General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
XX(X) 1–8

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

Keywords

methodology, motivated reasoning, publication, disclosure

Received 3/17/11; Revision accepted 5/23/11

Our job as scientists is to discover truths about the world. We generate hypotheses, collect data, and examine whether or not the data are consistent with those hypotheses. Although we aspire to always be accurate, errors are inevitable.

Perhaps the most costly error is a *false positive*, the incorrect rejection of a null hypothesis. First, once they appear in the literature, false positives are particularly persistent. Because null results have many possible causes, failures to replicate previous findings are never conclusive. Furthermore, because it is uncommon for prestigious journals to publish null findings or exact replications, researchers have little incentive to even attempt them. Second, false positives waste resources: They inspire investment in fruitless research programs and can lead to ineffective policy changes. Finally, a field known for publishing false positives risks losing its credibility.

In this article, we show that despite the nominal endorsement of a maximum false-positive rate of 5% (i.e., $p \leq .05$), current standards for disclosing details of data collection and analyses make false positives vastly more likely. In fact, it is unacceptably easy to publish "statistically significant" evidence consistent with *any* hypothesis.

The culprit is a construct we refer to as *researcher degrees of freedom*. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?

It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields "statistical significance," and to then report only what "worked." The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.

This exploratory behavior is not the by-product of malicious intent, but rather the result of two factors: (a) ambiguity in how best to make these decisions and (b) the researcher's desire to find a statistically significant result. A large literature documents that people are self-serving in their interpretation

Corresponding Authors:
Joseph P. Simmons, The Wharton School, University of Pennsylvania, 551 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104
E-mail: jsimmo@wharton.upenn.edu

Leif D. Nelson, Haas School of Business, University of California, Berkeley, Berkeley, CA 94720-1900
E-mail: leif_nelson@haas.berkeley.edu

Uri Simonsohn, The Wharton School, University of Pennsylvania, 548 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

of ambiguous information and remarkably adept at reaching justifiable conclusions that mesh with their desires (Babcock & Loewenstein, 1997; Dawson, Gilovich, & Regan, 2002; Gilovich, 1983; Hastorf & Cantril, 1954; Kunda, 1990; Zuckerman, 1979). This literature suggests that when we as researchers face ambiguous analytic decisions, we will tend to conclude, with convincing self-justification, that the appropriate decisions are those that result in statistical significance ($p \leq .05$).

Ambiguity is rampant in empirical research. As an example, consider a very simple decision faced by researchers analyzing reaction times: how to treat outliers. In a perusal of roughly 30 *Psychological Science* articles, we discovered considerable inconsistency in, and hence considerable ambiguity about, this decision. Most (but not all) researchers excluded some responses for being too fast, but what constituted "too fast" varied enormously: the fastest 2.5%, or faster than 2 standard deviations from the mean, or faster than 100 or 150 or 200 or 300 ms. Similarly, what constituted "too slow" varied enormously: the slowest 2.5% or 10%, or 2 or 2.5 or 3 standard deviations slower than the mean, or 1.5 standard deviations slower from that condition's mean, or slower than 1,000 or 1,200 or 1,500 or 2,000 or 3,000 or 5,000 ms. None of these decisions is necessarily incorrect, but that fact makes any of them justifiable and hence potential fodder for self-serving justifications.

How Bad Can It Be? A Demonstration of Chronological Rejuvenation

To help illustrate the problem, we conducted two experiments designed to demonstrate something false: that certain songs can change listeners' age. Everything reported here actually happened.[1]

Study 1: musical contrast and subjective age

In Study 1, we investigated whether listening to a children's song induces an age contrast, making people feel older. In exchange for payment, 30 University of Pennsylvania undergraduates sat at computer terminals, donned headphones, and were randomly assigned to listen to either a control song ("Kalimba," an instrumental song by Mr. Scruff that comes free with the Windows 7 operating system) or a children's song ("Hot Potato," performed by The Wiggles).

After listening to part of the song, participants completed an ostensibly unrelated survey: They answered the question "How old do you feel right now?" by choosing among five options (*very young*, *young*, *neither young nor old*, *old*, and *very old*). They also reported their father's age, allowing us to control for variation in baseline age across participants.

