The case for formal methodology in scientific reform

Berna Devezer[1], **Danielle J. Navarro**[2], **Joachim Vandekerckhove**[3], **Erkan Ozge Buzbas**[4]

***For correspondence:**
bdevezer@uidaho.edu (BD)

[1]Department of Business, University of Idaho; [2]School of Psychology, University of New South Wales; [3]Department of Cognitive Sciences and Department of Statistics, University of California, Irvine; [4]Department of Statistical Science, University of Idaho

Abstract Current attempts at methodological reform in sciences come in response to an overall lack of rigor in methodological and scientific practices in experimental sciences. However, some of these reform attempts suffer from the same mistakes and over-generalizations they purport to address. Considering the costs of allowing false claims to become canonized, we argue for more rigor and nuance in methodological reform. By way of example, we present a formal analysis of three common claims in the metascientific literature: (a) that reproducibility is the cornerstone of science; (b) that data must not be used twice in any analysis; and (c) that exploratory projects are characterized by poor statistical practice. We show that none of these three claims are correct in general and we explore when they do and do not hold.

Introduction

Widespread concerns about unsound research practices, lack of transparency in science, and low reproducibility of empirical claims have led to calls for methodological reform across scientific disciplines (***Begley and Ioannidis, 2015***; ***Donoho et al., 2008***; ***Ioannidis et al., 2009***; ***Open Science Collaboration, 2015***). The literature on this topic has been termed “meta-research” (***Ioannidis, 2018***) or “meta-science” (***Schooler, 2014***), and somewhat surprisingly this field has received little scrutiny itself. Policies are proposed without evidentiary backing and methods are suggested with no framework for assessing their validity or evaluating their efficacy (e.g., see policy and methods proposals in ***Chambers et al., 2015***; ***Hardwicke et al., 2018***; ***Lakens et al., 2018***; ***Munafò et al., 2017***; ***Nosek et al., 2012***; ***Wagenmakers et al., 2012***). This is a reason for concern: methodological reforms should be held to standards that are *at least* as rigorous as those we expect of empirical scientists. Should we fail to do so, we run the risk of repeating the mistakes of the past and creating new scientific processes that are no better than those they replace.

Methodologists have criticized empirical scientists for: (a) prematurely presenting unverified research results as facts (***McShane et al., 2019***); (b) overgeneralizing results to populations beyond the studied population (***Henrich et al., 2010***); (c) misusing or abusing statistics (***Gelman and Loken, 2013***; ***Simmons et al., 2011***); and (d) lack of rigor in the research endeavor that is exacerbated by incentives to publish fast, early, and often (***Munafò et al., 2017***; ***Nosek et al., 2012***). Regrettably, the methodological reform literature is affected by similar practices: prematurely claiming that untested methodological innovations will solve replicability/reproducibility problems; presenting conditionally true statements about methodological tools as unconditional, bold facts about scientific practice; presenting vague or misleading statistical statements as evidence for the validity of reforms; and an overall lack of rigor in method development that is exacerbated by incentives to find immediate solutions to the replication crisis. There is an uncomfortable symmetry to this, but also an opportunity: reformers are in an opportune position to take criticism and self-correct before allowing false claims to be canonized as methodological facts (***Nissen et al., 2016***).

In this paper we advocate for the necessity of statistically rigorous and scientifically nuanced arguments to make proper methodological claims in the reform literature. Toward this aim, we evaluate three examples of methodological claims that have been advanced and well-accepted (as implied by the large number of citations) in the reform literature:

1. Reproducibility is the cornerstone of, or a demarcation criterion for, science.
2. Using data more than once invalidates statistical inference.
3. Exploratory research uses "wonky" statistics.

Each of these claims suffers from some of the problems outlined earlier and as a result, has contributed to methodological half-truths (or untruths). We evaluate each claim using statistical theory against a broad philosophical and scientific background.

While we focus on these three claims, we believe our call for rigor and nuance can reach further with the following emphasis: Statistics is a *formal* science whose methodological claims follow from probability calculus. Methodological claims are either proved mathematically or by simulation before being advanced for the use of scientists. Most valid methodological advances are incremental, and they rarely ever provide simple prescriptions to complex inference problems. Norms issued on the basis of bold claims about new methods might be quickly adopted by empirical scientists as heuristics and might alter scientific practices. However, advancing such reforms in the absence of formal proofs is sacrificing rigor for boldness and can lead to unforeseeable scientific consequences. We believe that hasty revolution may hold science back more than it helps move it forward. We hope that our approach may facilitate scientific progress that stands on firm ground—supported by theory or evidence.

Claim 1: Reproducibility is the cornerstone of, or a demarcation criterion for, science.

A common assertion in the methodological reform literature is that reproducibility[1] is a core scientific virtue and should be used as a standard to evaluate the value of research findings (***Begley and Ioannidis, 2015***; ***Braude, 2002***; ***McNutt, 2014***; ***Open Science Collaboration, 2012***, ***2015***; ***Simons, 2014***). This assertion is typically presented without explicit justification, but implicitly relies on two assumptions: first, that science aims to discover regularities about nature and, second, that reproducible empirical findings are indicators of true regularities. This view implies that if we cannot reproduce findings, we are failing to discover these regularities and hence, we are not practicing science.

The focus on reproducibility of empirical findings has been traced back to the influence of falsificationism and the hypothetico-deductive model of science (***Flis, 2019***). Philosophical critiques highlight limitations of this model (***Leonelli, 2018***; ***Penders et al., 2019***). For example, there can be true results that are by definition not reproducible. Some fields aim to obtain contextually situated results that are subject to multiple interpretations. Examples include clinical case reports and participant observation studies in hermeneutical social sciences and humanities (***Penders et al., 2019***). Other fields perform inference on random populations resulting from path-dependent stochastic processes, where it is often not possible to obtain two statistically independent samples from the population of interest. Examples are inference on parameters in evolutionary systems or event studies in economics. There are also cases where observing or measuring a variable's value changes its probability distribution—a phenomenon akin to the observer effect. True replication may not be possible in these cases. In short, science does—rather often, in fact—make claims about

[1] Here we use reproducibility as in: "the extent to which consistent results are observed when scientific studies are repeated" (***Open Science Collaboration, 2012***, p.657). In ***Appendix 1*** we provide a technical definition of reproducibility which we use in obtaining our results. We limit our discussion to statistical reproducibility of results only, and exclude other types such as computational reproducibility.

non-reproducible phenomena and deems such claims to be true in spite of the non-reproducibility. In these instances what scientists do is to define and implement appropriate criteria for assessing the rigor and the validity of the results (***Leonelli, 2018***), without making a reference to replication or reproduction of an experimental result. Indeed, many scientific fields have developed their own qualitative and quantitative methods such as ethnography or event study methodology to study non-reproducible phenomena.

We argue that even in scientific fields that possess the ability to reproduce their findings in principle, reproducibility cannot be reliably used as a demarcation criterion for science because it is not necessarily a good proxy for the discovery of true regularities. To illustrate this, consider the following two unconditional propositions: (1) reproducible results are true results and (2) non-reproducible results are false results. If reproducibility serves as a demarcation criterion for science, we expect these propositions to be true: we should be able to reproduce all true results and fail to reproduce all false results with reasonable regularity. In this section, we provide statistical arguments to probe the unconditional veracity of these propositions and we challenge the role of reproducibility as a key epistemic value in science. We also list some *necessary* statistical conditions for true results to be reproducible and false results to be non-reproducible. We conclude that methodological reform first needs a mature theory of reproducibility to be able to identify whether *sufficient* conditions exist that may justify labeling reproducibility as a measure of true regularities.

1.1 Reproducibility rate is a parameter of the population of studies.

To examine the suitability of reproducibility as a demarcation criterion, a precise definition of what we mean by reproducibility of results is needed. In assessing the reproducibility of research results, literature refers to “independent replications” of a given study. Strictly, we cannot speak of statistical independence between an original study and its replications. If study B is a replication of study A, then many aspects of study B depends on study A. Rather, sequential replication studies should be assumed *exchangeable*, conditional on the results and the assumptions of the original study, in the sense that the group of results obtained from a sequence of replication studies are probabilistically equivalent to each other irrespective of the order in which these studies are performed. Assuming that exchangeability holds, probability theory shows that the results from replication studies become independent of each other, but *only* conditional on the background information about the system under investigation, model assumed, methods employed, and the decision process used in obtaining the result (see assumptions of idealized study in ***Appendix 1***). The commonly used phrase “independent replications” thus has little value in developing a theory of reproducibility unless one takes sufficient care to consider all these conditionalities.

This conditional independence of sequence of results immediately implies that irrespective of whether a result is true or false, there is a true reproducibility rate of *any given result*, conditional on the properties of the study. This true reproducibility rate is determined by three components of the study: The true model generating the data, the assumed model under which the inference is performed, and the methods with which the inference is performed (***Appendix 2***, Proposition 1.1). In this sense, the true reproducibility rate is a parameter of the population of studies.

1.2 True results are not necessarily reproducible.

Our first proposition is that true results are not always reproducible (***Appendix 2***, Proposition 1.2), in contrast to much of the reform literature that claims non-reproducible results are necessarily false. For example, ***Wagenmakers et al.*** (***2012***, p.633) assert that “Research findings that do not replicate are worse than fairy tales; with fairy tales the reader is at least aware that the work is fictional.” It is assumed that true results must necessarily be reproducible, and therefore non-reproducible results must be “fictional.”

A careful look at statistical theory paints a different picture. The mere fact that the true reproducibility rate is a parameter of the population of studies matters: this parameter is a probability and therefore, it takes values on the interval $[0, 1]$. This implies that for finite sample studies involv-

Box 1. Some necessary conditions to obtain true results that are reproducible and false results that are non-reproducible

- True values of the unknown and unobservable quantities for which inference is desired must be in the decision space (***Appendix 2***).

  Examples: (i) In model selection, selecting the true model depends on having an M-closed model space, which means the true model must be in the candidate set (***Clarke et al., 2013***). (ii) In Bayesian inference, converging on the true parameter value depends on the true parameter value being included in the prior distribution, as stated by Cromwell's rule (***Lindley, 2006***, p.90).

- If inference is performed under one assumed model, that model should correctly specify the true mechanism generating the data.

  Example: A simple linear regression model with measurement error misspecified as a simple linear regression model yields biased estimates of regression coefficients, which will affect reproducibility of true and false results (***Figure 1***, ***Figure 2***).

- The quantities that methods use to perform inference on unknown and unobservable components of the model must contain enough information about those components: If they are statistics, they cannot be only ancillary. If they are pivots that are a function of nuisance parameters, then the true value of those nuisance parameters should permit reproducibility of results (***Appendix 2***).

  Example: In a one sample z-test where the population mean is not equal to the hypothesized value under the null hypothesis, the test incorrectly fails to reject with large probability due to large population variance.

- If inference is about parameters, observables must carry enough discernible information about these parameters. That is, model parameters should be identifiable structurally and informationally. Even weak unidentifiability will reduce the reproducibility of true results.

  Example: The requirement that the Fisher information (***Lehmann and Casella, 1998***, p.115) about unknown parameters should be sufficiently large in likelihoodist frameworks.

- Free parameters of methods should be compatible with our research goals.

  Example: A hypothesis test in Neyman-Pearson framework with Type I error rate $\alpha \approx 1$ is a valid statistical procedure that rejects the null hypothesis almost always when it is true.

- Methods should be free of unknown bias.

  Example: Observer effect, where mere observation changes the system we study, may lead to false results that are reproducible.

- The sample on which inference is performed is representative of the population from which it is drawn.

  Example: Statistical methods assume probabilistic sampling and do not make any claims in a non-probabilistic sampling framework (***Meng et al., 2018***).

