Two Problematic Promises on Page One of the TAS Special Issue on Statistical Inference

Donald B. Macnaughton[*]

Abstract

The editorial in the March 2019 special issue of *The American Statistician* (TAS) promises several useful outcomes if we abandon using the concept of “statistical significance” in scientific research. Two of the promises are that (a) abandoning statistical significance will lead to fewer false-positive errors in scientific research, and (b) abandoning statistical significance will make it easier to replicate scientific research results. The present paper discusses the role of statistical significance as a gateway to publication in scientific journals. The paper then shows how abandoning statistical significance as a gateway to publication will lead to *more* false-positive errors in published research and will make published research results *harder* to replicate.

Keywords: Hypothesis test; *p*-value; False-negative error; Scientific publishing

1. Introduction

In March 2019, The American Statistical Association published a special issue of *The American Statistician* (TAS). The issue’s theme is “Statistical Inference in the 21st Century: A World Beyond $p < 0.05$”. The issue begins with a 19-page editorial that carefully summarizes the thinking in the 43 papers in the issue (Wasserstein, Schirm, and Lazar 2019).

The first sentence of the editorial identifies a key goal of both science and statistics, which is to separate the signal from the noise in data. This paper focuses on that goal.

Section 2 of the editorial is titled “Don’t Say ‘Statistically Significant’ ”. It is widely agreed among statisticians and experienced researchers that a scientific research result is deemed “statistically significant” if the properly computed *p*-value for the result is less than (or equal to) a chosen “critical” *p*-value, which is usually 0.05 or 0.01. As discussed below, the critical *p*-value *helps* us to separate the signal from the noise in scientific research data.

Section 2 of the editorial notes the important fact that the concept of statistical significance is widely misinterpreted by less experienced researchers and by students. It concludes that the misinterpretation has led the concept to become “meaningless”, and its original sensible interpretation is “irretrievably lost”. It also notes that (a) the concept of statistical significance can lead to erroneous beliefs and poor decision-making, (b) the concept doesn’t [directly] imply truth or importance, and (c) the concept prevents negative results from being published which, the section suggests, can “distort the [scientific] literature”. Section 2 concludes:

> For the integrity of scientific publishing and research dissemination, therefore, whether a *p*-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight (2019, p. 2).

The editorial promises that several positive outcomes will occur if we abandon using statistical significance (i.e., if we abandon critical *p*-values) in evaluating scientific research results. The following discussion considers two of the promised outcomes, which are both highly attractive. But both outcomes appear to be unattainable. And, if we abandon statistical significance, we will get the undesirable opposite outcomes.

2. A Practical Use of a Critical *p*-Value in Scientific Publishing

To prepare for discussion of the two outcomes, and in keeping with the editors’ theme of the integrity of scientific publishing, let us review the use of a critical *p*-value by a scientific journal as a *gateway to publication* of scientific papers submitted to the journal. This is a key practical use of a critical *p*-value. Let us consider a standard view of how the gateway works:

2.1 Discovering Effects in Populations

A paper that describes the results of an empirical (i.e., data-based) scientific research study generally reports the discovery of a new “effect” that the research has (apparently) discovered in the population of entities under study. We can often sensibly view an effect as being equivalent to the existence of a relationship between two or more variables.

For example, a medical research study may discover good evidence of a relationship between:

- the amount, x, of a new treatment (e.g., a drug) given to patients and
- a measure of the patients’ subsequent health, y.

If the side effects of the treatment are minimal, the discovery of a useful relationship between variables x and y in patients will help doctors to control (i.e., optimize) the values of y in new patients from the population. (Doctors do this by helping the patients to suitably adjust the values of x.) These basic ideas underlie most clinical research in medicine—how can we improve some variable y in patients by adjusting some other variable(s) x in them?

The preceding ideas readily generalize to all fields of scientific research. This is because every field is interested in

[*] Email: donmac@matstat.com

discovering and studying effects (usually relationships between variables) in entities in populations that are of interest to the field. For example, astronomers discover and study relationships between variables in the populations of celestial objects and celestial waves.

In any field of science, if we can find good evidence of a new relationship between variables, then we can use the knowledge of the relationship to predict or control the values of the "response" variable, y, in new entities from the population. We can also use the knowledge of the relationship to help us to understand how the entities in the population work. If the variables are chosen carefully, the abilities to predict, control, and understand based on a relationship between variables are often highly useful.

2.2 The Research Hypothesis and the Null Hypothesis

In many scientific research studies, the researcher will have a "research hypothesis". This hypothesis typically says that a relationship *exists* between a predictor variable, x, and a response variable, y, in the entities in the studied population of entities.

Following an old tradition, the research hypothesis is sometimes referred to as the "alternative hypothesis". However, that is a misnomer because it inappropriately downplays the vital importance of the research hypothesis.

In contrast to the research hypothesis, the "null hypothesis" says that *no* relationship exists between x and y in the entities. Thus, the null hypothesis is "empty"—implying that nothing (i.e., no relationship) is there.

For the sake of (a) logical sensibility, and (b) the principle of parsimony (Baker 2016), it is customary to begin the study of a new relationship between variables by *formally* assuming that the null hypothesis is true. That is, we assume that there is no relationship whatever between the predictor variable of interest and the response variable of interest in the entities in the population. This assumption helps us to avoid deceiving ourselves about a nonexistent relationship.

Of course, *informally*, we believe (hope) the opposite—we believe that our research hypothesis is true. And we believe that our study will find *good evidence* that the research hypothesis is true. This evidence will enable us to (tentatively) "reject" the null hypothesis and conclude that the research hypothesis is (likely) true in the population. The remainder of the present subsection discusses some important ramifications of obtaining good evidence that a research hypothesis is true. The next subsection discusses how we tell whether we have good evidence that a research hypothesis is true—how we tell whether we have good evidence that a relationship exists between variables.

If a research study obtains good evidence that its research hypothesis is true (e.g., good evidence of a new relationship between variables), then this is called a "positive" result. A positive result is gratifying for a researcher because it suggests that the postulated relationship (effect) exists in the population.

In contrast, if a research study *fails* to obtain good evidence that its research hypothesis is true, then this is called a "negative" result. A negative result is disappointing for a researcher because it implies that we must continue to assume that the empty null hypothesis is true. (As discussed in appendix A, we can never *prove* that a given null hypothesis is true, but it is efficient to formally *assume* each one is true until someone empirically proves otherwise.)

Unfortunately, negative results occur often in scientific research, more than half the time in research studies in some fields. This is because nature's secrets are hard to unlock, and thus many carefully-thought-out research hypotheses are untrue. However, this isn't a serious problem because experienced researchers understand that their research hypotheses are sometimes untrue. In cases of negative results, researchers understand that the relationship or effect that they thought existed may not exist (or at least it doesn't exist strongly enough for their present research to have detected it).

Positive results are much more interesting than negative results in scientific research because we can do reliable prediction or control with a correct positive result. For example, if medical researchers can find a usable relationship between (the amount of) a studied drug and (the amount of) a disease, then doctors can use the drug to treat the disease. But (with rare exceptions) we can't do much with a negative result. For example, if there is no good evidence of a relationship between a studied drug and a disease, then doctors can't do much with that.

So, forward-looking scientists find that negative results are boring, not telling us anything substantial beyond the null hypothesis, which we had already assumed to be true at the start. (A negative result also tells us that the research failed to find what it was looking for, but it generally can't say why, so a negative result is almost never definitive.) So, forward-looking scientific journals are eager to publish papers describing new convincing *positive* results, which give readers new information about the entities in the studied population. But journals almost never publish papers describing *negative* results because these papers tell readers nothing new about the entities.

