ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 632–638
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691612463078
http://pps.sagepub.com
SAGE

An Agenda for Purely Confirmatory Research

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit
University of Amsterdam, The Netherlands

Abstract

The veracity of substantive research claims hinges on the way experimental data are collected and analyzed. In this article, we discuss an uncomfortable fact that threatens the core of psychology's academic enterprise: almost without exception, psychologists do not commit themselves to a method of data analysis *before* they see the actual data. It then becomes tempting to fine tune the analysis to the data in order to obtain a desired result—a procedure that invalidates the interpretation of the common statistical tests. The extent of the fine tuning varies widely across experiments and experimenters but is almost impossible for reviewers and readers to gauge. To remedy the situation, we propose that researchers preregister their studies and indicate in advance the analyses they intend to conduct. Only these analyses deserve the label "confirmatory," and only for these analyses are the common statistical tests valid. Other analyses can be carried out but these should be labeled "exploratory." We illustrate our proposal with a confirmatory replication attempt of a study on extrasensory perception.

Keywords

confirmatory experiments, wonky statistics, ESP, Bayesian hypothesis test

> You cannot find your starting hypothesis in your final results. It makes the stats go all wonky.
>
> —Ben Goldacre (2009, p. 221)

Psychology is a challenging discipline. Empirical data are noisy, formal theory is scarce, and the processes of interest (e.g., attention, jealousy, loss aversion) cannot be observed directly. Nevertheless, psychologists have managed to generate many key insights about human cognition and behavior. For instance, research has shown that people tend to seek confirmation rather than disconfirmation of their beliefs—a phenomenon known as *confirmation bias* (Nickerson, 1998). Confirmation bias operates in at least three ways. First, ambiguous information is readily interpreted to be consistent with one's prior beliefs; second, people tend to search for information that confirms rather than disconfirms their preferred hypothesis; third, people more easily remember information that supports their position. We also know that people fall prey to hindsight bias, the tendency to judge an event as more predictable after it has occurred (Roese & Vohs, 2012).

In light of these and other biases[1] it would be naive to believe that, without special protective measures, the scientific research process is somehow exempt from the systematic imperfections of the human mind. For example, one indication that bias influences the research process is that researchers seek to confirm, not falsify, their main hypothesis (Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). The impact of bias is exacerbated in an environment that puts a premium on output quantity: When academic survival depends on how many papers one publishes, researchers are attracted to methods and procedures that maximize the probability of publication (Bakker, van Dijk, & Wicherts, 2012; John, Loewenstein, & Prelec, 2012; Neuroskeptic, 2012; Nosek, Spies, & Motyl, 2012). It should be noted that such behavior is ecologically rational in the sense that it maximizes the proximal goals of the researcher. However, when each researcher acts this way in an entirely understandable attempt at academic self-preservation, the cumulative effect on the field as a whole can be catastrophic. The primary concern is that many published results may simply be false, as they have been obtained partly by dubious or inappropriate methods of observation, analysis, and reporting (Jasny, Chin, Chong, & Vignieri, 2011; Sarewitz, 2012).

Corresponding Author:
Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychological Methods, Weesperplein 4, 1018 XA Amsterdam, The Netherlands
E-mail: EJ.Wagenmakers@gmail.com

Several years ago, Ioannidis (2005) famously argued that "most published research findings are false." And indeed, recent results from biomedical and cancer research suggest that replication rates are lower than 50%, with some as low as 11% (Begley & Ellis, 2012; Osherovich, 2011; Prinz, Schlange, & Asadullah, 2011). If the above results carry over to psychology, our discipline is in serious trouble (Carpenter, 2012; Roediger, 2012; Yong, 2012). Research findings that do not replicate are worse than fairy tales; with fairy tales the reader is at least aware that the work is fictional.

