

Beyond statistical significance: Five principles for the new era of data analysis and reporting

Michel Wedel¹  | David Gal² 

¹University of Maryland, College Park, Maryland, USA

²University of Illinois Chicago, Chicago, Illinois, USA

Correspondence

Michel Wedel, University of Maryland, College Park, MD, USA.
Email: mwedel@umd.edu

Abstract

A crisis of confidence in research findings in consumer psychology and other academic disciplines has led to various proposals to abandon, replace, strengthen, or supplement the null hypothesis significance testing paradigm. The proliferation of such proposals, and their often-conflicting recommendations, can increase confusion among researchers. We aim to bring some clarity by proposing five simple principles for the new era of data analysis and reporting of research in consumer psychology. We avoid adding to researchers' confusion and proposing more onerous or rigid standards. Our goal is to offer straightforward practical principles that are easy for researchers to keep in mind while analyzing their data and reporting their findings. These principles involve (1) interpreting *p*-values as continuous measures of the strength of evidence, (2) being aware of assumptions that determine whether one can rely on *p*-values, (3) using theory to establish the applicability of findings to new settings, (4) employing multiple measures of evidence and various processes to obtain them, but assigning special privilege to none, and (5) reporting procedures and findings transparently and completely. We hope that these principles provide researchers with some guidance and help to strengthen the reliability of the conclusions derived from their data, analyses, and findings.

KEYWORD

Statistics

Since the inception of the discipline, null hypothesis significance testing (NHST) has been the dominant paradigm for data analysis in consumer psychology. About as long as NHST has been used, it has also been critiqued. In the last two decades, these criticisms have proliferated in the wake of “replication crises” in the behavioral sciences (Open Science Collaboration, 2015) and other fields of academic research, and have been accompanied by numerous proposals to strengthen, supplement, replace, or abandon the NHST paradigm entirely.

At present, best practices for data analysis and reporting in consumer psychology, and in the behavioral

sciences (and some non-behavioral sciences) in general, are in a state of flux, potentially leading to confusion among researchers. The many proposals to improve our science have ushered in a new era of data analysis and reporting of findings, but nonetheless offer varied and sometimes conflicting recommendations. These include, for example, the need to do away with NHST entirely, to avoid labeling effects as “statistically significant,” to use power calculations to determine sample sizes, to use multiple comparisons, to report all studies, to preregister procedures, to independently replicate findings, to avoid dichotomizing measurements, and to report limits

Accepted by Lauren Block, Editor; Associate Editor, Joel Huber

Commentaries on “Beyond statistical significance: Five principles for the new era of data analysis and reporting” by By Norbert Schwarz, Fritz Strack, Andrew Gelman, Stijn van Osselaer, and Joel Huber. doi: [10.1002/jcpy.1378](https://doi.org/10.1002/jcpy.1378)

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Journal of Consumer Psychology* published by Wiley Periodicals LLC on behalf of Society for Consumer Psychology.

on generalizability (e.g., Amrhein et al., 2019; Amrhein & Greenland, 2018; Anderson et al., 2017; Benjamin et al., 2018; Friese & Frankenbach, 2020; Gelman, 2016; Hirschauer et al., 2020; McShane & Gal, 2017; Simmons et al., 2021; Yarkoni, 2022). Furthermore, people have pointed to the advantages of reporting confidence intervals, of using Bayesian inference, of conducting meta-analyses and meta-science, and to sample from non-“WEIRD” populations (e.g., Henrich et al., 2010; McShane & Böckenholt, 2017; Wasserstein & Lazar, 2016; Wedel & Dong, 2020).

Various of these recommendations have been mandated by some journals that consumer psychologists typically publish in, but not in others. Even for the same journal, some reviewers and editors demand adherence to some of these proposals whereas others require adherence to a different set of standards. It may be unclear to reviewers and editors, and consumers of research, what to request of researchers and how to evaluate data analyses and reporting.

Here, we aim to bring some clarity to this chaotic situation by proposing five simple principles for data analysis and reporting. We list those principles next and provide some background and context for each of them. In doing so, we reflect on several commonly used procedures and suggest alternatives that could be considered. We aim to avoid, as much as possible, adding to researchers' confusion and proposing more onerous or rigid standards. Instead, our goal is to offer straightforward practical principles that are relatively easy for researchers to keep in mind while conducting their research. These principles, collectively, apply primarily to research in which effects or theories are tested with the aim to draw conclusions that transcend the specific context of the study. They do not necessarily apply, for example, to research that aims to measure states or traits, or make population inferences. We hope that these principles provide researchers with some guidance and help to enhance the reliability of the conclusions derived from their data, analyses, and findings.