The preceding ideas imply that the omission of publication of negative results doesn't somehow distort the scientific research literature. This is because a negative result doesn't tell us anything substantial beyond the status quo—beyond the fact that the null hypothesis appears to be true, which we had already sensibly assumed anyway by default.

I discuss in a paper certain rare cases when negative results in scientific research are interesting and are therefore published (2018 app. L).

2.3 How Do We Tell Whether We Have a Positive Result?

So, how do we tell in a research study whether we have a positive result or a negative result? Similarly, how do we tell whether we have good evidence that a certain relationship exists between variables? Similarly, how do we tell whether we can reject the null hypothesis and conclude that the research hypothesis is (likely) true in the population? Similarly, how do we distinguish the signal from the noise in scientific research data? Each of these questions is the same question, but from a different point of view.

Many scientific journals use a critical-p-value gateway to *help* them to answer the question. This gateway says that if the (properly computed) p-value for an effect discovered in a research study is less than (or equal to) the journal's critical *p*-value, then the result is far enough above the noise to be a believable positive result *if* other important gateways are also successfully passed, as discussed below. Conversely, if the *p*-value is *greater* than the journal's critical *p*-value, then the result *isn't* far enough above the noise to be believable (*regardless* of whether the other gateways are passed), and thus the result is a negative result.

These ideas reflect the fact that the *p*-value is a measure of the "weight of evidence" that the effect under study exists in the population. The lower the *p*-value for a research result below the critical value, the greater the weight of evidence (in a reasonable statistical sense) that the effect exists, as explained in appendix B. The *p*-value is a reasonable measure of weight of evidence because its scale is designed to be directly (and reasonably) comparable from one research result to the next. Of course, the *p*-value is a *sensible* measure of the weight of evidence only if it is used properly, as discussed below.

The procedure of computing a *p*-value from research data and then comparing it to a critical *p*-value is often referred to as a "statistical test" of the research (or null) hypothesis. Many forms of statistical tests are available to enable us to test for the existence of the many forms of relationships and effects that are studied in scientific research.

Journals use a critical-*p*-value gateway because if the weight of evidence for an effect reported in a paper is insufficient, then the studied effect may not exist (at least in any detectable sense) in the population, and the paper may be reporting about mere statistical noise in the data. Journals don't want to waste space and promote misunderstanding by publishing papers that report about mere noise.

A journal using this gateway will say in its "Instructions to Authors" that a research paper submitted to the journal will be considered for publication only if the key *p*-value(s) in the paper is (are) less than the journal's critical *p*-value. Most such journals follow the convention of using a critical *p*-value of either 0.05 or 0.01.

2.4 False-Positive Errors

Unfortunately, the critical *p*-value (like all other similar statistical gateways, such as the critical Bayes factor) isn't perfect. Thus, sometimes when a critical *p*-value is used, the statistical test makes a "false-positive" error. (False-positive errors are also called "Type 1" errors, but that name is empty, so it is confusing to beginners.) A false-positive error occurs if a research study obtains a positive result, but (unbeknown to the researcher and the journal) the null hypothesis is actually true (or at least in effect true) in the underlying population—i.e., the studied effect (typically a relationship between variables) doesn't exist. In terms of the *p*-value, a false-positive error occurs if the *p*-value for a result is less than (or equal to) the critical *p*-value, but the null hypothesis is (unbeknown to us) either actually or in effect true.

A false-positive result in scientific research is misleading and costly because if the result is published, and if it is interesting, then it leads other researchers to try to replicate or use the nonexistent effect. The replications or use attempts of a false-positive result will invariably fail (because the effect doesn't exist), which is an unfortunate waste of resources. These failures are the source of the so-called "replication crisis" in scientific research, which is discussed later below.

False-positive errors in scientific research occur due to random chance and due to researcher errors. A journal can reduce the rate of false-positive errors due to random chance that are published in the journal by using a lower critical *p*-value—the lower the critical *p*-value, the lower the rate of false-positive errors published in the journal.

A researcher can reduce the rate of false-positive errors due to his or her own errors by ensuring that there is no reasonable alternative explanation (e.g., measurement errors, analysis errors, logic errors, failure to satisfy technical statistical assumptions, cherry-picking errors, confounding errors, data-entry errors, etc.) for their positive research results. Experienced researchers look carefully for reasonable alternative explanations of their results both in the vital design phase of the research and in the interpretation of the results. Researchers do this out of respect for their discipline and to avoid embarrassing false-positive errors.

Unfortunately, due to the whims of random chance, the publication of some false-positive errors in the scientific research literature is unavoidable. Fortunately, the investigative nature of science guarantees that a false-positive result will always be later exposed as being (likely) false if the result is of any importance.

False-positive errors are exposed through the process of independent replication of interesting results—a successful replication of a positive result greatly reduces the chance that the result is a false-positive error. Conversely, a *failed* replication of a positive result, though never definitive, somewhat *increases* the chance that the result is a false-positive error.

Replications are somewhat hidden in scientific research because, for the sake of moving forward, most replications involve significant enhancements or modifications to the original work. But in most scientific research there is a replication component in the background because science built out on existing ideas.

Some people incorrectly think that the rate of publication of false-positive errors in a scientific journal should equal the critical *p*-value used by the journal. However, the false-positive publication rate is generally somewhat higher than the critical *p*-value. For example, suppose that 20 percent of the research hypotheses that are studied in some field of scientific research are true. Therefore, the other 80 percent of the research hypotheses are, unfortunately, false—i.e., the corresponding null hypothesis is actually (or, at least, in effect) true in the population. (The 20 percent seems a reasonable guess in some fields of scientific research, such as in some areas of medical research.) And suppose that all the journals in this field use a critical-*p*-value gateway to publication of 0.05. If certain other sensible assumptions are satisfied, then it is easy to show mathematically that the long-run *rate of publication* of papers in the research literature of the field in which the key result is a false-positive error will be roughly 25 percent, as explained in appendix C.

The 25 percent false-positive publication rate in the research literature is based on the assumptions in the preceding paragraph, and the percentage will be different if the assumptions are different. However, the key point is that a significant percentage of new results published in scientific journals will be false-positive errors. Fortunately, we can easily reduce the rate of false-positive errors published in a journal by using a lower critical *p*-value, as noted above, though that has costs, as discussed below.

Interestingly, we can't *know* the actual rate of false-positive errors published in a field of scientific research because we generally don't know the percentage of research hypotheses that are true in the field, which is required to perform a reliable computation of the false-positive publication rate. If we wish to determine the percentage of research hypotheses that are true, then we would need to track negative results in scientific research. But science generally doesn't track negative results because tracking them reliably is difficult (because they generally aren't published, and most researchers don't care much about them) and tracking them is sensibly judged to be not worth the effort. Fortunately, we don't need to know the rate of false-positive errors in the published papers in a field of scientific research as long as we are aware that false-positive errors occur at a significant rate in every field, so we must be aware that any positive result could be a false-positive error.

The fact that a false-positive error is possible in any empirical research study explains why experienced researchers only *tentatively* draw conclusions in the papers that report their research studies. For example, a paper will typically say that the results of the research *suggest* (i.e., they don't *prove*) that such and such effect or relationship between variables exists in the studied population. This reflects the cautious logic of scientific reasoning—the conclusions of a research study must remain tentative until they are successfully replicated and accepted by the relevant research community.

