harm of misleading readers or distorting the interpretation of research findings were identified, although they fully meet our criteria for classification as QRPs. QRP = questionable research practice.

publications are perceived. Note that inflated credibility can be considered an indirect harm as well (e.g., as a consequence of biased statistical error rates that increase confidence in the results). "Compromised generalizability" was identified as a potential harm for 10 QRPs, referring to the tendency for these practices to artificially bolster the perceived applicability of research findings beyond their appropriate scope or context.

Several harms were associated with fewer QRPs but still warrant attention (referred to as "specific harms" in Table 5). These included "reduced validity of the measure," "inflated or deflated reliability of the measure," "publication bias," "restricted potential for secondary data analysis," and "lack of deserved credit for contributing authors."

As can be assumed from Table 5, certain harms are interconnected, with one potentially leading to or exacerbating another, and others arise from distinct, unrelated factors. For example, biased statistical error rates can undermine replicability, but reduced replicability may also result from separate issues, such as the "using biased measurement tools" or the "failure to disclose deviations from preregistered protocols." These latter practices can independently and directly affect the ability to replicate findings.

Preventive measures. We also collected preventive measures that may help to avoid engaging in specific QRPs and provide guidance to enhance research integrity. Although several preventive measures are rather unique to specific QRPs, common properties can be identified. Thus, we grouped the preventive measures into eight broader categories (see Table 6).

The most common category of preventive measures was transparent reporting, in 27 instances, underscoring its importance in ensuring research findings are communicated accurately and comprehensively. Preregistration emerged as a recommended practice in 19 cases, and blind data analysis was advocated in 15 instances. Following best practices was mentioned in 14 instances, emphasizing the value of adhering to established guidelines and standards to promote research integrity. Performing robustness checks, including sensitivity analysis, specification-curve analysis, and assumption checks, was suggested as a preventive measure in 12 cases. The use of psychometrically sound materials was recommended as a preventive measure in six instances, highlighting the importance of employing reliable and valid measures and stimuli. Sharing supplementary information was advocated as a preventive measure in three instances.

Other, more specific preventive measures were identified in five instances, indicating the presence of additional strategies tailored to addressing particular QRPs not covered by the aforementioned categories. Because these other remedies are typically specific to only one or two QRPs, we do not elaborate on them here. Instead, we direct interested readers to Table 3, which provides a comprehensive listing of each preventive measure.

Detectability and clues. As Table 7 shows, out of the total number of QRPs, we categorized 18 as detectable by readers possessing sufficient knowledge of the field if relying solely on the information presented in the publication. Conversely, three QRPs were deemed undetectable ("placing undue influence on participants," "publishing studies selectively," "declaring false authorship"). As for the remaining 19 QRPs, their detection is contingent on the availability of supplementary information, such as open data, preregistration records, or other relevant resources. We found that the clues used to identify a QRP exhibit a high degree of idiosyncrasy. In other words, the majority of clues are unique to each specific QRP and do not apply universally across different QRPs. We identified 97 separate clues, and only a few were associated with multiple QRPs. Note that observing any of the clues that we collected for each QRP does not automatically imply a questionable practice. Some well-justified actions in scientific practice might appear similar to QRPs only upon initial observation (Sacco et al., 2019).

Discussion

In this study, we aimed to define, collect, and categorize QRPs in quantitative psychological research. We used community consensus to develop a definition and a list of 40 QRPs and their aliases, examples, detectability, clues, associated harms, and preventive measures. We also implemented a higher-order categorization—QRP umbrella terms—to group conceptually related QRPs and specified the phase in the research process during which researchers may engage in QRPs. Our definition distinguishes QRPs from research misconduct and random error, and the QRPs bestiary specifies the actions that may produce misleading conclusions. The outcomes of QRPs are often in the interest of the researcher, but we leave discussions of intentionality aside. We hope that this bestiary will provide a useful resource for researchers and educators in and possibly beyond psychology.

Most of the previous studies investigating QRPs did not provide a definition and presented examples instead. There are two notable exceptions: a definition by Elsherif et al. (2025) and one by Gerrits et al. (2019), although the definition by Elsherif et al. considered only actions that distort data, and the definition by Gerrits et al. focused on reporting. We consider QRPs from a broader perspective (similar to the definition proposed by Banks, O'Boyle, et al., 2016), also referring to the collection and interpretation of data and the overall presentation of the research. Unlike the definitions by Lakens (2022) and Al-Marzouki et al. (2005), we explicitly

Table 6. Aggregated Preventive Measures Associated With Each QRP

QRP	Blind data analysis	Follow best practices	Perform robustness checks	Preregistration	Share supplementary information	Transparent reporting	Use psychometrically sound materials	Other preventive measures
Planning								
Choosing biased manipulations		X					X	
Choosing biased measurements		X					X	X
Choosing overlapping measures to find significant results						X	X	X
Preregistering after results are known (PARKing)				X		X		
Performing inappropriate power analysis		X				X		
Data collection								
Placing undue influence on participants		X					X	X
Mixing pilot- and main-study data						X		X
Optional stopping				X				
Selective sampling		X		X		X		X
Data processing								
Discretizing continuous variables	X		X			X	X	
Excluding data points	X		X	X		X		
Missing-data hacking	X		X	X				
Modifying measurements	X		X	X	X	X	X	
Redefining group-membership rules	X		X	X		X		
Using ad hoc exclusion criteria for participants	X		X	X		X		
Variable transformation fishing	X		X	X		X		
Data analysis								
Choosing a poor model specification	X	X	X			X		
Choosing unjustified *p*-value adjustment	X		X	X				
Neglecting assumptions for statistical models						X		
Selecting a favorable random-number-generator seed	X	X				X		
Using ad hoc covariates	X			X		X		

(continued)

Table 6. *(continued)*

QRP	Blind data analysis	Follow best practices	Perform robustness checks	Preregistration	Share supplementary information	Transparent reporting	Use psychometrically sound materials	Other preventive measures
Write-up								
Citing unreliable research		X						
Hypothesizing after the results are known (HARKing)	X		X	X		X		
Incorrect reporting of test statistics		X			X			
Making unsupported conclusions						X		
Not disclosing deviations from preregistration			X			X		
Omitting important details of the scientific process				X		X		
Selective citing		X						
Selective reporting of hypotheses				X		X		
Selective reporting of indicator variables	X			X		X		
Selective reporting of outcomes	X			X		X		
Selective test reporting	X		X	X		X		
Using irrelevant references		X						
Using unjustified references		X						
Visualizing data in a misleading way		X						
Publication								
Creating multiple publications from the same study				X		X		
Declaring false authorship						X		
Not linking the preregistration to the published study						X		
Not making data accessible		X			X			
Publishing studies selectively				X		X		

Note: For further discussion of preventive measures, see the How to Avoid QRPs section. QRP = questionable research practice.

