Received: 4 June 2018 | Revised: 30 October 2018 | Accepted: 30 December 2018

DOI: 10.1111/gbi.12333

PERSPECTIVE

WILEY gebiology

Statistical inference and reproducibility in geobiology

1 | INTRODUCTION

The late, great Karl Turekian would often joke about the number of new data points required for a geochemical paper. The answer was one: When combined with a previously published data point, a "best-fit" line could be drawn between the two points, and the slope of the line calculated, thereby giving rate. The joke had its roots in a seminar Dr. Turekian gave far earlier in his career, where the entire talk focused on the first data point of a novel measurement (a hard-won data point, but only one nonetheless). When an apparently exasperated audience member pointed this out, Turekian reportedly replied "well it's one more than anyone else has!" (Thiemens, Davis, Grossman, & Colman, 2013). This anecdote raises two important points about our field: First, establishing new analytical techniques is laborious, expensive, and requires considerable vision and skill, and second, at the nascent stages of any new record, the number of observations will be small.

Geobiology is now at the point where some mature data records (for instance $\delta^{13}C$ or $\delta^{34}S$) have had thousands or tens of thousands of data points generated. There also exist millions of individual gene sequences from environmental genomics surveys. In contrast, emerging proxies (for instance, selenium isotopes, I/Ca ratios, carbonate "clumped" isotopes), or biomarker studies and technical geomicrobiological experiments still face the issue Turekian described and have only a handful of measurements or are transitioning toward larger datasets. The question remains how best to interpret data-rich records on the one hand, and scattered, hard-won data points on the other hand. Here, using examples from within geobiology as well as the development of other fields such as ecology, psychology, and medicine, we argue that increased clarity regarding significance testing, multiple comparisons, and effect sizes will help researchers avoid false-positive inferences, better estimate the magnitude of changes in proxy records or experiments, and ultimately yield a richer understanding of geobiological phenomena.

2 | STATISTICS AND REPRODUCIBILITY IN OTHER FIELDS AND IN GEOBIOLOGY

We start by examining statistical practice and reproducibility outside of our field, before relating these broader themes back to geobiology. Science as a whole is currently described as facing a "crisis of reproducibility," with diverse studies in multiple fields failing to replicate published findings (see Baker, 2016 for a *Nature* survey of reproducibility across fields). The issue here is not technical reproducibility at the sample level (e.g., "If I put this same sample in a mass spectrometer, will I get the same result twice?") but rather at the level of the effects observed in the manuscript (e.g., a given treatment causes a specific outcome, or a predictor variable is correlated with a response variable). For example, independent efforts to reproduce "landmark papers" in cancer biomedical research by pharmaceutical companies Amgen (Begley & Ellis, 2012) and Bayer (Prinz, Schlange, & Asadullah, 2011) have only been able to replicate 11% and 20%–25% of published results, respectively. Similarly, a critical study of gene-by-environment (GxE) interactions in psychiatry found that only 27% of published replication attempts were positive (Duncan & Keller, 2011). Further, this small proportion of apparently positive replications was likely inflated; there were clear signatures of publication bias (the tendency to publish significant results more readily than non-significant results) among replication attempts in the GxE literature. In psychology, a large-scale replication effort found that only 39% of studies could be replicated (Open Science Collaboration, 2015), and similar problems plague much of neuroscience research (Button et al., 2013). The studies mentioned here are likely just the tip of the iceberg (Begley & Ioannidis, 2015; Ioannidis, 2005).

The causes of these low rates of replication are varied. There are likely different underlying causes for poor reproducibility across fields, and the severity of the problem undoubtedly varies as well. Our personal experience trying to implement published protocols without the highly detailed, laboratory-specific knowledge that is often omitted (generally for space reasons) from Materials and Methods sections suggests to us that at least some percentage of failed replications are caused by inadvertent methodological differences. Indeed, the Open Science Collaboration replication of psychology studies was attacked for poor adherence to original protocols (Gilbert, King, Pettigrew, & Wilson, 2016). In order to achieve precise replication of protocols, in extreme cases researchers have had to travel to other laboratories and work side-by-side to identify seemingly trivial methodological differences—vigorous stirring versus prolonged gentle shaking to isolate cells, for instance—that had an outsized effect on reproducibility (Hines, Su, Kuhn, Polyak, & Bissell, 2014; Lithgow, Driscoll, & Phillips, 2017). Methodological differences, though, can likely only explain a portion of the failed replication attempts. True scientific fraud is also something that makes headlines and erodes public trust, but again likely only accounts for a small proportion of replication failures (Bouter, Tijdink, Axelsen, Martinson, & ten Riet, 2016; Fanelli, 2009).

So, what causes such poor reproducibility? While recognizing again that the causes can vary across fields, clearly some of the most important factors are under-powered studies, a reliance on Null Hypothesis Significance Testing (NHST, e.g., $p < 0.05$) to determine "truth" or whether a paper should be published, and a lack of correction for implicit or explicit multiple comparisons (more broadly, "researcher degrees of freedom"). These issues were known within some fields, but were brought to more widespread attention through the publication of a provocative 2005 essay, "Why Most Published Research Findings Are False" (Ioannidis, 2005). Ioannidis identified the following causal factors as leading to the overall very low percentages of positive replication findings: (a) small sample sizes, (b) small effect sizes, (c) high numbers of tested relationships, (d) flexibility in designs and definitions, and what represents an unequivocal outcome, (e) conflicts of interest or prejudice within a field, thus increasing bias, and (f) "hot" fields of science, where more teams are simultaneously testing many different relationships. This tendency is exacerbated by the incentive structure for journals and authors to publish positive results and publish often.