PRINCIPLE 1: DESIGNATING EFFECTS AS “STATISTICALLY SIGNIFICANT” BASED ON A CUTOFF ON MEASURES OF EVIDENCE IS ILL-ADVISED; USE JUDGMENT TO INTERPRET CONTINUOUS MEASURES OF THE STRENGTH OF EVIDENCE FOR AN EFFECT

Current applications of the null-hypothesis significance testing (NHST) paradigm are an unfortunate amalgam of the procedures originally proposed by Fisher, Neyman, and Pearson (Gill, 1999), which has led to misleading interpretations and misuse of p -values (Nuzzo, 2014;

Wasserstein & Lazar, 2016). Greenland et al. (2016) document these misinterpretations extensively. At their core is that researchers have taken a p -value larger than 0.05 to imply that the null hypothesis is true, by which the absence of an effect is demonstrated. Conversely, $p < 0.05$ is taken to imply that an effect exists and that the probability of replicating the effect is larger than 0.95 (Gigerenzer, 2018). More than half of almost 800 articles across five journals were found to make these mistakes (Amrhein et al., 2019). McShane and Gal (2016, 2017) further document that this dichotomization of evidence is not only used as a shorthand summary for ease of reporting but in fact unduly influences the interpretation of the evidence.

Some of these common misinterpretations are caused by the hard cutoff value for “statistical significance,” which was initially proposed for mere convenience (Fisher, 1956; Pearson, 1935). While statisticians widely agree that “The widespread use of ‘statistical significance’ (generally interpreted as ‘ $p < 0.05$ ’) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process” (Wasserstein & Lazar, 2016), the practice persists widely. Some have argued for the p -value to be abolished entirely, whereas others have argued for a smaller cutoff value for “statistical significance,” such as $p < 0.005$ (Benjamin et al., 2018).

However, lowering the threshold does nothing to remove the misinterpretations of the p -value and may increase the file drawer problem whereby “insignificant” results are not reported or published. Many statisticians therefore believe that the best solution is to abandon any cutoff on the p -value to label an effect as “statistically significant” (Amrhein et al., 2019; McShane et al., 2019). Ideally, the term “statistically significant” (and others like it, including “marginally significant”) should be abolished, along with the “*-system” to indicate levels of significance. Similarly, the use of cutoffs on other statistics, including posterior probabilities, Bayes Factors, and effect sizes, to designate effects as practically important is misguided and in an ideal world should be avoided (McShane & Gelman, 2022). Instead, p -values and other statistical measures should be reported as continuous quantities, and interpreted as measuring the amount of evidence against the model that embeds the null hypothesis.

We caveat these ideals with the recognition of the need to be realistic. The reality is that despite many criticisms, having served as the main basis for evaluating evidence in the literature over many decades and as an essential part of scientists' training, the terminology of the NHST paradigm is entrenched. The use of the term “statistically significant” is deeply engrained in how we think about evidence, which stems from several facts, most of which should be apparent to consumer psychologists. First, its use is widespread and longstanding, and change is hard so that inertia often prevails (Gal, 2006). Second,

without a rule (such as $p < 0.05$ implies “statistical significance”) to qualify what counts as evidence in support of a hypothesis and what does not, uncertainty reigns, which people generally find uncomfortable. Third, while various recommendations have been made for reporting data without resorting to terms such as “statistically significant” or “no difference,” these recommendations often include unwieldy and lengthy qualitative descriptions that may confuse as much as they help. Finally, without that terminology, and its underlying rule, different researchers may adopt different standards of what qualifies as evidence for an effect, rendering the publication process more ambiguous and its outcome more difficult to predict. Completely dismantling the use of such NHST terms, at least in the short term, is therefore likely not realistic. Nonetheless because the rule $p < 0.05$, even if widely agreed upon, leads to misrepresentation of scientific evidence, it should be abandoned. At the same time, the terminology, “statistically significant,” may still be useful to loosely convey the statistical support for a theory, but it should be commonly understood that its interpretation cannot be linked to a specific cutoff and will depend on the research context.

Researchers must thus report and interpret continuous measures of evidence, including p -values. At the same time, it is critical that researchers, reviewers, and consumers of research should interpret terms such as “statistically significant,” and “no difference” (when used to describe “not statistically significant” results) as categorical shorthand terms that only very roughly and incompletely characterize the evidence. These terms should not be uniquely used in reference to a single rigid cutoff on continuous measures of evidence, such as $p < 0.05$. What effect can be labeled as “significant” will depend on the goal of the study, the sample size, plausible effect sizes, and the procedures used in collecting and analyzing data. An important implication is that lack of “statistical significance” cannot be used to disqualify any study. Ultimately, researchers' judgment, rather than any hard statistical decision rule, is required to assess the evidence for the hypothesis at hand.

We believe there is reason for optimism. Our perception is that while until recently researchers, reviewers, and consumers of research interpreted the cutoff for statistical significance rather strictly as a dichotomous indicator of whether a finding was true or false (i.e., supported a hypothesis or not), this is mostly no longer the case. Our experience is that researchers are more appreciative today than even a few years ago that the strength of evidence for a finding is a matter of degree that is best captured by multiple statistical measures, including p -values and effect sizes, and by the whole body of evidence presented by all the studies in a paper combined.