Appendix D discusses the new initiative in some fields of scientific research to preregister scientific research studies. If this initiative is successful, it will help to eliminate some false-positive errors in scientific research and will enable us to track negative results in a field.

2.5 False-Negative Errors

If we can reduce the false-positive error rate in the scientific research literature by using a lower critical *p*-value, then why don't we set the critical *p*-value at an extremely low value? The answer is that the lower we set the critical *p*-value, the *higher* the rate of false-*negative* errors "in" (i.e., omitted from) the literature. (False-negative errors are sometimes called "Type 2" errors.) A false-negative error occurs when we obtain a negative result in a research study but, unbeknown to us, the effect actually exists in reasonable strength in the population. A false-negative error is costly because it leads to a loss of knowledge for society and a loss of reward for the researcher.

False-negative errors are rarely discussed because they are by their nature hidden and therefore aren't published. That is, there is no report of all the recent false-negative errors in a research field because nobody knows anything about these errors except that they exist in a substantial number. (Even if we were to track negative results, this couldn't directly tell us which of them are *false* negative results.)

False-negative errors in scientific research are due to random chance, due to inefficient research design, and due to researcher errors. A researcher can reduce the rate of false-negative errors in his or her research due to random chance or due to inefficient research design by designing the research to maximize the "power" of the statistical tests in the research under the available resources. Methods for maximizing the power of statistical tests are described in statistics and data-science textbooks and can be very effective for increasing the chance of finding good evidence of a studied effect (assuming, of course, that the effect actually exists in the population). Easy-to-use software (e.g., nQuery or software included in some modern data-analysis software systems) is available to help researchers to design their research studies so that the statistical tests will have the maximum possible power to detect the sought-after effects under the available resources.

2.6 The Critical p-Value Balances the Rates of False-Positive Errors, False-Negative Errors, and Costs

The consistent use of a critical *p*-value by a scientific journal enables the journal to help to balance the rates of false-positive and false-negative errors in the published literature in the field served by the journal. The lower an editor sets the critical *p*-value for a journal, the lower the published rate of false-positive errors in the journal, but the higher the rate of false-negative errors that are omitted from the journal.

The cost of a scientific research study is another key variable in the present discussion. This is because we can always decrease the rate of false-positive and false-negative errors in scientific research simply by appropriately spending more money on the research (to increase the quality of the research in various senses). This will reduce the error rates. But researchers invariably have a limited research budget, so we must compromise between the error rates and the costs. The critical *p*-value helps us to do this.

2.7 Why Must We Draw a Line?

Although the 0.05 critical *p*-value is lenient (to reduce the rate of false-negative errors), it amounts to drawing a line. Arguably, such a line is *necessary* for efficient scientific publishing. This is because without the line, scientific journals would be swamped with submitted research papers whose results are weak. This is because, as suggested above, negative results occur often in scientific research. The line makes it easy for a journal to sensibly and fairly reject the multitude of submitted (or potentially submitted) research papers with negative (i.e., inconclusive) results.

2.8 The Optimal Critical p-Value for a Field of Scientific Research

How should a journal decide where to draw the line in choosing the value of the critical *p*-value? The editors and the researchers in a field would like to choose the critical *p*-value for a journal (or the critical range for a confidence interval, or the critical Bayes factor, etc.) so that the choice yields the maximum long-term social, scientific, or commercial payoff from the research in the field served by the journal. Through experience, editors and researchers have judged that critical *p*-values of 0.05 or 0.01 (or sometimes lower) seem approximately correct to maximize the payoff of research in a field. That is, these critical values seem to provide an acceptable mix of true positive results, false-positive errors, true negative results, and false-negative errors in the research papers that are either published or potentially published by the journal.

Unfortunately, the choice of the optimal critical *p*-value can't be *exact* in a formal cost-benefit sense to maximize the payoff of scientific research in a field. This is because, in a practical sense, we can't properly measure the relevant costs, benefits, and other features of ongoing scientific research in a field, which would be necessary to determine the current exact optimal critical *p*-value for the field. So, a research community chooses the critical value based on intuitive sensibility among experienced researchers and journal editors. The choice is influenced by historical precedence in the relevant field and by overall custom in scientific research. The choice also depends somewhat on the journal's prestige—a more prestigious journal can use a lower critical *p*-value (and thereby reduce the published rate of false-positive errors), but still get a good supply of high-quality submitted papers. It seems possible that, through a process of consensus, the popular critical *p*-values of 0.05 and 0.01 as gateways to publication are close to optimal to maximize the payoff of scientific research in a field.

Although we (apparently) can't choose the *exact* optimal critical *p*-value for a field of science, it is still highly sensible for statisticians to *model* the operation of the *p*-value gateway because this helps us to understand the scientific publishing process, which is a fulcrum of science. Campbell and Gustafson (2019) develop a sensible mathematical model of the operation of the gateway, revealing interesting facts about the operation of scientific publishing, as discussed in appendix E.

2.9 The Relationship Between the p-Value and Other Important Gateways to Publication

As noted, all journals with a critical-*p*-value gateway will also have several other important gateways that a paper must also successfully pass before it will be accepted for publication. These gateways pertain to the journal-readers' likely interest in the paper, the scientific importance of the paper's result, the quality of the research design behind the paper, the related prior evidence, the plausibility of the studied mechanism, the absence of reasonable alternative explanations for the research result, and so on, at the journal's discretion.

It is important to note that the critical-*p*-value gateway is in no way *superior* to the other gateways. This is because the critical-*p*-value gateway is merely a sensible way to tentatively eliminate "chance" as a reasonable alternative explanation of a scientific research result. "Chance" is merely one of many possible reasonable alternative explanations of a research result that we must eliminate before we can reasonably believe the result.

However, although the critical-*p*-value gateway isn't superior, journals that use this gateway will usually apply it to a paper *first* because, as noted, many journals sensibly believe it is a *necessary* gateway, and because it can be applied quickly, typically by an editor in under five minutes. If a submitted research paper fails to successfully pass the critical-*p*-value gateway, then the evidence provided by the paper isn't strong enough, and the editor can immediately reject the paper, which saves the journal from spending any more time on the paper.

In practice, the critical *p*-value gateway usually serves its function *before* a paper is submitted to a journal. This is because most researchers know that submitting a paper with a key *p*-value greater than a journal's critical *p*-value would be a waste of time because the paper would be rejected.

2.10 Setting the Critical p-Value on a Case-by-Case Basis

Why not have a journal editor choose the critical *p*-value for each submitted paper at his or her discretion? That is, the editor could set the critical *p*-value for a paper depending on other aspects of the research, such as the perceived importance of the paper's result. This approach is fully consistent with the idea of maximizing the payoff of scientific research. And the approach is clearly sensible in theory, especially in cases when it is easy for an editor to correctly assess the veracity and importance of a paper's result. And, in general, this approach is always open to a journal editor.

However, many editors will agree that the veracity and importance of a research paper's result often can't be reliably determined until several years after the paper has been published (after the relevant scientific community has had a chance to think about the paper, replicate its results, and comment on it). Also, a discretionary choice of the critical *p*-value would, to be impartial, require an editor to perform a somewhat detailed study of each submitted paper which, given the large number of potential papers with nonsignificant results, and given the high complexity of many papers, the editors usually don't have time to do.