Our field of geobiology is likely less besot by some of these issues related to, for instance, publication bias. The reason why is that data analysis is often accomplished through visual rather than statistical comparison; explicit *p*-values that would be used as the arbitrator of accept/reject decisions in other fields are not generated. Figure 1 documents the use of statistical testing in the journal *Geobiology* compared to a sister publication by Wiley, *Marine Ecology*. This comparison is not meant to single out *Geobiology* as a journal; we are certain the results would be similar in our other disciplinary journals. In this comparison, we considered the 100 most recent papers at the time of writing that reported new data (so review papers, commentaries, or papers that were descriptive in nature, such as describing stromatolite morphologies—basically any paper where no new numerical data were generated—were excluded). Papers were considered to have a "statistical analysis" simply if there was some effort to understand the possible influence of error and sampling on the precision of the inference, recognizing this can take many forms (e.g., formal NHST, bootstrapping, and Bayesian posterior probabilities). In the marine ecology journal, essentially every paper (97%) reporting new data used a statistical analysis (similar results were reported by Fidler, Burgman, Cumming, Buttrose, & Thomason, 2006, for other ecology journals). In *Geobiology*, the percentage with any sort of statistical analysis is far lower, at 38% (chi-squared test; $p = 2.0 \times 10^{-18}$). We recognize that some of this difference may be due to different data types and approaches, but our personal observation is that many studies in *Geobiology* are comparing groups of data visually rather than statistically. Further, the statistical analyses that are published in *Geobiology* are concentrated in molecular phylogenetic studies, where bootstraps or Bayesian posterior probabilities are commonly used to assess precision. Statistical testing is far less common in non-phylogenetic studies (e.g., many geomicrobiology, geochemistry, biomarker, and biomineralization studies). Comparing the most recent papers in *Geobiology* against the first 100 data-driven papers in the journal (first published in 2003) reveals

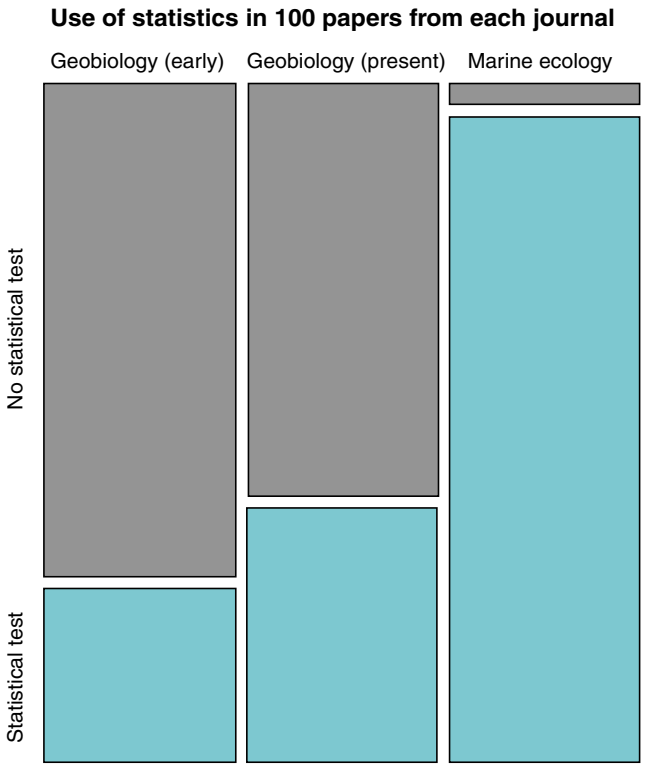

FIGURE 1 Use of statistical testing in the 100 most recent data-driven papers in *Geobiology* (present) versus *Marine Ecology*. The first 100 papers in *Geobiology* (early) were also compared. "Statistical tests" were broadly defined (for instance, bootstrapping, Bayesian approaches, etc., and not only formal Null Hypothesis Significance Testing). Papers without a "Materials and Methods" section or that did not present new numerical data (e.g., review/synthesis papers, modeling papers) were not included in the comparison

that the percentage with a statistical test has increased slightly (from 26% to 38%) but the difference between the two time intervals examined is not statistically significant (using $p < 0.05$ to declare statistical significance; *chi-sq* = 2.8, *df* = 1, $p = 0.095$).

We propose that increased emphasis on statistical testing as a critical step in data analysis is needed in the field of geobiology. Put simply, why build a scientific worldview based on studies where it has not been demonstrated that the observed differences are more than what would be expected from sampling variability? On the other hand, blind reliance on statistical testing will not be helpful either. As cogently argued by Fidler, Cumming, Burgman, and Thomason (2004), ecology as a field is mired in statistical fallacies: specifically, the erroneous beliefs that *p*-values are a direct index of effect size, and that the *p*-value represents the probability that the null hypothesis is true (or false). Consequently, simply running more statistical analyses is not a sufficient avenue to accurate, reproducible, and correctly interpreted findings. Here, we hope to use the opportunity provided by broader discussions about statistics and reproducibility in science to review three important concepts—(a) significance testing, (b) multiple comparisons, and (c) effect size—and translate them to our field of geobiology. This manuscript is intended

to be educational and to start a discussion regarding proper statistical analyses in geobiology. Our focus is on what we believe are the more familiar terms of formal NHST (i.e., a frequentist approach), but we discuss the goal of a more flexible approach to statistical thinking at the conclusion of the paper. In this spirit, we recognize that these topics will be familiar to many readers, and indeed aside from a geobiological spin, there is little intellectual territory here that has not been extensively covered in medical, psychological, and ecological journals. However, for readers less familiar with these topics we hope this essay may provide useful guidance for avoiding some of the problems of reproducibility that have plagued other fields. Geobiology ultimately has the opportunity to be among the select group of fields that do not have major reproducibility problems.

Take home message: There are acknowledged issues with Null Hypothesis Significance Testing (including publication bias and replicability in science at large), but studies should not rely solely on visual analysis alone. A formal examination of the size of the effect relative to sampling variation is a key tool in evaluating new scientific claims.

3 | NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST)

Despite its widespread use in science, the development of NHST has its roots in industrial applications. For instance, Ronald Fisher developed the analysis of variance working with experimental crop data, and William Gosset (nominated here as the patron saint of geobiological data analysis) developed the Student's *t* test working to increase yields at the Guinness Brewery (Student, 1908). When faced with two (or more) groups of samples, each with scatter, the pertinent question is whether the sets of samples may actually represent the same underlying data distribution, with observed variation between groups arising from sampling of this distribution (The null hypothesis, H_0, is that there is no difference between groups; in other words, there is only one underlying distribution). In NHST, this question is addressed by comparing the means or medians of the groups (and data variability) and calculating the probability that a result at least that extreme would be found if the groups were drawn from the same distribution or population. This probability is expressed as the *p*-value. Incorrectly rejecting the null hypothesis (a false positive) is referred to as Type I error, and incorrectly failing to reject the null hypothesis (a false negative) is Type II error. A variety of parametric (those that depend on a specified probability distribution from which the data are drawn) and nonparametric (those that do not) statistical analyses exist to make these comparisons between two or more groups. A full review of particular analyses is beyond the scope of this article and is best found in statistical textbooks and web resources.