In sum, rather than relying on a single cutoff on the p -value, researchers should present coherent, verifiable, and principled arguments that the evidence from a sequence of studies, as quantified by one or more statistical

measures, increases the confidence in their theoretical propositions.

PRINCIPLE 2: P -VALUES ARE SENSITIVE TO DEPARTURES OF MODEL ASSUMPTIONS; CHECK MODELING ASSUMPTIONS AND RELY ON CLASSICAL P -VALUES ESPECIALLY WHEN SAMPLE AND EFFECT SIZES ARE LARGE

Most statistical procedures are based on multiple assumptions. As such, no single measure derived from a model, including the p -value, is the be-all and end-all of the measurement of evidence for hypotheses, as several underlying assumptions may be violated for any particular data set.

To elaborate, statistical testing procedures assume that they encapsulate a valid representation of the mechanisms by which the data are generated. For example, for methods such as linear regression, ANOVA, and ANCOVA these assumptions are that the data are a random sample from a population, that the error terms follow a normal distribution, that data do not exhibit outliers, that treatment effects are additive, that variances in the treatment groups are equal, that observations are independent, and that all participants in an experiment respond identically to the experimental treatments (see e.g., Cochran, 1947; Eisenhart, 1947). Violations of these assumptions are not innocuous, because p -values, which are calculated to measure the extent to which the data are extreme under the null hypothesis, also measure the extent to which the data depart from these assumptions (Greenland et al., 2016). Ultimately, “all models are wrong, but some are useful” (Box, 1979). It therefore behooves researchers to select statistical tools that, as closely as possible, match the data generating process.

The Bayesian framework has been touted to offer such flexibility to match a statistical model to the characteristics of the data at hand (e.g., Wagenmakers et al., 2018), to accommodate treatment heterogeneity, and to coherently follow-up with floodlight and mediation analyses (e.g., Wedel & Dong, 2021). It may be unrealistic to expect that the entire field of consumer psychology will shift toward the use of such new methods (i.e., new to the researchers) of data analysis. Nonetheless, that it is possible is illustrated by the fact that the field has previously widely adopted the bootstrap, a randomization method for testing indirect effects in mediation analyses, in recognition that these effects do not follow a normal distribution (Preacher & Hayes, 2008). We thus hope researchers keep an open mind toward the application of statistical methods that better serve the purposes of their research and better match the measurement properties of the data they collected. Critical for the dissemination of alternative methods is the availability of easy-to-use

software. Probabilistic programming languages such as Stan (Carpenter et al., 2016) have greatly facilitated the use of Bayesian modeling. Nonetheless, programming in those languages may still be a challenge or burden to researchers in consumer psychology. R packages such as brms (Bürkner, 2017) and BANOVA (Wedel & Dong, 2020, 2021) provide flexible and easy-to-use interfaces to the Stan programming language that greatly reduce those challenges.

The good news for researchers adhering to the frequentist and NHST approaches is that for large samples, large effect sizes, and/or valid distribution assumptions, frequentist, randomization, and Bayesian procedures yield very similar results (Aczel et al., 2017; Albers et al., 2018; Huber & Train, 2001). While no statistical model is perfect, in many cases the results of the frequentist analyses typically performed by consumer psychologists will thus be “good enough” for providing internally valid inferences on the strength of evidence for a finding. By “good enough,” we mean that conclusions would not materially change if a technically more appropriate analysis was performed. Of note, by relinquishing a threshold for statistical significance (as we advocate in principle 1), conclusions will not change qualitatively depending on whether a p -value is just above or just below a particular cutoff, thus reducing the sensitivity of the conclusions to the specific statistical method used for the analysis.

In sum, attention to assumptions underlying statistical tools is warranted, but under the right conditions standard statistical methods are robust. The method that is ultimately applied should be as complex as needed to address the problem at hand, but not more complex than that.

PRINCIPLE 3: FORMAL POPULATION INFERENCE IS RARELY RELEVANT TO STUDIES IN CONSUMER PSYCHOLOGY; ESTABLISH APPLICABILITY OF FINDINGS TO NEW SETTINGS PRIMARILY VIA THEORY AND SOUND RHETORICAL ARGUMENTS

p -Values (and other inferential statistics) require a random sample from a well-defined population and assume that it is theoretically and practically possible to replicate that sample. The reason is that p -values are based on the distribution of a test statistic in repeated hypothetical samples from the same population under the counterfactual assumption that the null hypothesis holds. Therefore, it is inherent to the NHST paradigm that the results of an experiment are used to make inferences on the population from which the sample was drawn.

However, most experiments in consumer psychology do not aim at making inferences about a population.