Also, there is little point in devoting time to studying a paper (no matter how brilliant or potentially important the paper might be) if there is insufficient weight of evidence that the discovered effect is real in the population. (If the evidence is weak, the paper may be reporting mere noise.) Also, the number of interesting papers that successfully pass the weight-of-evidence gateway for a reputable journal is usually somewhat or substantially greater than there is room to publish in the journal, so the journal has a more-than-adequate pool of submitted papers with convincing weight of evidence from which to choose the papers to publish.

So, to save time and conserve effort, the journal stipulates a simple, fair, and efficient rule—your weight of evidence must be better than (or equal to) our specified critical value or we won't consider your paper for publication because you may be reporting mere noise. This saves the editorial board's time, reduces the pool of submitted papers to a manageable size, and reduces (though doesn't eliminate) the chance of publishing false-positive results.

2.11 Alternatives to the p-Value

As suggested above, various sensible alternatives to the *p*-value are available to measure the weight of evidence provided by a scientific research result. These include the *t*-value, the confidence interval, the likelihood ratio, the Bayes factor, the second-generation *p*-value, and several other sensible measures. All these measures have critical values that (conceptually) operate the same way as the critical *p*-value. I compare some of the measures in the 2018 paper.

2.12 Bending the Rules About p-Values

Virtually all researchers are highly interested in having their research papers published because (assuming a published paper is correct) this advances human knowledge and because publications help to advance a researcher's career. As noted, journals generally aren't interested in publishing negative results. Therefore, some less experienced researchers will bend the rules about *p*-values to obtain a lower *p*-value (i.e., below the critical *p*-value) and to thereby give their research paper a chance of being published, if the other gateways can be passed. There are many ways to bend the rules. The rule bending may be either due to misunderstanding or due to unscrupulous intent.

It is generally impossible for the editors and reviewers of a journal to detect rule bending in a submitted paper prior to publication (because the necessary information to enable them to detect the rule bending is generally hidden). The rule bending distorts the scientific research literature in the sense that it adds more false-positive errors to the literature. That is, many (though not all) cases of rule bending will be reporting false-positive results.

Of course, the rule bending isn't the fault of the measure of the weight of evidence (e.g., isn't the fault of the *p*-value) and isn't the fault of the use of the critical value for the measure. Instead, the rule bending is due to the researcher's incomplete knowledge of the damaging ramifications of rule bending to science and to the researcher's career.

Rule bending damages science by adding an inordinate number of false-positive errors to the literature. Rule bending damages a researcher's career because if a researcher publishes a report in which the key result is a false-positive error (whether obtained by rule bending or not), and if the result is at least somewhat important, then the error will always be later exposed in replication research, which tends to diminish the researcher's reputation. The desire to properly advance science and the desire to maintain their reputations leads experienced researchers to be scrupulously careful about observing all the rules about *p*-values and about scientific research in general.

2.13 Do p-Values Make Decisions?

Some people think a *p*-value together with a critical *p*-value *decides* whether an effect is real in the relevant population. Of course, that is a serious misunderstanding because *p*-values can make false-positive and false-negative errors. Instead of making decisions on its own, the *p*-value *helps* us (the scientific research community, the decision-makers) to make decisions.

The ultimate decision-maker about the existence of a new effect in a population is always the relevant research community (i.e., not the journal, not the individual researcher, and certainly not the critical *p*-value). Due to the possibility of false-positive errors, the relevant research community generally requires one or two successful convincing independent replications of a new effect (with no convincing replication failures) before they will agree that the effect probably exists in the population.

2.14 Implications of a Low-Enough p-Value

If a researcher computes a *p*-value for an effect, and if the *p*-value is only slightly below the critical *p*-value for a particular journal, then this isn't *strong* evidence that the effect exists in the population because there is a real chance that the null hypothesis is true (or is effectively true) and the result is a false-positive result. However, if the *p*-value is less than the journal's critical *p*-value, then this is judged to be enough weight of evidence to merit *consideration* for publication, with acceptance for publication contingent on successfully passing all the journal's other gateways.

If a paper successfully passes all the gateways, then it is sensible to publish the paper even though the result may be a false-positive error. This is because publication informs other researchers in the field about the putative effect so that (if they find it interesting) they can replicate it and study it further, thereby increasing our understanding of the effect. And if the result is a false-positive error, and if it is interesting then, as noted, the error will be quickly identified through failed replication attempts.

2.15 What About the "Lost" Effects?

A possible serious problem with the critical-*p*-value gateway concerns the "lost" effects—the false-negative errors that are *true* effects in the studied population but were rejected for publication by the gateway. If a journal uses the gateway, is it inappropriately hiding these real effects from the research community?

No, not in the important sense. This is because a researcher who performs a research study that yields a weak result almost always has a large vested interest in the effect, typically larger than anyone else. So, this researcher has a strong desire to see his or her research results published. And the researcher has both the knowledge and (usually) the means to address the problem of the weak evidence.

So, if a researcher performs a research study and obtains a key *p*-value that is *greater* than the critical values used by the journals in the relevant field, and if the researcher *continues to believe* in the existence and the importance of the studied effect, then the researcher should perform a new research study of the effect with a more powerful research design to see if they can find convincing publishable evidence of the existence of the effect. If they can find good evidence (e.g., a low-enough *p*-value under a sensible research design), then a journal will be pleased to consider publishing a research paper describing the result.

The preceding ideas are closely related to the so-called "file-drawer problem" in which negative results in scientific research aren't published. Researchers often omit writing papers about negative results because there is generally no payoff for writing such a paper. However, if a paper reporting a negative result is written, then generally the paper must be relegated to a (computer or physical) file drawer, never to be published (because journals generally don't publish negative results). Some people think that negative results in file drawers is a bad thing.

Of course, if a negative result in a file drawer is a *true* negative result, then it *shouldn't* be published because (with rare exceptions) nobody is interested in a true negative result (because it doesn't tell us anything beyond the status quo). In contrast, if a negative result in a file drawer is a *false*-negative result and thus the effect actually exists in the population, then we can hope that the researcher maintains faith in the effect and repeats his or her research with a more powerful research design in order to obtain a publishable positive result. This hope is sensible because thoughtful people generally don't give up immediately on an idea when they think they are right.

2.16 Negative Results as Failures to Replicate

A noteworthy exception to the idea that negative results aren't interesting occurs when a negative result reflects a failure to replicate a published positive result. If the research obtaining a failure to replicate an earlier result was carefully performed, then the negative result somewhat calls the positive result into question, suggesting that the positive result may reflect a false-positive error. However, although such negative results are important, they generally still aren't published because that generally isn't necessary. This is because word about convincing negative results gets around the relevant research community informally (in seminars, conventions, and social media) to the point where the positive result is properly called into question. This happens efficiently because scientists are highly interested in knowing the truth (unvarnished by false-positive results) about the entities in the populations that they study.

2.17 The Size and Importance of a Detected Effect

If we use a *p*-value in a research study to help us to discover the existence of an effect, and if the *p*-value is below the critical *p*-value, and if all the other important gateways are also successfully passed, then this tells us that the results of the study provide good evidence that the studied effect exists (i.e., good evidence that the effect is real) in the studied population. However, as many authors have pointed out, a low *p*-value tells us nothing about either the *size* or the *importance* of the putative effect.

Of course, the size of a studied effect is highly important in scientific research. Therefore, once we have confirmed that an effect is likely real, we must consider its size because the effect won't be useful (at least in a practical sense) unless it is big enough. Fortunately, we can use various statistical measures to tell us the estimated effect size, such as the "effect size", the correlation coefficient, the contingency coefficient, and so on, as appropriate. (For technical reasons, effect size estimates for newly discovered effects are often somewhat [e.g., 20%] too high, and lower estimated effect sizes will be obtained when the effect is properly replicated. However, this isn't a problem if we are aware of the phenomenon.)