Choosing the correct statistical analysis for a given set of data poses some issues, but the more important issue—the focus of this piece—is understanding the fundamental logic of statistical tests: What they do and do not tell you. It is errors in logic, rather than someone using a *t* test when they should have used a Wilcoxon, which are most problematic. One common fallacy regarding NHST is that the level of significance rejecting the null hypothesis (the *p*-value) is the probability that the null hypothesis is correct. There is a wide literature discussing this fallacy, with one of the best being Cohen's 1994 essay "The Earth is round ($p < 0.05$)." He notes that what we want to know, as researchers, is "given these data, what is the probability that H_0 is true?" But what NHST tells us is "given that H_0 is true, what is the probability of these (or more extreme) data?" In other words, rejecting a specific null hypothesis does not actually confirm any underlying truth or theory. Addressing this requires knowing the likelihood that a real effect exists in the first place, which may be difficult to calculate. Even if these odds can be calculated, the combined uncertainty results in more substantial murkiness about the results than the *p*-value alone indicate (Nuzzo, 2014, provides more information on what a *p*-value really tells you—in a form more palatable to the average reader than Cohen—as well as a historical discussion of how the $p = 0.05$ threshold came about).

Another common error in interpreting formal statistical tests regards "mechanical dichotomous decisions around a sacred 0.05 criterion" (Cohen, 1994). Obvious to most readers is that a 5.5% probability of generating results at least that extreme (e.g., $p = 0.055$) is basically no different, in an interpretive sense, than a 4.5% probability ($p = 0.045$). In other words, $p < 0.05$, $p < 0.01$, and $p < 0.005$ (Benjamin et al., 2018, recently advocated for the more stringent $p < 0.005$ statistical criteria) are all arbitrary criteria, although still useful conventions. A memorable quote by Rosnow and Rosenthal (1989) states:

> We want to underscore that, surely, God loves the 0.06 nearly as much as the 0.05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?

Yet while researchers inherently "know" $p = 0.045$ and 0.055 are really no different, one result is deemed "true" and publishable (positive publication bias) and one is deemed inconsequential and ignored. Or in Rosnow and Rosenthal's more dramatic prose:

> It may not be an exaggeration to say that for many PhD students, for whom the 0.05 alpha has acquired an almost ontological mystique, it can mean joy, a doctoral degree, and a tenure-track position at a major university if their dissertation p is < 0.05. However, if the p is > 0.05, it can mean ruin, despair, and their advisor's suddenly thinking of a new control condition that should be run.

Take home message: Null Hypothesis Significance Testing investigates the probability (expressed as *p*-values) of finding results as extreme as the observed data, given the null hypothesis. Of note, *p*-values cannot address the likelihood that a hypothesis is correct.

p-value thresholds represent arbitrary cutoffs rather than true/false criteria for accept/reject decisions.

4 | MULTIPLE TESTING AND RESEARCHER DEGREES OF FREEDOM

A colleague once described a conference—in the pre-PowerPoint days—in which geologists were instructed to bring an overhead transparency summarizing the record of their subfield through Earth history (tectonic events, fossil trends, geochemical proxy records, etc.) along the same *x*-axis temporal scale. The participants then took turns swapping different transparencies on and off the projector, looking for correlations. This strategy was innovative from a data exploration perspective—and indeed sounds intellectually stimulating—but ultimately is a nightmare regarding multiple comparisons. The multiple comparison problem essentially results from "more shots on goal." It can be illustrated with the following example—the probability of flipping any given coin as "heads" 10 times out of 10 is very low; however, if this is done over and over, the probability that one coin will be "heads" every time obviously increases. It would be incorrect, though, to conclude that that coin is different from the rest. Specifically for NHST, the probability of a false positive in a battery of tests will be $1 - (1 - \alpha)^k$, where alpha (α) is the significance level required (e.g., $p < 0.05$) and k is the number of tests performed (discussed in detail in Streiner, 2015). For 10 separate tests at the standard level of significance, the probability of a false positive is $1 - (1 - 0.05)^{10}$ or about 40%. This issue is perhaps even better illustrated by Figure 2 below (analyzing the effect of jelly bean colors).

Uncorrected multiple comparisons are one of the primary causes of replicability issues across many scientific fields. In some cases, these comparisons may have been done explicitly, such as with the green jelly beans. Many reports in the 1990s and 2000s about what genes "cause" specific effects in humans were the spurious result of trawling limited genetic data across comprehensive epidemiological datasets and looking for "significant" correlations. An analog in our field may be instances where biological data of interest (e.g., microbial abundance/ecological traits/gene expression) are collected from a modern environment, such as a hot spring, alongside environmental data. Which environmental parameters are correlated with the biological metric of interest? (leaving aside the broader question of correlation and causality). It would be inappropriate to simply conduct a pairwise comparison of all the environmental parameters with the biological metric without correction for multiple tests. By chance alone, some parameters might be significantly correlated: In null datasets, $p < 0.05$ occurs, by definition, 5% of the time.

Perhaps more insidiously, multiple comparisons can be done implicitly or unconsciously. For instance, a researcher may complete a compilation of fossil data from shelly invertebrates and then half-asleep in the shower mentally wander through all the different geological data records (sea level, temperature, redox proxies, strontium isotopes, etc.) before snapping awake after noting that the identified fossil trend looks very similar to a previously published record of calcium isotopes. A single statistical comparison is made, and voila: $p < 0.05$. Perhaps something in the calcium cycle is affecting the livelihood of these shell-forming organisms? Maybe. In this case, the researcher did not explicitly test each comparison with a *p*-value, as in the jelly bean or hot springs example, but the researcher still mentally conducted the equivalent of swapping out overhead transparencies: multiple comparisons until a match was found. A related problem is exploratory data analysis as data are being generated (the role of exploratory data analysis is discussed further below). A range of analyses might be conducted, with one predictor variable out of many yielding a significant correlation. When the full dataset is generated, "only one" explicit test is conducted on the final dataset and included in the publication, with the researcher honestly forgetting just how many analyses had been conducted. Or, a spuriously significant *p*-value is found early on, say for all available marine samples, which then disappears as more data are added. A second analysis, looking at individual ocean basins, and a third, looking by depth class, are conducted, perhaps excluding some extreme outliers, until another significant *p*-value reappears [so-called *p*-hacking, or "researcher degrees of freedom"; Simmons, Nelson, and Simonsohn (2011)]. It is then this sub-group analysis that is emphasized in the manuscript. More often than not, it is an unwitting error by a scientist excitedly analyzing their data. Remember Feynman's quote that "the first principle is that you must not fool yourself—and you are the easiest person to fool." This is not to discourage data exploration, but correctly accounting for these comparisons—or at least remaining cognizant of the issue—will ultimately be key to producing lasting insights.

