Journal of Experimental Social Psychology 66 (2016) 39–46

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Does the conclusion follow from the evidence? Recommendations for improving research☆

Andrew H. Hales

Department of Psychological Sciences, Purdue University, Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Received 15 August 2014
Revised 21 September 2015
Accepted 22 September 2015
Available online 24 March 2016

Keywords:
Replication
Publication bias
Type I error
Type II error
Type III error
Type IV error

ABSTRACT

Recent criticisms of social psychological research are considered in relation to an earlier crisis in social psychology. The current replication crisis is particularly severe because (1) psychologists are questioning the accuracy of findings rather than the meaning of findings, and (2) researchers are responding to real scientific failures, rather than hypothetical scientific failures. I present an expanded model of statistical decision making that can be used to help researchers draw more reliable conclusions. Based on the premise that drawing conclusions on relatively bad evidence is an error, Type III and IV errors are introduced as categories representing statistical decisions that align with reality, but do not follow from the available evidence. Treating these as errors helps researchers and evaluators of research to draw more reliable conclusions. From the perspective of this model I discuss procedures that researchers can use to not only produce more replicable results, but also conduct more powerful statistical tests.

1. Introduction

Psychology is experiencing a crisis of confidence in the reliability of published findings. Researchers are currently giving a remarkable amount of attention to bias within the publication process as well as the replicability of published results. Though there has been some disagreement as to the magnitude and justifiability of this concern (e.g., Cesario, 2014; Simons, 2014), there is little doubt that psychologists are now, more than ever, critically examining the way research is conducted and communicated.

The importance of replication and replicability in a scientific discipline such as psychology is difficult to overstate. The methods section of a research report is crucial in distinguishing science from other ways of understanding the universe (e.g., Hansson, 2013). It is not an accident that students are instructed to write method sections with enough detail for a reader to be able to replicate the procedure. Unlike tradition, religion, and other non-scientific ways of understanding the universe, science does not require that claims be accepted on authority. As students are instructed in methodology textbooks, “scientists do not accept on faith the pronouncements of anyone, regardless of that person's prestige or authority” (Cozby, 2009, p. 5). Ideally, any scientific claim can be verified by a skeptical reader of the research.

Unfortunately, there are many ways this ideal can go unmet. In the case of social psychology, many are concerned that researchers are not investing resources into verifying scientific findings through replication, and when they do, inconsistent results are obtained. If a body of scientific findings are not replicated, or worse, not replicable, then what is left to separate a scientific approach from other ways of understanding the world? Given the importance of replication to science broadly, and psychology specifically, there is understandable concern over issues of replication. While social psychology is certainly not the only area of science questioning its replicability (e.g., biotechonology; How Science Goes Wrong, 2013, October), we have an opportunity to change our research practices for the better.

This paper has two primary purposes. The first purpose is to bring historical context to recent replication discussions by considering the current crisis in comparison to an earlier crisis in social psychology. This historical review concludes, as many have recently observed, that methodological, not just theoretical innovation will be necessary to improve research. The second purpose follows this conclusion with the proposal of a reconceptualization of how decisions are made in psychological research that is sensitive to the methodological quality of a research claim. Through the lens of the new model I provide a system researchers can use to help identify research that is at risk of being in error, and promote practices that reduce errors.

2. Historical perspective

The current crisis is not unprecedented. Beginning towards the end of the 1960s social psychology experienced a “state of profound

☆ I am grateful to Christopher R. Agnew, William G. Graziano, Brad J. Sagarin, Kenneth Tan, and Erica D. Schneid, and Sara E. Branch for comments on earlier drafts of this manuscript. I also gratefully acknowledge those who organized and contributed to the Purdue Colloquia Series on Publication Bias including Gregory Francis, Anthony G. Greenwald, Uli Schimmack, Brian A. Nosek, and Uri Simonsohn.

E-mail address: ahales@purdue.edu.

http://dx.doi.org/10.1016/j.jesp.2015.09.011
0022-1031/© 2015 Elsevier Inc. All rights reserved.

intellectual disarray" (Ring, 1967, p. 119). This earlier crisis is similar to the current crisis, but with important differences. Researchers at the time identified a wide range of problems with social psychology in general, and the common paradigm of laboratory experimentation in particular (Elms, 1975; Jost & Mcguire, 2013; McGuire, 1967; McGuire, 1973; Ring, 1967). Some of these concerns were settled during the crisis and have not emerged since. For example there is currently little discussion of the ethics of deception, or the relative value of basic vs. applied research, two topics that were of concern at that time (McGuire, 1973). However, other concerns, such as those about the quality and meaningfulness of social psychology findings have returned in full force (Gergen, 1973; McGuire, 1973).

2.1. Comparing crises

2.1.1. External validity threat vs. statistical conclusion validity threat

Despite the obvious similarities (i.e., widespread skepticism directed at the value of scientific findings), there are two important differences between the earlier crisis and the concerns we face today. First, the problems faced in the 1960s and 1970s were problems of external validity; scholars were worried that findings would not replicate outside of tightly controlled laboratory situations. According to one critique calling for an increased use of field research, "what the experiment tests is not whether the hypothesis is true, but rather whether the experimenter is a sufficiently ingenious stage manager to produce in the laboratory conditions which demonstrate that an obviously true hypothesis is correct" (McGuire, 1973, p. 449). Additionally, some argued that social psychology should be classified as a historical discipline, because findings could represent artifacts of a particular historical period and cultural setting, and there was no guarantee that they could be replicated across different historical zeitgeists (Gergen, 1973). Both of these criticisms emphasize external validity: whether an established relationship between variables can be observed across different populations and situations (Shadish, Cook, & Campbell, 2002; Maner, 2016).