Similarly, the *importance* of an effect is obviously important. We decide whether a detected effect is scientifically or practically important by carefully considering its scientific or practical ramifications. Consideration of the importance of a newly discovered effect is the scientific bottom line of a research study, so experienced researchers consider the importance of a new effect in careful detail.

Clearly, both the size and the importance of an effect are of crucial importance in scientific research, so both must be carefully considered. However, there is generally no point in seriously considering either the size or the importance of an effect until we have first confirmed (e.g., with a relevant *p*-value) that the effect is (likely) real in the population.

2.18 Another Parallel View

The preceding discussion summarizes a general common view of the use of a critical *p*-value as a gateway to publication of a research paper in a scientific journal. Ioannidis (2019, 2019a) gives a parallel general view in support of the usefulness of the concept of statistical significance.

3. The Two Problematic Promises

With the preceding ideas in mind, let us now consider the two attractive outcomes promised in the TAS editorial if we abandon using a critical *p*-value (and if we abandon other equivalent approaches). As noted, the editorial refers to maintaining "the integrity of scientific publishing and research dissemination" (p. 2). Therefore, the promises should apply to the important case when a journal uses a critical *p*-value as a gateway to publication.

One promise in the editorial implies that if journals abandon using a critical-*p*-value gateway, then researchers will make fewer false-positive errors ("fewer false alarms", p. 1) and thus there will be fewer false-positive errors published in the scientific research literature. If this is true, it is an excellent reason to abandon critical *p*-values because (a) it is false-positive errors in the research literature that cause the "replication crisis" in scientific research and (b) false-positive errors are costly because they can lead to a substantial waste of resources as other researchers try to replicate or use the false finding. So, reducing the rate of false-positive errors in the scientific research literature is a very desirable goal.

Unfortunately, abandoning the critical-*p*-value gateway to publication won't lead to *fewer* published false-positive errors (in percentage terms), but will lead to more. This is because if the critical-*p*-value gateway is removed, then many less-conclusive research results will be submitted to journals for consideration for publication. And (due to the absence of the gateway) some of the less-conclusive results will be accepted for publication. A few of the papers with weak results will acknowledge that their results are essentially negative results. But most of the papers with weak results will present their results as positive results because positive results are more interesting, suggesting how we might predict or control the values of the response variable.

The rate of false-positive errors in the accepted less-conclusive "positive" results will be somewhat high because false-positive errors occur surprisingly often among the positive results in scientific research, and *more so* among the *weak* positive results. Therefore, if we abandon using a critical-*p*-value gateway to publication, then the body of new published scientific research papers (the repository of front-line scientific information) will unfortunately contain *more* false-positive errors (false alarms), not fewer.

I discuss the determinants of the rate of false-positive errors in a field of scientific research in the 2018 paper (app. B.10). Fricker, Burke, Han, and Woodall (2019) describe the apparent increase in the rate of false-positive errors in the journal *Basic and Applied Social Psychology* after the editors banned *p*-values and related ideas from the journal.

The other attractive promise in the editorial implies that if journals abandon using a critical-*p*-value gateway, then "researchers will [often] see their results more easily replicated" (p. 1). But if we abandon using a critical-*p*-value gateway, then more false-positive results will be published in the scientific research literature, as noted above. The rate of successful *replications* of the false-positive results will be zero (if we sensibly exclude any false-positive replications). Therefore, if journals abandon using a critical-*p*-value gateway, then researchers *won't* see their positive results more easily replicated. Instead, in percentage (success rate) terms, it will be more difficult to replicate positive scientific research results. This will make the "replication crisis" worse.

So, the two promises in the editorial appear to be unattainable, and we will obtain the opposite (less desirable) outcomes if we abandon using a critical-*p*-value gateway to publication. This raises two questions:

1. If we abandon using a critical-*p*-value gateway, what does this abandonment give us to justify the increase in false-positive errors and the decrease in successful replications in the scientific literature?
2. Does the central goal of objectivity in science *demand* a decisive impartial settable general gateway like the critical *p*-value to tell us whether a research signal is far enough above the noise to qualify for consideration for publication?

Acknowledgment

The author thanks Ross Macnaughton for insightful comments.

Appendix A: Some Ideas about the Null Hypothesis

Many authors (possibly beginning with Berkson, 1938, p. 527) have sensibly noted that the null hypothesis may never be *exactly* true in the real world in any research situation. Some authors go further and claim definitively that the null hypothesis *is* never exactly true in the real world in research.

Interestingly, although we can sometimes empirically prove (beyond a reasonable doubt) that a given null hypothesis is *false*, it appears to be impossible to ever empirically prove that a given null hypothesis is exactly *true*. This is because if we look for some effect and fail to find it, this doesn't imply that the effect doesn't exist, for it may exist but be too small to be detected by our current measuring instruments. So (unless we find a way to make our measuring instruments *perfect*), we can never empirically prove that a null hypothesis is *exactly* true.

Also, it appears to be impossible to empirically prove that the null hypothesis is *never* exactly true. This is because we can't examine every null hypothesis in the universe and somehow confirm that they are all false.

So even though a null hypothesis may or may not ever be *exactly* true, we (apparently) can't *know* whether this is so either in general or in any specific research situation. Of course, we do know that the null hypothesis is often "in effect" true in scientific research. And for certain relationships between variables there is (so far) no *detectable* relationship between the variables using our current best measuring methods and best statistical methods for detecting relationships. But in almost all such cases, a very weak relationship might still be present, and thus the null hypothesis may not be *exactly* true in the population.

Some authors conclude that since the null hypothesis may never be exactly true, therefore statistical hypothesis testing is illogical. However, whether the null hypothesis is ever exactly true is irrelevant for standard statistical hypothesis testing. This is because in standard hypothesis testing, we are trying to show that a null hypothesis is clearly *false*. In the relevant sense for this discussion, showing that a studied null hypothesis is false is a logically independent issue from showing whether a null hypothesis is ever true. In other words, in the case of showing that a null hypothesis is sometimes clearly false, it doesn't matter whether the null hypothesis is ever exactly true. If we can empirically show that a studied null hypothesis is clearly *false*, then (if the associated effect has been chosen carefully) this can lead to a substantial payoff—the knowledge of a new relationship between variables that we can use for reliable prediction or control.

In brief, it doesn't matter whether the null hypothesis is ever exactly true.

On a similar technical issue, some authors correctly note that the null value of the parameter whose value is tested by a statistical test is generally assumed to have "zero systematic error". These authors think it is implausible for a parameter to have zero systematic error, and this leads them to think that the null hypothesis is "implausible" and "uninteresting" (McShane, Gal, Gelman, Robert, and Tackett 2019, p. 236).

Of course, it is the true value of the relevant *population* parameter (not the value of the *estimate* of the population parameter) that is assumed to have zero systematic error. Under the standard frequentist approach to statistics, all population parameters (either of model equations or of the population directly) are assumed to have (unknown but estimable) *fixed* true values (possibly the null value) with zero systematic error. (In a few cases, the true value of a population parameter might vary "slowly" over time, but we generally view the value as being fixed in any analysis and fixed in most series of analyses across time.) So, under the frequentist approach to statistics, it is *the norm*, and thus isn't implausible, for a (fixed) population parameter to have zero systematic error. An *estimate* of the value of a parameter will have associated estimable error, but the unknown true value of the parameter in the population is sensibly viewed as being fixed with no error.