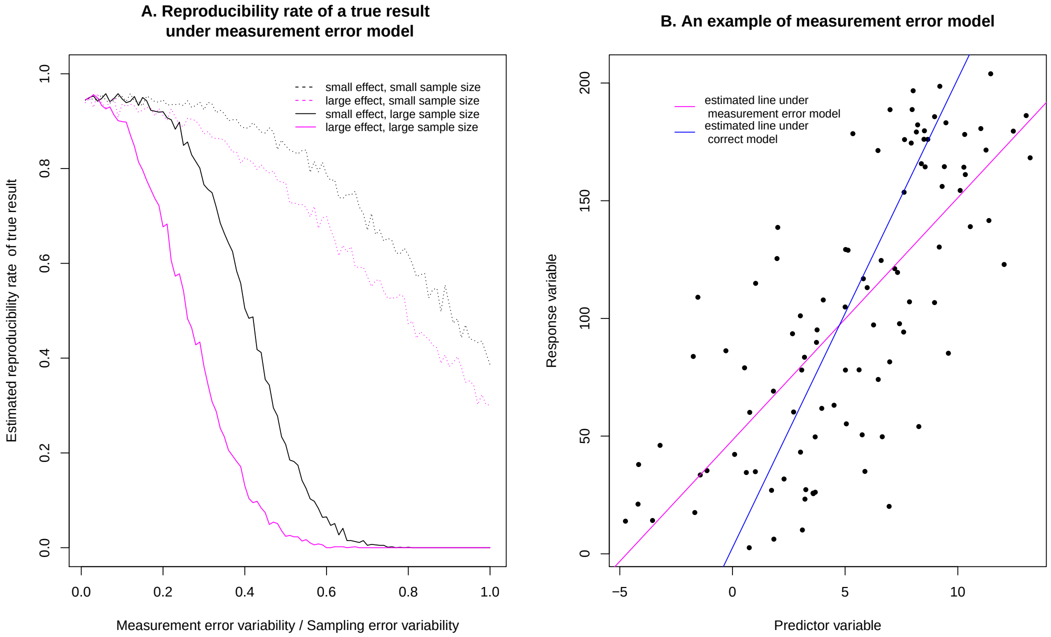


Figure 1. (A) Reproducibility rate of a true result decreases with measurement error in a misspecified simple linear regression model. Reproducibility rate is estimated by the proportion of times the 95% confidence interval captures the true effect. Sample sizes are 50 (small) and 500 (large). The true regression coefficient of the predictor variable is 2 (small effect) and 20 (large effect). Model details are given in ***Appendix 4***. (B) Example data (black points) generated under simple linear regression model $E(Y) = 2 + 20X$. Measurement and sampling error are normally distributed with standard deviations equal 3. Regression lines are fit under measurement error model (magenta line) and the correct model (blue line) with a sample size of 100. 95% confidence interval for the regression coefficient obtained under the measurement error model is $(7.94, 12.37)$, which does not include the true value 20. In contrast, 95% confidence interval for the regression coefficient obtained under the correct model, $(19.86, 20.21)$, includes the true value.

ing uncertainty, the true reproducibility rate must necessarily be smaller than one for any result. This point seems trivial and intuitive. However, it also implies that if the uncertainty in the system is large, true results can have reproducibility close to 0. Moreover, low uncertainty in the system is not a guarantee that true results will be reproducible. There are other necessary conditions related to the components of an idealized study that need to be met. Some of these conditions are listed in ***Box 1***.

For example, consider a scenario when the data are analyzed under a misspecified model, in this case a simple linear regression *measurement error model* in which the measurement error is unaccounted for (***Figure 1***). We are interested in the effect of measurement error on the reproducibility rate of a true effect. As the ratio of the measurement error variability in predictor to sampling error variability increases, the probability that an interval estimator of the regression coefficient (i.e., the effect size) at a fixed nominal coverage contains the true effect decreases. This is not simply an artifact of small sample sizes or small effects: the same pattern obtains for large sample sizes and large true effects. In fact, for large sample sizes, the reproducibility rate drops to zero at *lower* measurement error variability than for small sample sizes (also see ***Loken and Gelman, 2017***, for a similarly counter-intuitive effect of measurement error). Furthermore, the negative effect of measurement error on reproducibility rate of a true result actually grows with effect size, as ***Figure 1***A illustrates. Even in this relatively simple setting it is by no means a given that a true result will be reproducible. Measurement error is only one type of model misspecification. Other sources of misspecification and types of human error (e.g., questionable research practices) might further impair the reproducibility of true results.

When true reproducibility rate of a true result is low, the proportion of studies that fail to reproduce a true result will be high, even when methods being used have excellent statistical properties and the model is correctly specified. However, a true low reproducibility rate does not necessarily indicate a problem in the scientific process. As ***Heesen*** (***2018***) notes, low reproducibility in a given field or literature may be the result of there being few discoveries to be made in a given scientific system. When that is the case, a reasonable path to making scientific progress is to learn from non-reproducible results. Indeed, the history of science is full of examples of fields going through arduous sequence of experiments yielding failures such as non-reproducible results to eventually arrive at scientific regularities (***Barwich, 2019***; ***Chang, 2004***; ***Shiffrin et al., 2018***).

In an article that makes practical recommendations to improve the methodology of psychological science, ***Lakens and Evers*** (***2014***) argue that “One of the goals of psychological science is to differentiate among all possible truths” and suggest that one way to achieve this goal is to improve the statistical tools employed by scientists. Some care is needed when interpreting this claim. Statistical methods might indeed help us get close to the true data generating mechanism, if their modeling assumptions are met (thereby removing some of the reasons why true results can be non-reproducible). However, statistics’ ability to quantify uncertainty and inform decision making does not guarantee that we will be able to correctly specify our scientific model. Irrespective of reproducibility rates of results obtained with statistical methods, scientists attempting to model truth use theories developed based on their domain knowledge. Some of the problems raised in ***Box 1***, including model misspecification and decision spaces that exclude the true value of the unknown components, can only be addressed using a theoretical understanding of the phenomenon of interest. Without this understanding, there is no theoretical reason to believe that reproducibility rates will inform us about our proximity to truth.

It would be beneficial for reform narratives to steer clear of overly generalized sloganeering regarding reproducibility as a proxy for truth (e.g., reproducibility is a demarcation criterion or non-reproducible results are fairy tales). A nuanced view of reproducibility might help us understand why and when it is or is not desirable, and what its limitations are as a performance criterion.

1.3 False results might be reproducible.

Our second proposition is the converse of the first and considers the respects in which false results can sometimes be highly reproducible (***Appendix 2***, Proposition 1.3). In well-cited articles in methodological reform literature, high reproducibility of a result is often interpreted as evidence that the result is true (***Nosek et al., 2012***; ***Open Science Collaboration, 2015***; ***Pashler and Wagenmakers, 2012***). A milder version of this claim is also invoked, such as “Replication is a means of increasing the confidence in the truth value of a claim.” (***Nosek et al., 2012***, p.617). The rationale is that if a result is independently reproduced many times, it must be a true result.[2] This claim is not always true. To see this, it is sufficient to note that the true reproducibility rate of any result depends on the true model *and* the methods used to investigate the claim. We follow with two examples.

First, consider a valid hypothesis test in which the researcher unreasonably chooses to set $\alpha = 1$. Then, a true null hypothesis will be rejected with probability 1 and this decision will be 100% reproducible, assuming that replication studies also set the significance criterion (α) to 1. While we know better than to set our significance criterion so high, this example shows how reproducibility rate is not only a function of the truth but also our methods. Second, consider estimators that exploit the bias-variance trade-off by introducing a bias in the estimator to reduce its variance. These estimators have a higher reproducibility rate but for a false result by design. In this case, researchers deliberately choose false results that are reproducible when they prefer a biased estimator over a noisy one for usefulness. Next, we give a realistic example, in which we describe a *mechanism* for why reproducibility cannot serve as a demarcation criterion for truth.

[2] An epistemic claim that well-confirmed scientific theories and models capture (approximate) truths about the world is an example of *scientific realism*. The arguments for and against scientific realism are beyond the scope of this paper. Interested readers may follow up on discussions in the philosophical literature (***Chakravartty, 2017***).

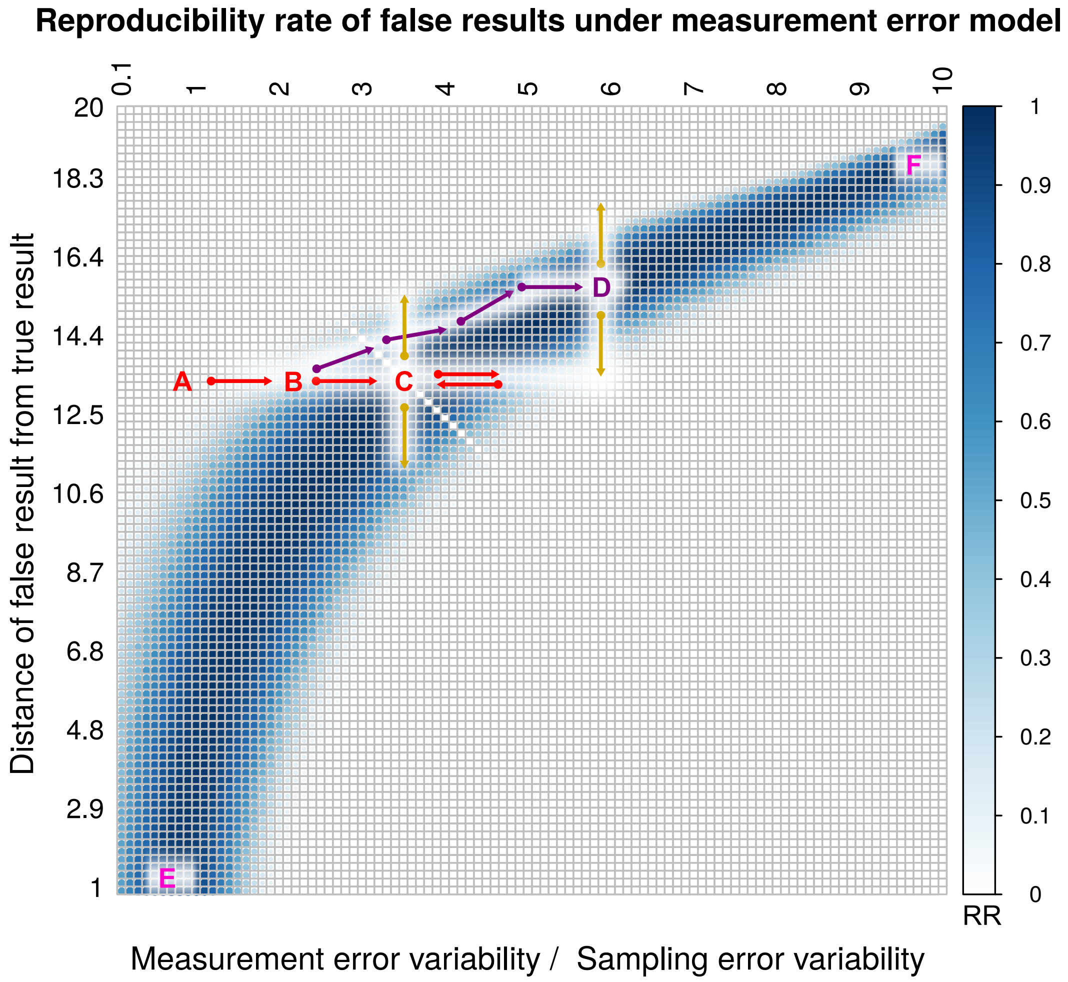


Figure 2. An example of almost perfectly reproducible false results in a misspecified simple linear regression model with measurement error. Color map shows reproducibility rate (RR). Darkest blue cells indicate perfect reproducibility rate (almost 100%) of false results at appropriate measurement error for each false effect size, shown by its distance from the true effect size on the vertical axis. The true regression coefficient of predictor variable (effect size) is 20. Details are given in ***Appendix 4***. For description of letters and arrows, refer to the text.

We consider model misspecification under a measurement error model in simple linear regression. Simple linear regression involves one predictor and one response variable, where the predictor variable values are assumed to be fixed and known. The measurement error model incorporates unobservable random error on predictor values. The blue belt in ***Figure 2*** shows that as measurement error variability grows with respect to sampling error variability, effects farther away from the true effect size become perfectly reproducible. At point F in figure ***Figure 2***, the measurement error variability is ten times as large as the sampling error variability, and we have perfect reproducibility of a null effect when the true underlying effect size is in fact large.