In contrast, the current crisis is much more troubling. Social psychologists are doubting the statistical conclusion validity of research findings: "the validity of inferences about covariation between two variables" (Roberts, 2012.; Shadish, et al., 2002, p. 512). Today there is a concern that the results of experiments reported in journals are describing phenomena that were not even produced in the initial research demonstrations, let alone in more generalized contexts (Simmons, Nelson, & Simonsohn, 2011). During the first crisis, the question for any given research claims was "does this phenomenon exist outside of the laboratory?". However, the question being asked today is much more unsettling: "does this phenomenon exist at all?".

At first it may seem that the current crisis is actually one of the external validity. After all, there is a current concern that research findings can be produced in one context (e.g., a specific lab or research team), but not other contexts. However, publications of failed replications are being interpreted as an indictment of the quality of the original research, and are being used not as evidence that certain effects are very delicate, but rather evidence that the effects are illusory or artifactual (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012). In fact, in some cases failed replications have been foreshadowed by analyses showing that the evidence reported in support of a finding can be implausibly positive. For example, multiple analyses have questioned whether findings in support of *precognition* (Bem, 2011) are too good to be obtained without using questionable research practices (Francis, 2012; Schimmack, 2012). In line with these analyses, researchers who have replicated Bem's procedures have not replicated his results (Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Given the analyses questioning the truth of the original research, failed replications such as these are no longer being disregarded as uninformative. Instead they are increasingly being interpreted as evidence that some findings may not be correct.

2.1.2. Real vs. hypothetical scientific failures

The second difference between the two crises relates to hypothetical versus actual failures of social psychological research. During the first crisis, researchers were appropriately concerned with the quality and meaning of the research they were producing. They were proactively seeking to improve scientific practices to optimize the accuracy of scientific knowledge and theory. In contrast, the current crisis was initiated by concrete instances of research practices leading to the publication of incorrect or fraudulent findings. The present crisis is a response to real, identified failures in the publication system. This has led to a much stronger push for formal top down changes in the scientific process in addition to the familiar but ignored calls for bottom up changes (e.g., improving power; Cohen, 1962).

Consensus seems to have emerged that the concerns about replicability are intertwined with current research and publication practices, and that to solve the current crisis we will need to change the way we conduct and communicate research. In the spirit of updating such practices, I will present a framework that elaborates on how statistical decisions are considered, with the goal of promoting replicability. The framework is based on two recent observations: social psychology primarily reward findings that are significant and interesting, and non-representative research can be identified.

2.2. Results must be significant and interesting

Recent criticisms have started with the observation that psychologists have an interest in obtaining significant results, which biases research practices (Nosek, Spies, & Motyl, 2012). Significant results are easier to publish than null results, largely because null results lead to ambiguous conclusions (Greenwald, 1975).

However, results must be more than significant to be published, they must also be *interesting*. The pressure to produce interesting results is a major contributor to the difficulties faced by psychologists who want to promote accurate conclusions while maintaining a successful career. If researchers pursued obvious questions, or even questions that are likely, but not guaranteed, the pressure to produce significant results would not necessarily lead to the irresponsible use of researcher degrees of freedom (the undisclosed use of statistical liberties to achieve desired test outcomes). However, social psychology has traditionally placed a very high premium on *interesting* findings (Elms, 1975). Common buzzwords used to congratulate psychologists who have achieved this in a research program include *counterintuitive*, *novel*, and *surprising*. Bem's (2011) argument in favor of precognition is perhaps the pinnacle of counter-intuitiveness; causes need not precede their effects. A familiar litmus test for whether findings have effectively defied intuition, is to ask whether your research claim is something that your grandmother could have told you. If so, your findings are uninteresting.[1] Social psychology has rewarded those who "conduct the most contrived, flamboyant, and mirth-producing experiments" with "the highest score on the kudometer" (Ring, 1967, p. 117).

The norm is not always to use Abelson's somewhat broad definition of *interesting results* as those that "change what scientists believe about important causal relationships" (Abelson, 1995, p. 158). Instead, research findings must clear a much higher hurdle; they must be counter-intuitive, yet still be grounded in predictions that could have been drawn from the existing literature. These features are not always compatible. In fact, they are often in direct competition. This creates pressure to engage in questionable research practices such as using

[1] Researchers interested in intelligence should study why the grandmothers of reviewers and editors tend to know so much more than the grandmothers of authors.

researcher degrees of freedom (Simmons et al., 2011), or *hypothesizing after results are known* (HARKing: Kerr, 1998).[2]

2.3. Identifying researcher degrees of freedom and HARKing

There is a growing interest in identifying situations where published results do not accurately and completely represent a program of research. Francis (2012) and Schimmack (2012) have used statistical procedures to assess the probability that a package of studies is in some way not a representative depiction of the research that was conducted.

According to this approach, it is possible for a set of studies to contain an implausible number of successful replications, and thereby undermine the evidence for a given phenomenon. In a typical research report of multiple tests of the same hypothesis, the number of studies that successfully reject a false null hypothesis is necessarily and calculably limited by their statistical power (probability of detecting a true effect). Even with well powered studies (which are rare in psychology; Maxwell, 2004), it is to be expected by the laws of probability alone that a certain number of the studies will fail to replicate the result. Rather than representing embarrassing evidence against the researcher's effect, the failures to replicate merely represent sampling error (see Giner-Sorolla, 2012). Further, it is highly implausible for an underpowered package of studies to uniformly reject a null hypothesis. In fact, the implausibility of such outcomes can be calculated using formulas and procedures provided by Francis (2012) and Schimmack (2012). A related approach to detecting implausibly positive results is to compare a distribution of reported p-values (perhaps bunching up right beneath .05) to what the distribution would look like if the effect was real and the studies were representative (Simonsohn, Nelson, & Simmons, 2014).