4.1 | Avoiding multiple comparison pitfalls

How then should one account for multiple comparisons? The best approach is early, explicit planning. The gold standard, as practiced in clinical trials, is pre-registration (for instance, on ClinicalTrials.gov run by the U.S. National Library of Medicine). In this strategy, the researcher publishes a white paper prior to starting the experiments, explicitly detailing how the data will be collected (including the stopping point), and the number and types of statistical analyses to be conducted. They are then held to this plan, or the trial is not considered valid. Such a strategy is often considered an unrealistic ideal outside of clinical trials, but notably an effort to publically post methodologies *a priori* has recently been initiated for molecular phylogenetics (Phylotocol; DeBiasse & Ryan, 2018), a field particularly susceptible to "researcher degrees of freedom." Whether or not such registries are appropriate for geobiology in the long term is a subject for debate, and certainly the approach ignores the fact that much of our science (especially field science) is truly "discovery driven" rather than "hypothesis driven." Nonetheless, increased pre-experiment effort put into planning statistical analyses will be effective in reducing multiple comparison "creep." This may be particularly useful to bring up during project planning with early-career researchers, as essentially all aspects of experimental design are

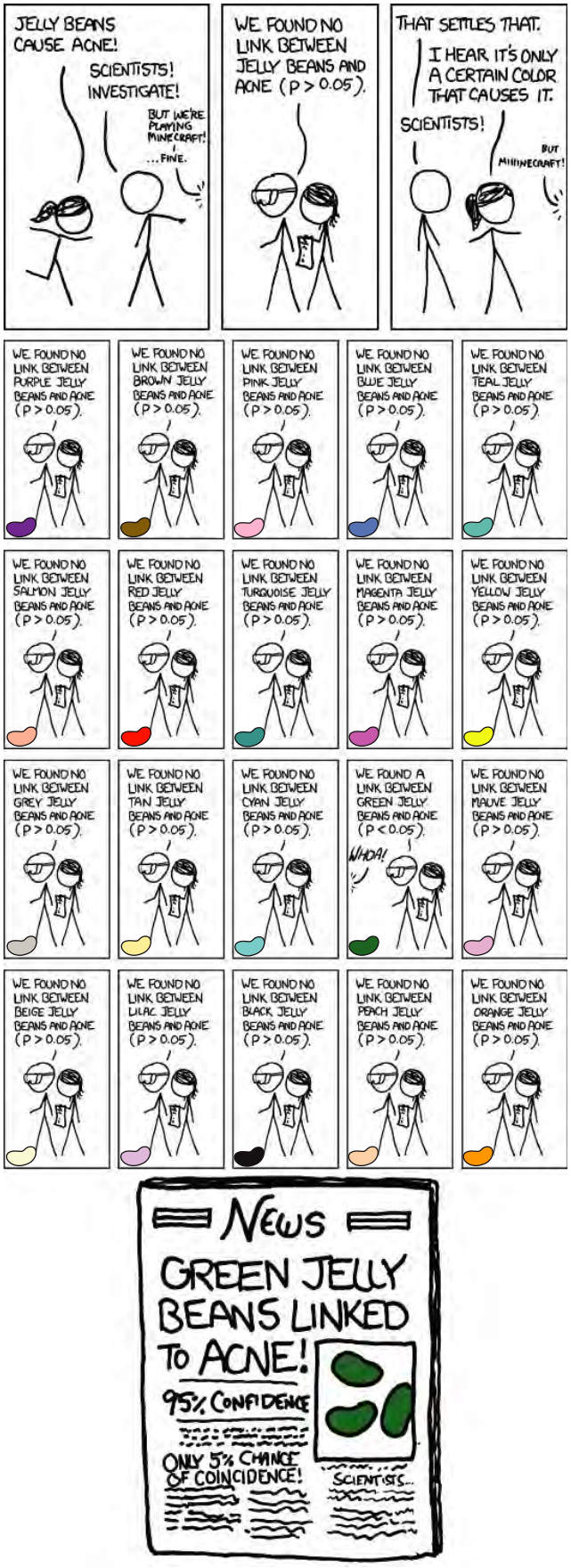

FIGURE 2 Figure modified from the webcomic XKCD.com. Multiple statistical comparisons increase the probability of a false-positive result ("more shots on goal"). Specifically, with twenty independent comparisons as shown in the comic, the probability of a false positive is ~65% (Streiner, 2015)

being taught at that point. Discussions of study-level reproducibility should also start making it into undergraduate and graduate geobiology curricula, as for instance is occurring in some psychology programs (Button, 2018).

Moving from pre-experiment awareness and planning to post-experiment data analysis, there are several techniques to account for multiple comparisons ("multiple testing correction" or "alpha inflation correction"). Certain Bayesian approaches may not require explicit correction (see Gelman, Hill, & Yajima, 2012), but in a frequentist context a common approach is to apply a direct correction that accounts for the increased family-wise error rate. The Bonferroni correction, for instance, divides the level of significance required (e.g., $\alpha = 0.05$) by the number of independent analyses conducted. So, a researcher investigating how brachiopod body size changed between four different stratigraphic formations would require $p < 0.008$ (an overall $\alpha = 0.05$ / 6 independent pairwise tests = 0.008) in order to achieve significance. Another common practice for such an analysis (multiple pairwise comparisons) is to conduct an omnibus test such as ANOVA, followed by a post hoc test such as the Tukey HSD test that directly accounts for increased family-wise error.

This general practice of alpha inflation correction has received some criticism, as the thresholds for significance can be overly conservative, leading to more false negatives and potentially restricting the path of future scientific curiosity (Moran, 2003; Rothman, 1990). Such criticisms commonly note that studies in their fields often have "a small number of replicates, high variability, and (subsequently) low statistical power" (Moran, 2003). Dr. Turekian might well relate! The argument put forth by such papers is that the increased incidence in false positives is not really an issue, because other researchers will repeat the experiments, be unable to replicate them, and the false claims will eventually disappear from the literature. However, careful study of replication attempts in psychiatry and psychology has demonstrated that (a) uncorrected multiple comparison testing does empirically lead to a morass of published false positives (Duncan & Keller, 2011), and (b) the original hypotheses do not simply disappear from the literature but have incredible persistence (Ioannidis, 2012). The causes of such persistence are varied but basically boil down to low incentives for journals or authors to publish negative replications (Ioannidis, 2012; Simmons et al., 2011). While these arguments on the stringency of multiple testing corrections are not meritless (see next paragraph), in our opinion widespread Type I errors (false positives) are a far greater hindrance to the advancement of science than Type II errors (false negatives).