In this article, we focus on what we believe to be the main "fairy-tale factor" in psychology today (and indeed in all of the empirical sciences): the fact that researchers do not commit themselves to a plan of analysis before they see the data. Consequently, researchers can fine tune their analyses to the data, a procedure that make the data appear to be more compelling than they really are. This fairy-tale factor increases the probability that a presented finding is fictional and hence non-replicable. We propose a radical remedy—preregistration—to ensure scientific integrity and inoculate the research process against the inalienable biases of human reasoning. We conclude by illustrating the remedy of preregistration using a replication attempt of an extrasensory-perception (ESP) experiment reported by Bem (2011).

Bad Science: Exploratory Findings, Confirmatory Conclusions

Science can be bad in many ways. Flawed design, faulty logic, and limited scholarship engender no confidence or enthusiasm whatsoever.[2] In this section, we discuss another important factor that reduces confidence and enthusiasm for a scientific finding: the fact that almost no psychological research is conducted in a purely confirmatory fashion[3] (e.g., Kerr, 1998; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; for a similar discussion in biology, see Anderson, Burnham, Gould, & Cherry, 2001). Only rarely do psychologists indicate, in advance of data collection, the specific analyses they intend to carry out. In the face of human biases and the vested interest of the experimenter, such freedom of analysis provides access to a Pandora's box of tricks that can be used to achieve any desired result (e.g., John et al., 2012; Simmons, Nelson, & Simonsohn, 2011; for what may happen to psychologists in the afterlife, see Neuroskeptic, 2012). For instance, researchers can engage in cherry picking: They can measure many variables (gender, personality characteristics, age, etc.) and only report those that yield the desired result, and they can include in their papers only those experiments that produced the desired outcome, even though these experiments were designed as pilot experiments that could be easily discarded had the results turned out less favorably. Researchers can also explore various transformations of the data, rely on one-sided *p* values, and construct post-hoc hypotheses that have been tailored to fit the observed data (MacCallum, Roznowski, & Necowitz, 1992). In the past decades, the development of statistical software has resulted in a situation in which the number of opportunities for massaging the data is virtually infinite.

True, researchers may not use these tricks with the explicit purpose to deceive—for instance, hindsight bias often makes exploratory findings appear perfectly sensible. Even researchers who advise their students to "torture the data until they confess"[4] are hardly evil geniuses out to deceive the public or their peers. Instead, these researchers may genuinely believe that they are giving valuable advice that leads the student to analyze the data more thoroughly and increases the odds of publication along the way. How could such advice be wrong?

In fact, the advice to torture the data until they confess is not wrong—just as long as this torture is clearly acknowledged in the research report. Academic deceit sets in when this does not happen and partly exploratory research is analyzed as if it had been completely confirmatory. At the heart of the problem lies the statistical law that, for the purpose of hypothesis testing, the data may be used only once. So when you turn your data set inside and out, looking for interesting patterns, you have used the data to help you formulate a specific hypothesis. Although the data may still serve many purposes after such fishing expeditions, there is one purpose for which the data are no longer appropriate—namely, for testing the hypothesis that they helped to suggest. Just as conspiracy theories are never falsified by the facts that they were designed to explain, a hypothesis that is developed on the basis of exploration of a data set is unlikely to be refuted by that same data. Thus, one always needs a fresh data set for testing one's hypothesis. This also means that the interpretation of common statistical tests in terms of Type I and Type II error rates is valid only if the data were used only once and if the statistical test was not chosen on the basis of suggestive patterns in the data. If you carry out a hypothesis test on the very data that inspired that test in the first place then the statistics are invalid (or "wonky", as Ben Goldacre put it). In neuroimaging, this has been referred to as "double dipping" (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Vul, Harris, Winkielman, & Pashler, 2009). Whenever a researcher uses double-dipping strategies, Type I error rates will be inflated and *p* values can no longer be trusted.