Consumer psychologists are instead often interested in the applicability of findings to new settings involving new populations, new stimuli, and new situations. For this purpose, formal population inference is not relevant, and consumer psychologists accordingly tend to use convenience samples instead of random samples from a clearly delineated population. Even for field experiments, in which participants are customers of a collaborating organization, it may be difficult to defend the assumption that the data arose from a random sample from a well-defined population. In addition, human cognition is sensitive to the context and to socio-cultural factors (Schwarz & Clore, 2016), which are often not sufficiently stable to allow exact replication of a sample from the same population at another point in time. The NHST statistical framework to compute p -values therefore does not technically apply in most studies in consumer psychology, because the hypothetical counterfactual is infeasible (Hirschauer et al., 2020).

Alternative statistical frameworks do not rely on the assumption of random population sampling. Randomization tests (cf., Ernst, 2004) only assume that participants are randomly assigned to the treatment conditions. Hypotheses are tested by randomly reassigning data to the treatment conditions, which enables the calculation of the distribution of a test statistic under the null hypothesis. In the Bayesian approach (cf., Edwards et al., 1963) inferences are conditional upon the specific data set at hand. The researcher's uncertainty about a parameter is characterized via a (prior) probability distribution that is updated after the data are observed. This results in a posterior probability distribution that captures the uncertainty about a hypothesis given the data that were collected. The Bayesian framework thereby enables one to make probability statements about hypotheses, even null hypotheses. Neither the randomization approach nor the Bayesian approach assumes random sampling from a population. These frameworks are therefore formally better suited to the analysis of theory-testing experiments in consumer psychology. In applying these frameworks to convenience samples, however, one sacrifices the ability to make formal inferences on a population, and one needs to rely on qualitative arguments to generalize findings (Ludbrook & Dudley, 1998).

If statistical extrapolation to a population is desired, frequentist, randomization and Bayesian approaches all require a random sample from that population. Statistical generalization of results from a sample to a population, and to other samples obtained from it, is possible only when the population boundaries have been carefully established, a probability sample has been obtained, and the population is stable (e.g., Kukull & Ganguli, 2012; Mullinix et al., 2015). But one is rarely if ever interested, even in principle, in statistically extrapolating findings from consumer psychology experiments exclusively to the exact population from which the sample should have been obtained. Even in a field experiment where the

experimental manipulation is applied to the entire population of customers from the participating organization, statistical inference has little to say about applicability to other organizations, customers, and situations.

The reality for research in consumer psychology is that one almost always must rely on theoretical reasoning to extrapolate findings to a different sample, population, or context, and not on statistical extrapolation of the findings of the experiment to those different situations, regardless of the statistical paradigm one works in. In applying the findings of studies to different contexts and populations, consumer psychologists must rely on theory, logic, judgment, and arguments based on the evidence, as “in a court of law” (cf. Calder et al., 1982; Gilboa et al., 2014), to determine whether and how the sample selection process, the presence of treatment heterogeneity, and moderating variables could have influenced the results. Establishing the causal mechanism of an effect, its heterogeneity among participants, and its moderators can help build stronger theory, and theory, in turn, allows one to predict if and how effects will occur in new settings (Gal et al., 2023). However, theories likely will not comprehensively encapsulate heterogeneity of treatment effects (cf., Levitt & List, 2007), and standard mediation analysis in between-subject experiments, being essentially correlational, does not unequivocally support causality of the underlying mechanism (MacKinnon et al., 2007; Selig & Preacher, 2009). These problems may hamper the development of theory, and thereby the basis for the applicability of results to new contexts.

The previous arguments point to some of the limitations of replication studies as well (cf., Gigerenzer, 2018). Even replication studies that are intended to be exact are often conducted on a sample from a different population and at a different time. Furthermore, considering the correct NHST interpretation of an “insignificant” finding as a mere failure to reject the null hypothesis and not a confirmation of it, reveals that failure to replicate findings cannot prove that the original finding is invalid (Maxwell et al., 2015). The only way to statistically assess whether a particular effect also occurs in a new setting is to empirically replicate the study in that setting. But the idea that we can establish whether the results of an experiment generalize broadly, never mind *universally*, by replicating the findings in another population or setting is false. Finding an effect in two populations (as opposed to one) does not mean it will apply to a third population, or to those same populations at a different time or under different conditions. Discussions of “limits to generalizability” are therefore often ill-conceived because they presume that we have a comprehensive theory of treatment heterogeneity and that we can precisely circumscribe the boundaries of an effect or theory (Gal et al., 2023).

We advocate that rather than fruitlessly attempting to circumscribe the bounds of effects, we should build and

test broad theories, and rely on them and on sound judgment to infer the applicability of findings to alternate contexts that support further theorizing and are relevant for practice.