That said, determining the correct balance between the likelihood of false negatives and false positives, as well as encouraging cutting-edge methods development that must start with small datasets

(Turekian's point), is obviously a complicated endeavor. Several other methods, such as the Holm and Hochberg methods (reviewed by Streiner, 2015), offer protection against alpha inflation but are not as conservative as a strict Bonferroni correction. Such corrections should also follow common sense, as they can degenerate into the absurd (García, 2004; Streiner, 2015). For instance, how many truly independent comparisons are really being conducted? Oceanographic factors such as temperature, oxygen, pH, light, and pressure can all be broadly correlated with depth. Comparing all of these against microbial abundances might result in multiple significant correlations that become (inappropriately) non-significant after correction for multiple comparisons. Another issue to consider is whether there may be pre-existing hypotheses that motivated the study. For the brachiopodologist studying body size across four formations, they may be testing previous hypotheses of body size evolution during a specific time period (e.g., Heim, Knope, Schaal, Wang, & Payne, 2015). The real test may be between the stratigraphically lowest and highest formations, and it may be unduly stringent to require a very low value *p*-value resulting from the multiple pairwise comparisons.

This difference has been discussed by Streiner (2015) as the difference between hypothesis testing (which may not require correction) and hypothesis generating (which should be reported as tentative and/or exploratory results). Or in plainer language, exploratory data analysis can be good, and explicit hypothesis testing can be good, but the approach being used should be clear. Indeed, there are many powerful techniques (including new machine learning techniques) to understand which of multiple predictor variables might best explain variance in the response variable of interest (a situation we often face in geobiology). The results of such analyses, though, cannot be turned around and presented as an *a priori* hypothesis complete with a significant *p*-value. Dr. Brian McGill's blog post provides a cogent "defense" of exploratory data analysis: (https://dynamicecology.wordpress.com/2013/10/16/in-praise-of-exploratory-statistics). He closes, "I use exploratory statistics and I'm proud of it! And if I claim something was a hypothesis it really was an *a priori* hypothesis. You can trust me because I am out and proud about using exploratory statistics."

The key thread running through these statistical commentaries regarding multiple comparisons is conscientiousness—personally with respect to how many analyses have been conducted, but also with respect to how the full scope of statistical procedures is described in the paper. Simmons et al. (2011) provide excellent guidelines in this regard for both authors and reviewers. Notably, while promoting strict reporting guidelines, these authors also advocate for increased tolerance of statistical "imperfection" by reviewers and editors in well-documented papers: "one reason researchers exploit research degrees of freedom is the unreasonable expectation we often impose as reviewers for every data pattern to be (significantly) as predicted. Under-powered studies with perfect results are the ones that should invite extra scrutiny."

Finally, at the end of this discussion, it is worthwhile to ask yourself the basic question of whether multiple testing is an issue in your particular research area. If you are working solely with a single proxy record, or an isolated genetic system, the answer might be no. The issue arises if you want to understand what correlates with your data—if there is a large constellation of possible correlates, there is an equally large likelihood of spurious correlations. Learning from the abysmal record of replication in other fields, caution against alpha inflation in these cases will be a cornerstone to a robust and healthy field of geobiology.

Take home message: Vigilance against alpha inflation starts with the individual researcher, ideally during the pre-experimental design phase. Data exploration is encouraged, but trawling through data for significant correlations that are then presented as an *a priori* hypothesis must be avoided: Papers should explicitly state if they are to be viewed as exploratory. The full sweep of data collected, statistical tests conducted, and choices of data inclusion/exclusion (or other "degrees of freedom") by a researcher should be made clear in publication in order to avoid cherry-picking (Simmons et al., 2011).

5 | EFFECT SIZE

So let us say you have found a significant difference between two groups of data, and through attentive practice and analysis, you have determined it is not a chance result based on numerous "shots on goal." The question now is—does the result matter? This last question seems silly, but a mistaken focus on significant *p*-values as the be-all/end-all, instead of on effect size, has hampered progress in fields such as ecology (Fidler et al., 2004). Simply, effect size is a quantitative measure of the magnitude of a phenomenon. Statistical power is the likelihood that an analysis will detect a real effect (as determined by a significant *p*-value), and is governed by the size of the effect, the variation present in the groups, the number of samples in the analysis, and the threshold for significance (α). Thus, two groups with widely separated means (large effect size) and little within-group variation will require relatively few samples for a well-powered study.

The flip side of this—and why *p*-values must be regarded as the statistical likelihood a result that extreme would be found by chance, rather than how "*important*" a result is—is that given enough samples, literally any effect size, no matter how small, can be detected. This has received considerable attention in the statistical literature:

> [The null hypothesis] can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal in rejecting it?
>
> (Cohen, 1990)

At this point, you may be asking, "since what we are really interested in is effect size, and given Cohen's statement that the null

hypothesis is always false, do we even need to run statistical tests?" Yes! Without a significant result, there is no reason to believe the observed results may not be due to sampling variation. In other words, significance is the jumping off point, the license to start talking about effect size from a position of confidence. For further reading, the relationship between sample size, *p*-values, and accuracy in inferring effect size is intelligently dissected by Halsey, Curran-Everett, Vowler, and Drummond (2015).

What constitutes an important effect will vary by field. Returning to the coin toss example, given millions or billions or trillions of flips—whatever the required sample size may be—the differing weights of the coin sides will ultimately cause one side to land up significantly more times than the other. Yet, this does not matter when two friends are choosing a restaurant: It is still ~50:50, and the miniscule effect size is irrelevant in this instance. This is the difference between a *significant* effect and an *important* effect. Nonetheless, this fallacy is often committed in the literature:

> All psychologists know that *statistically significant* does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results sections studded with asterisks implicitly becomes in the Discussion section highly significant or very highly significant, important, big!
>
> (Cohen, 1994)

Sometimes, though, small effect sizes really do matter. As an example, human geneticists have learned that the majority of phenotypes are not controlled by a single gene/locus (such as the case for Huntington's disease). Rather, characteristics like height, and the risk of diseases such as heart disease, Type II diabetes, and depression are caused by the (largely) additive effects of thousands of genetic loci. In the case of psychiatric disorders, each individual locus explains far <1% of variance in risk for a psychiatric disorder (i.e., small individual effects) and yet total genetic contributions explain 40%–80% of the variance in risk for these disorders (large summed effect) (Duncan et al., 2017; Ripke et al., 2014; Wray et al., 2018). Thus, as our questions move toward datasets with hundreds or thousands of measurements, the focus must be on geobiologically important effect sizes (which will vary by question) and confidence intervals rather than simply statistical significance.