Appendix B: The Logic of the *p*-Value

Mathematically, the *p*-value is the *probability* of obtaining a result as "extreme" or more extreme as obtained in the current research situation *given that* the null hypothesis is true (and given that certain sensible assumptions are properly satisfied). The extent to which a result is "extreme" is measured in terms of how much the result *deviates* from the ideal result that we would get if the null hypothesis is or were true. We generally compute the *p*-value probability in terms of the estimated value of a *parameter* of an appropriate model equation. We compute the probability of the parameter estimate deviating as far as it has or farther from its null value if the null hypothesis is or were true.

For example, consider the relationship between continuous variables x and y in some population of entities. And consider a simple linear model equation for the relationship:

$$y = b_0 + b_1 x.$$

If there is *no* relationship between x and y, then the equation implies that the true (but unknown) value of parameter b_1 in the population will be zero, which is the null value for the parameter in this example. And if there *is* a relationship between the two variables, then the true value of b_1 will usually be different from zero in the population. So, we can generally determine whether a relationship exists by determining whether b_1, as estimated from relevant research data, is different from zero.

Statisticians have invented three main methods to estimate the values of parameters of model equations from relevant research data: the least-squares method, the maximum-likelihood method, and the Bayesian methods. Each of these methods is based on different theoretical principles, but in general, and when applicable, they all give the same or highly similar estimates of the values of the parameters for any given non-pathological set of data and any appropriate model equation.

These ideas are made more complicated by the fact that there will always be noise in the data. So, even if there is no relationship whatever between x and y, the estimated value of b_1 from relevant data will almost never be *exactly* equal to zero. So, we must determine whether the estimated value is *far enough* (i.e., "significantly") away from zero to enable us to reject the null hypothesis.

The simplest way to do this is to simply look at the estimated numeric value of b_1 and decide whether it is far enough from zero for us to believe that a relationship exists. However, this approach fails to take account of all the available and useful information because it ignores the available information about the estimated standard error of b_1. So, this isn't a sensible approach.

A sensible simple approach is to measure the distance of b_1 from zero as a multiple of its estimated standard error. Of course, this is the familiar *t*-statistic, which is used sometimes in the physical sciences, sometimes with a critical value of 2.0. (In rare cases when the standard error is *known*, and thus needn't be *estimated*, researchers sensibly use the closely related *Z*-statistic.)

Thus, in the example, if the estimate of b_1 is more than 2.0 (estimated) standard errors away from zero, physicists or chemists will say that this satisfies the 2σ criterion where σ stands for the estimated standard error, and therefore the effect is "significant", and therefore it likely exists in the population. (To be on the safe side, some research studies in the physical sciences use a higher critical *t*-value, sometimes as high as 5σ, which greatly reduces the chance of a false-positive error, but also greatly increases the cost of the research.) Of course, regardless of where we set the critical *t*-value, false-positive errors are always possible.

It is easy to show that in standard situations if the relevant sample size is greater than 30 or so, a critical *t*-value of 2.0 is roughly equivalent to a critical p-value of 0.05 in deciding whether a research result is positive or negative.

A technical problem with the *t*-statistic is that its distribution varies slightly depending on its "degrees of freedom", which depends on the sample size and other aspects of the research. This means that the value of the *t*-statistic isn't strictly comparable from one research situation to the next. In a rough-and-ready sense, this fact is ignorable in many research situations. However, in science it is sensible to be as strictly correct as sensibly possible. Therefore, it is sensible to compute the *fraction of the time* that the *t*-statistic will deviate as far as it does in the present case or farther from the null value if the null hypothesis is or were true. This computation is easy to do, and it can take proper account of the degrees of freedom and thus is (when the underlying assumptions are adequately satisfied) strictly comparable from one research situation to the next.

Of course, the fraction of the time that the *t*-statistic will deviate as far as it does or farther from the null value if the null hypothesis is or were true is simply the relevant *p*-value. Thus, the *p*-value is simply a measure of the deviation of the parameter estimate from the null value on a scale that is meaningfully comparable from one research situation to the next. (In cases where the *t*-statistic is appropriate, the *p*-value is a mathematically well-understood monotonic decreasing deterministic function of both [a] the absolute value of the *t*-statistic, and [b] the relevant degrees of freedom.) So, a critical value on the *p*-value scale is sensibly comparable among all research situations when the underlying assumptions of the *p*-value are adequately satisfied. (The assumptions are often adequately satisfied in carefully planned scientific research.)

The idea of testing whether a parameter is significantly different from its null value readily generalizes from the preceding simple example to other types of model equations and other types of relationships between variables. These ideas are a sensible, simple, and widely accepted approach for detecting relationships between variables and other effects in research data. The ideas are imperfect because they make false-positive and false-negative errors, but (it appears) nobody has found a better approach.

It is worth repeating that a low *p*-value supports the idea that a relationship (or other studied effect) exists *only if* there is no reasonable alternative explanation for the low *p*-value, where the set of reasonable alternative explanations is open-ended. That is, *any* alternative explanation is viable if it is "reasonable". The relevant scientific research community decides what is "reasonable".

Appendix O of the 2018 paper discusses whether the parameters of a model equation of a relationship between variables are "real".

Appendix C: The Rate of Publication of False-Positive Results in Scientific Research

The estimate earlier in this paper that approximately 25 percent of the published research results in a field will be false-positive results is based on the following assumptions:

- the percentage of research hypotheses that are true in the field is 20 percent
- the average critical *p*-value used in research in the field is 0.05
- the average power of statistical tests used in the field is 0.6
- the underlying assumptions of the statistical tests used in the field are always adequately satisfied
- all positive results in the field are published, and
- researcher errors in the field are negligible.

The estimate of approximately 25 percent is taken from figure B.1, which shows the percentage of false-positive errors as a function of the percentage of research hypotheses that are true in a field under the second through last assumptions above.

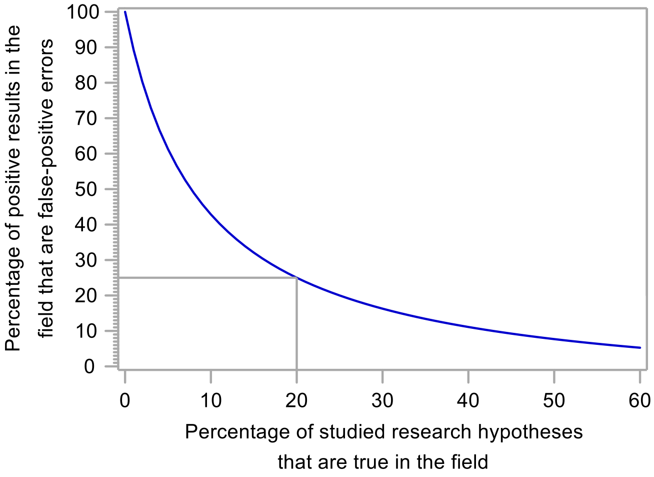

Figure B.1. The percentage of positive results in a field of scientific research as a function of the percentage of studied research hypotheses that are true in the field under the assumptions discussed in the text.

If the average critical *p*-value used in a field is less than 0.05, then the curving blue line on figure B.1will bulge proportionately closer toward the lower left corner of the graph, decreasing the rate of false-positive results. If the average power used in a field is less than 0.6, then the curving blue line will move proportionately further away from the lower left corner, increasing the rate of false-positive results. The R program (text file) that generates the graph explains the simple mathematics behind the graph (Macnaughton 2019). You needn't have or understand the R software to understand the logic in the program. It is easy to modify the program to plot the graph under different scenarios.