An analysis of covariance (ANCOVA) revealed the predicted effect: People felt older after listening to "Hot Potato" (adjusted $M = 2.54$ years) than after listening to the control song (adjusted $M = 2.06$ years), $F(1, 27) = 5.06$, $p = .033$.

In Study 2, we sought to conceptually replicate and extend Study 1. Having demonstrated that listening to a children's song makes people feel older, Study 2 investigated whether listening to a song about older age makes people *actually* younger.

Study 2: musical contrast and chronological rejuvenation

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either "When I'm Sixty-Four" by The Beatles or "Kalimba." Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father's age. We used father's age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted $M = 20.1$ years) rather than to "Kalimba" (adjusted $M = 21.5$ years), $F(1, 17) = 4.92$, $p = .040$.

Discussion

These two studies were conducted with real participants, employed legitimate statistical analyses, and are reported truthfully. Nevertheless, they seem to support hypotheses that are unlikely (Study 1) or necessarily false (Study 2).

Before detailing the researcher degrees of freedom we employed to achieve these "findings," we provide a more systematic analysis of how researcher degrees of freedom influence statistical significance. Impatient readers can consult Table 3.

"How Bad Can It Be?" Simulations

Simulations of common researcher degrees of freedom

We used computer simulations of experimental data to estimate how researcher degrees of freedom influence the probability of a false-positive result. These simulations assessed the impact of four common degrees of freedom: flexibility in (a) choosing among dependent variables, (b) choosing sample size, (c) using covariates, and (d) reporting subsets of experimental conditions. We also investigated various combinations of these degrees of freedom.

We generated random samples with each observation independently drawn from a normal distribution, performed sets of analyses on each sample, and observed how often at least one of the resulting p values in each sample was below standard significance levels. For example, imagine a researcher who collects two dependent variables, say liking and willingness to

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three *t* tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one *t* test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a *t* test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender × Condition interaction was significant. Results for Situation D were obtained by conducting *t* tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = −1, medium = 0, high = 1).

pay. The researcher can test whether the manipulation affected liking, whether the manipulation affected willingness to pay, and whether the manipulation affected a combination of these two variables. The likelihood that one of these tests produces a significant result is at least somewhat higher than .05. We conducted 15,000 simulations of this scenario (and other scenarios) to estimate the size of "somewhat."[2]

We report the results of our simulations in Table 1. The first row shows that flexibility in analyzing two dependent variables (correlated at $r = .50$) nearly doubles the probability of obtaining a false-positive finding.[3]

The second row of Table 1 shows the results of a researcher who collects 20 observations per condition and then tests for significance. If the result is significant, the researcher stops collecting data and reports the result. If the result is nonsignificant, the researcher collects 10 additional observations per condition, and then again tests for significance. This seemingly small degree of freedom increases the false-positive rate by approximately 50%.

The third row of Table 1 shows the effect of flexibility in controlling for gender or for an interaction between gender and the independent variable.[4] Such flexibility leads to a false-positive rate of 11.7%. The fourth row of Table 1 shows that running three conditions (e.g., low, medium, high) and reporting the results for any two or all three (e.g., low vs. medium, low vs. high, medium vs. high, low vs. medium vs. high) generates a false-positive rate of 12.6%.

The bottom three rows of Table 1 show results for combinations of the situations described in the top four rows, with the bottom row reporting the false-positive rate if the researcher uses all of these degrees of freedom, a practice that would lead to a stunning 61% false-positive rate! A researcher is more likely than not to falsely detect a significant effect by just using these four common researcher degrees of freedom.

As high as these estimates are, they may actually be conservative. We did not consider many other degrees of freedom that researchers commonly use, including testing and choosing among more than two dependent variables (and the various ways to combine them), testing and choosing among more than one covariate (and the various ways to combine them), excluding subsets of participants or trials, flexibility in deciding whether early data were part of a pilot study or part of the experiment proper, and so on.

A closer look at flexibility in sample size

Researchers often decide when to stop data collection on the basis of interim data analysis. Notably, a recent survey of behavioral scientists found that approximately 70% admitted to having done so (John, Loewenstein, & Prelec, 2011). In conversations with colleagues, we have learned that many believe this practice exerts no more than a trivial influence on false-positive rates.