Table 7. Detectability of QRPs

Detectability	QRPs	*N*
Yes	Choosing a poor model specification, choosing unjustified *p*-value adjustment, citing unreliable research, discretizing continuous variables, incorrect reporting of test statistics, making unsupported conclusions, neglecting assumptions for statistical models, not disclosing deviations from preregistration, not making data accessible, preregistering after results are known (PARKing), performing inappropriate power analysis, selective citing, selective sampling, using biased measurements, using irrelevant references, using measurement overlap to find significant results, using unjustified references, visualizing data in a misleading way	18
No	Declaring false authorship, placing undue influence on participants, publishing studies selectively	3
Maybe	Creating multiple publications from the same study, excluding data points, hypothesizing after the results are known (HARKing), missing-data hacking, modifying measurements, not linking the preregistration to the published study, omitting important details of the scientific process, optional stopping, redefining group-membership rules, mixing pilot- and main-study data, selecting a favorable random-number-generator seed, selective reporting of hypotheses, selective reporting of indicator variables, selective reporting of outcomes, selective test reporting, using ad hoc covariates, using ad hoc exclusion criteria for participants, using biased manipulations, variable transformation fishing	19

Note: QRP = questionable research practice.

distinguish research misconduct and fraud from QRPs, which legitimizes using the term "questionable" in the first place. Misconduct and fraud by definition can never be questionable and should therefore be viewed and treated differently from QRPs. In our view, questionable measurement practices, as suggested by Flake and Fried (2020), are a subset of QRPs that specifically deal with QRPs related to the measurement of constructs.

We identified 40 QRPs corresponding to six research phases from research planning to publication. Prior studies that listed QRPs did not collect them systematically and, rather, used those found in published articles, originating from the first study by John et al. (2012). Lakens (2022) collected 15 studies attempting to estimate the prevalence of QRPs and found a union of 12 items. This difference in number indicates that a substantial portion of the QRPs we described were previously unlisted even though they are generally recognized as QRPs (e.g., "grooming participants," "visualizing data in a misleading way," "citing unreliable research"). Note that we decided to drop "falsifying data," a practice listed by John et al. as a QRP, because we believe it is research misconduct and therefore outside our definition of QRPs.

While trying to define widely used terms such as "p-hacking" and "cherry-picking," we realized that they actually cover several otherwise distinguishable QRPs. By breaking up these and other broad terms into smaller units, we were able to clarify definitions, harms, and preventive measures more precisely. Concurrently, we introduced the concept of QRP umbrella terms to keep these existing labels without compromising the distinctiveness of individual QRPs. Keeping these widely used terms as broader categories emphasizes that it is important to be specific about the actual QRP implied when using an umbrella term, such as "p-hacking," in the future.

We also identified six potential harms (e.g., biased statistical error rates, reduced replicability, inflated credibility) associated with several different QRPs and many more that were specific to individual QRPs. Determining harms may help to clarify why QRPs pose a threat to scientific integrity. However, as Stefan and Schönbrodt (2023) pointed out in their simulation study, QRPs vary in the extent to which they can bias the interpretation of results, such as inflating statistical error rates. Although the methodology of our study is adequate to identify the potential harms, it is up to future studies to empirically test and quantify the extent of associated harms.

Our findings indicate that several QRPs may be detectable based on the original publication- and domain-specific knowledge. Other QRPs might be detected if supplementary materials, such as preregistration, open data, analysis code, and so on, are provided. We collected several clues that can indicate—but not unequivocally prove—if a study's authors potentially engaged in a specific QRP. This list of clues can be used to improve the peer-review process either by providing a checklist for human reviewers or by providing prompts to artificial-intelligence systems for automatic risk assessment.

How to avoid QRPs

In addition to cataloging QRPs and their attributes, we also assembled a list of preventive measures that may

help researchers to prevent engaging in QRPs. In this section, we explore a few general strategies that have emerged as promising approaches to thwart QRPs.

One of the fundamental ways to avoid QRPs is by embracing best practices and using comprehensive checklists. Researchers should adopt rigorous and transparent methodologies, adhering to established guidelines and protocols specific to their field (Appelbaum et al., 2018). By following standardized procedures and employing systematic checklists, researchers can ensure that critical steps are not omitted, minimizing the potential for QRPs to influence their findings (e.g., see checklist provided by Aczel et al., 2020).

Preregistration has gained significant attention as a powerful tool in reducing QRPs (Nosek et al., 2018). By publicly registering their research plans, including hypotheses, methodologies, and analysis strategies, before data collection, researchers establish a predetermined framework that guards against selective reporting, data manipulation, and HARKing. Preregistration enhances transparency, encourages hypothesis-driven research, and helps to demarcate confirmatory from exploratory analyses, thereby mitigating the impact of QRPs on study outcomes (Sarafoglou et al., 2022; van den Akker, Bakker, et al., 2024; van den Akker, van Assen, et al., 2024). Although preregistration does not eliminate QRPs, as demonstrated by van den Akker, Bakker, et al. (2024), it is an important preventive tool.

Extending preregistration, registered reports offer additional safeguards for maintaining high research quality. This approach divides the scientific process into two distinct stages: a peer-reviewed planning phase and an execution stage (Chambers et al., 2015; Chambers & Tzavella, 2022). This separation ensures that the research design is rigorously assessed before data collection begins. By doing so, registered reports enhance the credibility of subsequent findings and contribute to overall research robustness and transparency. This methodology helps to identify potential biases and methodological issues early on, leading to more reliable knowledge generation. Our bestiary might help researchers identify QRPs during the preregistration before engaging in them.