As illustrated in Figure 1, psychological studies can be placed on a continuum from purely exploratory, where the hypothesis is found in the data, to purely confirmatory, where the entire analysis plan has been explicated before the first participant is tested. Every study in psychology falls somewhere along this continuum; the exact location may differ depending on the initial outcome (i.e., poor initial results may encourage exploration), the clarity of the research question (i.e., vague questions allow more exploration), the amount of data collected (i.e., more dependent variables encourage more exploration), the a priori beliefs of the researcher (i.e., strong belief in the presence of an effect encourages exploration when the initial result is ambiguous), and so on. Hence, the amount of exploration, data dredging, or data torture may differ widely from one study to the next; consequently, so does

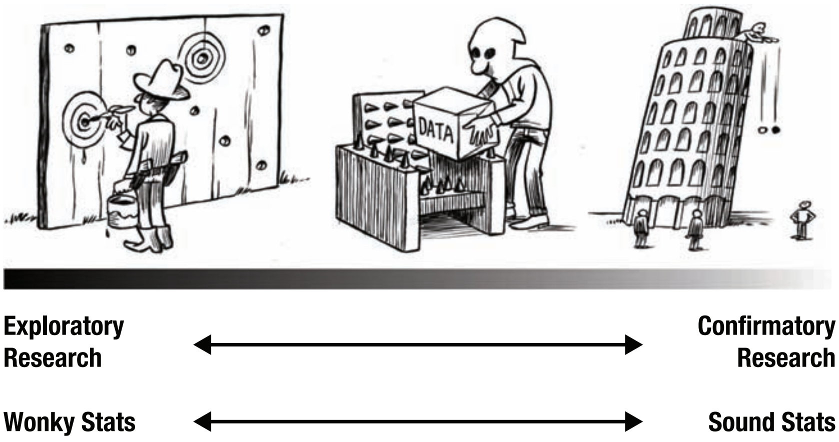

Fig. 1. A continuum of experimental exploration and the corresponding continuum of statistical wonkiness. On the far left of the continuum, researchers find their hypothesis in the data by post-hoc theorizing, and the corresponding statistics are "wonky," dramatically overestimating the evidence for the hypothesis. On the far right of the continuum, researchers preregister their studies such that data collection and data analyses leave no room whatsoever for exploration, and the corresponding statistics are "sound" in the sense that they are used for their intended purpose. Much empirical research operates somewhere in between these two extremes, although for any specific study the exact location may be impossible to determine. In the grey area of exploration, data are tortured to some extent, and the corresponding statistics are somewhat wonky. Figure downloaded from Flickr, courtesy of Dirk-Jan Hoek.

the reliability of the statistical results. It is important to stress again that we do not disapprove of exploratory research as long as its exploratory character is openly acknowledged. If fishing expeditions are sold as hypothesis tests, however, it becomes impossible to judge the strength of the evidence reported.

Together with other fairy-tale factors, the pervasive confusion between exploratory and confirmatory research threatens to unravel the very fabric of our field. This special issue features several papers that propose remedies to right what is wrong, such as changes in incentive structures (Nosek et al., 2012) and an increased focus on replicability (Bakker et al., 2012; Frank & Saxe, 2012; Grahe et al., 2012). In the next section, we stress a radical remedy that holds great promise, not just for the state of the entire field but also for researchers individually.

Good Science: Confirmatory Conclusions Require Preregistration

Science can be good in many ways, but a key characteristic is that the researcher is honest. Unfortunately, an abstract call for more honesty is unlikely to change anything. Blinded by confirmation bias and hindsight bias, researchers may be convinced that they are honest even when they are not. We therefore focus on a more concrete objective: separating exploratory experiments from confirmatory experiments.