In real life, the above list of assumptions won't be satisfied for a branch of science. And in real life we can't *know* the value of the variable on the horizontal axis for a branch of science. We can't know this value because, as noted, we generally (and arguably sensibly) don't track negative results in scientific research. Similarly, in real life we can't know the average power of the statistical tests in a field of scientific research, though it seems clear in some cases with small sample sizes that the average power is less than 0.6, which increases the false-positive publication rate.

Although in real life we can't use the graph (or similar graphs), each branch of science will still have its own point somewhere below a downward-sloping diagonal straight line on the graph [running from coordinates (0,100) down to (100,0)]. Knowing that each branch of science has its own point somewhere on the graph helps us to understand the occurrence of false-positive errors in scientific research.

Appendix D: The Initiative to Preregister Scientific Research Studies

There is presently an initiative in some branches of scientific research to "preregister" scientific research studies. Under this initiative, a copy of the plan for a research study is submitted to a central repository (i.e., the plan is "registered") before the research is begun (or perhaps shortly after the research is begun, to allow for last-minute changes in the research design). This approach may include peer review of the research design at the time of registration (Center for Open Science, 2019) or it may only include registration with no review (Nosek, Ebersole, DeHaven, and Mellor 2018). If research studies are registered when they begin, this enables external auditing of the research, which can identify and discourage some types of researcher errors (in particular, unwarranted deviations from the research plan in search of positive results, which tends to lead to false-positive errors). Preregistration also enables tracking of negative results and has the important benefit of encouraging researchers to better plan their research studies.

The initiative for preregistration comes mainly from the so-called "replication crisis" that exists in scientific research—the fact that replication attempts fail in a substantial number of cases in some areas of research. These failures to replicate earlier results occur for two reasons. First, as discussed in appendix C, there will be a substantial number of false-positive errors published in journals in a field of scientific research due to the intentional leniency of statistical tests coupled with random chance. (As discussed in sections 2.5–2.8, the leniency reduces the rate of false-negative errors while maintaining sensible research costs.) Let us call the false-positive errors due to the leniency of statistical tests and random chance "natural" false-positive errors. Second, some of the false-positive errors in a field will be due to various kinds of researcher errors.

As discussed in appendix C, it is difficult to determine the rate of natural false-positive errors in scientific research. Likewise, it is presumably difficult (impossible?) to empirically separate the rates of natural false-positive errors from false-positive errors due to researcher errors.

It is noteworthy that requiring research preregistration will help to reduce the rate of false-positive errors due to researcher errors. But preregistration will have no effect on the natural false-positive errors. The researcher errors will only be a certain percentage (perhaps small) of the overall set of false-positive errors that occur in scientific research. Therefore, it is unclear whether the initiative to require preregistration of research studies will provide enough benefit to justify its cost.

Also, there is a sensible belief that researchers operate on their honor because they are professionals, though there are exceptions. So, there is a belief that an auditing system is inconsistent with the principle of honor in science.

Also, due to the investigative nature of science, even without the principle of honor, experienced researchers scrupulously strive to avoid errors. This is because, as noted, experienced researchers know that if their research is sufficiently important, then critical errors in the research will always be exposed in later replicating research, and exposure of such errors tends to diminish a researcher's reputation. So farsighted researchers, who seek the truth (as opposed to merely a positive result), do their best to eliminate the possibility of false-positive errors.

Instead of having a policing function through preregistration, it might be more sensible to attack the problem of researcher errors at its root cause. The cause lies in weak training which leads to researchers' incomplete knowledge of the damaging ramifications of researcher errors both to science and to a researcher's career. Thus, it might be better to spend the resources needed for the policing function on better statistical and data-science education. This would help researchers to understand that knowing the rules for proper scientific research and scrupulously following them is in their best interest.

Despite the preceding points, it is still conceivable that mandatory research preregistration would improve scientific research. As with any scientific question, the usefulness of preregistration must be determined through appropriate empirical research. If preregistration can demonstrably improve the quality of science on a sensible measure of quality, and if the improvement is cost-efficient, then clearly preregistration should be required.

The proponents of preregistration are diligence scientists, so empirically demonstrating the value of preregistration is very much their goal. However, in view of the difficulty of separating natural false-positive errors from researcher-induced false-positive errors, they may find this is a difficult task. However, if they are correct about the usefulness of preregistration, they will likely find a way.

Appendix E: Campbell and Gustafson (2019) Implications

Scientific researchers have a strong incentive to increase the count (quantity) of their published papers because we often use the count to determine a researcher's prestige, promotions, and salary. Of course, publication *quality* is scientifically more important than publication *quantity*. But quality is hard to measure. As noted in section 2.10, quality is hard to measure because it is generally impossible to reliably *directly* measure the quality of an individual research paper until several years after the paper has been published (when the ramifications and the importance of the paper can be determined through citation research).

In contrast, quantity is easy to measure (by merely asking a researcher for his or her publication list, which most researchers sensibly maintain). And a researcher's publication quantity somewhat reflects his or her research quality (in a cumulative sense) because each published paper has achieved at least the reasonably high level of quality required for publication.

So, publication quantity is often somewhat sensibly used as a surrogate for quality in evaluating researchers for prestige, promotions, and salaries. For example, it is easy to imagine a faculty salary committee at a research-oriented university noting that associate professor A published ten acceptable though not outstanding papers over the last two years, but associate professor B published only one acceptable though not outstanding paper, and awarding raises in salary accordingly. The people who produce more results get greater rewards.

Thus, to maximize prestige, promotions, and salary, it is sensible for researchers to conduct their research in a way that will maximize their publication count. Campbell and Gustafson (2019) show (through a sensible mathematical model and under reasonable assumptions) that, to maximize his or her publication count, a researcher should perform many research studies of long-shot novel effects using low statistical power for each study. The reason for performing long-shot studies is that positive results in such studies are intriguing and are therefore generally likely to be accepted for publication if all the relevant gateways are passed. The reason for using low power is to reduce resource requirements for each study and to thereby enable the researcher to increase the number of studies that he or she can perform under the resource budget, which will lead to more positive results.

Unfortunately, although this "scattershot" approach tends to generate many publications for a researcher, it also leads (under the Campbell and Gustafson model) to a relatively large number of false-positive errors in the researcher's research. Campbell and Gustafson show this fact in the small light-blue and green boxes in lower-left large square in their information-rich figure 1. So, the scattershot approach is good for researchers in the short term, but bad for science in the sense of yielding a substantial number of published false-positive errors. Of course, in the long term, the scattershot approach is generally bad for a researcher because, as noted, if the research is sufficiently important, then the false-positive errors will be exposed in replication studies, which will tend to diminish the researcher's reputation.

Also, the scattershot approach isn't sensible for a researcher who has a "cherished" research hypothesis that he or she wishes to demonstrate is true and wishes to test well. In that case, the researcher should design the research to have a high-power test to increase the chance that the research will obtain a positive result (assuming, of course, that the studied effect exists in enough strength in the population).

Whether a researcher should follow a cherished-hypothesis approach or a scattershot approach in his or her research depends on the relative quality of the cherished and scattershot research hypotheses that the researcher can generate. (Perhaps this matter is irrelevant for some or many researchers because they can only generate one type of research hypothesis, which is a hypothesis that seems possibly correct to them, so is worth studying.) It also depends on whether other researchers will likely try (and therefore fail) to replicate the researcher's false-positive results (under either approach), and it depends on the cumulative cost of these failures to the researcher's reputation.