Blind data analysis is another emerging trend that can help prevent several QRPs (MacCoun & Perlmutter, 2015). In blind analysis, researchers carry out data processing and analysis without access to the labels of conditions (or key outcomes). These labels are added to the data set only once the exact data-analysis strategy and code are finalized. By eliminating awareness of conditions and correlations with the main outcomes in the first stage, researchers prevent conscious or subconscious manipulation of the analytical process to yield desired results. This approach can safeguard against QRPs, such as *p*-hacking, cherry-picking results, or adjusting methodologies to fit preconceived notions, and help to adhere to the preregistered analysis (Dutilh et al., 2021; Sarafoglou et al., 2023).

Robust statistical methods and error checking play a vital role in combating QRPs. Researchers should employ rigorous statistical techniques, such as appropriate power analysis, multiple comparison corrections, and effect-size estimation, to ensure that their findings are statistically sound. In addition, implementing thorough error-checking procedures, such as data and result validation, helps to identify and rectify potential errors or anomalies that could be indicative of QRPs, for example, impossible values in summary statistics or when test statistics do not match *p* values (Brown & Heathers, 2017; Nuijten, 2022; Nuijten et al., 2016).

Replication can be another way to counteract and limit the harms of QRPs. By demonstrating consistent results across multiple replications, researchers enhance the reliability and generalizability of their findings. Replications can be done using a separate data collection or using an unobserved subset of the data as a test set (i.e., holdout data set) to verify the results (Weston et al., 2019; Yarkoni & Westfall, 2017).

Sensitivity analysis involves systematically varying key parameters and assumptions in a study to assess the robustness of the findings. This process helps researchers understand the potential impact of different choices on their results, thereby enhancing transparency and reducing the likelihood of QRPs. Specification-curve analysis (also called "multiverse analysis"; see Steegen et al., 2016), a more recent development, takes sensitivity analysis a step further by exploring a wide range of modeling specifications to create a comprehensive view of how different analytical choices influence outcomes (Simonsohn et al., 2020). This approach aids in uncovering the potential effects of undisclosed decisions that might have otherwise remained hidden, reducing the temptation for researchers to employ QRPs that lead to positive results.

Collaborative validation, such as adversarial collaborations (Rakow, 2022), multianalyst approaches (Aczel et al., 2021; Silberzahn & Uhlmann, 2015; Wagenmakers et al., 2022), and red teaming (Lakens, 2020), provides an innovative, although resource-intensive, approach to identifying and addressing potential QRPs. In this collaborative setup, multiple independent analysts with diverse perspectives and expertise critically evaluate and replicate each other's work. By subjecting research to rigorous scrutiny and incorporating dissenting viewpoints, collaborative validation fosters intellectual honesty and exposes potential QRPs, resulting in more reliable and robust scientific conclusions (Aczel et al., 2021).

Finally, fostering open-science practices contributes significantly to the mitigation of QRPs. Researchers should actively embrace the principles of openness by sharing their data, research materials, and code with the

scientific community. By making these resources openly accessible, other researchers can verify and replicate the findings, increasing transparency and accountability. Although the open sharing of supplementary information and data is listed as a preventive measure against QRPs only sparingly, its real value lies in facilitating the detection of questionable practices. Practices such as "born open data," in which data sets are automatically archived at the time of creation, represent an even more transparent approach (Kekecs et al., 2023; Rouder, 2016).

Although our list of preventive measures offers valuable strategies to prevent QRPs, this compilation may not be exhaustive. QRPs can manifest in diverse ways across research contexts, and addressing them requires ongoing vigilance and adaptability. Nonetheless, the provided preventive measures serve as a starting point to promote transparency, rigor, and the cultivation of trustworthy scientific outcomes. These preventive measures and safeguards provide a foundation for increasing the reliability and reproducibility of scientific findings, ultimately advancing the pursuit of knowledge and benefiting society as a whole. The first step to any prevention of QRPs is increasing awareness and ability to recognize potential QRPs, for which this bestiary provides a helpful starting point.

Concepts similar to QRPs

In other areas of research, scholars have sought to define and catalog threats to scientific integrity independently from QRPs. Wicherts et al. (2016) provided a wide definition for researcher degrees of freedom that was accompanied by a list of 34 items. Researcher degrees of freedom are choice points in which researchers make decisions about how to proceed in the scientific process. Although some of these choices can be justified, others may be questionable. For instance, the researcher degrees of freedom of "measuring a dependent variable in several alternative ways" becomes a QRP only if not all outcomes are reported ("selective reporting of outcomes" in our bestiary), potentially leading to false conclusions.

A recent preprint introduced Seaboat (https://www.seaboat.io/), a consensus-based tool that evaluates the most common threats to the validity of empirical research (Schiavone et al., 2023). Seaboat was designed to help reviewers evaluate the quality of quantitative empirical research and make the peer-review process more systematic. It focuses on four types of validities and 32 potential threats to these. The list consists of five threats to construct validity (e.g., insufficient information about operationalization), eight threats to internal validity (e.g., order effects in within-persons designs not ruled out), six threats to external validity (e.g., overgeneralizations), and 13 threats to statistical-conclusion validity (e.g., data exclusion not justified). As shown in the examples, threats to validity and QRPs overlap considerably, but these constructs are not identical. QRPs are actions, whereas threats to validity represent attributes of the article and/or the research. However, some of these threats can be a result of QRPs that have been part of the research process, whereas others (e.g., the neglect of order effects) are arguably more likely to stem from insufficient statistical rigor and may not introduce systematic bias.

The Catalog of Bias is an ongoing collaboration of health researchers to compile an up-to-date and exhaustive list of biases "which may distort the design, execution, analysis, and interpretation of research" (Sackett, 1979, p. 51). The catalog not only collects biases but also provides definitions and potential impacts on empirical research. As a result, Sackett (1979) listed 35 biases regarding sampling and measurement in clinical trials and 56 biases that may potentially affect case-control and cohort studies. Throughout the years, the list has changed, and currently, the catalog consists of 65 biases (https://catalogofbias.org/). Examples include "selection bias," "allocation bias," and "detection bias." We consider such biases as harmful research outcomes, whereas QRPs are all specific actions.