The articles by Simmons et al. (2011) and John et al. (2012) suggest to us that considerable care needs to be taken before researchers are allowed near their own data: They may well torture them until a confession is obtained, even if the data are perfectly innocent. More important, researchers may then proceed to analyze and report their data as if these had undergone a spa treatment rather than torture. Psychology is not the only discipline in which exploratory methods masquerade as confirmatory, thereby polluting the field and eroding public trust (Sarewitz, 2012). In his fascinating book *Bad Science*, Ben Goldacre discusses several fairy tale factors in public health science and medicine, and concludes:

> What's truly extraordinary is that almost all of these problems—the suppression of negative results, data dredging, hiding unhelpful data, and more—could largely be solved with one very simple intervention that would cost almost nothing: a clinical trial register, public, open, and properly enforced (…) Before you even start your study, you publish the 'protocol' for it, the methods section of the paper, somewhere public. This means that everyone can see what you're going to do in your trial, what you're going to measure, how, in how many people, and so on, *before you start*. The problems of publication bias, duplicate publication and hidden data on side-effects—which all cause unnecessary death and suffering—would be eradicated overnight, in one fell swoop. If you registered a trial, and conducted it, but it didn't appear in the literature, it would stick out like a sore thumb. (Goldacre, 2009, pp. 220–221)

We believe this idea has great potential for psychological science as well (see also Bakker et al., 2012; Nosek et al., 2012, and the Neuroskeptic blog)[5] By preregistering the study design and the analysis plan, psychology's main fairy tale factor (i.e., presenting and analyzing exploratory results as if they were confirmatory) is eliminated in its entirety. To some, preregistering an experiment may seem a draconian measure. To us, this response only highlights how exceptional it is for psychologists to commit to a specific method of analysis in advance of data collection. Also, we wish to emphasize that we have nothing against exploratory work per se. Exploration is an essential component of science and is key to new discoveries and scientific progress; without exploratory studies, the scientific landscape is sterile and uninspiring. However, we do believe that it is important to separate exploratory from confirmatory work, and we do not believe that researchers can be trusted to observe this distinction if they are not forced to.[6] Hence, in the first stage of a research program, researchers should feel free to conduct exploratory studies and do whatever they please: turn the data inside out, discard participants and trials at will, and enjoy the fishing expedition. However, exploratory studies cannot be presented as strong evidence in favor of a particular claim; instead, the focus of exploratory work should be on describing interesting aspects of the data, on determining which tentative findings are of particular interest, and on proposing efficient ways in which future studies may confirm or disconfirm the initial exploratory results.

In the second stage of a research program, a purely confirmatory approach is desired. This requires the psychological science community to begin using online repositories such as the one that has recently been set up by the Open Science Framework at http://openscienceframework.org/.[7] Before a single participant is tested, the researcher submits to the online repository a document that details what dependent variables will be collected and how the data will be analyzed (i.e., which hypotheses are of interest, which statistical tests will be used, and which outlier criteria or data transformations will be applied). When *p* values are used, the researcher also needs to indicate exactly how many participants will be tested. When researchers wish to claim that their studies are confirmatory, the online document then becomes part of the review process.

An attractive implementation of this two-step procedure is to collect the data all at once and then split the data in an exploratory and a confirmatory subset.[8] For example, researchers can decide to freely analyze only the even-numbered participants, exploring the data however they like. In the next stage, however, the favored hypothesis can be tested on the odd-numbered participants in a purely confirmatory fashion. To enforce academic self-discipline, the second stage still requires preregistration. Although it is always possible for researchers to cheat, the main advantage of preregistration is that it removes the effects of confirmation bias and hindsight bias. In addition, researchers who cheat with respect to preregistration of experiments are well aware that they have committed a serious academic offense.

What we propose is a method to ensure academic honesty: there is nothing wrong with exploration as long as it is explicitly acknowledged as such. The only way to safeguard academics against fooling themselves, their readers, reviewers, and the general public, is to demand that confirmatory results are clearly separated from work that is exploratory. In a way, our proposal is merely a matter of common sense, and we have not met many colleagues who wish to argue against it; nevertheless, we know of almost no research in experimental psychology that follows this procedure.

Example: Precognitive Detection of Erotic Stimuli?