6 | INTERPRETATIVE EXAMPLES

The relationship between *p*-values (significance), effect size, and sample size is illustrated in Figure 3. Note there is no relationship implied between the left and right panels, they are simply illustrative examples of these concepts. In Panel A, a researcher has identified relationships that appear interesting to pursue, with the Group B mean ~50% higher than Group A in the *t* test example (left side). On the right side, the predictor variable accounts for 27% of the variation (R^2 value) in the regression example. But the results are not statistically significant. Especially in light of recent claims that $p < 0.05$ is too lax a standard (Benjamin et al., 2018), there is a strong likelihood that this result—specifically the observed size and direction of the effect—is due to sampling effects (Halsey et al., 2015). If this were your hard-won data, it is important to avoid the temptation of knowingly or unknowingly manipulating the data (*p*-hacking) to achieve a "significant result." For instance, simply removing the lowest data point in Group B results in $p = 0.03$, significant! Maybe, there is an imminently logical reason to exclude that data point—perhaps it is from a different and inappropriate lithology, or the sampling methodology was actually different. The goal though is to avoid inventing post hoc justifications for data manipulation, as it is so easy to fool yourself, particularly if removing that point provides support for long-held ideas (confirmation bias). If such data exclusions are made, they should be noted clearly in the manuscript, and both sets of analyses, with a reasoned explanation, should be included (Simmons et al., 2011). Moving toward shared transparency within scientific subfields for data collection, reporting and presentation methods will be instrumental in helping researchers present reasonable results with less threat of conscious/unconscious *p*-hacking. Panel B depicts essentially the same analysis as in Panel A, but with more samples. In this case, the result is significant, and the researcher can feel more confident describing the size and direction of the effect. Note that if Panel B were an extension of the study in Panel A, the best practice (sometimes difficult to achieve but nonetheless the best practice) is actually to establish a pre-determined stopping point for data collection (Simmons et al., 2011). Continuing to add bits of data, with the analysis rerun each time, effectively represents multiple tests.

Panel C depicts how with larger sample sizes the power to identify small but statistically significant effects also increases. In the comparison of groups A and B, the means only differ by ~3%, but the result is highly significant ($p = 0.007$). In the regression analysis, the predictor variable only describes 4% of the variation in the response variable, but nonetheless, the result is significant by traditional measures ($p = 0.047$). Looking at the scatter in this plot is instructive, as it appears as a cloud of points with no correlation, but statistically, a small correlation does exist. In other words, it visually illustrates the quotes from Cohen: Effect size, not significance alone, is the end goal. As also previously discussed, the ultimate interpretation of importance is question- and field-specific. To reiterate the point, robust increases in crop yields of 4% may feed millions, whereas a change of 4% in modern marine sulfate levels (e.g., ~1 mM) may have little relevance to current research questions regarding sulfur biogeochemistry.

7 | BEST PRACTICES MOVING FORWARD

We start this "best practices" section from a humble position, as we are by no means trained statisticians but rather enthusiastic advocates for increased statistical rigor in geobiology. We have made (or will make)

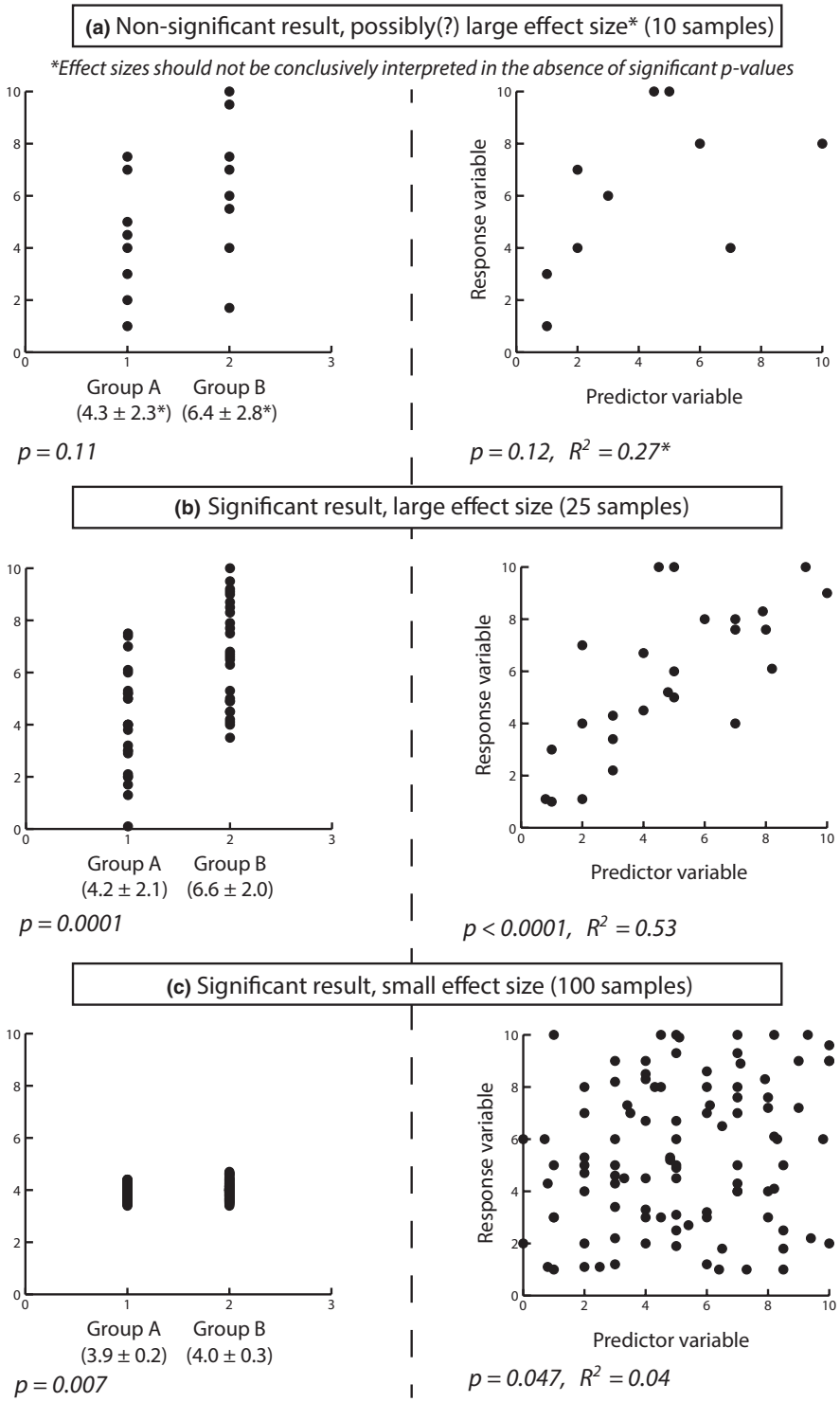

FIGURE 3 The relationship between significance, effect size, and sample size shown with hypothetical data points. This is depicted for measurements sampled from two different groups (left side) and correlation between a predictor variable and response variable (right side); note there is no strict relationship between the data in the two panels. (a) Small sample sizes may reveal a large effect, but the result is not significant and should be considered an intriguing finding rather than a confident conclusion. (b) With increased sampling, a researcher may (or may not) reveal a significant effect. The sign (direction) of this effect is likely correct, while increased sampling will lead to better accuracy of the effect's magnitude. (c) Given enough sampling, even the smallest effect size will become statistically significant. The right side plot in c is visually informative in this regard—what appears to be a cloud of points are, statistically speaking, correlated. For illustrative purposes, we have described effects here as "large" and "small," but as discussed in the text, this distinction will be field- and question-specific