Furthermore, FORRT (Parsons et al., 2022) introduced another compilation of QRPs accompanied by definitions. Their methodology involves presenting a QRP alongside a definition from a specific published finding, leaving ambiguity in the selection process when multiple interpretations exist. In contrast, we have created a systematic list based on community consensus, providing a more objective perspective. Our bestiary can be used as a peer-reviewed compendium of QRPs and their respective definitions, clues, harms, and preventive measures.

Strengths, limitations, and future directions

We chose to use the community-consensus method to identify as many QRPs as possible without the constraints of relying only on existing lists of QRPs that usually focused on QRPs that were creating false-positive findings. This allowed us to acknowledge that in some fields, QRPs may be used to produce false null findings. For example, researchers may want to downplay the side effects of an intervention or want to maintain the status quo to keep getting funding. Our approach considered this possibility for each QRP. In addition, we aimed to provide evidence for the listed research practices as being questionable; therefore, we collected references for each. This approach enriched the community

consensus with additional support from the existing literature.

Despite concerns about the reliability of community consensus (or expert consensus) as a form of evidence (see Minas & Jorm, 2010), we chose this method specifically because we believed it might be better suited for finding lesser known QRPs, particularly when including collaborators who are more interested in research methodology. Moreover, in this instance, finding a QRP that is seldom used may cause less harm than not finding an important QRP. Therefore, we prioritized comprehensiveness even if it meant including rare QRPs. In fact, some practices that were previously not recognized as QRPs (e.g., "selecting a favorable random-number-generator seed" or "visualizing data in a misleading way") have references that mention them, proving that other researchers have encountered these.

Our particular approach to community consensus may also raise concerns about bias or representativeness because of its deviation from more established protocols, such as the Delphi method (Dalkey & Helmer, 1963). Unlike the Delphi method, we did not adhere to certain key aspects of expert selection and anonymity in our process. Note that these deviations could potentially affect the validity of our findings. However, our study did incorporate several essential elements commonly associated with consensus-building methods. First, we held the initial hackathon at a conference that focused on metascience and methodological advancement in psychological science, attracting a diverse group of 37 highly engaged experts, and 19 collaborators remained after the hackathon. Multiple rounds of feedback allowed collaborators to iteratively refine the contributions to the bestiary. Although our approach might differ from the strict procedures of the Delphi method, the inclusion of these fundamental elements allowed us to engage collaborators in a constructive dialogue. This dialogue facilitated the exchange of perspectives and ideas, ultimately contributing to a more informed decision-making process. In response to a reviewer's suggestion, we also conducted a thorough review of published literature specifically addressing QRPs and their broader categories (Banks, O'Boyle, et al., 2016; Banks, Rogelberg, et al., 2016; Ravn & Sørensen, 2021; Suter, 2020; Y. Xie et al., 2021). This review allowed us to cross-reference our list of QRPs and identify any practices that we might have inadvertently omitted. We did not identify any further QRPs that would fit our definition.

Although our list is probably the most comprehensive list of QRPs to date, it is likely incomplete. As Stefan and Schönbrodt (2023) already noted, compiling a truly exhaustive list of QRPs seems like an impossible endeavor, especially because many highly specialized fields of research—both within and outside of psychological science—have very specific ways of designing studies and collecting, processing, and analyzing data. However, we hope that our list contains most of the QRPs used in psychology and the social sciences in general. Moreover, we see our work as more than just a list of QRPs because it also represents a systematic approach to QRPs (e.g., describing actions and behaviors related to specific QRPs, providing a definition, identifying the research phase and associated harms and remedies). This systematic approach is likely more enduring than the items on the list.

Note that our bestiary was developed exclusively for quantitative research methods, although some of the QRPs (e.g., placing undue influence on participants or QRPs associated with the write-up phase) might be relevant for qualitative research as well. We believe that our work might inform similar endeavors aimed at improving research practices in the qualitative-research community.

The aim of this work was to create an extended list of QRPs in terms of specific behaviors that may result in specific harms. Because we are providing a new list of QRPs, their prevalence and the impact of associated harms are largely unknown and require further research to quantify. In addition, even for the more prominent QRPs, only a few potential harms have been investigated (effects on statistical error rate, estimated effect size, and replicability), and other harms are more difficult to quantify (e.g., effect on credibility and reproducibility). With this work, we hope to spark future research interest aimed at estimating the prevalence of engagement in QRPs, the effectiveness of preventive measures, and the severity of QRP-associated harms.

Conclusion

Our goal with this bestiary is to help researchers avoid QRPs, thus ultimately raising the standard of psychological research. This community-consensus study showed that a proper definition of QRPs is essential and achievable, as is their collection and categorization. Among the 40 QRPs identified in our study, many were previously unrecognized. Moreover, we found that some formerly known QRPs (e.g., *p*-hacking) actually comprised several well-distinguishable QRPs—we named these "QRP umbrella terms." We concluded that most QRPs can be detected, especially when journals and institutions embrace open-science practices, such as data sharing and preregistration. Moreover, specific harms were identified for each QRP along with preventive measures that can help to avoid or mitigate the detrimental effects of QRPs on the integrity and validity of research findings. Finally, this work can be used for further research on QRPs, (self-)education, and peer review.

Transparency

Action Editor: Katie Corker
Editor: David A. Sbarra

Author Contributions

T. Nagy and J. Hergert are shared first authors.

Tamás Nagy: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Jane Hergert: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Validation; Writing – original draft; Writing – review & editing.

Mahmoud M. Elsherif: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Writing – review & editing.

Lukas Wallrich: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Kathleen Schmidt: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Tal Waltzer: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Jason W. Payne: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Biljana Gjoneska: Conceptualization; Data curation; Writing – review & editing.

Yashvin Seetahul: Conceptualization; Data curation; Investigation; Writing – review & editing.