According to Francis, in addition to researcher degrees of freedom and HARKing committed by authors, a set of studies might also reject null hypotheses beyond chance levels because reviewers or editors chose to exclude failed replications. Francis and Schimmack have used this method to serially target and critique articles that report implausibly successful sets of studies. In this way, to some extent, the effects of researcher degrees of freedom and HARKing can be assessed at the micro level.

Applying these methods to individual packages of studies does not lead to conclusions about the reliability of social psychological findings in general, especially when the individual packages to be evaluated are selected in a non-random way (Simonsohn, 2012). However, these methods are extremely valuable to individual researchers interested in evaluating the plausibility of certain findings. In addition to the inherent value in being able to more accurately judge the quality of a potentially meaningful research claim, these methods can help guide future research. A researcher who is considering investing time and resources into a line of research will be interested in knowing the likelihood that a seminal package of studies was the result of biased analysis or reporting (or at least overestimates an effect size; see Fabrigar and Wegener, 2016, for a discussion of conclusions to be drawn from these analyses).

To summarize, there is a pressure to produce interesting findings. This can make inappropriate research practices look attractive. However, tools currently exist to identify cases where these practices have been put to use. In other words even though we do not know if a specific research claim is true or false, we may be able to discover whether the research itself shows signs of resulting from inappropriate practices. This is useful information that ought to influence how we interpret statistical conclusions. Next I introduce and outline a model that accounts for this valuable information.

3. Restoring confidence in psychological research

3.1. An expanded model of statistical decision making

It has been observed that individual researchers are motivated to seek truth, but are also motivated to amass publications for career advancement. To the extent that the most efficient ways to publish articles do not promote truth (because questionable research practices enhance the chances of publication), individuals are tempted to show reduced concern for the truth of published findings (Nosek & Bar-Anan, 2012; Nosek et al., 2012). Critically, a research claim may be true, in that it aligns with reality, yet still not follow from the available evidence. These are claims that do not show sufficient concern with truth, but nonetheless happen to align with reality.

This type of research claim results from the use of researcher degrees of freedom or HARKing to test research claims that are actually correct. These practices are not used to actively deceive an audience. The researcher does not know whether their hypothesis is true or not, and is therefore not lying when engaging in these practices. However, these practices do prioritize finding significant results over finding results that accurately describe reality. In this sense, conclusions drawn from these research practices are errors.

Because a statement may be technically true (in that it aligns with reality) but still be the result of research practices that do not prioritize the truth, we can hypothetically restructure the classic four-outcome statistical testing scheme. In the classic scheme a null hypothesis is identified. This hypothesis is acknowledged to be either true or false, and it cannot be known with certainty which is the case.[3] The researcher calculates the probability of obtaining the available observations given that the null hypothesis is true, and makes a decision to either reject or not reject the null hypothesis. This produces four possible outcomes of a statistical decision: correctly rejecting a false null hypothesis, correctly retaining a true null hypothesis, incorrectly rejecting a true null hypothesis (a Type I error), and incorrectly retaining a false null hypothesis (a Type II error).

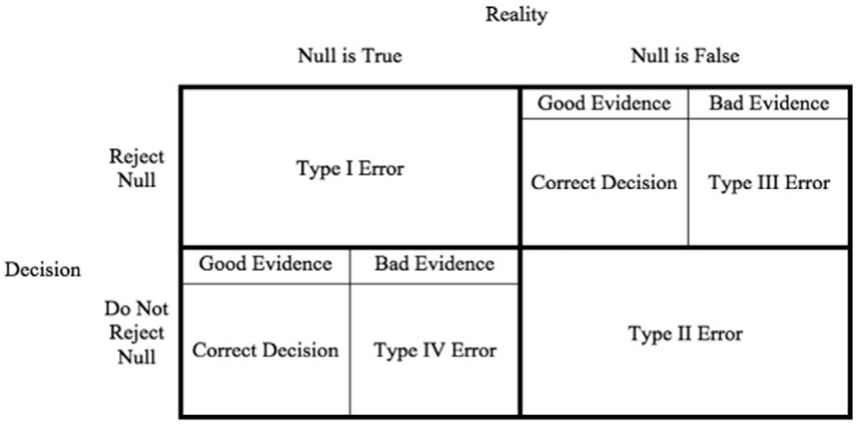
Notably, these four outcomes are not sensitive to the quality of the evidence used in support of the research claim. Statistical tests that have been contaminated by questionable research practices are treated the same as tests that are uncontaminated and purely confirmatory. By distinguishing between research claims based on the quality of the underlying evidence, we can draw more nuanced statistical conclusions.

We can imagine a model of statistical decisions that is expanded to include Type III and Type IV errors (see Fig. 1). A Type III error is committed when a researcher (or a consumer of research) correctly rejects a false null hypothesis, but does so on the basis of evidence gathered using researcher degrees of freedom, HARKing, or the suppression of studies that do not support the hypothesis. A Type IV error is committed when one correctly fails to reject a true null hypothesis, but does so on the basis of bad evidence such as a small sample size, or poor measurement. These are cases where concern for truth, at least in part, is replaced by concern for finding significant results.

The idea of articulating additional types of statistical errors goes back as far as 1948 (Mosteller, 1948). In the past, researchers have used Type III error to refer either to getting the right answer to the wrong question (Schwartz & Carpenter, 1999), or to mistaking the direction of an effect (Shaffer, 2002). Additionally, Type IV errors have been defined as "the incorrect interpretation of a correctly rejected hypothesis" (Marascuilo & Levin, 1970, p. 398). While these existing meanings have much to offer, the current use of the term is meant to be separate and distinct from these meanings. Here I consider Type III errors as instances in which the researcher is mistaken, not in the

[2] It should be noted that while some questionable research practices constitute threats to statistical conclusion validity, HARKing does not have a clear place in the Campbell causality model. It threatens reliability, but seems to be something more then simply misapplied use of statistics.