Now consider a scientist who takes reproducibility rate as a demarcation criterion. Assume she starts at point A and she performs a study which lands her at point B—which might happen by knowingly or unknowingly choosing noisier measures or by reducing sampling error variability. The reproducibility of her results has increased (from white to inside the blue belt) and to increase it further, she performs another study by further tweaking the design, which then lands her at point C. If she were to move horizontally to the right with her future studies, the reproducibility of results will decrease, and she will turn back to C, which ultimately will be a stable equilibrium of maximal reproducibility. Further, this is just one of the possible paths that she could take to achieve maximal reproducibility. When at point B, she might perform a study that follows the purple path, always increasing the reproducibility of her results ending up at point D, which is another stable equilibrium point of maximal reproducibility. In fact, any sequence of studies that increases reproducibility will end at one of the points that corresponds to the darkest blue color in the belt. At this point, however, we note that going from point A to point C, our researcher started with a false result where the estimated slope was some ≈ 13 units off the true value (y axis, point A) and arrived at the same false result (y axis, point C), even though she has maximized the reproducibility of her results. Worse, when she arrived at point D, the estimated slope is now some ≈ 15 (y axis, point D) units away from the true value, even though she still maximized the reproducibility of her results.

Taking a step back, we note that to approach the true result, one needs to move to the origin in this plot. However, that approach is controlled by the vertical axis, and not the horizontal. Unless we know that we are committing a model misspecification error, we get no feedback when we perform studies that move us randomly on the vertical axis (yellow arrows). For example, points C and D have similar reproducibility of results but at C we are closer to truth then D. In fact, consider points E and F: we get high reproducibility of results at both points, but estimates obtained at point E are much closer to the true value than estimates obtained at point F. The mechanistic explanation of this process is that reproducibility-as-a-criterion can be optimized by the researcher *independently of the underlying truth of their hypothesis*. That is, optimizing reproducibility can be achieved without getting any closer to the true result. This is not to say that reproducibility is not useful, but it means that it cannot be used as a demarcation criterion for science.

While we advance a statistical argument for the reproducibility of false results, the truth value of reproducible results from laboratory experiments has also been challenged for non-statistical reasons (***Hacking, 1992***, p.30). Hacking notes that mature laboratory sciences sometimes construct an irrefutable system by developing theories and methods that are “mutually adjusted to each other”. As a result, these sciences become what Hacking calls “self-vindicating”. That is:

> “The theories of the laboratory sciences are not directly compared to ’the world’; they persist because they are true to phenomena produced or even created by apparatus in the laboratory and are measured by instruments we have engineered.”

Hacking concludes that “[h]igh level theories are not ‘true’ at all.” They can be viewed as a summary of the collection of laboratory operations to which they are adapted, but if that set of operations is selected to match a particular theory, its evidentiary value may be limited. Hacking’s description of what makes mature laboratory sciences highly reproducible is consistent with our definition of reproducibility rate as a function of true model, assumed model, and methods.

An example of a theory from laboratory sciences that is not directly compared to ’the world’ comes from cognitive science. One high level theory that has become prominent in this field over the last two decades is the “probabilistic” or “Bayesian” approach to describing human learning and reasoning (***Oaksford and Chater, 1998***; ***Chater et al., 2008***). As the paradigm rose to prominence, questions were raised as to whether claims of the Bayesian theory of the mind held any truth value at all, in either a theoretical or empirical sense (***Bowers and Davis, 2012***).

Within a specific framework, a particular experimental result may have value in connection to a theoretical claim without being tied to the world. For instance, ***Hayes et al.*** (***2019***) presented several experiments that appear to elicit the “same” phenomenon in different contexts, and an accompanying Bayesian cognitive model that renders these results interpretable within that framework. It is less clear — even to the authors of the original study — what relationship the robust empirical results have to the true mechanisms underpinning human reasoning; the experiments were designed from and adapted to the Bayesian framework and the results can be given a clear interpretation *within* that theoretical perspective, but it is not easy to justify stronger claims.

As this example illustrates, Hacking’s observations about the “mutual tuning” between theoretical claims and laboratory manipulations are observed in practice, in cognitive science and potentially in other disciplines. Our measurement error example shown in ***Figure 2*** provides just one possible realization for Hacking's conjecture (see also ***Flake and Fried, 2019***, for a detailed discussion on measurement practices that might exacerbate measurement error). Other forms of inference under model misspecification might present different scenarios under which this mutual tuning may take place—for example, the inadvertent introduction of an experimental confound or an error in a statistical computation have the potential to create and reinforce perfectly reproducible *phantom* effects. The possibility of such tuning renders suspect the idea that reproducibility is a good proxy for assessing the truth potential of a result.

If the heuristic that reproducibility is a demarcation criterion were to take hold in scientific discourse, false results might get treated as true, irreversibly altering the course of scientific progress with implications for broader society.

Claim 2: Using data more than once invalidates statistical inference.

A well-known claim in the methodological reform literature regards the (in)validity of using data more than once, which is sometimes colloquially referred to as *double-dipping* or *data peeking*. For instance, ***Wagenmakers et al.*** (***2012***, p.633) decry this practice with the following rationale: "Whenever a researcher uses double-dipping strategies, Type I error rates will be inflated and p values can no longer be trusted." The authors further argue that "At the heart of the problem lies the statistical law that, for the purpose of hypothesis testing, the data may be used only once." Similarly, ***Kriegeskorte et al.*** (***2009***, p.535) define double dipping as "the use of the same data for selection and selective analysis" and add the qualification that it would invalidate statistical inference "whenever the test statistics are not inherently independent of the selection criteria under the null hypothesis." This rationale has been used in reform literature to establish the necessity of preregistration for "confirmatory" statistical inference (***Nosek et al., 2018***; ***Wagenmakers et al., 2012***).

In this section, we provide examples to show that it is incorrect to make these claims in overly general terms. The reform literature is not very clear on the distinction between "exploratory" and "confirmatory" inference. We will revisit these concepts in the next claim but for now, we evaluate the claim that using data multiple times invalidates statistical inference. For that, we will steer away from the exploratory-confirmatory dichotomy and focus on the validity of statistical inference specifically.

At the outset, we note that the phrases *double-dipping, data peeking,* and *using data more than once* do not have a formal definition and thus cannot be the basis of any *statistical law*. These verbally stated terms are ambiguous and create a confusion that is non-existent in statistical theory. Many well-known valid statistical procedures use data more than once (see ***Darnieder, 2011***, for a detailed analysis in the context of data dependent priors). For example, a one sample t-test for testing whether the population mean is μ_o uses the test statistic $\left(\bar{X} - \mu_o\right) \Big/ \left(s/\sqrt{n}\right)$, where n, $\bar{X}$, and s are the sample size, the sample mean, and the sample standard deviation respectively. Clearly, the test statistic uses the data three times: once to get n, a second time to get $\bar{X}$, and a third time to get s. In fact, a valid statistical test can be built by using the data to obtain *almost all aspects of a hypothesis test that are not specifically user defined*, including the hypotheses themselves. The key to validity is not how many times the data are used, but appropriate application of the correct conditioning as dictated by probability calculus (***Lindley, 2000***). Furthermore, under many cases, the conditioning does not affect the validity of the test of interest, and therefore can be dropped, freeing the data from its prison for use prior to test of interest (***Buzbas, 2019***).

When conditioning on prior activity on the data is indeed needed to make a test valid, overlooking that a procedure should be modified to accommodate this prior activity might lead to an erroneous test. However, this situation only arises if we disregard the elementary principles of statistical inference such as correct conditioning, sufficiency, completeness, and ancillarity. Conditional inferences are statistically valid when their interpretation is properly conditioned on the information extracted from the observed data, which are sufficient for model parameters. Therefore, unconditionally stating that *double-dipping*, *data peeking*, or *using data more than once* invalidates inference does not make statistical sense. In contrast with common reform narratives, one can use the data many times in a valid statistical procedure. Below, we describe the conditions under which this validity is satisfied. We also discuss why preregistration cannot be a prerequisite for valid statistical inference, confirmatory or otherwise.

2.1. Valid conditional inference is well-established.

Imagine we aim to confirm a scientific hypothesis of interest which can be formulated as a statistical hypothesis and be tested using a chosen a test of interest. Suppose we perform some statistical activity on the data until we begin the test of interest. This activity may comprise informal or formal analyses on the data. To assess the effect of this activity on the validity of the test of interest, we assume that the information obtained from prior analyses can be summarized by a statistic.

First, we categorize the amount of information contained in the test statistic of interest. This statistic may contain anywhere from *no information* to *all information* in the data about the parameter of interest. Further, it can satisfy some statistical optimality criterion, in which case it is identified as the best statistic with respect to this criterion. The case of no information is trivial and not interesting. The case of all information is well known.[3] For many commonly used models, an optimal statistic is also well known[4] (first column in left and right blocks, ***Box 2***). Other cases include partial information (second column in left and right blocks, ***Box 2***).

Second, the statistic that summarizes the analyses performed on the same data prior to the test of interest may also contain anywhere from no information to all information in the data (rows in left and right blocks, ***Box 2***). However, here the case of no information is *also* of interest[5].

If the statistic summarizing the prior analysis is used in a subsequent analysis for the test of interest, the validity of the test is guaranteed by conditioning the subsequent analysis on this statistic, using probability calculus. A relatively simple case may involve only conditioning on the statistic obtained from prior analysis (left block, ***Box 2***). In this case, no quantity exogenous to the model generating the data is introduced into the test of interest. If the test of interest uses an optimal statistic (which is the case for many well-known models), the conditioning is irrelevant because the validity of the test is not affected by the prior information (left block first column in ***Box 2***). The same result with the same validity is obtained *as if* we did not perform any activity on the data, previous to the test of interest. Hence, one can freely use information prior to performing the test of interest without any modification in the test of interest. If the test of interest does not use an optimal statistic, then conditioning will maintain the validity and often improve the performance of the test (left block second column in ***Box 2***). This is a manifestation of Rao-Blackwellization of the test statistic to reduce its variance. We reproduce an example by ***Mukhopadhyay*** (***2006***) of estimating the parameter of a normal distribution whose mean and standard deviation are equal using a randomly sampled single observation in ***Figure 3***. We give a statistical justification in Proposition 2.1, ***Appendix 3***. Therefore, Claim 2 is false for this case.

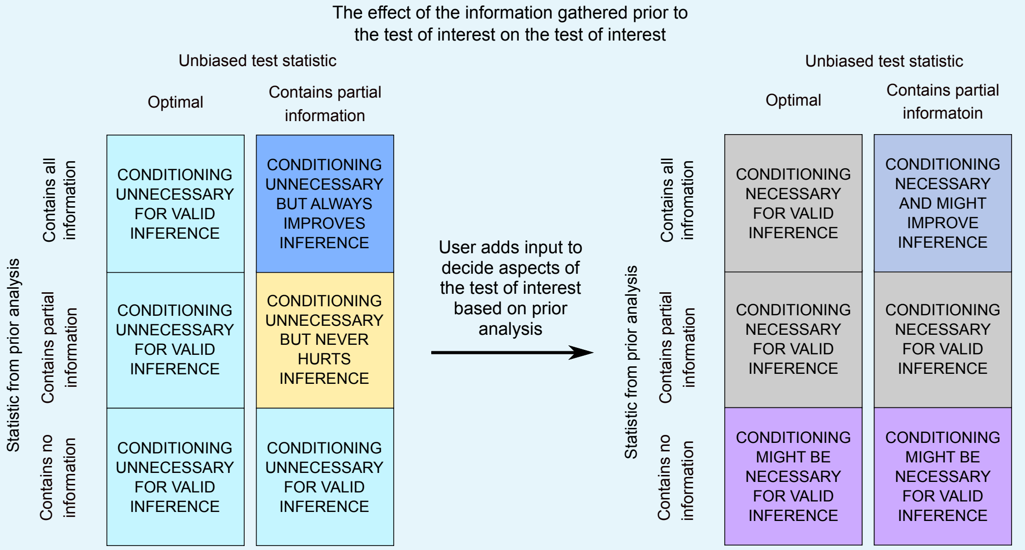
A more complicated case occurs when one not only obtains a statistic from prior analysis, but also makes a decision to redefine the test of interest based on the observed value of that statistic—a decision that depends on an exogenous criterion and alters the set of values the test statistic of interest is allowed to take (right block, ***Box 2***). For example, an exogenous criterion might be *to perform the test only if the statistic from prior analysis satisfies some condition*. Subgroup analyses or determining new hypotheses based on the results of prior analysis (HARKing) are other examples (***Rubin, 2017***). Conditional quantities which make the test of interest valid are now altered because conditioning on *a statistic* and conditioning on *whether a statistic obeys an exogenous criterion* have different statistical consequences. If this criterion affects the distribution of the test statistic of interest, then conditioning is necessary. The correct conditioning will modify the test in such a way that the distribution of the test statistic under the null hypothesis is derived, critical values for the