technical and logical errors in our published papers and almost certainly have made unconscious "researcher degrees of freedom" decisions that impacted results (Simmons et al., 2011). Even Jacob Cohen, whose thoughtful papers we have cited numerous times, was called out for incorrect statistical logic (Oakes, 1986). Learning correct statistical practice and logic is thus a career-long endeavor for most scientists.

As a first step, we suggest that both reviewers and authors insist on some form of statistical analysis as a best practice approach. The critical concept here is that your particular set of measurements is not a static truth about a group. Rather, they are values drawn from a distribution, and repeated draws may yield very different results—especially at low sample sizes (Halsey et al., 2015). We do recognize that many geobiological studies will not lend themselves to formal statistical tests and that there is a healthy tradition of discovery-based geobiological science—especially in field studies—that should not be stifled. Nonetheless, if a paper is reporting observed differences between groups or correlations between variables, it is scientific due-diligence to investigate the degree to

which these differences would be expected given the null hypothesis (or more broadly, the uncertainty of the result).

In this respect, we note that we have focused primarily here on formal NHST, but there are certainly other strategies such as model-based approaches, information theoretic approaches, and/or bootstrapping that achieve the same general goal of understanding true relationships (as distinguished from spurious results that are simply due to sampling variability). For instance, maximum likelihood and Bayesian approaches to assess the impact of random error are the common practice in molecular phylogenetics. The debate about whether formal NHST should be retained and improved or done away with is decades old, rages still, and is not something we can adequately address here (Benjamin et al., 2018; Cohen, 1994; Falk & Greenbaum, 1995; Fidler et al., 2004; Halsey et al., 2015). And as Cohen (1994) noted, there is no magic alternative to NHST. Certainly, though, all approaches discussed above are more likely to lead to correct inference than visual inspection of data.

As statistical tests are increasingly implemented in geobiology, the next step is to learn from the mistakes of other fields and help each other not fall foul of the fallacies described in this essay. Specifically, the points to avoid are:

1. Not correctly accounting for multiple testing.
2. Considering the *p*-value to be the probability that the hypothesis is correct.
3. Considering $p = 0.045$ to be a dramatically stronger rejection of the null hypothesis than $p = 0.055$ ("mechanical dichotomous decisions").
4. Not explicitly reporting "researcher degrees of freedom" (Simmons et al., 2011).
5. Considering every "significant" result to be an "important" result. Rather, significance should typically be the requirement for positing a scientific effect, which must then be put into appropriate context.

Of these, multiple testing and "researcher degrees of freedom" are likely the most problematic with respect to reproducibility, especially as they can often be done unconsciously. Explicit planning at the start of a study is crucial in this regard. Also important at the planning stages is power analysis. Ioannidis (2005) notes that in terms of achieving long-lasting scientific insights, fewer well-powered studies are vastly preferable to many low-powered studies, and certainly the field benefits from not chasing false leads. Further, as demonstrated by Halsey et al. (2015), larger samples sizes and well-powered studies also more precisely estimate the effect size, which is after all what we are interested in. The challenge here is an obvious conflict between incentive structures for the field as a whole and individual researchers, specifically early-career researchers. Such considerations should ideally play into evolving discussions on how post-docs, faculty positions, and tenure are evaluated. Fortunately, in geobiology we are often addressing first-order questions with large effects, and the required increase in sample size to achieve a well-powered study is often not that large (Sterne & Smith, 2001).

As one final note regarding reproducibility, correctly documenting original data and metadata in accessible supplementary documents or data repositories is key to allowing researchers to test results and build on previous studies in meta- and mega-analyses (see for instance Ioannidis et al., 2009). Original code used for analyses must also be adequately curated in a public repository; this step is common in biological fields such as ecology (Cooper et al., 2017; Ram, 2013), but is not across geobiology.

To avoid feeding a publication bias monster, and to encourage new developments in a manner Karl Turekian would be proud of, we do not advocate a strict requirement of significance for publication. In our opinion, the results in Figure 3, Panel A, if from an emerging proxy record, would be quite suggestive and should be considered for publication, but the reader should be notified how likely the data are (given the null hypothesis). Power analysis is also helpful in this regard (Cohen, 1992). As an example, an influential genetics paper was published in *Nature*, despite having null results for the primary hypothesis (The International Schizophrenia Consortium, 2009). This paper provided strong evidence that significant results would be detectable with larger sample sizes in the near future, and it was the application of a recently developed statistical technique (polygenic risk scoring) that made it worthy of publication in *Nature*.

Moving to the longer-term view, the "best practices" described above are hopefully useful, but the goal for the next generation of geobiologists should not be best practices lists taped to the side of cubicles. Put simply, even the most well-intentioned of "best practices" lists can lead to a "cookbook" view of data analysis, where there is a right and a wrong way to do things, and statistics are a computer button to push after data acquisition. Rather the goal should be a situation where, for many, computational reasoning and data science are a natural, integrated part of our science alongside field and laboratory skills. The main need for this is simply that many of our scientific questions do not readily conform to classic statistical tests. As an example, Keller and Schoene (2012) investigated how igneous geochemistry has changed through Earth history, and recognized that these rocks are not evenly sampled in space and time—some plutons are heavily sampled, while others were sampled rarely or not at all. In other words, samples are not independent. To address this issue of sampling heterogeneity, they utilized a re-weighted bootstrapping approach, with bootstrap weights for a given sample related to the spatial and temporal proximity to other samples. Paleontologists have also addressed the same issue of sampling heterogeneity, but using different methods appropriate to the data archive of that field (e.g., Alroy, 2010). Neither of these solutions came from a statistics "cookbook," and achieving the flexibility to design the most appropriate test (be it frequentist, likelihood, or Bayesian), or to perform numerical experiments testing different scenarios, will require a foundation in statistics but also, importantly, computational thinking (Weintrop et al., 2016). Geobiology has had considerable success in breaking field boundaries and educating students who are as comfortable with a rock hammer as a pipette; integrating a computational and statistical perspective into this training will be the next step to drawing robust insights from ever-larger geobiological datasets.