In 2011, Bem published an article in the *Journal of Personality and Social Psychology*, the flagship journal of social psychology, in which he claimed that people can look into the future (Bem, 2011; but see Galak, LeBoeuf, Nelson, & Simmons, in press; Ritchie, Wiseman, & French, 2012). In his first experiment, "precognitive detection of erotic stimuli," participants were instructed as follows: "(…) on each trial of the experiment, pictures of two curtains will appear on the screen side by side. One of them has a picture behind it; the other has a blank wall behind it. Your task is to click on the curtain that you feel has the picture behind it. The curtain will then open, permitting you to see if you selected the correct curtain." In the experiment, the location of the pictures was random and chance performance is therefore 50%. Nevertheless, Bem's participants scored 53.1%, significantly higher than chance; however, the effect was present only for erotic pictures, and not for neutral pictures, positive pictures, negative pictures, and romantic-but-not-erotic pictures. Bem also claimed that the psi effects were more pronounced for extraverts and that women showed psi for certain erotic pictures but men did not.

To illustrate our proposal we set out to replicate Bem's experiment in a purely confirmatory fashion. First, we detailed our method, design, and planned analyses in a document that we posted online before a single participant was tested.[9] As outlined in the online document, our replication focused on Bem's key findings; therefore, we tested only women, used only neutral and erotic pictures, and included a standard extraversion questionnaire. We also tested each participant in two contiguous sessions. Each session featured the same pictures but presented in a different random order. The idea is that individual differences in psi—if these exist—would lead to a positive correlation between performance in Session 1 and Session 2. Performance is quantified by the proportion of times that the participant chooses the curtain that hides the picture. Each session featured 60 trials, with 45 neutral pictures and 15 erotic pictures.

A vital part of the online document concerns the a priori specification of our statistical analyses. We decided in advance not to compute *p* values, as their main drawbacks include the inability to quantify evidence in favor of the null hypothesis (e.g., Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009), the sensitivity to optional stopping (e.g., Dienes, 2011;

Wagenmakers, 2007), and the tendency to overestimate the support in favor of the alternative hypothesis (e.g., Edwards, Lindman, & Savage, 1963; Sellke, Bayarri, & Berger, 2001; Wetzels et al., 2011). Instead, our main analysis tool is the Bayes factor (e.g., Hoijtink, Klugkist, & Boelen, 2008; Jeffreys, 1961; Kass & Raftery, 1995). The Bayes factor BF_{01} quantifies the evidence that the data provide for the null hypothesis (H_0) vis-a-vis an alternative hypothesis (H_1). For instance, when $BF_{01} = 10$, the observed data are 10 times as likely to have occurred under H_0 than under H_1. When $BF_{01} = 1/5 = .20$, the observed data are 5 times as likely to have occurred under H_1 than under H_0. An additional bonus of using the Bayes factor is that it eliminates the problem of optional stopping. As noted in the classic article by Edwards et al. (1963), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" (p. 193; see also Kerridge, 1963).

Hence, we outlined the details of our Bayes factor calculation in the online document:

> Data analysis proceeds by a series of Bayesian tests. For the Bayesian t-tests, the null hypothesis H_0 is always specified as the absence of a difference. Alternative hypothesis 1, H_1, assumes that effect size is distributed as Cauchy (0,1); this is the default prior proposed by Rouder et al. (2009). Alternative hypothesis 2, H_2, assumes that effect size is distributed as a half-normal distribution with positive mass only and the 90th percentile at an effect size of 0.5; this is the "knowledge-based prior" proposed by Bem et al. (submitted).[10] We will compute the Bayes factor for H_0 vs. H_1 (BF_{01}) and for H_0 vs. H_2 (BF_{02})."

The details of how the two alternative hypotheses were specified are not important here, save for the fact that these hypotheses were constructed a priori, based on general principles (the default prior) or substantive considerations (the knowledge-based prior).

Next, we outlined a series of six hypotheses to test. For instance, the second analysis was specified as follows:

> "(2) Based on the data of session 1 only: Does performance for erotic pictures differ from chance (in this study 50%)? To address this question we compute a one-sample t-test and monitor BF_{01} and BF_{02} as the data come in."