Y. Andre Wang: Conceptualization; Data curation; Investigation; Writing – review & editing.

Daniel Scharfenberg: Conceptualization; Data curation; Investigation; Writing – review & editing.

Gabriella Tyson: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Yu–Fang Yang: Conceptualization; Data curation; Investigation; Writing – review & editing.

Aleksandrina Skvortsova: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Samuel Alarie: Conceptualization; Investigation; Writing – review & editing.

Katherine Graves: Conceptualization; Data curation; Investigation; Writing – review & editing.

Lukas K. Sotola: Conceptualization; Data curation; Investigation; Writing – review & editing.

David Moreau: Conceptualization; Data curation; Investigation; Writing – review & editing.

Eva Rubínová: Conceptualization; Data curation; Investigation; Methodology; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

Kathleen Schmidt was supported by the John Templeton Foundation; Mahmoud M. Elsherif was supported by the Leverhulme Early Career Researcher Grant; David Moreau was supported by the Marsden Grant from Royal Society NZ; Gabriella Tyson was supported by the NIHR Maudsley Biomedical Research Centre; Tal Waltzer was supported by the NSF SPRF-FR# 2104610; Tamás Nagy was supported by the Research Fund of the National Research, Development and Innovation Office (PD 131954), the University Excellence Fund of Eötvös Loránd University, Budapest, Hungary (ELTE), and the János Bolyai research fellowship of the Hungarian Academy of Sciences; Y. Andre Wang was supported by a Social Sciences and Humanities Research Council (SSHRC) Insight Development Grant (430-2022-00087).

Open Practices

This article has received the badge for Open Data Badge. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Tamás Nagy https://orcid.org/0000-0001-5244-0356
Jane Hergert https://orcid.org/0009-0000-4079-9927
Mahmoud M. Elsherif https://orcid.org/0000-0002-0540-3998
Lukas Wallrich https://orcid.org/0000-0003-2121-5177
Kathleen Schmidt https://orcid.org/0000-0002-9946-5953
Jason W. Payne https://orcid.org/0000-0001-9176-9795
Biljana Gjoneska https://orcid.org/0000-0003-1200-6672
Yashvin Seetahul https://orcid.org/0000-0001-7487-3398
Y. Andre Wang https://orcid.org/0000-0002-5729-7373
Yu-Fang Yang https://orcid.org/0000-0001-9089-6020
Lukas K. Sotola https://orcid.org/0000-0001-8120-394X
David Moreau https://orcid.org/0000-0002-1957-1941

Notes

1. In this article, we use "credibility" in the sense of perceived accuracy and trustworthiness.
2. Hackathons are collaborative events during which individuals with diverse skills come together to address a specific problem or create innovative solutions in a short time frame.

References

Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N. N., Donkin, C., van Doorn, J. B., . . . Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, *10*, Article e72185. https://doi.org/10.7554/eLife.72185

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

Adam, D. (2019, May 23). A solution to psychology's reproducibility problem just failed its first test. *Science*. https://doi.org/10.1126/science.aay1207

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLOS ONE*, *12*(3), Article e0172792. https://doi.org/10.1371/journal.pone.0172792

Al-Marzouki, S., Roberts, I., Evans, S., & Marshall, T. (2008). Selective reporting in clinical trials: Analysis of trial protocols accepted by *The Lancet*. *Lancet*, *372*(9634), 201. https://doi.org/10.1016/S0140-6736(08)61060-0

Al-Marzouki, S., Roberts, I., Marshall, T., & Evans, S. (2005). The effect of scientific misconduct on the results of clinical trials: A Delphi survey. *Contemporary Clinical Trials*, *26*(3), 331–337. https://doi.org/10.1016/j.cct.2005.01.011

American Psychological Association. (2023). *Journal article reporting standards (JARS)*. https://apastyle.apa.org/jars

Anderson, S. F., & Liu, X. (2023). Questionable research practices and cumulative science: The consequences of selective reporting on effect size bias and heterogeneity. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000572

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *The Journal of Clinical Psychiatry*, *82*(1), Article 20f13804. https://doi.org/10.4088/JCP.20f13804

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *The American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Artino, A. R., Jr., Driessen, E. W., & Maggio, L. A. (2019). Ethical shades of gray: International frequency of scientific misconduct and questionable research practices in health professions education. *Academic Medicine*, *94*(1), 76–84. https://doi.org/10.1097/ACM.0000000000002412

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)*. OSF. https://doi.org/10.31219/osf.io/bnh7p

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421.

Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *The Journal of Communication*, *71*(5), 715–738. https://doi.org/10.1093/joc/jqab031

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Bakker, M., & Wicherts, J. M. (2014a). Outlier removal and the relation with reporting errors and quality of psychological research. *PLOS ONE*, *9*(7), Article e103360. https://doi.org/10.1371/journal.pone.0103360

Bakker, M., & Wicherts, J. M. (2014b). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. *Psychological Methods*, *19*(3), 409–427. https://doi.org/10.1037/met0000014

Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., Vist, G. E., Falck-Ytter, Y., Meerpohl, J., Norris, S., & Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, *64*(4), 401–406. https://doi.org/10.1016/j.jclinepi.2010.07.015

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*(1), 5–20. https://doi.org/10.1177/0149206315619011

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

Becker, T. E., Atinc, G., Breaugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2016). Statistical control in correlational studies: 10 essential recommendations for organizational researchers. *Journal of Organizational Behavior*, *37*(2), 157–167. https://doi.org/10.1002/job.2053

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—When and how? *Journal of Clinical Epidemiology*, *54*(4), 343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

Bergkvist, L. (2020). Preregistration as a way to limit questionable research practice in advertising research. *International Journal of Advertising*, *39*(7), 1172–1180. https://doi.org/10.1080/02650487.2020.1753441

Bespalov, A., Barnett, A. G., & Begley, C. G. (2018). Industry is more alarmed about reproducibility than academia. *Nature*, *563*(7733), Article 626. https://doi.org/10.1038/d41586-018-07549-w

Boué, S., Byrne, M., Hayes, A. W., Hoeng, J., & Peitsch, M. C. (2018). Embracing transparency through data sharing. *International Journal of Toxicology*, *37*(6), 466–471. https://doi.org/10.1177/1091581818803880