[3] A null hypothesis is never strictly true in the sense that variables being compared are perfectly equated or unrelated in the population (Cohen, 1994). When I refer to a true null hypothesis I mean a situation in which two variables are not *appreciably* different (or related, depending on the test).



<table>
<tr><td></td><td></td><td colspan="4">Reality</td></tr>
<tr><td></td><td></td><td colspan="2">Null is True</td><td colspan="2">Null is False</td></tr>
<tr><td rowspan="2">Decision</td><td>Reject Null</td><td colspan="2">Type I Error</td><td>Good Evidence
Correct Decision</td><td>Bad Evidence
Type III Error</td></tr>
<tr><td>Do Not Reject Null</td><td>Good Evidence
Correct Decision</td><td>Bad Evidence
Type IV Error</td><td colspan="2">Type II Error</td></tr>
</table>

Fig. 1. Expanded model of statistical decision making.

accuracy of their conclusion, but in that their conclusion is based on poor evidence.

Type III and IV errors[4] can be illustrated by considering the differing claims of Bem (2011) and those who have performed failed replications (Galak et al., 2012; Ritchie et al., 2012; Wagenmakers et al., 2012) under different hypothetical assumptions about the null hypothesis. Imagine, for the sake of example, that precognition is a real phenomenon (a false null hypothesis). In terms of statistical decision making, Bem committed a Type III error by concluding that differences between experimental conditions were not due to chance, and instead reflect precognition. He happens to be correct, in that precognition is real. However, the nine studies used as evidence were implausibly consistent given their power, suggesting that questionable research practices were used (Francis, 2012; Schimmack, 2012). Therefore, Bem's rejection of the null hypothesis is based on bad evidence, so he committed a Type III error. On the other hand, still assuming that precognition is real (H_o = false), those who have replicated Bem's procedures but failed to detect effects have committed the traditional Type II error of failing to reject a false null hypothesis (perhaps because the studies were not sufficiently powered, or did not satisfy certain known or unknown boundary conditions).

Conversely, we can imagine that precognition is not a real phenomenon (a true null hypothesis). In this case Bem's conclusion that precognition is real represents the traditional Type I error of rejecting a true null hypothesis. Those who fail to replicate Bem would be correct in concluding that the effect is not real.

Cases in which conclusions are drawn on bad evidence – Type III and IV errors – are properly considered mistakes in the sense that the conclusions do not follow from the evidence. In other words, it is an error to draw conclusions on faulty evidence, regardless of the incidental truthfulness of the conclusion. By analogy, most people would consider it an error to place an unwise bet in a game of chance, even if the winning outcome happens to occur.[5] The expanded model broadens the definition of "error" from the narrow use it receives in statistical testing and alerts researchers to the mistake of believing a true claim based on bad evidence.

This model is proposed as a tool for thinking about how psychologists draw conclusions based on data, not as a replacement to null hypothesis testing. Unlike other proposed statistical approaches, such as the use of Bayes' Theorem (Kruschke, 2010), or effect size estimation (Cumming, 2014), the expanded model does not require statistical retraining or major departures from current analytic practices. Although these Bayesian and effect size estimation approaches have many advantages, the consideration of Type III and IV errors can be immediately used to increase the reliability of research claims. Of course Type III and IV errors cannot be formally calculated. Their utility lies in raising consciousness about what constitutes good reasons for endorsing a research claim, and providing a language to consider claims that are theoretically plausible, very likely to be true, and yet only supported by evidence that shows signs of contamination.

How can it be a mistake to believe a true claim? Recall that according to the logic of null hypothesis statistical testing, one is never certain of the truth of a null hypothesis. Rather, the researcher calculates the probability of obtaining the evidence at hand, assuming the null hypothesis is true. But the actual truth of the null hypothesis remains unknowable. On the other hand, the quality of the evidence is knowable in principle (even if it is frequently not disclosed in practice). Researchers and consumers of research at least have some access to knowledge regarding the quality of evidence in support of a claim (e.g., whether an analysis was run both with and without outliers). In other words, the truth of the null hypothesis is unknowable, whereas the quality of the evidence at hand is knowable.

This is an attractive feature of the expanded model of statistical decision making over the traditional model. One can determine whether an error has been made without knowing the state of the universe. One only needs to know the quality of the evidence at hand. For example, if a researcher rejects a null hypothesis and it can be determined that the researcher engaged in optional stopping (without an a priori sequential analysis plan; Lakens, 2014), one can conclude that the researcher has committed some type of error. If the null hypothesis is true, the researcher committed a Type I error. If the null hypothesis is false, the researcher committed a Type III error; their decision aligns with reality, but does not follow from the available evidence (or is at least overstated given the evidence). In this case the quality of evidence does not justify the researcher's claim, and they have committed the error of believing (and promoting) a claim on bad evidence (see Fig. 2 for a flowchart representing these decisions).

A second attractive feature of the expanded model is that it broadens the list of agents who participate in the decision making processes. In the traditional four-outcome model, a researcher is the sole agent who makes a decision. The researcher then shares that decision with other interested scientists, who may of course harbor skepticism. Under the expanded model, those conducting research may commit any type of error, but those evaluating the research may commit only Type III and IV errors. That is, a researcher can be wrong about whether a null hypothesis is true or false (traditional Type I and II errors), and they may also make the error of concluding that a null hypothesis is true or not true despite poor evidence (Types III and IV errors). However, *second generation* errors are also possible. An editor, reviewer, or reader may fail to notice that researchers drew a claim on bad evidence and thus

[4] Because researchers rarely take nonsignificant *p* values to be evidence in favor of a null hypothesis, Type IV errors are most likely less relevant than Type III errors.