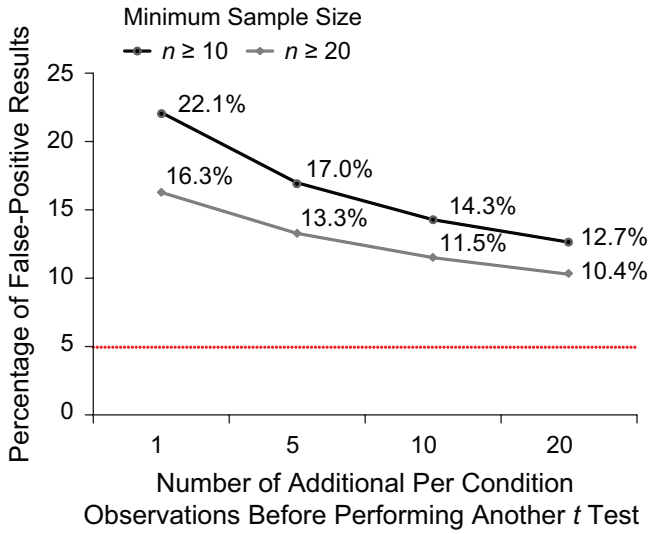

Fig. 1. Likelihood of obtaining a false-positive result when data collection ends upon obtaining significance ($p \leq .05$, highlighted by the dotted line). The figure depicts likelihoods for two minimum sample sizes, as a function of the frequency with which significance tests are performed.

Table 2. Simple Solution to the Problem of False-Positive Publications

Requirements for authors

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Guidelines for reviewers

1. Reviewers should ensure that authors follow the requirements.
2. Reviewers should be more tolerant of imperfections in results.
3. Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions.
4. If justifications of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication.

Contradicting this intuition, Figure 1 shows the false-positive rates from additional simulations for a researcher who has already collected either 10 or 20 observations within each of two conditions, and then tests for significance every 1, 5, 10, or 20 per-condition observations after that. The researcher stops collecting data either once statistical significance is obtained or when the number of observations in each condition reaches 50.

Figure 1 shows that a researcher who starts with 10 observations per condition and then tests for significance after every new per-condition observation finds a significant effect 22% of the time. Figure 2 depicts an illustrative example continuing sampling until the number of per-condition observations reaches 70. It plots p values from t tests conducted after each pair of observations. The example shown in Figure 2 contradicts the often-espoused yet erroneous intuition that if an effect is significant with a small sample size then it would necessarily be significant with a larger one.

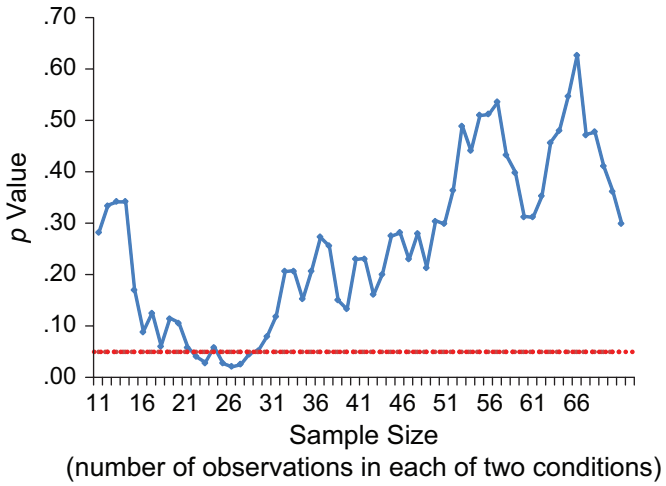

Fig. 2. Illustrative simulation of p values obtained by a researcher who continuously adds an observation to each of two conditions, conducting a t test after each addition. The dotted line highlights the conventional significance criterion of $p \leq .05$.