[3] sufficient statistic

[4] complete sufficient statistic

[5] ancillary statistic

Box 2. Valid inference using data multiple times


The effect of the information gathered prior to the test of interest on the test of interest

Statistic from prior analysis	Unbiased test statistic: Optimal	Unbiased test statistic: Contains partial information
Contains all information	CONDITIONING UNNECESSARY FOR VALID INFERENCE	CONDITIONING UNNECESSARY BUT ALWAYS IMPROVES INFERENCE
Contains partial information	CONDITIONING UNNECESSARY FOR VALID INFERENCE	CONDITIONING UNNECESSARY BUT NEVER HURTS INFERENCE
Contains no information	CONDITIONING UNNECESSARY FOR VALID INFERENCE	CONDITIONING UNNECESSARY FOR VALID INFERENCE

User adds input to decide aspects of the test of interest based on prior analysis

Statistic from prior analysis	Unbiased test statistic: Optimal	Unbiased test statistic: Contains partial informatoin
Contains all infromation	CONDITIONING NECESSARY FOR VALID INFERENCE	CONDITIONING NECESSARY AND MIGHT IMPROVE INFERENCE
Contains partial information	CONDITIONING NECESSARY FOR VALID INFERENCE	CONDITIONING NECESSARY FOR VALID INFERENCE
Contains no information	CONDITIONING MIGHT BE NECESSARY FOR VALID INFERENCE	CONDITIONING MIGHT BE NECESSARY FOR VALID INFERENCE

We assume a test based on an unbiased test statistic generates valid inference, in the sense of achieving its nominal Type I error probability, under its assumptions within the Neyman-Pearson hypothesis testing paradigm. Information extracted from the data prior to the test of interest is represented by a statistic from prior analysis. Cells describe the necessity and/or the outcome of conditioning the test of interest on this statistic from prior analysis, for varying levels of information captured. Some technical clarifications for special cases are discussed in ***Appendix 3***.

Left: The statistic from prior analysis is not used in decision making, for example, by combining it with a user defined criterion which might affect aspects of the test of interest. Many commonly used linear models fall in the first column where procedures are based on an optimal test statistic and therefore, using the information from prior analysis does not affect the validity of the test of interest. However, even if the statistic for the test of interest is not optimal, conditioning on statistic from prior analysis is not necessary for validity of inference. Further, conditioning never hurts the validity of inference and improves the performance in most cases. Details of the conditional analyses in this block are provided in Propositions 2.1 and 2.2 in ***Appendix 3***

Right: The statistic from prior analysis is combined with a user defined criterion to affect aspects of the test of interest through a decision. An example is using the data to determine which subsamples to compare. The validity of the test of interest is maintained when inference is conditioned on this decision if the statistic from prior analysis contains at least some information about the parameter to be tested.

The change in corresponding cells between left block and right block shows the effect of using this user defined criterion on conditional statistical inference.

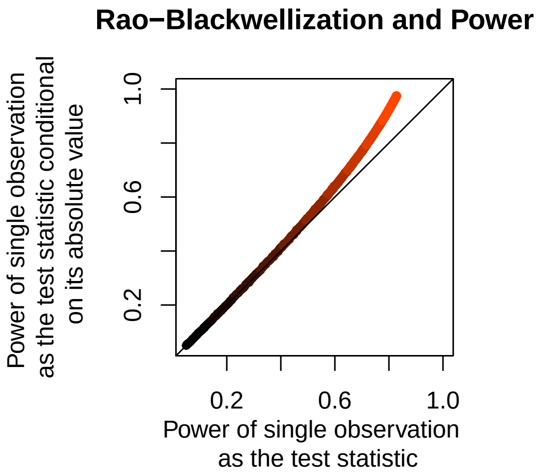


Figure 3. For a normally distributed variable with equal mean and variance, we randomly sample a single observation from the population. We plan to use this observation as a test statistic for the common parameter. However, prior to this test we observe the absolute value of the sample and we decide to perform the test using the information in both the observation and its absolute value, therefore, using the unsigned part twice. The plot compares power of the test based on the single observation and on the single observation conditioned on its absolute value. Conditioning improves inference by reducing the variance of the test statistic. This case corresponds to left block, second row, first column in ***Box 2***. Lighter shades represent larger true parameter values. Technical details are given in ***Appendix 4***.

test are re-adjusted, and desired nominal error rates are achieved. A general algorithm to perform statistically valid conditional analysis in this sense is provided in ***Appendix 5***. Adhering to correct conditioning, then, guarantees the validity of the test, making Claim 2 false again.

Figure 4 provides an example of how conditioning can be used to ensure that nominal error rates are achieved. We aim to test whether the mean of Population 1 is greater than the mean of Population 2, where both populations are normally distributed with known variances. An appropriate test is an upper-tail two-sample z-test. For a desired level of test, we fix the critical value at z, and the test is performed without performing any prior analysis on the data. The sum of the dark green and dark red areas under the black curve is the nominal Type I error rate for this test. Now, imagine that we perform some prior analysis on the data and use it only if it obeys an exogenous criterion: We do not perform our test unless “the mean of the sample from Population 1 is larger than the mean of the sample from Population 2.” This is an example of us deriving our alternative hypothesis from the data. The test can still be made valid, but proper conditioning is required. If we do not condition on the information given within double quotes and we still use z as the critical value, we have inflated the observed Type I error rate by the sum of the light green and light red areas because the distribution of the test statistic is now given by the red curve. We can, however, adjust the critical value from z to z^* such that the sum of the light and dark red areas is equal to the nominal Type I error rate, and the conditional test will be valid. This case corresponds to the right block, first row, first column in ***Box 2***. Technical details are provided in ***Appendix 4***.

Although caution with regard to double dipping is sometimes justified, the claim that it invariably invalidates statistical inference is unsupported. In fact, the opposite is true since all cells in ***Box 2*** yield valid tests. Clearly, proper conditioning solves a statistical problem. However, the garden of forking paths applies to problems of scientific importance as well, since our conclusions become dependent on decision we make in our analysis. Statistical rigor is the prerequisite of a successful solution, but we should ask: Solution to which problem? Statistical validity does not necessarily imply scientific validity (***Navarro, 2019***). The connection between statistical and scientific models

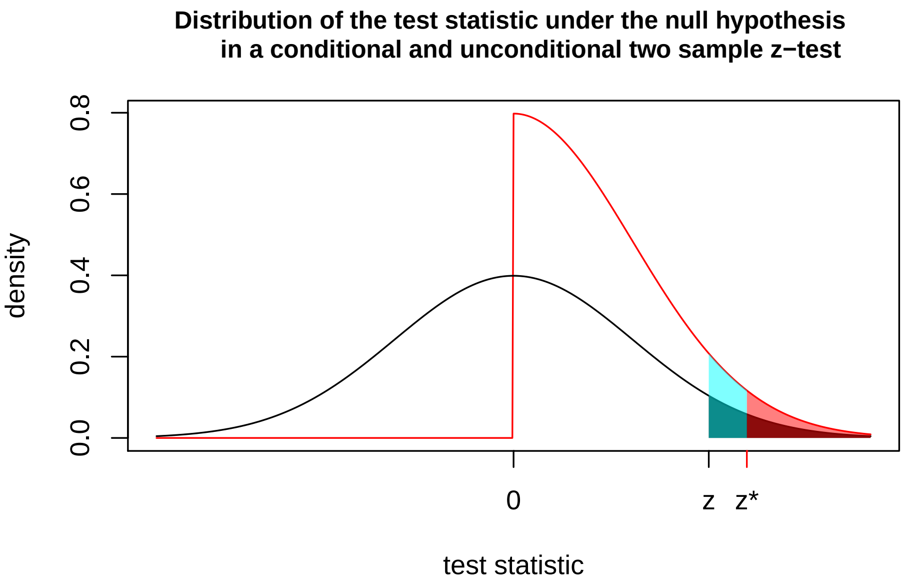


Figure 4. For a two sample z-test, we display rejection regions for an unconditional test and a conditional test, setting the alternative hypothesis in the direction of the observed effect. The black curve shows the distribution of the unconditional test statistic, with the critical value given by z. Red curve shows the distribution of the conditional test statistic, with the adjusted critical value given by z^*.

might be weak—a problem that cannot be fixed by statistical rigor.[6] Further, valid inference by proper conditioning entails maintaining the same conditioning for correct interpretation of scientific inference. Viable alternatives to multiple or sequential hypothesis testing include multilevel modeling (***Gelman et al., 2012***; ***Gelman and Loken, 2013***) and multiverse analysis (***Steegen et al., 2016***). The key to successfully implement these solutions is a good understanding of statistical theory and a careful interpretation of results under clearly stated assumptions.

2.2. Preregistration is not necessary for valid statistical inference.

Nosek et al. (***2018***) claim that “Standard tools of statistical inference assume prediction.”[7] ***Nosek et al.*** (***2018***) intend to convey that in hypothesis testing, the analytical plan needs to be determined (i.e., preregistered) prior to data collection or observing the data for statistical inference to have diagnostic value, that is, to be valid. In other words, “Confirmatory conclusions require preregistration” (***Wagenmakers et al., 2012***, p.634). According to the methodological reform, any inferential procedure that is not preregistered is categorized as *postdiction* or *exploratory* analysis, and should not be used to arrive at *confirmatory* conclusions.

In this section, we first clarify the *statistical* problem which preregistration aims to address. Then we assess what preregistration cannot statistically achieve under its strict and flexible interpretation.

[6] Testing hypotheses with no theory to motivate them is a fishing expedition regardless of methodological rigor. See ***Gervais*** (***2020***); ***Guest and Martin*** (***2020***); ***MacEachern and Van Zandt*** (***2019***); ***Muthukrishna and Henrich*** (***2019***); ***Oberauer and Lewandowsky*** (***2019***); ***Szollosi and Donkin*** (***2019***); ***Szollosi et al.*** (***2019***); ***van Rooij*** (***2019***) and ***van Rooij and Baggio*** (***2020***) for discussions on scientific theory.

[7] Prediction here is not used in statistical sense but refers to “the acquisition of data to test ideas about what will occur” (***Nosek et al., 2018***, p.2600). To clarify, statistics uses sample quantities (observables) to perform inference on population quantities (unobservables). Inference, therefore, is about unobservables. Statistical prediction, on the other hand, is defined as predicting a yet unobserved value of an observable and therefore, is about observables. The quote refers to a procedure about unobservables and hence “prediction” is not used in a statistical sense. Instead it is used to demarcate the timing of hypothesis setting and analytical planning with regard to data collection or observation. The authors also specifically refer to null hypothesis significance testing procedure as *the standard tool for statistical inference* referenced in this quote. While the statement itself can be misleading because of these local definitions and assumptions, our aim is to critique the intended meaning not the idiosyncratic use of statistical terminology.

We argue how preregistration can harm statistical inference while trying to solve its intended problem. After showing that preregistration is not necessary for valid statistical inference, we describe what it can achieve statistically.