Brachem, J., Frank, M., Kvetnaya, T., Schramm, L. F. F., & Volz, L. (2022). Replikationskrise, *p*-hacking und Open Science. *Psychologische Rundschau*, *73*(1), 1–17. https://doi.org/10.1026/0033-3042/a000562

Broad, W. J. (1981). The publishing game: Getting more for less. *Science*, *211*(4487), 1137–1139. https://doi.org/10.1126/science.7008199

Brookes, S. T., Whitley, E., Peters, T. J., Mulheran, P. A., Egger, M., & Davey Smith, G. (2001). Subgroup analysis in randomised controlled trials: Quantifying the risks of false-positives and false-negatives. *Health Technology Assessment*, *5*(33), 1–56. https://doi.org/10.3310/hta5330

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022

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

Chuard, P. J. C., Vrtílek, M., Head, M. L., & Jennions, M. D. (2019). Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? *PLOS Biology*, *17*(1), Article e3000127. https://doi.org/10.1371/journal.pbio.3000127

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), Article 211037. https://doi.org/10.1098/rsos.211037

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 247–253. https://doi.org/10.1177/014662168300700301

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

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., Waldorp, L. J., & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647. https://doi.org/10.3758/s13423-015-0913-5

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9*(3), 458–467. http://www.jstor.org/stable/2627117

de Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, *28*(3), 795–812. https://doi.org/10.3758/s13423-020-01803-x

DeCoster, J., Gallucci, M., & Iselin, A.-M. R. (2011). Best practices for using median splits, artificial categorization, and their continuous alternatives. *Journal of Experimental Psychopathology*, *2*(2), 197–209. https://doi.org/10.5127/jep.008310

DeepChecks. (n.d.). *Selective sampling*. https://deepchecks.com/glossary/selective-sampling/

Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2021). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*(23), 5745–5772. https://doi.org/10.1007/s11229-019-02456-7

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. *Journal of Clinical Epidemiology*, *88*, 92–101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Ellis, S. E., & Leek, J. T. (2018). How to share data for collaboration. *The American Statistician*, *72*(1), 53–57. https://doi.org/10.1080/00031305.2017.1375987s

Elsherif, E., Kalandadze, T., Ngiam, W., Parsons, S., Paul, M., Rinke, E. M., Roettger, T., & Azevedo, F. (2025, February 7). *Questionable Research Practices or Questionable Reporting Practices (QRPs)*. Framework for Open and Reproducible Research Training. Retrieved June 10, 2025, from https://forrt.org/glossary/english/questionable_research_practices_or_questionable_reporting_practices/

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLOS ONE*, *12*(12), Article e0187394. https://doi.org/10.1371/journal.pone.0187394

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, *13*(7), Article e0200303. https://doi.org/10.1371/journal.pone.0200303

Freuli, F., Held, L., & Heyard, R. (2023). Replication success under questionable research practices—A simulation study. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *38*(4), 621–639. https://doi.org/10.1214/23-sts904

Gelman, A., & Loken, E. (n.d.). *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*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gernsbacher, M. A. (2018). Writing empirical articles: Transparency, reproducibility, clarity, and memorability. *Advances in Methods and Practices in Psychological Science*, *1*(3), 403–414. https://doi.org/10.1177/2515245918754485

Gerrits, R. G., Jansen, T., Mulyanto, J., van den Berg, M. J., Klazinga, N. S., & Kringos, D. S. (2019). Occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: A structured assessment of publications authored by researchers in the Netherlands. *BMJ Open*, *9*(5), Article e027903. https://doi.org/10.1136/bmjopen-2018-027903

Gopalakrishna, G., Ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE*, *17*(2), Article e0263023. https://doi.org/10.1371/journal.pone.0263023

Götz, M., O'Boyle, E. H., Gonzalez-Mulé, E., Banks, G. C., & Bollmann, S. S. (2021). The "Goldilocks zone": (Too) Many confidence intervals in tests of mediation just exclude zero. *Psychological Bulletin*, *147*(1), 95–114. https://doi.org/10.1037/bul0000315

Gøtzsche, P. C. (2022). Citation bias: Questionable research practice or scientific misconduct? *Journal of the Royal Society of Medicine*, *115*(1), 31–35. https://doi.org/10.1177/01410768221075881

Gould, E., Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Hamilton, D. G., Abbey-Lee, R. N., Abbott, J. K., Aguirre, L. A., Alcaraz, C., Aloni, I., Altschul, D., Arekar, K., Atkins, J. W., Atkinson, J., Baker, C., Barrett, M., . . . Zitomer, R. A. (2023). *Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology*. EcovoRxiv. https://ecoevorxiv.org/repository/view/6000/

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), Article e1002106. https://doi.org/10.1371/journal.pbio.1002106

Heckman, M. G., Davis, J. M., 3rd, & Crowson, C. S. (2022). Post hoc power calculations: An inappropriate method for interpreting the findings of a research study. *The Journal of Rheumatology*, *49*(8), 867–870. https://doi.org/10.3899/jrheum.211115

Hilgard, J., Sala, G., Boot, W. R., & Simons, D. J. (2019). Overestimation of action-game training effects: Publication bias and salami slicing. *Collabra: Psychology*, *5*(1), Article 30. https://doi.org/10.1525/collabra.231

Hodson, G. (2021). Construct jangle or construct mangle? Thinking straight about (nonredundant) psychological constructs. *Journal of Theoretical Social Psychology*, *5*(4), 576–590. https://doi.org/10.1002/jts5.120

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

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

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694

Joe, S., & Leif, N. (2020, March 10). *Data Replicada #4: The problem of hidden confounds*. Data Colada. https://datacolada.org/85

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

Kaiser, M., Drivdal, L., Hjellbrekke, J., Ingierd, H., & Rekdal, O. B. (2021). Questionable research practices and misconduct among Norwegian researchers. *Science and Engineering Ethics*, *28*(1), Article 2. https://doi.org/10.1007/s11948-021-00351-4