Solution

As a solution to the flexibility-ambiguity problem, we offer six requirements for authors and four guidelines for reviewers (see Table 2). This solution substantially mitigates the problem but imposes only a minimal burden on authors, reviewers, and readers. Our solution leaves the right and responsibility of identifying the most appropriate way to conduct research in the hands of researchers, requiring only that authors provide appropriately transparent descriptions of their methods so that reviewers and readers can make informed decisions regarding the credibility of their findings. We assume that the vast majority of researchers strive for honesty; this solution will not help in the unusual case of willful deception.

Requirements for authors

We propose the following six requirements for authors.

1. **Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.** Following this requirement may mean reporting the outcome of power calculations or disclosing arbitrary rules, such as "we decided to collect 100 observations" or "we decided to collect as many observations as we could before the end of the semester." The rule itself is secondary, but it must be determined ex ante and be reported.

2. **Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.** This requirement offers extra protection for the first requirement. Samples smaller than 20 per cell are simply not powerful enough to detect most effects, and so there is usually no good reason to decide in advance to collect such a small number of observations. Smaller samples, it follows, are much more likely to reflect interim data analysis and a flexible termination rule. In addition, as Figure 1 shows, larger minimum sample sizes can lessen the impact of violating Requirement 1.
3. **Authors must list all variables collected in a study.** This requirement prevents researchers from reporting only a convenient subset of the many measures that were collected, allowing readers and reviewers to easily identify possible researcher degrees of freedom. Because authors are required to just list those variables rather than describe them in detail, this requirement increases the length of an article by only a few words per otherwise shrouded variable. We encourage authors to begin the list with "only," to assure readers that the list is exhaustive (e.g., "participants reported only their age and gender").
4. **Authors must report all experimental conditions, including failed manipulations.** This requirement prevents authors from selectively choosing only to report the condition comparisons that yield results that are consistent with their hypothesis. As with the previous requirement, we encourage authors to include the word "only" (e.g., "participants were randomly assigned to one of only three conditions").
5. **If observations are eliminated, authors must also report what the statistical results are if those observations are included.** This requirement makes transparent the extent to which a finding is reliant on the exclusion of observations, puts appropriate pressure on authors to justify the elimination of data, and encourages reviewers to explicitly consider whether such exclusions are warranted. Correctly interpreting a finding may require some data exclusions; this requirement is merely designed to draw attention to those results that hinge on ex post decisions about which data to exclude.
6. **If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.** Reporting covariate-free results makes transparent the extent to which a finding is reliant on the presence of a covariate, puts appropriate pressure on authors to justify the use of the covariate, and encourages reviewers to consider whether including it is warranted. Some findings may be persuasive even if covariates are required for their detection, but one should place greater scrutiny on results that do hinge on covariates despite random assignment.

Guidelines for reviewers

We propose the following four guidelines for reviewers.

1. **Reviewers should ensure that authors follow the requirements.** Review teams are the gatekeepers of the scientific community, and they should encourage authors not only to rule out alternative explanations, but also to more convincingly demonstrate that their findings are not due to chance alone. This means prioritizing transparency over tidiness; if a wonderful study is partially marred by a peculiar exclusion or an inconsistent condition, those imperfections should be retained. If reviewers require authors to follow these requirements, they will.
2. **Reviewers should be more tolerant of imperfections in results.** One reason researchers exploit researcher degrees of freedom is the unreasonable expectation we often impose as reviewers for every data pattern to be (significantly) as predicted. Underpowered studies with perfect results are the ones that should invite extra scrutiny.
3. **Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions.** Even if authors follow all of our guidelines, they will necessarily still face arbitrary decisions. For example, should they subtract the baseline measure of the dependent variable from the final result or should they use the baseline measure as a covariate? When there is no obviously correct way to answer questions like this, the reviewer should ask for alternatives. For example, reviewer reports might include questions such as, "Do the results also hold if the baseline measure is instead used as a covariate?" Similarly, reviewers should ensure that arbitrary decisions are used consistently across studies (e.g., "Do the results hold for Study 3 if gender is entered as a covariate, as was done in Study 2?").[5] If a result holds only for one arbitrary specification, then everyone involved has learned a great deal about the robustness (or lack thereof) of the effect.
4. **If justifications of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication.** If a reviewer is not persuaded by the justifications for a given researcher degree of freedom or the results from a robustness check, the reviewer should ask the author to conduct an exact replication of the study and its analysis. We realize that this is a costly solution, and it should be used selectively; however, "never" is too selective.