What is the statistical problem that preregistration aims to address? Preregistration is offered as a solution to the statistical problem of using data multiple times (**Lindsay et al., 2016**; **Nosek et al., 2018**; **Wagenmakers et al., 2012**). Once a hypothesis and an analytical plan is preregistered, the idea is that researchers would be prevented from performing analyses that were not preregistered and subsequently, from presenting them as "confirmatory". We have shown that using data multiple times per se does not present a statistical problem. The problem arises if proper conditioning on prior information or decisions is skipped. The reform literature misdiagnoses the problem as an ordinal issue regarding the order of: hypothesis setting, decisions on statistical procedures, data collection, and performing inference. Preregistration locks this order down for an analysis to be called "confirmatory". Our examples of valid tests in **Box 3** show that the problem is not ordinal but one of statistical rigor. Prediction and postdiction—as proposed by **Nosek et al.** (**2018**)—do not have technical definitions in their intended meaning that reflects on statistical procedures. Further, the reform literature does not present any theoretical results to show the effects of this dichotomy on statistical inference. All well-established statistical procedures deliver their claims when their assumptions are satisfied. Other non-mathematical considerations are irrelevant for the validity of a statistical procedure. A valid statistical procedure can be built either before or after observing the data, in fact, even after using the data if proper conditioning is followed. Therefore, the validity of statistical inference procedures cannot depend on whether they were preregistered.

How can preregistration (strict or flexible) harm statistical inference? Preregistration may interfere with valid inference because nothing prevents a researcher from preregistering a poor analytical plan. Preregistering invalid statistical procedures does not on its own ensure the validity of inference (see also **Rubin, 2017**), while it does add a superficial veneer of rigor.

Assume hypotheses, study design, and an analysis plan are preregistered, and the researchers follow their preregistration to a T. Many hypothesis tests make parametric assumptions and not all are robust to model misspecification. **Dennis et al.** (**2019**) show that under model misspecification, the Neyman-Pearson hypothesis testing paradigm might lead to Type I error probabilities approaching 1 asymptotically with increasing sample sizes. Model misspecification is suspected to be common in scientific practice (**Box, 1976**; **Navarro, 2019**; **Szollosi et al., 2019**). Since the validity of a statistical inference procedure depends on the validity of its assumptions, performing assumption checks (if possible) to choose and proceed with the model and method whose assumptions hold is sound practice. Assumption checks are performed *after* data collection and on the data, but *before specifying a model and a method for analysis*. To accommodate assumption checks under preregistration philosophy, an exception would need to be made to the core principle because they necessitate using data multiple times. Indeed such exceptions are often made (**Lindsay et al., 2016**; **Nosek et al., 2018**) and it has been suggested that assumption checks and contingency plans should be preregistered. However, no statistical reasoning is provided to define the boundaries of such deviations from preregistration.

A common reform slogan states that "preregistration is a plan, not a prison[8]," offering an escape route from undesirable consequences of rigidity. **Nosek et al.** (**2018**, p.2602) suggest that compared to a researcher who did not preregister their hypotheses or analyses, "preregistration with reported deviations provides substantially greater confidence in the resulting statistical inferences." This statement has no support from statistical theory. On the other hand, the claim may make researchers feel justified in changing their preregistered analyses as a result of practical problems in data collection or analysis, without accounting for the conditionality in their decisions, leading to invalid statistical inference.

[8] While not part of our core argument this particular slogan is underspecified. It is not clear how the argument for the necessity of preregistration for statistically valid inference should be reconciled with the proposed flexibility of preregistrations. In any case, this line of thinking is moot from our perspective since the underlying premise itself does not hold.

A study of 16 *Psychological Science* papers with open preregistrations shows that research often deviated from preregistration plans (***Claesen et al., 2019***). Hence, in practice, preregistration fails to lock researchers in an analytical plan. Deviating from a preregistered plan might prevent a statistically flawed procedure from being implemented, and hence, might improve statistical validity of conclusions. On the other hand, it is possible to deviate from a plan by introducing more sequential decisions and contingency to data analysis, which if not accounted for, would invalidate the statistical inference. A strict interpretation of preregistration may also lead to invalid inference by locking researchers in a faulty plan. As such, preregistration or deviations from preregistration have little say over the diagnosticity of p-values or error control. Statistical rigor can neither be ensured by preregistration nor would be compromised by not preregistering a plan.

What can preregistration achieve statistically? Strict preregistration might work as a behavioral sanction that prevents researchers from doing any statistical analysis that involves conditioning on data, valid or invalid. This way, preregistration can prevent using data multiple times without proper conditioning by preventing proper conditioning procedures along with it. Nevertheless, as we show in ***Box 2***, conditioning on data may improve inference. On the other hand, a flexible interpretation of preregistration that allows for deviations in the plan so long as they are labeled as "exploratory" rather than "confirmatory" has no bearing on statistical outcomes. It remains unclear why these labels should be preferred over more direct descriptors such as "preregistered" or "not preregistered". If proper conditioning is performed, analyses that are referred to as "exploratory" in the reform literature might observe strict error control and if it is not, analyses currently being labeled "confirmatory" might be statistically uninterpretable.

There exist other social advantages to preregistration of empirical studies, such as the creation of a reference database for systematic reviews and meta-analysis that is relatively free from publication bias. While these represent genuine advantages and good reasons to practice preregistration, they do not affect the interpretation or validity of the statistical tests in a particular study. We demonstrate some of the points discussed in this section with examples in ***Box 3***. The statistical theory behind these examples show that the benefits of preregistration —in promoting systematic documentation and transparent reporting of hypotheses, research design, and analytical procedures— should not be mistaken for a technical capacity for ensuring statistical validity. If and only if a statistically appropriate analytical plan has been preregistered and performed, would preregistration have a chance of ensuring the meaningfulness of statistical results. Yet a well-established statistical procedure always returns valid inference, preregistered or not.

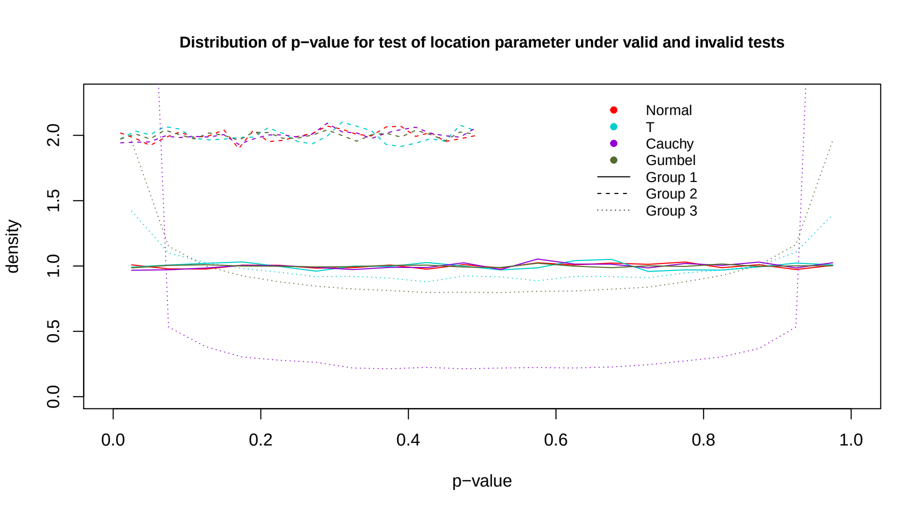
Claim 3: Exploratory Research Uses "Wonky" Statistics

A large body of reform literature advances the exploratory-confirmatory research dichotomy from an exclusively statistical perspective. ***Wagenmakers et al.*** (***2012***) argue that purely exploratory research is one that finds hypotheses in the data by post-hoc theorizing and using inferential statistics in a "wonky" manner where p-values and error rates lose their meaning: "In the grey area of exploration, data are tortured to some extent, and the corresponding statistics is somewhat wonky." The reform movement seems to have embraced ***Wagenmakers et al.*** (***2012***)'s distinction and definitions, and this dichotomy has been emphasized in required documentation for preregistrations (***van't Veer and Giner-Sorolla, 2016***) and registered reports (***McIntosh, Robert D., 2017***; ***Nosek and Lakens, 2014***).

We start by discussing why the exploratory-confirmatory dichotomy is not tenable from a purely statistical perspective. The reform literature does not provide an unambiguous definition for what is considered "confirmatory" or "exploratory". There are many possible interpretations including: (1) Formal statistical procedures such as null hypothesis significance testing are confirmatory, informal ones are exploratory. (2) Only preregistered hypothesis tests are confirmatory, non-preregistered ones are exploratory. (3) Only statistical procedures that deliver their theoretical claims (e.g., error control) are confirmatory, invalid ones are exploratory. These three dichotomies are not consistent with each other and lead to confusing uses of terminology. One can speak of formal statistical

Box 3. Validity of statistical analyses under strict, flexible, and no preregistration

We show how a strict interpretation of preregistration and a failure to use proper statistical conditioning may hinder valid statistical inference with a simulation example. Our simulations consist of 10^6 replications of hypothesis tests for the difference in the location parameter between two populations. We build the distribution of p-values under the null hypothesis of no difference for three cases and four true data generating models. In addition to the Normal distribution with exponentially bounded tail, we use Cauchy and T distributions for heavy tail, and Gumbel distribution for light tail. By a well-known result, the distribution of p-value under the null hypothesis is standard uniform for a valid statistical test.



- Hypothesis tests in Group 1 (solid lines) were performed using the following procedure:
  1. Collect data with no specification of hypothesis, model, or method (no preregistration).
  2. Calculate the sample medians. Set the alternative hypothesis so that the median of the population corresponding to the larger sample median is larger than the median of the other population (using the data to determine the hypotheses).
  3. Build the *conditional* reference distribution of the test statistic by permuting the data (reusing the data to determine the method).
  4. Calculate the test statistic from the data to compare with the reference distribution (reusing the data to calculate observed value of the test statistic).

  The tests in Group 1 derive almost all their components from the data by reusing them multiple times. The distribution of the p-values show that these tests are valid since they follow the standard uniform distribution (solid lines).
- Hypothesis tests in Group 2 (dashed lines) demonstrate a situation that may arise under either flexible preregistration (assumption checks allowed) or no preregistration, when proper statistical conditioning is not performed in step 3. This is akin to HARKing without statistical controls. In this case, the distribution of p-values is uniform on $(0, 0.5)$. These tests are not valid, since $\mathbb{P}(p \leq \alpha | H_0) = 2\alpha$ for some significance thresholds α.
- Hypothesis tests in Group 3 (dotted lines) demonstrate a situation that may arise under a strict preregistration protocol (altering the preregistered model or methods not allowed) when there is model misspecification. The preregistered model is Normal, but the data are generated under other models. These tests are not valid, since $\mathbb{P}(p \leq \alpha | H_0) > \alpha$ for some significance thresholds α.

procedures such as significance tests, and informal procedures such as data visualization, or valid and invalid statistical inference, but there is no mathematical mapping from these to exploratory or confirmatory research, especially when clear technical definitions for the latter are not provided. Moreover, the general usefulness and relevance of this dichotomy has also been challenged for theoretical reasons (***Oberauer and Lewandowsky, 2019***; ***Szollosi and Donkin, 2019***). In this section, we sidestep issues with the dichotomy but argue against the core claim presented by (***Wagenmakers et al., 2012***) regarding the nature of exploratory research specifically, advancing the following points:

- Exploratory research aims to facilitate scientific discovery, which requires a broader approach than statistical analysis alone,
- Exploratory data analysis is a tool for performing exploratory research and uses methods that only answer to their assumptions to be valid,
- Using “wonky” inferential statistics does not facilitate and probably hinders exploration, and
- Exploratory research needs rigor to serve its intended aim to facilitate scientific discovery.

Scientific exploration is the process of attempting to discover new phenomena (***Swedberg, 2018***). Outside of the methodological reform literature, exploratory research is typically associated with hypothesis generation and is contrasted with hypothesis testing—sometimes referred to as confirmatory research. Exploratory research may lead to serendipitous discoveries. However, it is not synonymous with serendipity but is a deliberate and systematic attempt at discovering generalizations that help us describe and understand an area about which we have little or no knowledge (***Stebbins, 2001***). In this sense, it is analogous to topographically mapping an unknown geographical region. The purpose is to create a complete map until we are convinced that there is no element within the region being explored that remains undiscovered. This process may take many forms from exploration of theoretical spaces (i.e., theory development; ***van Rooij, 2019***; ***van Rooij and Baggio, 2020***) and exploration of model spaces (***Devezer et al., 2019***; ***MacEachern and Van Zandt, 2019***) to conducting qualitative exploratory studies (***Reiter, 2017***) and designing exploratory experiments (***Arabatzis, 2013***; ***Waters, 2007***), and finally to exploratory data analysis (***Behrens, 1997***; ***Gelman, 2003***; ***Tukey, 1980***).