PRINCIPLE 4: MORE STATISTICAL MEASURES OF EVIDENCE AND MORE DATA ARE BETTER THAN FEWER; DO NOT GET HUNG UP ON SPECIFIC MEASURES, POWER CALCULATIONS, OR META-ANALYSIS

Sample size matters: the lack of power resulting from small samples (Cohen, 1962; Marszalek et al., 2012; Maxwell, 2004) negatively affects the reliability of study results (Greenland, 2012). Conversely, the very large sample sizes one now often sees in field studies tend to make statistical testing of null hypotheses vacuous, as they result in very small *p*-values for minute and substantively irrelevant deviations of the null hypothesis. Both are reasons to use multiple measures of evidence (Nuzzo, 2014; Wasserstein & Lazar, 2016). Along with point estimates, effect sizes, *p*-values, confidence (or credible) intervals (CIs), posterior probabilities, likelihood ratios, and Bayes factors all provide summary measures of the evidence for a hypothesized effect. Note, however, that default 95% CIs are also based on the arbitrary 0.05 cutoff for the *p*-value and therefore suffer from the same drawbacks as when using that cutoff to label effects as “statistically significant.” Sometimes it is therefore insightful to inspect and/or report multiple CIs.

One-sided tests could be considered to test the directional hypotheses frequently formulated in consumer psychology (Cho & Abe, 2013). Theories that predict that an effect is not equal to zero but are ignorant about its direction are often substantively implausible. However, in exploratory studies and in cases where effects in either direction are expected, two-sided tests are appropriate (van Osselaer & Janiszewski, 2021). In exploratory studies that employ two-sided tests, an effect in an unexpected direction can often be explained by including an omitted variable or a moderator. Several one-sided tests (e.g., *t*-tests, *Z*-tests) simply involve lowering the cutoff for “statistical significance,” because of which they have unfortunately been associated with *p*-hacking (Pillemer, 1991). However, as the sample size becomes larger (e.g., in large field experiments and other high-powered studies) the hypothesis of a zero effect is ever more likely to be rejected by a two-sided test, but a one-sided test establishes the direction of the effect with increasing certainty (Marsman & Wagenmakers, 2017). Because even for one-sided tests the power to detect small, possibly substantively irrelevant, effects increases as the sample size gets larger, equivalence tests (Lakens, 2017) may be useful. These tests are used to examine effects

that are large enough to be deemed interesting (e.g., the two one-sided test). One-sided and equivalence tests have uniformly better power than standard tests of the null hypothesis of a zero effect (Schuirmann, 1987).

Assessing power accurately is generally tricky given that, as discussed previously, each study is ultimately *sui generis* and often not derived from a random sample (an assumption that also underlies power calculations). Moreover, in *a-priori* power calculations the true effect size is unknown and needs to be guessed. Using the empirical estimate of the effect size from a prior study for power calculations leads to underpowered studies because effect sizes are subject to sampling variation (Gelman, 2019; McShane et al., 2020; Yuan & Maxwell, 2005). Investigating the power for a range of plausible effect sizes that reflect meaningful variation in the effect size can be useful (Dallow & Fina, 2011; Maxwell et al., 2015; Yuan & Maxwell, 2005). Thus, while power calculations may provide useful insights, in most cases they do not produce a definitive conclusion on the sample size. Subjective judgments and budget or other practical constraints can provide compelling motivations for selecting a particular sample size.

Meta-analysis is often advocated as an important tool for measuring the strength of evidence, across published research, and across experiments in a single study (McShane & Böckenholt, 2017). Under the right conditions, a meta-analysis provides more power than a single study and a more precise summary of the evidence than a narrative review. Clearly, reliance on any one study may lead to spurious findings or erroneous conclusions due to sample size, noise or other particularities associated with the specific study (Ledgerwood & Sherman, 2012). But, because overall power decreases as a function of the number of statistical tests, conducting multiple studies increases the probability of “nonsignificant” findings (Schimmack, 2012). Therefore, the (unrealistic) expectation that the *p*-value for each experiment in a sequence indicates “significance” increases the file drawer problem, whereby studies without “significant” results are not reported (Sterling, 1959).

A common recommendation is thus that researchers report all studies. While ideally this prescription should be followed, it is often neither realistic nor practical. Some studies are exploratory or, in retrospect, might have poorly operationalized constructs, ineffective manipulations, other design flaws, or turn out to be of questionable relevance to the research at hand. Notably, if a key construct is not successfully operationalized, the study does not provide a valid test of the theory in question. It is then unrealistic and ill-advised to expect researchers to report such studies or reviewers to evaluate them (for a discussion, see Schwarz & Clore, 2016). Although selective

reporting of studies does not necessarily severely bias the research record (Simonsohn, 2016), reporting all pertinent studies with valid results is essential when estimating an effect using meta-analysis. On the one hand, the file drawer problem (and other types of selective reporting) may induce bias in meta-analyses, but on the other hand, mechanical inclusion of all studies without regard to their validity may introduce bias as well (Simonsohn et al., 2022). Furthermore, researchers often surmise that meta-analysis, through its synthesis of complete information, will yield “significant” results and conclusive evidence on an effect or effect size. However, no statistical method eliminates uncertainty (Gelman, 2016), and no number of studies can yield definitive conclusions or unambiguous support for a theoretical proposal (Amrhein et al., 2019). Analyzing quality and validity of data and methods, sources of bias, and effect heterogeneity is essential to the use of meta-analysis. In addition, a single estimated summary effect produced by a meta-analysis ignores treatment heterogeneity among studies with different samples and designs (Finckh & Tramèr, 2008). All else equal, larger sample sizes and more statistical measures of evidence are better than less, and meta-analysis can be helpful, but it has no undisputed status in establishing evidence. It performs poorly especially when there is selective reporting of studies, and when the number of studies, effect sizes, and sample sizes are small, and effect heterogeneity is large (Stanley, 2017).