In closing, we do not expect a statistical revolution in geobiology overnight. The common implementation of statistical analyses in fields like ecology took decades. In fact, Gosset ("Student") wrote to Fisher regarding the *t* test that "I am sending you a copy of Student's Tables as you are the only [person] that's ever likely to use them!" (cited in Box, 1981). We hope this Perspective helps start a dialogue regarding statistical practice in geobiology, while also recognizing it is heavily colored by our perspective—we look forward to seeing commentary from other perspectives (different subfields of geobiology, Bayesianists, etc.). Ultimately, this will help us avoid the problems with reproducibility present in other fields, be more confident in our results, and as a field move more quickly toward deeper geobiological understanding.

ACKNOWLEDGMENTS

We thank David Johnston, Jon Payne, Matt Clapham, and an anonymous reviewer for comments on a previous version of this manuscript, and James Farquhar, Anne Dekas, Joe Ryan, David Evans, and Alex Bradley for helpful discussion. We thank Randall Munroe of XKCD.com for permission to reproduce Figure 2. EAS and SAT were funded by a Sloan Research Fellowship.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Erik A. Sperling https://orcid.org/0000-0001-9590-371X

Erik A. Sperling[1]
Sabrina Tecklenburg[1]
Laramie E. Duncan[2]

[1]*Department of Geological Sciences, Stanford University, Stanford, California*
[2]*Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California*

Correspondence
Erik A. Sperling, Department of Geological Sciences, Stanford University, Stanford, CA.
Email: esper@stanford.edu

REFERENCES

Alroy, J. (2010). The shifting balance of diversity among major marine animal groups. *Science*, *329*, 1191–1194. https://doi.org/10.1126/science.1189910

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

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533. https://doi.org/10.1038/483531a

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*, 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bouter, L. M., Tijdink, J., Axelsen, N., Martinson, B. C., & ten Riet, G. (2016). Ranking major and minor research misbehaviors: Results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review*, *1*, 17. https://doi.org/10.1186/s41073-016-0024-5

Box, J. F. (1981). Gosset, Fisher, and the t distribution. *The American Statistician*, *35*, 61–66.

Button, K. (2018). Reboot undergraduate courses for reproducibility. *Nature*, *561*, 287. https://doi.org/10.1038/d41586-018-06692-8

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cooper, N., Hsing, P.-Y., Croucher, M., Graham, L., James, T., Krystalli, A., & Primeau, F. (2017). *A guide to reproducible code in ecology and evolution*. BES Guides to Better Science. London, UK: British Ecological Society.

DeBiasse, M. B., & Ryan, J. F. (2018). Phylotocol: Promoting transparency and overcoming bias in phylogenetics (No. e26585v1). *Systematic Biology*, *syy090*. https://doi.org/10.1093/sysbio/syy090

Duncan, L. E., & Keller, M. C. (2011). A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *The American Journal of Psychiatry*, *168*, 1041–1049. https://doi.org/10.1176/appi.ajp.2011.11020191

Duncan, L., Yilmaz, Z., Gaspar, H., Walters, R., Goldstein, J., Anttila, V., … Bulik, C. M. (2017). Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *American Journal of Psychiatry*, *174*, 850–858. https://doi.org/10.1176/appi.ajp.2017.16121402

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75–98. https://doi.org/10.1177/0959354395051004

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

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology*, *20*, 1539–1544. https://doi.org/10.1111/j.1523-1739.2006.00525.x

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics*, *33*, 615–630. https://doi.org/10.1016/j.socec.2004.09.035

García, L. V. (2004). Escaping the Bonferroni iron claw in ecological studies. *Oikos*, *105*, 657–663. https://doi.org/10.1111/j.0030-1299.2004.13046.x

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211. https://doi.org/10.1080/19345747.2011.618213

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*, 1037. https://doi.org/10.1126/science.aad7243

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle *P* value generates irreproducible results. *Nature Methods*, *12*(3), 179–185. https://doi.org/10.1038/nmeth.3288

Heim, N. A., Knope, M. L., Schaal, E. K., Wang, S. C., & Payne, J. L. (2015). Cope's rule in the evolution of marine animals. *Science*, *347*, 867–870. https://doi.org/10.1126/science.1260065

Hines, W. C., Su, Y., Kuhn, I., Polyak, K., & Bissell, M. J. (2014). Sorting out the FACS: A devil in the details. *Cell Reports*, *6*, 779–781. https://doi.org/10.1016/j.celrep.2014.02.021

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

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645–654. https://doi.org/10.1177/1745691612464056

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., … van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*, 149–155. https://doi.org/10.1038/ng.295

Keller, C. B., & Schoene, B. (2012). Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago. *Nature*, *485*, 490–493. https://doi.org/10.1038/nature11024

Lithgow, G. J., Driscoll, M., & Phillips, P. (2017). A long journey to reproducible results. *Nature News*, *548*, 387. https://doi.org/10.1038/548387a

Moran, M. D. (2003). Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos*, *100*, 403–405. https://doi.org/10.1034/j.1600-0706.2003.12010.x

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature News*, *506*, 150. https://doi.org/10.1038/506150a

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: Wiley.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science 349*, aac4716.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712. https://doi.org/10.1038/nrd3439-c1

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, *8*, 7. https://doi.org/10.1186/1751-0473-8-7

Ripke, S., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. A., Lee, P., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421–427.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*, 43–46. https://doi.org/10.1097/00001648-199001000-00010

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Sterne, J. A. C., & Smith, G. D. (2001). Sifting the evidence—what's wrong with significance tests? *British Medical Journal*, *322*, 226–231. https://doi.org/10.1136/bmj.322.7280.226

Streiner, D. L. (2015). Best (but oft-forgotten) practices: The multiple problems of multiplicity-whether and how to correct for many statistical tests. *The American Journal of Clinical Nutrition*, *102*, 721–728. https://doi.org/10.3945/ajcn.115.113548

Student (1908). The probable error of a mean. *Biometrika 6*, 1–25.

The International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*, 748–752.

Thiemens, M. H., Davis, A. M., Grossman, L., & Colman, A. S. (2013). Turekian reflections. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 16289–16290. https://doi.org/10.1073/pnas.1315804110

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, *25*, 127–147. https://doi.org/10.1007/s10956-015-9581-5

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., … Sullivan, P. F. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*, 668–681. https://doi.org/10.1038/s41588-018-0090-3