This process of hypothesis generation is notoriously hard to formalize, as ***Russell*** (***1945***, p.544) so clearly laid out:

> As a rule, the framing of hypotheses is the most difficult part of scientific work, and the part where great ability is indispensable. So far, no method has been found which would make it possible to invent hypotheses by rule. Usually some hypothesis is a necessary preliminary to the collection of facts, since the selection of facts demands some way of determining relevance. Without something of this kind, the mere multiplicity of facts is baffling.

Informally, hypothesis generation requires creativity, flexibility, and open-mindedness to allow for ideas to emerge (***Stebbins, 2001***; ***Swedberg, 2018***). The inferential approach employed during exploration cannot be described as deduction or induction since it requires adding something new to known facts. This process of generating explanatory hypotheses is known as *abduction proper*[9] (***Peirce, 1974***), which involves studying the facts and generating a theory to explain them (***Peirce, 1974***, p.90). Abduction proper requires scientists to absorb and digest all known facts about a phenomenon, mull them over, use introspection and common sense (***Good, 1983***), evaluate them against their background knowledge (***van Rooij and Baggio, 2020***), and add something as of yet unknown, with the intention of providing new insight or understanding that would not have been possible without abduction (***Peirce, 1974***). Hypothesis generation, therefore, cannot be reduced down to formal statistical inference, whose methods are deductively derived and used inductively

[9] Abductive inference involves both the process of making inference to the best explanation based on a set of candidate hypotheses and the process of generating that set of hypotheses. The latter process, which is of interest to our discussion, is specifically known as *abduction proper* (***Blokpoel et al., 2018***; ***van Rooij and Baggio, 2020***). Abduction proper is then a way to meaningfully reduce the search space for possible hypotheses.

in application. In fact, meticulous exploration via abduction proper would improve our statistical inference by facilitating the first two conditions mentioned in ***Box 1*** by constraining our search space in a theoretically meaningful fashion.

That said, exploratory data analysis (EDA) can be instrumental in hypothesis generation. ***Tukey*** (***1980***) suggests that EDA is not a bundle of formal inferential techniques and that it requires extensive use of data visualization with a flexible approach. EDA is usually an iterative process of model specification, residual analysis, examination of assumptions, and model respecification (***Behrens, 1997***; ***MacEachern and Van Zandt, 2019***) to find patterns and reveal data structure. If inferential statistics are employed for the purposes of data exploration, we can prioritize minimizing the probability of failing to reject a false null hypothesis (***Goeman et al., 2011***; ***Jaeger and Halliday, 1998***) as opposed to minimizing false positives because priority is given to not missing true discoveries. Nonetheless, other methods than hypothesis testing are often more closely associated with EDA due to their flexibility in revealing patterns, such as graphical evaluation of data (***Behrens, 1997***; ***Tukey, 1980***), exploratory factor analysis (***Behrens, 1997***), principal components regression (***Massy, 1965***), and Bayesian methods to generate EDA graphs (***Gelman, 2003***, ***2004***).

Whichever method is selected for EDA; however, it needs to be implemented rigorously to maximize the probability of true discoveries while minimizing the probability of false discoveries. As ***Behrens*** (***1997***, p.134) observes:

> A researcher may conduct an exploratory factor analysis without examining the data for possible rogue values, outliers, or anomalies; fail to plot the multivariate data to ensure the data avoid pathological patterns; and leave all decision making up to the default computer settings. Such activity would *not* be considered EDA because the researcher may be easily misled by many aspects of the data or the computer package. Any description that would come from the factor analysis itself would rest on too many unassessed assumptions to leave the exploratory data analyst comfortable.

The implication is that using “wonky” statistics cannot be a recommended practice for data exploration. The reason is that by repeatedly misusing statistical methods, it is possible to generate an infinite number of patterns from the same data set but most of them will be what ***Good*** (***1983***, p.290) calls a *kinkus*—“a pattern that has an extremely small prior probability of being potentially explicable, given the particular context”. If the process of hypothesis generation yields too many such kinkera (plural of kinkus), it can neither be considered a proper application of abduction principle nor would serve the ultimate goal of exploratory research: making true discoveries. Relying on statistical abuse in the name of scientific discovery will easily lead to well-known statistical problems such as increasing false positives by multiple hypothesis testing (***Benjamini and Hochberg, 1995***) or by failing to use proper conditioning.

If exploratory research needs to satisfy a certain level of rigor to be effective, what criteria should we use to assess its quality? Since the process of exploration is elusive and informal, it may not be possible to derive some minimum standards all exploratory studies need to meet. Nonetheless some desirable qualities can be inferred from successful implementation of exploratory approaches in different fields. (1) As suggested by Russell’s quote, exploration needs to start with subject matter expertise or theoretical background, and hence, cannot be decontextualized, free of theory, or completely dictated by the data (***Behrens, 1997***; ***Blokpoel et al., 2018***; ***Good, 1983***; ***van Rooij and Baggio, 2020***; ***Reiter, 2017***; ***Waters, 2007***). (2) The key for running successful exploratory studies is the richness of data (***Reiter, 2013***). Random data sets that are uninformative about the area to be explored will likely not yield important discoveries. (3) Exploration requires robust methods that are insensitive to underlying assumptions (***Behrens, 1997***). As such, rather than misusing or abusing standard procedures for inferential statistics, using robust approaches such as multiverse analysis (***Steegen et al., 2016***) or metastudies (***Baribault et al., 2018***) could be more appropriate for exploration purposes. (4) Exploratory work needs to be done in a structured, systematic, honest, and transparent manner using a deliberately chosen methodology appropriate for the task (***Lee*** *et al., 2019*; *Reiter, 2013*).

The above discussion should make two points clear, regarding Claim 3: First, exploratory research cannot be reduced to exploratory data analysis and thereby to the absence of a preregistered data analysis plan, and second, when exploratory data analysis is used for scientific exploration, it needs rigor. Describing exploratory research as though it were synonymous with or accepting of "wonky" procedures that misuse or abuse statistical inference not only undermines the importance of systematic exploration in the scientific process but also severely handicaps the process of discovery.

Conclusion

Our call for rigor and nuance encompasses all claims regarding scientific practice and policy changes. Rigor requires attention to detail, precision, clarity in statements and methods, and transparency. Nuance necessarily means moving away from speculative, sweeping claims and not losing sight of the context of inference. Simple fixes to complex scientific problems rarely exist. Simple fixes motivated by speculative arguments, lacking rigor and proper scientific support might appear to be legitimate and satisfactory in the short run, but may prove to be counter-productive in the long run. It is instructive to remember how taking $p < 0.05$ as a sign of scientific relevance or even truth has proved to be detrimental to scientific progress.

Recent developments in methodological reform have already been impactful in inducing behavioral and institutional changes. However, as *Niiniluoto* (*2019*) suggests, impact of research "only shows that it has successfully 'moved' the scientific community in some direction. If science is goal-directed, then we must acknowledge that movement in the wrong direction does not constitute progress." Advancing robust methodological tools, carefully documenting limitations of these tools, providing precise, unambiguous definitions of concepts these tools rely on, and stating our claims about these tools with transparency and under clearly stated assumptions would aid us in making *positive* contributions to scientific progress, rather than just having impact.

Acknowledgements

The authors thank Iris van Rooij for her generous and insightful feedback on a previous draft of the manuscript, and John T. Ormerod for the discussions that informed some of our ideas. BD and EOB were supported by NIGMS of the NIH under award #P20GM104420. JV was supported by grants #1850849 and #1658303 from the National Science Foundation's Cognitive Neuroscience panel.

References

Arabatzis T. Experiment. In: *The Routledge Companion to Philosophy of Science* Routledge; 2013.p. 223–234.

Baribault B, Donkin C, Little DR, Trueblood JS, Oravecz Z, Van Ravenzwaaij D, White CN, De Boeck P, Vandekerckhove J. Metastudies for robust tests of theory. Proceedings of the National Academy of Sciences. 2018; 115(11):2607–2612.

Barwich AS. The Value of Failure in Science: The Story of Grandmother Cells in Neuroscience. Frontiers in neuroscience. 2019; 13:1121.

Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circulation research. 2015; 116(1):116–126.

Behrens JT. Principles and procedures of exploratory data analysis. Psychological Methods. 1997; 2(2):131.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995; 57(1):289–300.

Berry AC. The accuracy of the Gaussian approximation to the sum of independent variates. Transactions of the american mathematical society. 1941; 49(1):122–136.

Blokpoel M, Wareham T, Haselager P, Toni I, van Rooij I. Deep analogical inference as the origin of hypotheses. Journal of Problem Solving. 2018; 11(1):1–24.

Bowers JS, Davis CJ. Bayesian just-so stories in psychology and neuroscience. Psychological bulletin. 2012; 138(3):389.

Box GE. Science and statistics. Journal of the American Statistical Association. 1976; 71(356):791–799.

Braude SE. ESP and psychokinesis: A philosophical examination. Universal-Publishers; 2002.

Buzbas E. Need of mathematical formalism in proposals for robust modeling. Computational Brain & Behavior. 2019; 2(3-4):197–199.

Casella G, Berger R. Statistical inference second edition; 2002.

Chakravartty A. Scientific Realism. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*, summer 2017 ed. Metaphysics Research Lab, Stanford University; 2017.

Chambers CD, Dienes Z, McIntosh RD, Rotshtein P, Willmes K. Registered reports: realigning incentives in scientific publishing. Cortex. 2015; 66:A1–A2.

Chang H. Inventing temperature: Measurement and scientific progress. Oxford University Press; 2004.

Chater N, Oaksford M, et al. The probabilistic mind: Prospects for Bayesian cognitive science. OUP Oxford; 2008.

Claesen A, Gomes SLBT, Tuerlinckx F, et al. Preregistration: Comparing Dream to Reality. . 2019; https://osf.io/n3axs/download.

Clarke JL, Clarke B, Yu CW, et al. Prediction in M-complete Problems with Limited Sample Size. Bayesian Analysis. 2013; 8(3):647–690.

Darnieder WF. Bayesian methods for data-dependent priors. PhD thesis, The Ohio State University; 2011.

Dennis B, Ponciano JM, Taper ML, Lele SR. Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. Frontiers in Ecology and Evolution. 2019; 7:372.

Devezer B, Nardin LG, Baumgaertner B, Buzbas EO. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PloS one. 2019; 14(5):e0216125.

Donoho DL, Maleki A, Rahman IU, Shahram M, Stodden V. Reproducible research in computational harmonic analysis. Computing in Science & Engineering. 2008; 11(1):8–18.

Dvoretzky A, Kiefer J, Wolfowitz J, et al. Sequential decision problems for processes with continuous time parameter. Testing hypotheses. The Annals of Mathematical Statistics. 1953; 24(2):254–264.

Esseen CG. On the Liapunov limit error in the theory of probability. Ark Mat Astr Fys. 1942; 28:1–19.

Flake JK, Fried EI, Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. PsyArXiv; 2019. psyarxiv.com/hs7wm, doi: 10.31234/osf.io/hs7wm.

Flis I. Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. Theory & Psychology. 2019; 29(2):158–181.

Gelman A. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. International Statistical Review. 2003; 71(2):369–382.

Gelman A. Exploratory data analysis for complex models. Journal of Computational and Graphical Statistics. 2004; 13(4):755–779.

Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness. 2012; 5(2):189–211.

Gelman A, Loken E. 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. . 2013; https://osf.io/n3axs/download.

Gervais WM, Practical Methodological Reform Needs Good Theory. PsyArXiv; 2020. psyarxiv.com/jcs6e, doi: 10.31234/osf.io/jcs6e.

Goeman JJ, Solari A, et al. Multiple testing for exploratory research. Statistical Science. 2011; 26(4):584–597.

Good IJ. The philosophy of exploratory data analysis. Philosophy of science. 1983; 50(2):283–295.

Guest O, Martin AE, How computational modeling can force theory building in psychological science. PsyArXiv; 2020. psyarxiv.com/rybh9, doi: 10.31234/osf.io/rybh9.