Examining the effect of an intervention on a specific outcome variable is less often the purpose of consumer psychology experiments. Consumer psychologists are primarily interested in explaining theoretical relationships among constructs rather than in examining the effect of one specific, experimentally manipulated, variable on another. In addition, effects of those interventions are less stable in psychology than the sciences due to greater inherent variability, treatment heterogeneity, and context sensitivity (Fanelli & Glänzel, 2013). For example, as Gal and Rucker (2022) point out, “whether a deficit in savings is aided by a nudge will depend heavily on how the nudge is construed by the individual (e.g., if it is seen as manipulative) at a particular time and in a particular situation.” Although several of the principles mentioned here apply to intervention studies as well, in this context meta-analysis is of more limited relevance in consumer psychology than in other fields that focus on establishing the effects of relatively more stable interventions.

In sum, while larger samples, more studies, and more measures of evidence are, *ceteris paribus*, preferred to less, no specific measures of evidence (even *p*-values) or processes to obtain the evidence ought to be privileged or mandated.

PRINCIPLE 5: WHETHER QUESTIONABLE RESEARCH PRACTICES ARE QUESTIONABLE DEPENDS ON THE STATISTICAL METHOD USED; REPORTING PROCEDURES AND FINDINGS IN A TRANSPARENT AND COMPLETE MANNER SHOULD TAKE PRECEDENCE

The key recommendation advocated here, reiterating principles of the “Open Science” (OS) movement (see e.g., Nosek et al., 2015; Vicente-Saez & Martinez-Fuentes, 2018), is that researchers report as honestly, transparently, and as completely as possible their findings and the methods by which they obtained them. The selection and application of data collection, processing, and analysis methods almost always involves subjective judgments with which other researchers may disagree (Gelman & Loken, 2014). To the extent that it is practical, research reports should therefore describe in detail the full sequence of steps that have led to the findings and conclusions of a study, including the population, sample size, study design, stimuli, measured variables, raw and processed data, exclusion criteria, variable selection, and the sequence of statistical analyses that were conducted. Making data and computer code available, as is required, or encouraged by several journals, is an important aspect of OS. Unfortunately, in our field, confidentiality (for laboratory studies) and NDA (for field studies) are reasons often cited as barriers.

The Open Science Framework (OSF) provides the infrastructure that enables one to disclose the entire research process, from the research idea to the final reporting (Foster & Deardorff, 2017). Some have even argued that journals' complete review processes should be disclosed as well (Wicherts, 2016), which is sensible because reviewers' opinions and assumptions are frequently reflected in the final analyses and conclusions, are an essential part of the research dialogue, and should be open to critical peer evaluation. Careful documentation opens to rigorous scrutiny the entire process by which researchers arrive at the findings and recommendations, and the motivations for each step taken. It thereby allows other researchers to see for themselves whether the judgments made in collecting, analyzing, and reporting the data are defensible, whether procedures were used that could be labeled as “questionable,” to carry out their own alternate analyses to establish the robustness of the results, and to judge for themselves the strength of evidence for the theoretical proposal.

The sequence of decisions made in data collection and analysis require judgment and flexibility, that is, “researcher degrees of freedom,” which absent strict adherence to preregistration will be data dependent (Gelman & Loken, 2014; Greenland et al., 2016). Flexibility in

the selection of statistical methods enables researchers to use different methods depending on features of the data that reveal themselves post-hoc, on ad-hoc decisions made to produce a data set that can be analyzed, and on insights obtained in exploratory analyses. This holds even more for exploratory research (van Osselaer & Janiszewski, 2021). Flexibility in statistical modeling is useful and needed, even when it affects “statistical significance” and the interpretation of the p -value.

Questionable research practices (QRPs) are data collection and analysis procedures that violate one or more of the assumptions of the testing procedures that are used. Some of these include selective reporting of studies, selecting which dependent measures to report, collecting more data after determining whether the results were significant, reporting unexpected findings as having been predicted, formulating hypotheses after data analysis, carrying out multiple analyses with different covariates as ancillary variables, and excluding data on individuals, groups or factor levels post-hoc (John et al., 2012; Simmons et al., 2011). The use of QRPs has increased false positive findings (Bakker & Wicherts, 2011; John et al., 2012), but their prevalence has likely been overestimated (Fiedler & Schwarz, 2016). Moreover, QRPs are questionable only if they violate assumptions of the testing procedures and thereby affect the interpretation of the p -value. This holds in particular if a fixed cutoff on the p -value is used to designate effects as “significant,” because then QRPs may increase false positives. In most cases, the relatively simple statistical methods that are applied in consumer psychology (t -tests, ANOVA, regression) ignore these practices. If statistical procedures are used that properly reflect them (such as sequential testing, multiple comparisons, multivariate testing, and regression with outliers, missing values, or truncated variables), QRPs are no longer questionable. Following the ideals of transparent and complete reporting allows for beneficial flexibility in data analysis and reporting, while preserving a record that allows other researchers to identify whether data handling procedures are properly reflected in the analyses, and to conduct alternative analyses.