[5] Of course, virtually all bets are unwise in the sense that the expected value favors the house. But a reckless bet does not retroactively become wise when an unlikely winning outcome is realized.

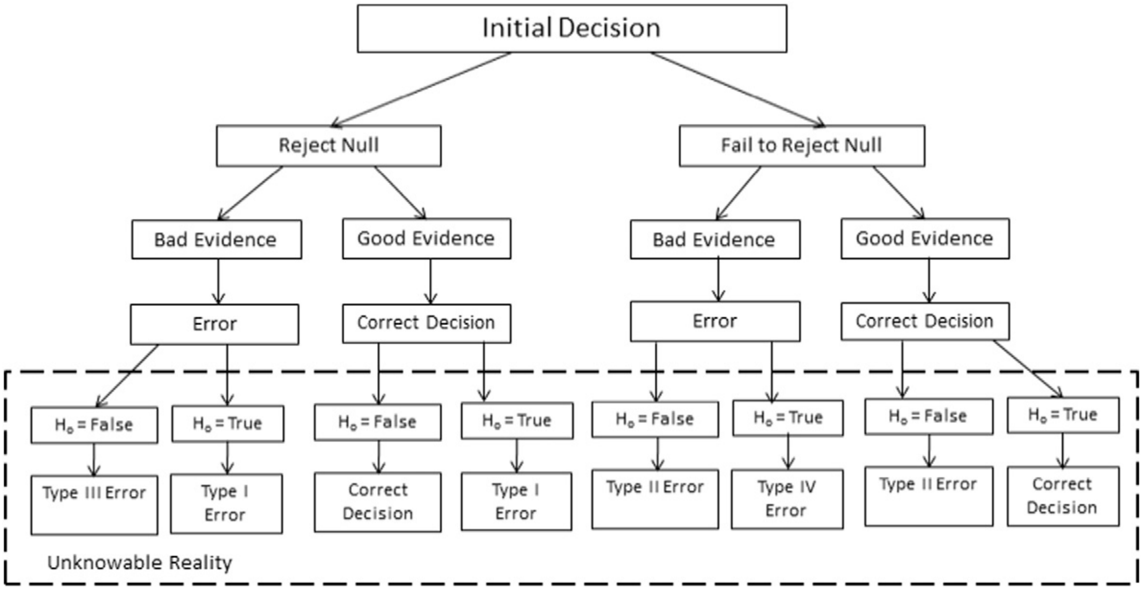


Fig. 2. Flowchart of decision-making with the expanded model.

commit the Type III error of believing the research claim, even though it aligns with reality, when the evidence is bad.

This reasoning raises important questions about what is to be regarded as good and bad evidence, and how information regarding the quality of the evidence is to be shared with consumers. Of course, evidence in not simply bad or good, but instead varies along a continuum from bad to good. Questions about how to weigh and interpret evidence from different research practices are complicated and it is unlikely that a uniform consensus will emerge. However, much of the recent work has made substantial progress towards understanding what constitutes relatively bad evidence and how consumers of research can be better able to identify such evidence.

A final noteworthy feature of the expanded model is that it narrows the universe of outcomes that researchers are motivated to reach. In terms of surface area, Fig. 1 shows that less space is available for correct decisions compared to the amount of space for correct decisions