And the sixth analysis was specified as follows:

> "(6) Same as (2), but now for the combined data from sessions 1 and 2."

Readers curious to know whether people can look into the future are invited to examine the results for all six hypotheses in the online appendix at http://pps.sagepub.com/supplemental.[11] In this article, we only present the results from our sixth hypothesis. Figure 2 shows the development of the Bayes factor as the data accumulate. It is clear that the evidence in favor of H_0 increases as more participants are tested and the number of sessions increases. With the default prior, the data are 16.6 times more likely under H_0 than under H_1; with the "knowledge-based prior" from Bem, Utts, and Johnson (2011), the data are 6.2 times more likely under H_0 than under H_1. Because our analysis uses the Bayes factor, we did not have to indicate in advance that we were going to test 100 participants. We calculated the Bayes factor two or three times as the experiment was running, and after 100 participants we inspected

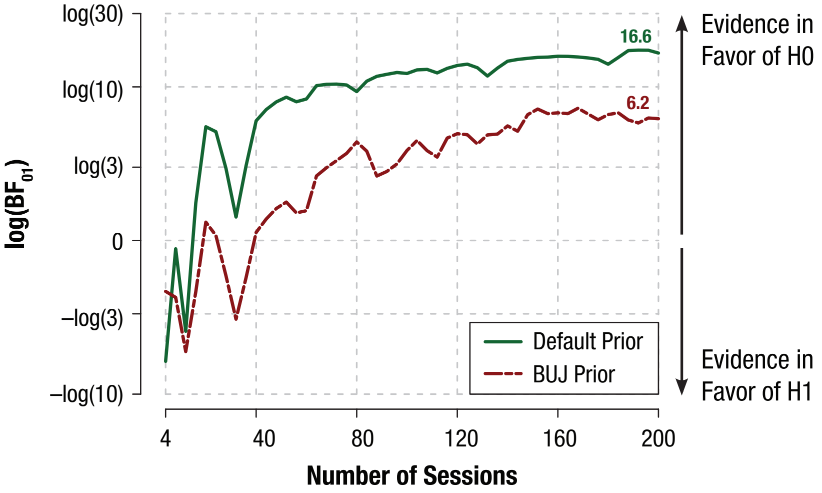

Fig. 2. Results from a purely confirmatory replication test for the presence of precognition. The intended analysis was specified online in advance of data collection. The evidence (i.e., the logarithm of the Bayes factor) supports H_0 ("performance for erotic stimuli does not differ from chance"). Note that the evidence may be monitored as the data accumulate.

Figure 2 and decided that the results were sufficiently compelling for the present purposes. Also note how the Bayes factor can be used to quantify evidence in favor of the null hypothesis.

The results reported here are purely confirmatory—absolutely everything that we have done here was decided before we saw the data. In this respect, these results are exceptional in experimental psychology, a state of affairs that we hope will change in the future.

Naturally, it is possible that our data might have shown something unexpected and interesting or that we could have forgotten to include an important analysis in our preregistration document. It is also possible that reviewers of this article could have asked for additional information (e.g., a credible interval for effect size). How should we deal with such alterations of the original data-analysis scheme? We suggest that, rather than walking the fine line of trying to decide which alterations are appropriate and which are not, *all* such findings and analyses should be mentioned in a separate section entitled "exploratory results." When such exploratory results are analyzed, it is important to realize that the data have been used more than once and that the inferential statistics may therefore to some extent be wonky.

Preregistration of our study was suboptimal. The key document was posted on Eric-Jan Wagenmakers's website and a purpose-made blog, and therefore the file would have been easy to alter, remove, or ignore.[12] With the online resources of the current day, however, the field should find it easy to construct a professional repository to push academic honesty to greater heights. We believe that researchers who use preregistration will quickly realize how different this procedure is from what is now standard practice. The extra work involved in preregistering an experiment is a small price to pay for a large increase in evidentiary impact. Top journals could facilitate the transition to more confirmatory research by implementing a policy to reward empirical manuscripts that feature at least one confirmatory experiment; for instance, these manuscripts could be published in a separate section explicitly containing confirmatory research. We hope that our proposal will increase the transparency of the scientific process, diminish the proportion of false findings, and improve the status of psychology as a rigorous scientific discipline.