Hacking I. The Self-Vindication of the Laboratory Sciences. In: Pickering A, editor. *Science as Practice and Culture* University of Chicago Press; 1992.p. 29–64.

Hardwicke TE, Mathur MB, MacDonald K, Nilsonne G, Banks GC, Kidwell MC, Hofelich Mohr A, Clayton E, Yoon EJ, Henry Tessler M, et al. Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. Royal Society open science. 2018; 5(8):180448.

Hayes BK, Banner S, Forrester S, Navarro DJ. Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. Cognitive Psychology. 2019; 113. doi: https://doi.org/10.1016/j.cogpsych.2019.05.003.

Heesen R. Why the reward structure of science makes reproducibility problems inevitable. The Journal of Philosophy. 2018; 115(12):661–674.

Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behavioral and brain sciences. 2010; 33(2-3):61–83.

Ioannidis JP. Meta-research: Why research on research matters. PLoS biology. 2018; 16(3):e2005468.

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, et al. Repeatability of published microarray gene expression analyses. Nature genetics. 2009; 41(2):149.

Jaeger RG, Halliday TR. On confirmatory versus exploratory research. Herpetologica. 1998; p. S64–S66.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nature neuroscience. 2009; 12(5):535.

Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MA, Argamon SE, Baguley T, Becker RB, Benning SD, Bradford DE, et al. Justify your alpha. Nature Human Behaviour. 2018; 2(3):168–171.

Lakens D, Evers ER. Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. Perspectives on Psychological Science. 2014; 9(3):278–292.

Lee MD, Criss AH, Devezer B, Donkin C, Etz A, Leite FP, Matzke D, Rouder JN, Trueblood JS, White CN, Vandekerckhove J. Robust modeling in cognitive science. Computational Brain & Behavior. 2019; 2:141–153. doi: 10.1007/s42113-019-00029-y.

Lehmann EL, Casella G. Theory of point estimation, 2nd Ed. Springer-Verlag: New York; 1998.

Leonelli S. Rethinking Reproducibility as a Criterion for Research Quality. In: *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise* Emerald Publishing Limited; 2018. p. 129–146.

Lindley DV. Philosophy of statistics. Journal of The Royal Statistical Society: Series D. 2000; 49(3):293–337.

Lindley DV. Understanding Uncertainty. Wiley; 2006.

Lindsay DS, Simons DJ, Lilienfeld SO. Research preregistration 101. APS Observer. 2016; 29(10).

Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017; 355(6325):584–585.

MacEachern SN, Van Zandt T. Preregistration of modeling exercises may not be useful. Computational Brain & Behavior. 2019; 2(3-4):179–182.

Massy WF. Principal components regression in exploratory statistical research. Journal of the American Statistical Association. 1965; 60(309):234–256.

McIntosh, Robert D . Exploratory reports: A new article type for Cortex. Cortex. 2017; 96:A1–A4. https://doi.org/10.1016/j.cortex.2017.07.014.

McNutt M, Journals unite for reproducibility. American Association for the Advancement of Science; 2014.

McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. The American Statistician. 2019; 73(sup1):235–245.

Meng XL, et al. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics. 2018; 12(2):685–726.

Mukhopadhyay N. MVUE for the mean with one observation: Normal with same mean and variance. The American Statistician. 2006; 60(1):71–74.

Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, du Sert NP, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JP. A manifesto for reproducible science. Nature Human Behaviour. 2017; 1: 0021.

Muthukrishna M, Henrich J. A problem in theory. Nature Human Behaviour. 2019; p. 1.

Navarro DJ. Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. Computational Brain & Behavior. 2019; 2(1):28–34.

Niiniluoto I. Scientific Progress. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*, winter 2019 ed. Metaphysics Research Lab, Stanford University; 2019.

Nissen SB, Magidson T, Gross K, Bergstrom CT. Publication bias and the canonization of false facts. Elife. 2016; 5:e21451.

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proceedings of the National Academy of Sciences. 2018; p. 201708274.

Nosek BA, Lakens D. A method to increase the credibility of published results. Social Psychology. 2014; 45(3):137–141.

Nosek BA, Spies JR, Motyl M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science. 2012; 7(6):615–631.

Oaksford M, Chater N. Rational models of cognition. Oxford University Press Oxford; 1998.

Oberauer K, Lewandowsky S. Addressing the theory crisis in psychology. Psychonomic bulletin & review. 2019; 26(5):1596–1618.

Open Science Collaboration. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. Perspectives on Psychological Science. 2012; 7:657–660.

Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349(6251):aac4716.

Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives on Psychological Science. 2012; 7(6):528–530.

Peirce CS, The Collected Papers of Charles S. Peirce. Eds. C. Hartshorne, P. Weiss, and A.W. Burks. Cambridge, MA: Harvard University Press; 1974.

Penders B, Holbrook JB, de Rijcke S. Rinse and repeat: Understanding the value of replication across different ways of knowing. Publications. 2019; 7(3):52.

Reiter B. The epistemology and methodology of exploratory social science research: Crossing Popper with Marcuse. Government and International Affairs Faculty Publications. 2013; (Paper 99).

Reiter B. Theory and methodology of exploratory social science research. International Journal of Science and Research Methodology. 2017; 5(4):129.

van Rooij I, Psychological science needs theory development before preregistration; 2019. https://featuredcontent.psychonomic.org/psychological-science-needs-theory-development-before-preregistration/.

van Rooij I, Baggio G, Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. PsyArXiv; 2020. psyarxiv.com/7qbpr, doi: 10.31234/osf.io/7qbpr.

Rubin M. An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. Review of General Psychology. 2017; 21(4):321–329.

Russell B. A History of Western philosophy: Collectors edition. NY: Simon and Schuster; 1945. Fourth Printing.

Schooler JW. Metascience could rescue the 'replication crisis'. Nature. 2014; 515(7525):9–9.

Shiffrin RM, Börner K, Stigler SM. Scientific progress despite irreproducibility: A seeming paradox. Proceedings of the National Academy of Sciences. 2018; 115(11):2632–2639.

Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science. 2011; 22(11):1359–1366.

Simons DJ. The value of direct replication. Perspectives on Psychological Science. 2014; 9(1):76–80.

Stebbins RA. Exploratory research in the social sciences, vol. 48. Sage; 2001.

Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science. 2016; 11(5):702–712.

Swedberg R. On the uses of exploratory research and exploratory studies in social science. Canada: Connell University. 2018; .

Szollosi A, Donkin C. Arrested theory development: The misguided distinction between exploratory and confirmatory research. . 2019; https://psyarxiv.com/suzej/.

Szollosi A, Kellen D, Navarro DJ, Shiffrin R, van Rooij I, Van Zandt T, Donkin C. Is preregistration worthwhile? Trends in Cognitive Sciences. 2019; .

Tukey JW. We need both exploratory and confirmatory. The American Statistician. 1980; 34(1):23–25.

van't Veer AE, Giner-Sorolla R. Pre-registration in social psychology—A discussion and suggested template. Journal of Experimental Social Psychology. 2016; 67:2–12.

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. An agenda for purely confirmatory research. Perspectives on Psychological Science. 2012; 7:632–638.

Waters CK. The nature and context of exploratory experimentation: An introduction to three case studies of exploratory research. History and Philosophy of the Life Sciences. 2007; p. 275–284.

Appendix 1

Notation, assumptions, definitions

Regularity conditions and notation. We assume some regularity conditions for all random variables:

- Distribution functions $F \equiv F(w) = \mathbb{P}(W \leq w)$, are absolutely continuous and non-degenerate, endowed with the density function $f(w) = dF(w)/dw$.
- $\{\mathbb{E}(|W|^n) < \infty, \forall n, \}$, $\mathbb{E}(W^2) > 0$, where $\mathbb{E}(W) = \int_{-\infty}^{\infty} f(w)dw$, and $\mathbb{V}(W) = \mathbb{E}(W^2) - [\mathbb{E}(W)]^2$.
- We make frequent use of the indicator function: $\mathbf{I}_{\{A\}} = 1$ if A, and 0 otherwise.

Assumptions of idealized study. We build on the notion of *idealized study* (***Devezer et al., 2019***), obeying the following assumptions:

A1. There exists a true probability model M_T, completely specified by F_T of random variable X, which is the observable for a phenomenon of interest.

A2. Some known background knowledge K partially specifies M_T up to property $\theta \in \Theta$, which denotes unknown and unobservable components of M_T. For notational economy, K is often dropped, with the understanding that all statements are conditional on K.

A3. A statement that is in principle testable via statistical inference using a simple random and finite sample $\mathbf{X_n} = (X_1, X_2, \cdots, X_n)$, where $X_i \sim F_T$ is made about θ.

A4. Candidate mechanisms M_i, inducing distribution functions F_i are formulated.

A5. A fixed and known function S, is used to extract the information in $\mathbf{X_n}$ pertinent to M_i. S evaluated at $\mathbf{X_n}$ returns $\mathbf{S_n}$, with non-degenerate distribution function $\mathbb{P}(\mathbf{S_n} \leq s)$.

A6. Formal statistical inference returns a *result* $\{R = d(\mathbf{S_n}, c),\ R \subset \Theta\}$, where c is a user-defined known quantity, and $d(\cdot, \cdot)$ is a fixed and known non-constant decision function which formalizes the statistical inference (by inducing a frequency assessment for a result).

Definitions.

- $\xi = (M_i, \theta, \mathbf{X_n}, S, K, d)$ is an idealized study.
- $\xi^{(i)}$ which differs from ξ only in $\mathbf{X_n}^{(i)}$ generated independently from $\mathbf{X_n}$, is a replication experiment.

Appendix 2

Relationship between true results and reproducible results

Proposition 1. Let R_o be a result. If $\mathbf{I}_{\{R^{(i)}=R_o|R_o\}} = 1$ we say that R_o is reproduced by $R^{(i)}$. Else, we say that R_o failed to reproduce by $R^{(i)}$.

1.1 Conditional on R_o, the relative frequency of reproduced results $\phi_N \to \phi \in [0,1]$, as $N \to \infty$. Further, $\phi = 1$ only trivially.

1.2 There exists true results $R_o = R_T$, whose true reproducibility rate ϕ_T is arbitrarily close to 0.

1.3 There exists false results $R_o = R_F$ whose true reproducibility rate ϕ_F is arbitrarily close to 1.

Proof. $R^{(i)}$ are $\{0,1\}$ exchangeable random variables since $\xi^{(i)}$ are invariant under permutation of labels. By De Finetti's representation theorem for $\{0,1\}$ variables, there exists a ϕ such that $R^{(i)}$ are conditionally independent given ϕ. For a finite subsequence $R^{(1)}, R^{(2)}, \cdots, R^{(N)}$, and the relative frequency of reproduced results defined by $\phi_N = N^{-1}\sum_{i=1}^{N} \mathbf{I}_{\{R^{(i)}=R_o|R_o\}}$, we have $\lim_{N\to\infty} \phi_N = \phi$, almost surely by the Strong Law of Large Numbers.

By definition $\phi \geq 0$, since it is a probability. It follows by contradiction that $\phi = 1$ only in trivial cases: Assume $\phi = 1$. We have $\phi = \mathbb{E}(\mathbf{I}_{\{R=R_o|R_o\}}) = \mathbb{P}(R = R_o|R_o) = 1$, which implies that $\mathbf{I}_{\{R^{(i)}=R_o|R\}} = 1$ for all i. Therefore, $d(\mathbf{S_n}, c)$ in **A6** must return a singleton (R_o) for all values of $\mathbf{S_n}$. This can happen in three ways: $\mathbf{X_n}$ is non-stochastic, which contradicts **A1**, or $\mathbf{S_n}$ is non-stochastic, which contradicts **A5**, or R_o is not a proper subset of Θ, which contradicts **A6**.

The truth of 1.2 implies 1.3 and vice versa: if a result is not true, then it is false because $\phi_T + \phi_F = 1$. To see that ϕ_T can be arbitrarily close to zero (and ϕ_F arbitrarily close to 1), fix R_T. Choose S such that $d(\mathbf{S_n}, c)$ does not return R_T with probability $1 - \phi_T$. A simple example is a biased estimator of a parameter in a probability distribution. We also note that by Proposition 1.1, ϕ_T must have positive probability for every point on its support for some ξ, which includes values arbitrarily close to 0.