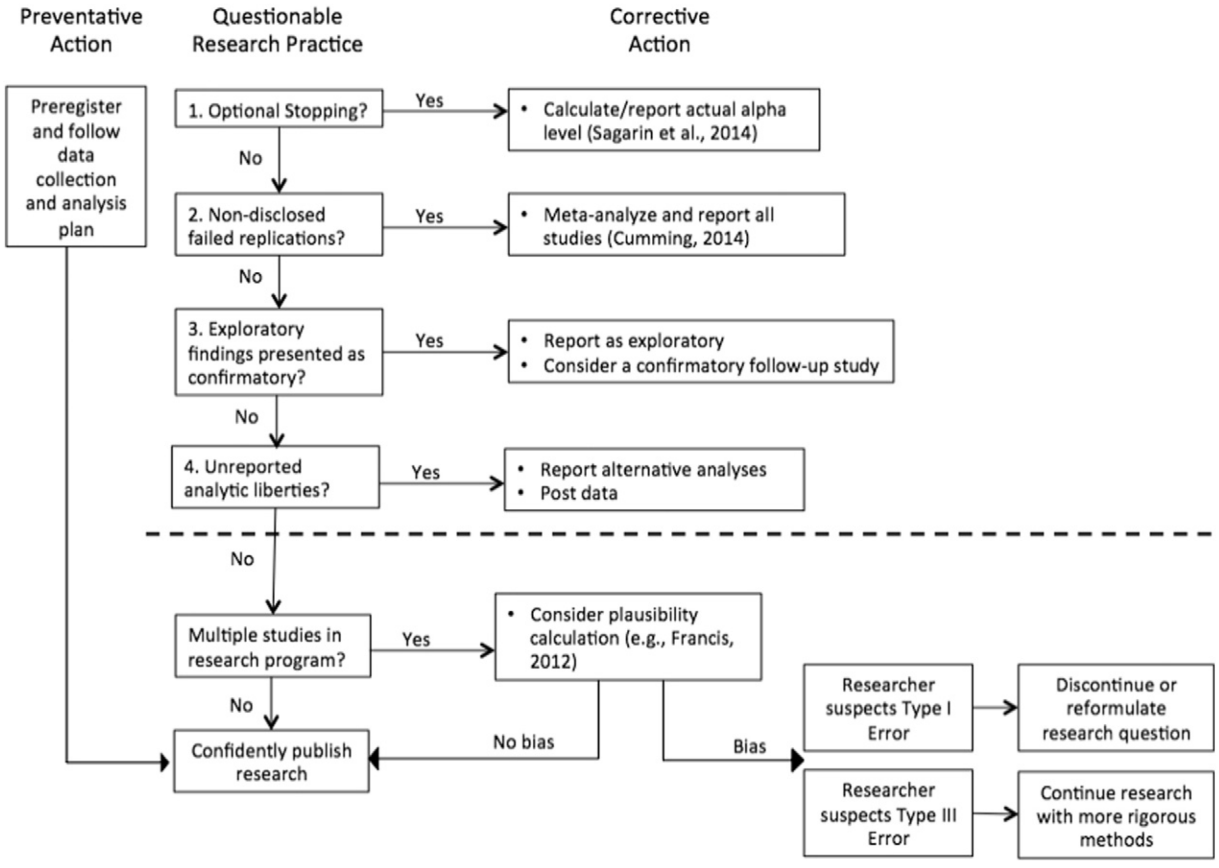

Preventative Action
Questionable Research Practice
Corrective Action
Preregister and follow data collection and analysis plan
1. Optional Stopping?
Yes
• Calculate/report actual alpha level (Sagarin et al., 2014)
No
2. Non-disclosed failed replications?
Yes
• Meta-analyze and report all studies (Cumming, 2014)
No
3. Exploratory findings presented as confirmatory?
Yes
• Report as exploratory
• Consider a confirmatory follow-up study
No
4. Unreported analytic liberties?
Yes
• Report alternative analyses
• Post data
No
Multiple studies in research program?
Yes
• Consider plausibility calculation (e.g., Francis, 2012)
No
Confidently publish research
No bias
Bias
Researcher suspects Type I Error
Discontinue or reformulate research question
Researcher suspects Type III Error
Continue research with more rigorous methods

Fig. 3. Guidelines for minimizing Type I and Type III errors.

provided by the traditional 2 × 2 grid. This model reduces Type I and II errors by proposing that a statistical claim only be regarded as correct if it is based on good evidence (similar to Francis's assertion that research showing signs of publication bias be regarded as anecdotal; Francis, 2012).

4. Applying the new model to increase the accuracy of research claims

Fig. 3 outlines specific guidelines that researchers can follow to minimize errors of both types. The middle column lists practices that can increase Type I and Type III errors. The right column lists corresponding strategies that can be used to minimize the effects of these practices, and impose transparency so that readers are able to accurately determine how well research claims are supported by evidence. The left column depicts how obeying a well-specified preregistered analysis plan can allow researchers to confidently publish their findings (discussed below).

4.1. Determining whether an error occurred

Although Type III and Type IV errors cannot be precisely calculated (since we never know whether an effect exists), researchers have two tools at their disposal to discern whether *some sort* of error has occurred.

4.1.1. Tracking questionable research practices

First, the questionable research practices listed in Fig. 3 have all been documented to increase alpha levels. While the underlying truth of the null hypothesis is unknowable, researchers can know whether they engaged in these practices. Accordingly, researchers can use boxes 1–4 in the “Research Practice” column to tally the number of “counts” of questionable research practices. That is, they can count the number of times they engaged in each of the four practices listed and use the sum as an Error Likelihood Score. Notice that this number may be greater than four if any of the research practices were used more than once. After determining the initial Error Likelihood Score, the researcher's goal should be to reduce that score by taking the corrective actions suggested in Fig. 3. With each corrective action, the score approaches zero, transparency is increased, and consumers of the research will be better able to determine for themselves how well conclusions follow from the evidence (thereby addressing Type III errors). If the researcher does not bring the score down to 0 through corrective actions, there is some reason to suspect that either the effect does not exist (Type I error) or is not strongly supported by the evidence (Type III error).

Tabulating counts of questionable research practices is a tool that researchers can use to determine whether they should be worried about committing an error; if questionable research practices have gone unaddressed then there is cause for concern. The system is not statistically precise, but it provides a useful heuristic for knowing when the evidence for a research claim has been faithfully depicted so that editors, reviewers, and ultimately readers can judge how well the evidence supports that claim. Such a system provides structure and guidance to researchers who seek to minimize errors.

4.1.2. Plausibility calculations

Second, researchers can perform calculations to assess the degree to which their own data show signs of biased reporting. As discussed, researchers are typically motivated to find significant results. It is easy to imagine how motivated reasoning can obscure even the most well-intentioned researchers' judgments of their own research practices (e.g., trying to recall months or years later whether a specific analysis was planned a priori; Nosek et al., 2012). Given this, even after exhausting the categories of questionable research practices shown in Fig. 3, the thorough researcher may be interested in proactively calculating and reporting statistical information about the diagnosticity of their evidence using one of numerous recently developed procedures (Francis, 2012; Schimmack, 2012; Simonsohn et al., 2014). By using these procedures researchers can check themselves against their own potential bias.

Just as experimenter bias during data collection is a plausible alternative explanation for findings, so too is experimenter bias during data analysis. If it is appropriate to report information about how experimenters avoid letting their expectations bias results during data collection (e.g., double-blind procedures), then it is also appropriate to report information about how experimenters avoided letting their expectations bias results during analysis (e.g., a priori with preregistration, or post hoc with p-curve analysis; Simonsohn et al., 2014). By doing so, researchers can publish research that is more persuasive, as it rules out the potential alternative explanation that results were due to *downstream* experimenter bias during analysis.