The principle of reporting data analysis as transparently and completely as is practical also has implications for the practice of preregistration. This topic was recently featured in a Research Dialogue in the *Journal of Consumer Psychology* which debated whether preregistration could mitigate the selective reporting of study results based on their statistical significance (e.g., Krishna, 2021). Preregistration intends to take away researcher degrees of freedom that enable p -hacking (Simmons et al., 2011). In our view, however, researcher degrees of freedom are often *desirable* because they enable one to match the properties of statistical methods more accurately to the measurement properties of the data and to the procedures used to produce a data set that can be analyzed. Preregistration is ultimately a

rearguard measure against an undesirable consequence of the misuse of the NHST paradigm (i.e., p-hacking). Its main benefit may be to give other researchers confidence that the results were not selectively reported to support a hypothesis, and thus that the reported results can be relied upon. At the same time, there are hazards to considering it a panacea or mandating it. For example, it may only become clear post-hoc that certain assumptions of the intended statistical procedures are violated. Adapting these procedures could be at odds with preregistration. We do not contest the usefulness of preregistration, as it can surely have benefits in specific situations, such as in replication studies. But following the OS ideals should take precedence. OS provides researchers with more flexibility in analyzing and reporting data, since any deviations from analysis and reporting standards, or preregistered procedures, can be recognized, and supplemented or corrected, if deemed necessary, by other researchers.

In sum, we hope that OS practices will become more widely accepted in consumer psychology.

CONCLUSION

Our five principles for the new era of data analysis and reporting are meant to help improve the quality of inferences and the reliability of findings in consumer psychology. They are not intended to provide a straitjacket for researchers to prevent p-hacking and other undesirable practices that detract from replicability of findings. Rather we hope they enhance flexibility in the application of research methods while at the same time making it clear where researchers' judgment has entered data analysis procedures, interpretations of findings, and conclusions drawn from them. Our recommendations are also not intended to eliminate or reduce fraud. No statistical measure or procedure is impervious to fraud, including *p*-values (Simmons et al., 2011), posterior distributions, Bayes factors (Simonsohn, 2014), and even preregistration (Yamada, 2018). Fraud, erroneous, or uninformed application of statistical methods can be detected, and those efforts have become a valuable component of the scientific discourse (e.g., Ioannidis & Trikalinos, 2007; Schimmack, 2012; Simonsohn et al., 2014). Our aim, instead, is to provide a few simple principles that help researchers test their hypotheses in a way that enhances reliability and credibility of their research findings, by making the judgments needed to arrive at those findings transparent, while recognizing that no amount of data or statistical method can eliminate uncertainty or provide unequivocal support for our theories.

ORCID

Michel Wedel  <https://orcid.org/0000-0002-1244-4923>

David Gal  <https://orcid.org/0000-0002-3446-9304>

REFERENCES

- Aczel, B., Palfi, B., & Szasz, B. (2017). Estimating the evidential value of significant results in psychological science. *PLoS One*, 12(8), e0182651.
- Albers, C. J., Kiers, H. A., & van Ravenzwaaij, D. (2018). Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, 4(1), 31.
- Amrhein, V., & Greenland, S. (2018). Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, 2(1), 4.
- Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, 567, 305–307.
- Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 28(11), 1547–1562.
- Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, 43, 666–678.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., de Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10.
- Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.
- Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.
- Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research*, 9(3), 240–244.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 1.
- Cho, H. C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, 66(9), 1261–1266.
- Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 3(1), 22–38.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145.
- Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, 10, 311–317.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*, 3(1), 1–21.
- Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, 1(1), 676–685.
- Fanelli, D., & Glänzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. *PLoS One*, 8(6), e66938.
- Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1), 45–52.
- Finckh, A., & Tramèr, M. R. (2008). Primer: Strengths and weaknesses of meta-analysis. *Nature Clinical Practice Rheumatology*, 4(3), 146–152.
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Hafner Publishing Co.
- Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association*, 105(2), 203.
- Friese, M., & Frankenbach, J. (2020). p-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, 25(4), 456.
- Gal, D. (2006). A psychological law of inertia and the illusion of loss aversion. *Judgment and Decision Making*, 1(1), 23–32.