Kekecs, Z., Palfi, B., Szaszi, B., Szecsi, P., Zrubka, M., Kovacs, M., Bakos, B. E., Cousineau, D., Tressoldi, P., Schmidt, K., Grassi, M., Evans, T. R., Yamada, Y., Miller, J. K., Liu, H., Yonemitsu, F., Dubrov, D., Röer, J. P., Becker, M., . . . Aczel, B. (2023). Raising the value of research studies in psychological science by increasing the credibility of research reports: The transparent Psi project. *Royal Society Open Science*, *10*(2), Article 191375. https://doi.org/10.1098/rsos.191375

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

KNAW, NFU, NWO, TO2-federatie, Hogescholen, V., & VSNU. (2018). *Nederlandse gedragscode wetenschappelijke integriteit* [Dutch code of conduct for scientific integrity] [Data set]. Data Archiving and Networked Services (DANS). https://doi.org/10.17026/DANS-2CJ-NVWU

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Kovacs, M., Hoekstra, R., & Aczel, B. (2021). The role of human fallibility in psychological research: A survey of mistakes in data management. *Advances in Methods and Practices in Psychological Science*, *4*(4). https://doi.org/10.1177/25152459211045930

Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., & Aczel, B. (2022). SampleSizePlanner: A tool to estimate and justify sample size for two-group studies. *Advances in Methods and Practices in Psychological Science*, *5*(1). https://doi.org/10.1177/25152459211054059

Kravitz, D., & Mitroff, S. (2020). *Quantifying, and correcting for, the impact of questionable research practices on false discovery rates in psychological science*. PsyArXiv. https://doi.org/10.31234/osf.io/fu9gy

Kravitz, D. J., & Mitrof, S. R. (2023). Quantifying, and correcting for, the impact of questionable research practices on false discovery rates in psychological science. *Journal for Reproducibility in Neuroscience*. https://doi.org/10.36850/jrn.2023.e44

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2010). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Lakens, D. (2020, May 11). *Pandemic researchers — Recruit your own best critics*. Nature, *581*(7807), Article 121. https://doi.org/10.1038/d41586-020-01392-8

Lakens, D. (2022a). *Improving your statistical inferences*. Zenodo. https://doi.org/10.5281/zenodo.6409077

Lakens, D. (2022b). Sample size justification. *Collabra: Psychology*, *8*(1), Article 33267. https://doi.org/10.1525/collabra.33267

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, *10*(1), Article 117094. https://doi.org/10.1525/collabra.117094

Lang, S., Armstrong, N., Deshpande, S., Ramaekers, B., Grimm, S., de Kock, S., Kleijnen, J., & Westwood, M. (2019). Clinically inappropriate post hoc exclusion of study

participants from test accuracy calculations: The ROMA score, an example from a recent NICE diagnostic assessment. *Annals of Clinical Biochemistry*, *56*(1), 72–81. https://doi.org/10.1177/0004563218782722

Lee, D. K. (2020). Data transformation: A focus on the interpretation. *Korean Journal of Anesthesiology*, *73*(6), 503–508. https://doi.org/10.4097/kja.20137

Letrud, K., & Hernes, S. (2019). Affirmative citation bias in scientific myth debunking: A three-in-one case study. *PLOS ONE*, *14*(9), Article e0222213. https://doi.org/10.1371/journal.pone.0222213

Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review*, *7*(3), 471–479. https://doi.org/10.1111/j.1740-8784.2011.00222.x

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*(7572), 187–189. https://doi.org/10.1038/526187a

Marchetti, S., & Schellens, J. H. M. (2007). The impact of FDA and EMEA guidelines on drug development in relation to Phase 0 trials. *British Journal of Cancer*, *97*(5), 577–581. https://doi.org/10.1038/sj.bjc.6603925

Marks, E. S. (1947). Selective sampling in psychological research. *Psychological Bulletin*, *44*(3), 267–275. https://doi.org/10.1037/h0061812

McCambridge, J., de Bruin, M., & Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PLOS ONE*, *7*(6), Article e39116. https://doi.org/10.1371/journal.pone.0039116

Mehregan, M. (2022). Scientific journals must be alert to potential manipulation in citations and referencing. *Research Ethics*, *18*(2), 163–168. https://doi.org/10.1177/17470161211068745

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, Article 104092. https://doi.org/10.1016/j.jml.2020.104092

Midway, S., Robertson, M., Flinn, S., & Kaller, M. (2020). Comparing multiple comparisons: Practical guidance for choosing the best multiple comparisons test. *PeerJ*, *8*, Article e10387. https://doi.org/10.7717/peerj.10387

Minas, H., & Jorm, A. F. (2010). Where there is no evidence: Use of expert consensus methods to fill the evidence gap in low-income countries and cultural minorities. *International Journal of Mental Health Systems*, *4*, Article 33. https://doi.org/10.1186/1752-4458-4-33

National Academy of Sciences [US], National Academy of Engineering [US], & Institute of Medicine [US] Panel on Scientific Responsibility and the Conduct of Research. (1992). *Responsible science: Ensuring the integrity of the research process*. National Academies Press. https://doi.org/10.17226/1864

National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Board on Research Data and Information; Division on Engineering and Physical Sciences; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Analytics; Division on Earth and Life Studies; Nuclear and Radiation Studies Board, & Division of Behavioral and Social Sciences and Education. (2019). *Reproducibility and replicability in science*. National Academies Press.