Acknowledgments

We thank Adam Sasiadek, Boris Pinksterboer, Esther Lietaert Peerbolte, and Rebecca Schild for their help with data collection.

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.

Funding

This research was supported by Vidi grants from the Dutch Organization for Scientific Research (NWO).

Notes

[1] For an overview, see http://en.wikipedia.org/wiki/List_of_cognitive_biases.

[2] We are indebted to an anonymous reviewer of a different paper for bringing this sentence to our attention.

[3] Note the distinction between confirmation bias, which drives researchers to fine tune their analyses to the data, and confirmatory research, which prevents researchers from such fine tuning because the analysis steps have been specified in advance of data collection.

[4] The expression is attributed to Ronald Coase. Earlier, Mackay (1852/1932) made a similar statement, one that is perhaps even more apt: "When men wish to construct or support a theory, how they torture facts into their service!" (p. 552).

[5] See in particular http://neuroskeptic.blogspot.co.uk/2008/11/ registration-not-just-for-clinical.html, http://neuroskeptic.blogspot.co.uk/2011/05/how-to-fix-science.html, and http://neuroskeptic.blogspot.co.uk/2012/04/ fixing-science-systems-and-politics.html.

[6] This should not be taken personally: We distrust ourselves as well. In his cargo cult address, Feynman (1974) famously argued that the first principle of scientific integrity is that "(…) you must not fool yourself—and you are the easiest person to fool" (p. 12).

[7] The feasibility of this suggestion is evident from the fact that some other fields already use such registers—see, for instance, http://isrctn.org/ or http://clinicaltrials.gov/.

[8] This procedure is conceptually similar to cross-validation.

[9] See http://confrep.blogspot.nl/ and http://dl.dropbox.com/u/1018886/Advance_Information_on_Experiment_and_Analysis.pdf.

[10] This paper has since been published (i.e., Bem, Utts, & Johnson, 2011).

[11] Available from the first author's webpage or directly from https://dl.dropbox.com/u/1018886/Appendix_PoPS_WagenmakersEtAl.pdf.

[12] Some protection against this is offered by automatic archiving programs such as the Wayback Machine at http://archive.org/web/web.php.

References

Anderson, D. R., Burnham, K. P., Gould, W. R., & Cherry, S. (2001). Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin*, *29*, 311–316.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*, 531–533.

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.

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*, 716–719.

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558–1560.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Feynman, R. P. (1974). Cargo cult science. *Engineering & Science*, *37*, 10–13.

Frank, M. C., & Saxe, R. (2012). Teaching replication to promote a culture of reliable science. *Perspectives on Psychological Science*, *7*, 600–604.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (in press). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

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

Grahe, J., Reifman, A., Herman, A., Walker, M., Oleson, K., Nario–Redmond, M., & Wiebe, R. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science*.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. New York, NY: Springer.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again. *Science*, *334*, 1225.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth–telling. *Psychological Science*, *23*, 524–532.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *The Annals of Mathematical Statistics*, *34*, 1109–1110.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*, 535–540.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504.

Mackay, C. (1932). *Extraordinary popular delusions and the madness of crowds* (2nd ed.). Boston, MA: Page. (Original work published 1852)

Neuroskeptic. (2012). The nine circles of scientific hell. *Perspectives on Psychological Science*, *7*, 643–644.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

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.

Osherovich, L. (2011). Hedging against academic risk. *Science–Business eXchange*, *4*. doi:10.1038/scibx.2011.416

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.

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*, e33423.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*(2), 9, 27–29 .

Roese, N., & Vohs, K. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*, 411–426.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature*, *485*, 149.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

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.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

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

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*, 752–760.

Yong, E. (2012). Bad copy. *Nature*, *485*, 298–300.