- Gal, D., & Rucker, D. D. (2022). Experimental validation bias limits the scope and ambition of applied behavioural science. *Nature Reviews Psychology*, 1(1), 5–6.
- Gal, D., Sternthal, B., & Calder, B. J. (2023). Confidence in research findings depends on theory. *Behavioral and Brain Sciences*.
- Gelman, A. (2016). The problems with p -values are not just with p -values. *The American Statistician*, 70(10).
- Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, 269(1), e9–e10.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—A “garden of forking paths”—Explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6), 460.
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2), 198–218.
- Gilboa, I., Postlewaite, A., Samuelson, L., & Schmeidler, D. (2014). Economic models as analogies. *The Economic Journal*, 124(578), F513–F533.
- Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52(3), 647–674.
- Greenland, S. (2012). Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology*, 22(5), 364–368.
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p -values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83.
- Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C., & Jantsch, A. (2020). Can p -values be meaningfully interpreted without random sampling? *Statistics Surveys*, 14, 71–91.
- Huber, J., & Train, K. (2001). On the similarity of classical and Bayesian estimates of individual mean partworths. *Marketing Letters*, 12, 259–269.
- Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4, 245–253.
- 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.
- Krishna, A. (2021). The need for synergy in academic policies: An introduction to the dialogue on pre-registration. *Journal of Consumer Psychology*, 31(1), 146–150.
- Kukull, W. A., & Ganguli, M. (2012). Generalizability: The trees, the forest, and the low-hanging fruit. *Neurology*, 78(23), 1886–1891.
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.
- Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, 7, 60–66.
- Levitt, S. D., & List, J. A. (2007). On the generalizability of lab behaviour to the field. *Canadian Journal of Economics*, 40(2), 347–370.
- Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, 52(2), 127–132.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593.
- Marsman, M., & Wagenmakers, E. J. (2017). Three insights from a Bayesian interpretation of the one-sided p value. *Educational and Psychological Measurement*, 77(3), 529–539.
- Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2012). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills*, 112, 331–348.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9, 147–163.
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70(6), 487–498.
- McShane, B. B., & Böckenholt, U. (2017). Single paper meta-analysis: Benefits for study summary, theory-testing, and replicability. *Journal of Consumer Research*, 43(6), 1048–1063.
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2020). Average power: A cautionary note. *Advances in Methods and Practices in Psychological Science*, 3(2), 185–199.
- McShane, B. B., & Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, 62(6), 1707–1718.
- McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519), 885–895.
- McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(supl), 235–245.
- McShane, B. B., & Gelman, A. (2022). Selecting on statistical significance and practical importance is wrong. *Journal of Information Technology*, 2022, 02683962221086297.
- Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2), 109–138.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., & Contestabile, M. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425.
- Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506, 150–152.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Pearson, K. (1935). Letter to the editor: Statistical tests. *Nature*, 136, 550.
- Pillemer, D. B. (1991). One-versus two-tailed hypothesis tests in contemporary educational research. *Educational Researcher*, 20(9), 13–17.
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879–891.
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17(4), 551–566.
- Schuurmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680.
- Schwarz, N., & Clore, G. L. (2016). Evaluating psychological research requires more than attention to the N : A comment on Simonsohn's (2015) “small telescopes”. *Psychological Science*, 27(10), 1407–1409.
- Selig, J. P., & Preacher, K. J. (2009). Mediation models for longitudinal data in developmental research. *Research in Human Development*, 6, 144–164.
- 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.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration: Why and how. *Journal of Consumer Psychology*, 31(1), 151–162.
- Simonsohn, U. (2016). *The file-drawer problem is unfixable, and that's OK*. Datacolada.
- Simonsohn, U. (2014). Posterior-hacking: Selective reporting invalidates Bayesian results also. Available at SSRN 2374040.
- 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.

- Simonsohn, U., Simmons, J., & Nelson, L. D. (2022). Above averaging in literature reviews. *Nature Reviews Psychology*, 1(10), 551–552.
- Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8(5), 581–591.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance— Or vice versa. *Journal of the American Statistical Association*, 54(285), 30–34.
- van Osselaer, S. M., & Janiszewski, C. (2021). A recipe for honest consumer research. Available at SSRN 3786989.
- Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88, 428–436.
- Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25, 35–57.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133.
- Wedel, M., & Dong, C. (2020). BANOVA: Bayesian analysis of experiments in consumer psychology. *Journal of Consumer Psychology*, 30(1), 3–23.
- Wedel, M., & Dong, C. (2021). A tutorial on the analysis of experiments using BANOVA. *Psychological Methods*.
- Wicherts, J. M. (2016). Peer review quality and transparency of the peer-review process in open access and subscription journals. *PLoS One*, 11(1), e0147913.
- Yamada, Y. (2018). How to crack pre-registration: Toward transparent and open science. *Frontiers in Psychology*, 9, 1831.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45(e1), 1–78. <https://doi.org/10.1017/S0140525X20001685>
- Yuan, K.-H., & Maxwell, S. E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30(2), 141–167.

How to cite this article: Wedel, M., & Gal, D. (2024). Beyond statistical significance: Five principles for the new era of data analysis and reporting. *Journal of Consumer Psychology*, 34, 177–186. <https://doi.org/10.1002/jcpsy.1379>