Nguyen, V. T., Jung, K., & Gupta, V. (2021). Examining data visualization pitfalls in scientific publications. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1), Article 27. https://doi.org/10.1186/s42492-021-00092-y

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

Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W. S., Bürgi, E., Scherer, M., Altman, D. G., & Jüni, P. (2009). The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. *BMJ*, *339*, Article b3244. https://doi.org/10.1136/bmj.b3244

Nuijten, M. B. (2022). Assessing and improving robustness of psychological research findings in four steps. In W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding questionable research practices in applied psychology* (pp. 379–400). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_17

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Nuijten, Michele, 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*. PsyArXiv. https://doi.org/10.31234/osf.io/tcxaj

O'Boyle, E. H., Jr, Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, *43*(2), 376–399. https://doi.org/10.1177/0149206314527133

Olejnik, S. F., & Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. *Journal of Educational Statistics*, *12*(1), 45–61. https://doi.org/10.3102/10769986012001045

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

Ord, A. S., Ripley, J. S., Hook, J., & Erspamer, T. (2016). Teaching statistics in APA-accredited doctoral programs in clinical and counseling psychology. *Teaching of Psychology*, *43*(3), 221–226. https://doi.org/10.1177/0098628316649478

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), Article 6. https://doi.org/10.7275/qf69-7k43

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

Penders, B. (2018). Ten simple rules for responsible referencing. *PLOS Computational Biology*, *14*(4), Article e1006036. https://doi.org/10.1371/journal.pcbi.1006036

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424–432. https://doi.org/10.3102/0013189X13507104

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

Rabelo, A. L. A., Farias, J. E. M., Sarmet, M. M., Joaquim, T. C. R., Hoersting, R. C., Victorino, L., Modesto, J. G. N., & Pilati, R. (2020). Questionable research practices among Brazilian psychological researchers: Results from a replication study and an international comparison. *International Journal of Psychology: Journal International de Psychologie*, *55*(4), 674–683. https://doi.org/10.1002/ijop.12632

Rakow, T. (2022). Adversarial collaboration. In W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding questionable research practices in applied psychology* (pp. 359–377). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_16

Ravn, T., & Sørensen, M. P. (2021). Exploring the gray area: Similarities and differences in questionable research practices (QRPs) across main areas of research. *Science and Engineering Ethics*, *27*(4), Article 40. https://doi.org/10.1007/s11948-021-00310-z

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

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior Research Methods*, *48*(3), 1062–1069. https://doi.org/10.3758/s13428-015-0630-z

Rubin, M. (2023). Questionable metascience practices. *Journal of Trial & Error*, *4*(1). https://doi.org/10.36850/mr4

Sacco, D. F., Brown, M., & Bruton, S. V. (2019). Grounds for ambiguity: Justifiable bases for engaging in questionable research practices. *Science and Engineering Ethics*, *25*(5), 1321–1337. https://doi.org/10.1007/s11948-018-0065-x

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*(1–2), 51–63. https://doi.org/10.1016/0021-9681(79)90012-2

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2023). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*, *6*(1). https://doi.org/10.1177/25152459221128319

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(7), Article 211997. https://doi.org/10.1098/rsos.211997

Schiavone, S. R., Quinn, K. A., & Vazire, S. (2023). *A consensus-based tool for evaluating threats to the validity of empirical research*. PsyArXiv. https://doi.org/10.31234/osf.io/fc8v3

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339. doi:10.1037/met0000061

Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*(7572), 189–191. https://doi.org/10.1038/526189a

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

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). 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. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Stack Exchange. (2018). *Choosing the "correct" seed for reproducible research/results*. https://stats.stackexchange.com/questions/335936/choosing-the-correct-seed-for-reproducible-research-results

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stefan, A., & Schönbrodt, F. D. (2022). *Big little lies: A compendium and simulation of p-hacking strategies*. PsyArXiv. https://doi.org/10.31234/osf.io/xy2dk

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), Article 220346. https://doi.org/10.1098/rsos.220346

Suter, W. N. (2020). Questionable research practices: How to recognize and avoid them. *Home Health Care Management & Practice*, *32*(4), 183–190. https://doi.org/10.1177/1084822320934468

Teixeira da Silva, J. A., & Vuong, Q.-H. (2021). The right to refuse unwanted citations: Rethinking the culture of science around the citation. *Scientometrics*, *126*(6), 5355–5360. https://doi.org/10.1007/s11192-021-03960-9

Ulrich, R., & Miller, J. (2020). Questionable research practices may have little effect on replicability. *eLife*, *9*, Article e58237. https://doi.org/10.7554/eLife.58237

van den Akker, O. R., Bakker, M., van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F. M., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., & Wicherts, J. M. (2024). The

potential of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000687

van den Akker, O. R., van Assen, M. A. L. M., Bakker, M., Elsherif, M. M., Wong, T. K., & Wicherts, J. M. (2024). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods 56*(6), 5424–5433. https://doi.org/10.3758/s13428-023-02277-0

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*(11), 1473–1480. https://doi.org/10.1038/s41562-021-01211-8

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Wang, Y. A., & Eastwick, P. W. (2020). Solutions to the problems of incremental validity testing in relationship science. *Personal Relationships*, *27*(1), 156–175. https://doi.org/10.1111/pere.12309

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D., & Milic, N. M. (2019). Reveal, don't conceal: Transforming data visualization to improve transparency. *Circulation*, *140*(18), 1506–1518. https://doi.org/10.1161/circulationaha.118.037777

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214–227. https://doi.org/10.1177/2515245919848684

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*, Article 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika*, *81*(1), 27–32. https://doi.org/10.1007/s11336-015-9445-1

Wilke, C. O. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures*. O'Reilly Media.

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

Wislar, J. S., Flanagin, A., Fontanarosa, P. B., & Deangelis, C. D. (2011). Honorary and ghost authorship in high impact biomedical journals: A cross-sectional survey. *BMJ*, *343*, Article d6128. https://doi.org/10.1136/bmj.d6128

Witt, J. K. (2019). Insights into criteria for statistical significance from signal detection analysis. *Molecular Pathology*, *3*. https://doi.org/10.15626/MP.2018.871

Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (2024). Best practices for addressing missing data through multiple imputation. *Infant and Child Development*, *33*(1), Article e2407. https://doi.org/10.1002/icd.2407

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095823

Xie, J. S., & Ali, M. J. (2023). To slice or perish. *Seminars in Ophthalmology*, *38*(2), 105–107. https://doi.org/10.1080/08820538.2023.2172813

Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics*, *27*(4), Article 41. https://doi.org/10.1007/s11948-021-00314-9

Yamada, Y. (2018). How to crack pre-registration: Toward transparent and Open Science. *Frontiers in Psychology*, *9*, Article 1831. https://doi.org/10.3389/fpsyg.2018.01831

Yarkoni, T. (2020). The generalizability crisis. *The Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393