Remark. ϕ_N should not be misinterpreted as an estimator with less than ideal properties. Quite the opposite: By Central Limit Theorem, $(\phi_N - \phi)/[\phi(1-\phi)]$ converges to the standard normal distribution and ϕ_N has excellent statistical properties as an estimator of ϕ (***Dvoretzky et al., 1953***; ***Berry, 1941***; ***Esseen, 1942***).

Remarks for some cases in *Box 1*.

Bullet 1. Fix c such that $\epsilon(c) > 0$. Consider a model selection problem where $d(\mathbf{S_n}, c)$ returns a model between two candidate models M_1 and M_2, which are different from the true model M_T. The selected model M_1 or M_2 is false with probability 1 independent of how well S performs. Yet, M_1 and M_2 can be chosen so that the divergence or metric on which the model selection measure S is based satisfy selecting M_1 over M_2 with probability $\phi_F = 1 - \epsilon(c)$.

Bullet 3. Let θ_o be the parameter of interest of F_T and θ_o' be nuisance parameters. Assume that the true value of θ_o is in Θ. We let $d(\mathbf{S_n}, c)$ to return $\mathbf{S_n}$ as an estimator of parameter θ_o where $\mathbb{E}(\mathbf{S_n})$ is not equal to the true value. $\mathbf{S_n}$ is often a pivotal quantity. We consider two cases: If further, $\mathbf{S_n}$ is a statistic then it is ancillary for θ_o. Let $\mathbb{V}(\mathbf{S_n}) = \epsilon(c)^2$. By Chebychev's inequality we have $|\mathbf{S_n} - \mathbb{E}(\mathbf{S_n})| \leq \epsilon(c)$ with probability 1. Thus, the result returned is false and $\phi_F > 1 - \epsilon(c)$. Else if, $\mathbf{S_n}$ is not a statistic, but depends on θ_o', choosing the value of θ_o' suitably yields the result.

Appendix 3

Conditional analysis

Definition. Let $S \sim \mathbb{P}(S|\theta)$ be a test statistic such that it is: 1) a function of an unbiased estimator of θ, and 2) fixed prior to seeing the data. Let $U \sim \mathbb{P}(U|\theta)$ be a statistic obtained from the data, after seeing the data. If U is complete sufficient for θ, it is denoted by U_s, and if U is ancillary for θ, it is denoted by U_a.

Proposition 2.1. Let $S' = \mathbb{E}(S|U_s)$. For an upper tail test, define $\alpha = \mathbb{P}(S \geq s_\alpha|H_o) = \mathbb{P}(S' \geq s'_\alpha|H_o)$. Then, $s_\alpha \geq s'_\alpha$ and $\mathbb{P}(S' \geq s_\alpha|h_o) < \alpha$. Parallel arguments hold for lower and two tail tests.

Proof 2.1. By Chebychev's inequality we have $\mathbb{P}\{|S-\theta| \leq \sqrt{\mathbb{V}(S)}/\alpha\} \leq \alpha^2$ and $\mathbb{P}\{|S'-\theta| \leq \sqrt{\mathbb{V}(S')}/\alpha\} \leq \alpha^2$, where $\mathbb{V}(S)/\alpha$ and $\mathbb{V}(S')/\alpha$ are critical values of the two tests. We have $0 \leq \mathbb{V}(S') \leq \mathbb{V}(S)$ by Rao-Blackwell Theorem (***Casella and Berger, 2002***, p.342). It follows that $s_\alpha \geq s'_\alpha$ and $\mathbb{P}(S' \geq s_\alpha|H_o) < \alpha$.

Proposition 2.2. Let $H_o : \theta \in \Theta_o$ such that $\Theta_o = g(U_a)$, where g is a known function and U_a is a function of the data. Then, the upper tail test $\mathbb{P}(\mathbf{S_n} \geq s|H_o) \leq \alpha$ is a valid level α test. Parallel arguments hold for lower and two-tailed tests.

Proof 2.2. By ancillarity we have $\mathbb{P}(U_a|\theta) = \mathbb{P}(U_a)$, implying $\mathbb{P}(U_a|\mathbf{S_n}, \theta) = \mathbb{P}(U_a|\mathbf{S_n})$. The sampling distribution of S given θ can be written as:

$$\mathbb{P}(\mathbf{S_n}|\theta) = \mathbb{P}(\mathbf{S_n}|U_a, \theta)\mathbb{P}(U_a|\theta)/\mathbb{P}(U_a|\mathbf{S_n}, \theta) = \mathbb{P}(\mathbf{S_n}|U_a, \theta)\left[\mathbb{P}(U_a)/\mathbb{P}(U_a|\mathbf{S_n})\right],$$

where the second equality follows by substituting for $\mathbb{P}(U_a|\theta)$ and $\mathbb{P}(U_a|\mathbf{S_n}, \theta)$. The term within the brackets is independent of θ, so that a test based on $\mathbf{S_n}$, and a test based on $\mathbf{S_n}|U_a$ yield the same result. Therefore, using U_a to inform H_o does not affect the validity of the test.

Remarks for some cases in *Box 2*.

Left block, 1st row, 1st column. If S is not complete sufficient and U_s is minimally sufficient, then for an upper tail test, then $\mathbb{P}(S \geq s|U_s, H_a) \geq \mathbb{P}(S \geq s|H_a)$ for some s is possible, where H_a is the alternative hypothesis. That is, the test conditional on a statistic from prior analysis can be more powerful. Parallel arguments hold for lower and two-tailed tests.

Left block, 1st row, 2nd column. Rao-Blackwellization guarantees that $\mathbb{V}(S|U) \leq \mathbb{V}(S)$. See ***Figure 3*** for an example.

Right block, 1st row, 1st column. Conditioning on a decision based on user defined criterion might alter the support of the sampling distribution of S. In these cases, conditioning is necessary for a valid test. See ***Figure 4*** for an example.

Right block, 3rd row. U_a and S might be dependent (see ***Casella and Berger*** (***2002***, p.284–285) for an example). Applying a decision with a user defined criterion and U_a might affect the support of the sampling distribution of S. In these cases, conditioning on the decision regarding U_a is necessary for a valid test.

Appendix 4

Details of models used in Figures

Figure 1A. The simple linear regression model is given by $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where the errors obey Gauss-Markov conditions: $\mathbb{E}(\epsilon_i) = 0,\ \mathbb{V}(\epsilon_i) = \sigma^2_\epsilon,\ \forall i,$ and $Cov(\epsilon_i, \epsilon_j) = 0,\ \forall (i, j)$. The x_i are assumed fixed and known. The errors $\epsilon_i \sim \mathrm{Nor}(0, \sigma_\epsilon)$. The measurement error model is the true model when there is stochastic measurement error in x making it a random variable X. We assume $X_i = x_i + \eta_i$, where $\eta_i \sim \mathrm{Nor}(0, \sigma_\eta)$. The assumed (incorrect) model under which inference is performed is the simple linear regression model, which corresponds to $\sigma_\eta = 0$. Specific values used in the plot are: $x \sim \mathrm{Unif}(0, 10)$, $\beta_0 = 2$, $\beta_1 \in \{2, 20\}$, $\sigma_\epsilon = 1$, $\sigma_\eta \in \{0.01, 0.02, \cdots, 1.0\}$, and the sample size is 50.

Figure 2. The model is the same as in ***Figure 1***A, except that the values plotted are $\sigma_\eta \in \{0.01, 0.02, \cdots, 10\}$, and the true value is $\beta_1 = 20$. The vertical axis shows the distance between $\hat{\beta}_1$ and β_1.

Figure 3. This example is from ***Mukhopadhyay*** (***2006***). Let $X \sim \mathrm{Nor}(\mu, \mu),\ \mu > 0$. The data is a single observation X_1, which is an unbiased estimator of μ. Using Rao-Blackwellization, $|X_1|$ is a sufficient statistic for μ and the mean of X_1 conditional on the value $|X_1|$ improves the power of a test while maintaining its validity.

Figure 4. Let $X_i \sim \mathrm{Nor}(\mu_X, \sigma^2_X)$ and $Y_i \sim \mathrm{Nor}(\mu_Y, \sigma^2_Y),\ i = 1, 2, \cdots, n$ independent samples with known population variances σ^2_X and σ^2_Y. Let the null and the alternative hypotheses be $H_o : \mu_X = \mu_Y,\quad H_a : \mu_X > \mu_Y$ respectively. An appropriate test statistic for level $\alpha = \mathbb{P}(Z \geq z_\alpha | H_o)$ test is the z-score: $Z = (\bar{X} - \bar{Y})/(\sigma_X/\sqrt{n} + \sigma_Y/\sqrt{n})$, which follows a standard normal distribution under H_o. Assume we perform the test is *only if* we observe $\bar{X} - \bar{Y} > 0$. Define: $U(c) = \bar{X} - \bar{Y}$ if $\bar{X} > \bar{Y}$, and $U(c) = 0$ otherwise. Here, $U(c)$ is the statistic $U = \bar{X} - \bar{Y}$ whose nonzero values are constrained by the user defined criterion c, given by $\bar{X} > \bar{Y}$. The conclusion of the test depends on $U(c)$ since when $\bar{X} > \bar{Y}$, larger the value of U, larger the value of Z. The distribution of the conditional test statistic $Z|U(c), H_o$ is not standard normal and therefore the level of the test is not necessarily α for the critical value z_α, as is with the test statistic Z. However, if the distribution of $Z|U(c), H_o$ is available then the correct critical value, can be chosen to perform a level α test. We let $W = Z\mathbf{I}_{\{\bar{X}>\bar{Y}\}}$, the standard normal random variable with support on non-negative real line (folded at zero), properly normalized. This is known as the standard half-normal distribution.

We see that $\mathbb{P}(W > z_\alpha | H_o) = 2\alpha$. For the level of the conditional test to be α, we adjust the critical value as $z^* = z_{\alpha/2}$ and have $\mathbb{P}(W > z^* | H_o) = \alpha$.

Appendix 5

A simulation-based method to sample the conditional distribution of the test statistic

If the distribution of the conditional test statistic under H_o is not available as a closed form solution, an appropriate simulation-based method can be used to sample it. Here, we give an example for the unconditional test statistic $\mathbf{S_n}$ with distribution $\mathbb{P}(\mathbf{S_n}|H_o)$, where $H_o : \theta = \theta_o$. We aim to sample M values from the conditional distribution of $\mathbf{S_n}|U(c), H_o$ where $U(c)$ is a statistic obtained from the data constrained by a user defined criterion c.

Algorithm.

```
Initialize: Set $M$ (large desired number), and $i = 0$.
Begin While $i < M$, do:
    1. Simulate $X_j \sim \mathbb{P}(X_i|\theta_o),\ j = 1, 2, \cdots, n$ independently of each other. Set $\mathbf{X_n}^{(i)} = (X_1, X_2, \cdots, X_n)$.
    2. Calculate $\mathbf{S_n}^{(i)} = S(\mathbf{X_n}^{(i)})$ and $U^{(i)} = U(\mathbf{X_n}^{(i)})$.
    3. If $U^{(i)}$ obeys $c$ accept $\mathbf{S_n}^{(i)}$ as a draw from the distribution of the conditional test statistic and set $i = i + 1$. Else discard $(\mathbf{X_n}^{(i)}, \mathbf{S_n}^{(i)}, U^{(i)})$.
End While
```

The accepted values $\mathbf{S_n}^{(1)}, \mathbf{S_n}^{(2)}, \cdots, \mathbf{S_n}^{(M)}$ is a sample from the distribution $\mathbf{S_n}|U(c), H_o$. A valid level α test can be built by finding the relevant sample quantile. This method is precise up to a Monte Carlo error which vanishes as $M \to \infty$.

Sometimes it may not be possible to condition on the exact value of statistic $U(c)$, for example when c involves an equality (instead of inequality) and U is continuous random variable. In these cases, the algorithm given above can be modified to build an approximate test using an approximate simulation method such as a likelihood free method. The error rates in approximation can be estimated by simulation.