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked ~~20~~ 34 University of Pennsylvania undergraduates to listen only **to either "When I'm Sixty-Four" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age**, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants**.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), $F(1, 18) = 1.01$, $p = .33$.

The Solutions in Action: Revisiting Chronological Rejuvenation

To show how our solutions would work in practice, we return to our Study 2, which "showed" that people get younger when listening to The Beatles, and we report it again in Table 3, following the requirements we have proposed. The merits of reporting transparency should be evident, but three highlights are worth mentioning.

First, notice that in our original report, we redacted the many measures other than father's age that we collected (including the dependent variable from Study 1: feelings of oldness). A reviewer would hence have been unable to assess the flexibility involved in selecting father's age as a control. Second, by reporting only results that included the covariate, we made it impossible for readers to discover its critical role in achieving a significant result. Seeing the full list of variables now disclosed, reviewers would have an easy time asking for robustness checks, such as "Are the results from Study 1 replicated in Study 2?" They are not: People felt older rather than younger after listening to "When I'm Sixty-Four," though not significantly so, $F(1, 17) = 2.07$, $p = .168$. Finally, notice that we did not determine the study's termination rule in advance; instead, we monitored statistical significance approximately every 10 observations. Moreover, our sample size did not reach the 20-observation threshold set by our requirements.

The redacted version of the study we reported in this article fully adheres to currently acceptable reporting standards and is, not coincidentally, deceptively persuasive. The requirement-compliant version reported in Table 3 would be—appropriately—all but impossible to publish.

General Discussion

Criticisms

Criticism of our solution comes in two varieties: It does not go far enough and it goes too far.

Not far enough. Our solution does not lead to the disclosure of all degrees of freedom. Most notably, it cannot reveal those arising from reporting only experiments that "work" (i.e., the file-drawer problem). This problem might be addressed by requiring researchers to submit all studies to a public repository, whether or not the studies are "successful" (see, e.g., Ioannidis, 2005; Schooler, 2011). Although we are sympathetic to this suggestion, it does come with significant practical challenges: How is submission enforced? How does one ensure that study descriptions are understandably written and appropriately classified? Most notably, in order for the repository to be effective, it must adhere to our disclosure policy, for it is impossible to interpret study results, whether successful or not, unless researcher degrees of freedom are disclosed. The repository is an ambitious long-term extension of our recommended solution, not a substitute.

In addition, a reviewer of this article worried that our solution may not go far enough because authors have "tremendous disincentives" to disclose exploited researcher degrees of freedom. Although researchers obviously have incentives to publish, if editors and reviewers enforce our solution, authors will have even stronger incentives to accurately disclose their methodology. Our solution turns inconsequential sins of omission (leaving out inconvenient facts) into consequential, potentially career-ending sins of commission (writing demonstrably false statements). Journals implementing our disclosure requirements will create a virtuous cycle of transparency and accountability that eliminates the disincentive problem.

Too far. Alternatively, some readers may be concerned that our guidelines prevent researchers from conducting exploratory research. What if researchers do not know which dependent measures will be sensitive to the manipulation, for example, or how such dependent measures should be scored or combined? We all should of course engage in exploratory research, but we should be required either to report it as such (i.e., following the six requirements) or to complement it with

(and possibly only report) confirmatory research consisting of exact replications of the design and analysis that "worked" in the exploratory phase.

Nonsolutions

In the process of devising our solution, we considered a number of alternative ways to address the problem of researcher degrees of freedom. We believe that all solutions other than the one we have outlined are less practical, less effective, or both. It might be worth pursuing these other policy changes for other reasons, but in our view, they do not address the problem of researcher degrees of freedom. The following are four policy changes we considered and rejected.

Correcting alpha levels. Á la Bonferroni, one may consider adjusting the critical alpha (α) level as a function of the number of researcher degrees of freedom employed in each study, as is supposed to be done with multiple-hypothesis testing. Something like this has been proposed for medical trials that monitor outcomes as the study progresses (see, e.g., Pocock, 1977).