Two caveats are necessary here. First, these types of analyses require multiple studies in order to be meaningfully interpreted, so this approach may not be possible in all cases. Second, just like traditional hypothesis testing, the results of these types of analyses are subject to error, and making strong dichotomous decisions between “definitely biased” and “definitely not biased” is not recommended (as noted by Schimmack, 2012, p. 555). Instead, the tests may be judiciously examined to inform the researcher's degree of confidence in whether the results show indications of implausible success given the available power. While the tracking of questionable research practices can only realistically be conducted by the researchers themselves, the post hoc plausibility calculations can be put to use by anyone interested in the research.

4.2. Preregistration and transparency

Preregistering hypotheses, planned analyses, and exclusion criteria for a study is an inexpensive and immediately implementable way to decrease the problem of exploratory findings masquerading as confirmatory findings (Goldacre, 2009; Wagenmakers et al., 2012). If a researcher publicly commits to certain research procedures they will be less likely to depart from those procedures, and any departures will be visible to consumers of the research. Preregistration clarifies exactly which analyses are exploratory and which are confirmatory so that any confusion of the two will be transparent.

Preregistration decreases the chances of committing a Type III error by making the quality of the evidence more visible. Planned analyses constitute better evidence than unplanned analyses. Preregistration makes it possible to know whether analyses were planned. Therefore, preregistration has the effect of taking decisions that would have been false positives and instead placing them in the top right box in Fig. 1, regarding them as Type III errors. It is also possible that as preregistration becomes more common, researchers will self-police, and Type III errors will quietly begin to disappear.

Similar proposals will have the same effect. These include recent calls for posting raw data (Perrino et al., 2013; Simonsohn, 2013), and greater transparency in disclosing the use and effects of a wide range of research practices (Funder et al., 2014; Sagarin, Ambler, & Lee, 2014; Simmons, Nelson, & Simonsohn, 2012). Journals are increasingly asking researchers to actively state that they have disclosed important research details (Eich, 2014). As researchers are compelled to honestly disclose whether or not they have reported all variables, conditions, analyses, excluded cases, and stopping criteria, the result will be more information available regarding the quality of evidence, and a greater ability to detect Type III and IV errors (and thereby reduce Types I and II errors).

4.2.1. Preregistration's benefits to the researcher

At this point one might get the impression that acknowledging Type III and IV errors requires reforms that lead to uniformly stricter standards for significance, which risks sacrificing interesting findings in the service of controlling positive error rates (Fiedler, Kutzner, &

Krueger, 2012). This impression would be mistaken. In fact, acknowledging Type III and IV errors sets the stage for researchers to invest in valuable Type III error insurance: the preregistration of analyses. This insurance pays dividends in the form of license to more boldly and powerfully test hypotheses about which a researcher is confident. This is an underappreciated benefit of preregisteration that is actually enjoyed by researchers themselves. Responsible and clear preregistration licenses researchers to eat the forbidden fruit of one-tailed hypothesis testing.

The primary objection to one-tailed hypothesis tests is that they invite Type I and III errors. A marginally significant finding can be turned into a tidy significant finding if a researcher switches out a two-tailed test for a one-tailed test. As one journal editor observed, “no one except the experimenter is likely to know whether the decision to use a one-tailed test was made before or after the data were collected” (Levitt, 1994, p. 4). As a consequence, people are generally distrusting of one-tailed tests, with both researchers and students being cautioned against using them (Burke, 1953; Field, 2013; Levitt, 1994). The result is that researchers very rarely perform one-tailed tests.

Despite this, it is easy to think of scenarios in which a principled researcher has good reasons for choosing to concentrate their alpha in a single tail. A researcher might be testing a legitimately directional hypothesis in which an effect in the opposite direction is uninteresting, uninterpretable, or simply unlikely to occur in the researcher's judgment (see Kimmel, 1957). Or a particular test may be a replication of an already discovered effect (Wike, 1971). Some journals routinely publish packages of up to seven or more studies. Should these researchers be required to use half of their alpha for each replication on the possibility that the earlier studies were falsely significant *in the wrong direction*? In these cases, if the researcher is correct about the direction of the effect, then the one-tailed test is much more powerful, resulting in fewer Type II and IV errors.

For many researchers who are either testing legitimately directional hypotheses, or conducting replications of already discovered effects, the current situation is that a full 50% of their alpha is languishing off in the distance where it serves no purpose but to satisfy reviewers and editors trained to react to one-tailed tests as a warning sign that researchers have taken inappropriate statistical liberty. Of course, this skepticism is also understandable. Without some assurance that the researcher would not have drawn conclusions given an effect of equal magnitude in the opposite direction then it is not unimaginable that the one-tailed test was selected as a tool out of the arsenal of unreported researcher degrees of freedom.

Preregistration solves this dilemma. By preregistering one-tailed hypothesis tests researchers can unlock the untapped source of statistical power, while assuring others that the test was planned a priori. The objection to one-tailed tests – that they are not to be trusted because people can use them to scoot marginal effects below $p = .05$ – only applies in a world where researchers have no way to transparently declare their a priori directional hypotheses. Preregistration offers a way to do just this. According to this view, the alpha level is the property of the researcher, and they can use it however they choose, provided they clearly document their analysis plan. If a researcher has the courage to call *eight-ball-in-the-corner-pocket*, then critics of the research must respect that the alpha level has been preserved. In situations where an effect is very strong but in the non-predicted direction, researchers must honor their analysis plan and retain the null hypothesis. But they are still free to conduct follow-up research, to test an updated hypothesis.