If a researcher can generate cherished research hypotheses that are correct a "good percentage" of the time, then the researcher should use the cherished-hypothesis approach, which will provide a good percentage of true positive results and will tend to minimize the absolute number of false-positive errors that the researcher's research makes. The numeric value of "good percentage" could be modeled in theory, but the mathematical model would almost certainly be much too slippery to reliably estimate the required minimum percentage. So, we must apparently fall back on intuitive sensibility, which is difficult here, but appears to be all we have. Of course, ethical considerations point toward the cherished-hypothesis approach.

Concerned about the high rate of false-positive errors under the scattershot approach, Campbell and Gustafson sensibly consider in section 4 of their article the implications of journals imposing a statistical power requirement on research studies to reduce the rate of false-positive errors. They show that, under their model, the number of false-positive results under the scattershot research approach could be reduced by requiring that research studies have, say, at least 50 percent a priori power for the key statistical test for the expected effect. By "a priori power" they mean power that was computed before the research was begun based on assumptions about the likely effect size and its likely standard error (as opposed to rarely useful "retrospective power" that is computed based on the effect size and its standard error that are actually estimated in the research).

However, Campbell and Gustafson show in their table 5 that (under their model) although the power requirement will decrease the rate of false-positive errors, this approach will also decrease the rate of publication of "breakthrough" discoveries in scientific research. Also, as Campbell and Gustafson note, an a priori power requirement would be less reliable because it is difficult to compute a priori statistical power reliably (because one must make usually hard-to-justify but easy-to-game assumptions about the likely effect size and the likely standard error).

As Campbell and Gustafson also observe, although statisticians and researchers have discussed statistical power many times over the years, the concept has never caught on in the sense of being enforced in practical scientific research. That is, although we often see journals with a critical-*p*-value gateway, few journals (if any) have a statistical-power gateway. This may be because the concept of the *absolute* power of a statistical test is a chimera because it can't be measured ahead of time unless we make some speculative assumptions, which give the exercise an air of arbitrariness, which science generally avoids.

Although measuring *absolute* power generally isn't reliable, it is important to note that the concept of *relative* power of two statistical tests of the same research hypothesis under two equally costly research designs is very useful when a researcher is designing a scientific research study. This is because (if other things are equal) the design with the greater relative power is preferred. The greater power makes it more likely that the research will find the effect it is looking for (if the effect exists in the population).

Instead of requiring that statistical tests in research studies have a particular power, a sensible similar gateway would be to require that sample sizes in research be "sufficiently" large, as indirectly suggested by Campbell and Gustafson. This is sensible because (unlike power) sample sizes are directly measurable. For example, for a research study that compares different groups, a journal might require that at least 30 independent measurements of the response variable be made for each group. Of course, research studies are free to use many more measurements of the response variable per group, such as 100, 500, or more (to achieve the required statistical power in cases when the effect is likely to be weak). A minimum permissible sample size of 30 measurements per group would ensure that research projects comparing groups aren't ridiculously underpowered.

Due to the many different possible research designs, specifying a general and fair sample-size gateway for publication in a journal might be difficult, but it would help to ensure that statistical tests have enough power.

Alternatively, as noted earlier, it is possible that the current standard system for evaluating the weight of evidence with a critical *p*-value, but with no power or sample size requirement, is roughly optimal, possibly having achieved near-optimality through a form of consensus in journals over the more than 90 years and many millions of times that *p*-values have been used. If so, this would further attest to the perceptiveness of Sir Ronald Aylmer Fisher, who recommended the informal use of a critical *p*-value of 0.05 more than 90 years ago (1925, chap. IV, sec. 20).

It is instructive to note that Fisher later softened his view, saying that

> no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects [null] hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas (1973, p. 45).

This softening of Fisher's ideas may have been a reaction to criticism about the apparent arbitrariness of 0.05. However, Fisher was discussing the use of a critical *p*-value by an individual scientific worker to evaluate his or her results, and he wasn't discussing the use of a critical *p*-value as a sensible gateway to publication in a scientific journal where the use of a fixed critical *p*-value makes more sense.

As Campbell and Gustafson suggest in their "main recommendation" (2019, p. 369), further mathematical modeling of the scientific publication process may help us to determine the optimal approach, and (they sensibly suggest) such modeling is advisable before we adopt any policy changes about statistical hypothesis testing. (It would be odd if statisticians, who are arguably the world's best general modelers, would omit the use of models in deciding to reject the use of the concept of statistical significance.) In such modeling, a sensible goal is to determine the approach to scientific publishing that maximizes the long-term overall payoff of scientific research in a field—a property that, sadly, is difficult or impossible to reliably directly measure, which makes it hard to maximize. This may imply that we must again fall back on intuitive sensibility in the research community. However, due to the wisdom of the crowd, it seems a reasonable bet that intuitive sensibility over millions of *p*-values and thousands of editors over the years is reliable.

References

Baker, A. (2016), "Simplicity," *The Stanford Encyclopedia of Philosophy* (Winter 2016, ed. E. N. Zalta). https://plato.stanford.edu/archives/win2016/entries/simplicity/

Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test", *Journal of the American Statistical Association*, 33 (203), 527–536. https://doi.org/10.1080/01621459.1938.10502329

Campbell, H., and Gustafson, P. (2019), "The World of Research Has Gone Berserk: Modeling the Consequences of Requiring 'Greater Statistical Stringency' for Scientific Publication", *The American Statistician*, 73:sup1, 358–373. https://doi.org/10.1080/00031305.2018.1555101

Center for Open Science (2019), "Registered Reports: Peer-review before results are known to align scientific values and practices", https://cos.io/rr/

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd. https://psychclassics.yorku.ca/fisher/methods/index.htm

Fisher, R. A. (1973), *Statistical Methods and Scientific Inference* (3rd ed.). New York: Hafner. **In Fisher (1991)**.

Fisher, R. A. (1991), *Statistical Methods*, *Experimental Design, and Scientific Inference*. Oxford, UK: Oxford University Press.

Fricker, R. D., Jr., Burke, K., Han, X., and Woodall, W. H. (2019), "Assessing the Statistical Analyses Used in *Basic and Applied Social Psychology* After Their *p*-Value Ban", *The American Statistician*, 73:sup1, 374–384. https://doi.org/10.1080/00031305.2018.1537892

Ioannidis, J. P. A. (2019), "The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance", *Journal of the American Medical Association*. 321 (21), 2067–2068. https://doi.org/10.1001/jama.2019.4582

Ioannidis, J. P. A. (2019a), "Retiring significance: A free pass to bias" (letter to editor), *Nature*, 567, 461. https://doi.org/10.1038/d41586-019-00969-2

Macnaughton, D. B. (2018), "The *p*-Value Is Best to Detect Effects". https://matstat.com/macnaughton2018d.pdf

Macnaughton, D. B. (2019), pctfp.R (text file containing an annotated computer program in the R language). https://matstat.com/pctfp.r

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019), "Abandon Statistical Significance," *The American Statistician*, 73:sup1, 235–245. https://doi.org/10.1080/00031305.2018.1527253

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

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019), "Moving to a World Beyond '$p < 0.05$'", *The American Statistician*, 73:sup1, 1–19. https://doi.org/10.1080/00031305.2019.1583913