First, given the broad and ambiguous set of degrees of freedom in question, it is unclear which and how many of them contribute to any given finding, and hence what their effect is on the false-positive rate. Second, unless there is an explicit rule about exactly how to adjust alphas for each degree of freedom and for the various combinations of degrees of freedom (see the bottom three rows in Table 1), the additional ambiguity may make things worse by introducing new degrees of freedom.

Using Bayesian statistics. We have a similar reaction to calls for using Bayesian rather than frequentist approaches to analyzing experimental data (see, e.g., Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Although the Bayesian approach has many virtues, it actually increases researcher degrees of freedom. First, it offers a new set of analyses (in addition to all frequentist ones) that authors could flexibly try out on their data. Second, Bayesian statistics require making additional judgments (e.g., the prior distribution) on a case-by-case basis, providing yet more researcher degrees of freedom.

Conceptual replications. Because conceptual replications, in contrast to exact replications, do not bind researchers to make the same analytic decisions across studies, they are unfortunately misleading as a solution to the problem at hand. In an article with a conceptual replication, for instance, authors may choose two of three conditions in Study 1 and report one measure, but choose a different pair of conditions and a different measure in Study 2. Indeed, that is what we did in the experiments reported here.

Posting materials and data. We are strongly supportive of all journals requiring authors to make their original materials and data publicly available. However, this is not likely to address the problem of interest, as this policy would impose too high a cost on readers and reviewers to examine, in real time, the credibility of a particular claim. Readers should not need to download data, load it into their statistical packages, and start running analyses to learn the importance of controlling for father's age; nor should they need to read pages of additional materials to learn that the researchers simply dropped the "Hot Potato" condition.

Furthermore, if a journal allows the redaction of a condition from the report, for example, it would presumably also allow its redaction from the raw data and "original" materials, making the entire transparency effort futile.

Concluding Remarks

Our goal as scientists is not to publish as many articles as we can, but to discover and disseminate truth. Many of us—and this includes the three authors of this article—often lose sight of this goal, yielding to the pressure to do whatever is justifiable to compile a set of studies that we can publish. This is not driven by a willingness to deceive but by the self-serving interpretation of ambiguity, which enables us to convince ourselves that whichever decisions produced the most publishable outcome must have also been the most appropriate. This article advocates a set of disclosure requirements that imposes minimal costs on authors, readers, and reviewers. These solutions will not rid researchers of publication pressures, but they will limit what authors are able to justify as acceptable to others and to themselves. We should embrace these disclosure requirements as if the credibility of our profession depended on them. Because it does.

Acknowledgments

All three authors contributed equally to this article. Author order is alphabetical, controlling for father's age (reverse-coded). We thank Jon Baron, Jason Dana, Victor Ferreira, Geoff Goodwin, Jack Hershey, Dave Huber, Hal Pashler, and Jonathan Schooler for their valuable comments.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. Our goal was to pursue a research question that would not implicate any particular field of research. Our concerns apply to all branches of experimental psychology, and to the other sciences as well.
2. We conducted simulations instead of deriving closed-form solutions because the combinations of researcher degrees of freedom we considered would lead to fairly complex derivations without adding much insight over simulation results.
3. The lower the correlation between the two dependent variables, the higher the false-positive rate produced by considering both. Intuitively, if $r = 1$, then both variables are the same; if $r = 0$, then the two tests are entirely independent.

4. We independently assigned each observation a gender of 1 (50% probability) or 0 (50% probability); "gender" is a placeholder for any covariate with similar properties.

5. It is important that these alternatives be reported in the manuscript (or in an appendix) rather than merely in a private response to reviewers, so that the research community has access to the results.

References

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives*, *11*, 109–126.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin*, *28*, 1379–1387.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, *44*, 1110–1126.

Hastorf, A. H., & Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology*, *49*, 129–134.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. Retrieved from http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124

John, L., Loewenstein, G. F., & Prelec, D. (2011). *Measuring the prevalence of questionable research practices with incentives for truth-telling*. Manuscript submitted for publication.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*, 191–199.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*, 437.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *100*, 426–432.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, *47*, 245–287.