Preregistration is a powerful way to increase transparency, and reduce Type III errors. But many researchers likely view it as yet more work with little reward. Why would researchers take the initiative to perform the extra and often risky work of preregistering their analysis? My hope is that the doubled power from responsible, selective, and transparent one-tailed testing will provide a big enough carrot to encourage researchers to preregister analyses. The broad research community will enjoy greater transparency, and individual researchers will enjoy more efficient use of statistical power without increasing error rates.

5. Conclusion

Social psychology has successfully navigated a crisis of confidence in the past, and is making impressive progress in emerging from the current crisis as a stronger and more rigorous discipline. Guided by the observation that research claims based on relatively bad evidence may actually be true, I proposed a reconceptualization of statistical decision making and offered a framework researchers can use to help minimize errors of all types. By tracking and correcting research practices that produce errors researchers will be able to draw conclusions that more closely align with reality, and social psychology will continue to thrive.

References

Abelson, R. P. (1995). *Statistics as principled argument.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. http://dx.doi.org/10.1037/a0021524.

Burke, C. J. (1953). A brief note on one-tailed tests. *Psychological Bulletin*, *50*, 384–387.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*, 40–48. http://dx.doi.org/10.1177/1745691613513470.

Cohen, J. (1962). Statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. http://dx.doi.org/10.1037/h0045186.

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

Cozby, P. (2009). *Methods in behavioral research* (10th ed.). New York: McGraw-Hill.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. http://dx.doi.org/10.1177/0956797613504966.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind but whose mind? *PloS One*, *7*, 1–7. http://dx.doi.org/10.1371/journal.pone.0029081.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6. http://dx.doi.org/10.1177/0956797613512465.

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist*, *30*, 967–976. http://dx.doi.org/10.1037/0003-066X.30.10.967.

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669. http://dx.doi.org/10.1177/1745691612462587.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London: Sage Publications.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156. http://dx.doi.org/10.3758/s13423-012-0227-9.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3–12. http://dx.doi.org/10.1177/1088868313507536.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*, 933–948. http://dx.doi.org/10.1037/a0029709.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*, 309–320.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*, 562–571. http://dx.doi.org/10.1177/1745691612457576.

Goldacre (2009). *Bad science.* London England: Fourth Estate.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20. http://dx.doi.org/10.1037/h0076157.

Hansson, S. O. (2013). Defining pseudoscience and science. In M. Pigliucci, & M. Boudry (Eds.), *Philosophy of pseudoscience* (pp. 61–77). Chicaco, IL: The University of Chicago Press.

How science goes wrong (2013, October 19). *The Economist.*

Jost, J. T., & McGuire, W. J. (2013). An additional future for psychological science. *Perspectives on Psychological Science*, *8*, 414–423.

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

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*, 351–353. http://dx.doi.org/10.1037/h0046737.

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial with R and BUGS.* Burlington, MA: Academic Press.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701–710. http://dx.doi.org/10.1002/ejsp.2023.

Levitt, E. E. (1994). The one-tailed test: A statistical editorial. *International Journal of Clinical and Experimental Hypnosis*, *42*, 4–6. http://dx.doi.org/10.1080/00207149408409336.

Maner, J. K. (2016). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*, *66*, 100–106.

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post Hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of type IV errors. *American Educational Research Journal*, *7*, 397–421.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163. http://dx.doi.org/10.1037/1082-989X.9.2.147.

McGuire, W. J. (1967). Some impending reorientations in social psychology: Some thoughts provoked by Kenneth Ring. *Journal of Experimental Social Psychology*, *3*, 124–139. http://dx.doi.org/10.1016/0022-1031(67)90017-0.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, *26*, 446–456. http://dx.doi.org/10.1037/h0034345.

Mosteller, F. (1948). A *k*-sample slippage test for an extreme population. *Ann. Math. Stat.*, *19*, 58–65.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*, 217–243. http://dx.doi.org/10.1080/1047840X.2012.692215.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. http://dx.doi.org/10.1177/1745691612459058.

Perrino, T., Howe, G., Sperling, A., Beardslee, W., Sandler, I., Shern, D., ... Brown, C. (2013). Advancing science through collaborative data sharing and synthesis. *Perspectives on Psychological Science*, *8*, 433–444. http://dx.doi.org/10.1177/1745691613491579.

Ring, K. (1967). Experimental social psychology: some sober questions about some frivolous values. *Journal of Experimental Social Psychology*, *3*, 113–123. http://dx.doi.org/10.1016/0022-1031(67)90016-9.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PloS One*, *7* http://dx.doi.org/10.1371/journal.pone.0033423.

Roberts, B. (2012, May 9). Minimal and incomplete methodological literacy in personality psychology. Retrieved from http://pigee.wordpress.com/minimal-and-incomplete-methodological-literacy-in-personality-psychology/

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*http://dx.doi.org/10.1177/1745691614528214.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566. http://dx.doi.org/10.1037/a0029487.

Schwartz, S., & Carpenter, K. M. (1999). The right answer for the wrong question: Consequences of type III error for public health research. *American Journal of Public Health*, *89*, 1175–1180.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shaffer, J. P. (2002). Multiplicity, directional (type III) errors, and the null hypothesis. *Psychological Methods*, *7*, 356–369.

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. http://dx.doi.org/10.1177/0956797611417632.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue*, *26*, 4–7.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. http://dx.doi.org/10.1177/1745691613514755.

Simonsohn, U. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspectives on Psychological Science*, *7*, 597–599. http://dx.doi.org/10.1177/1745691612463399.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875–1888. http://dx.doi.org/10.1177/0956797613480366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. http://dx.doi.org/10.1037/a0033242.

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. http://dx.doi.org/10.1177/1745691612463078.

Wike, E. L. (1971). *Data analysis: A statistical primer for psychology students.* Chicago: Aldine- Atherton.