Bayesian model selection with applications in social science
Wetzels, R.M.

**Citation for published version (APA):**
Wetzels, R. M. (2012). Bayesian model selection with applications in social science

**General rights**
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
8 An Agenda for Purely Confirmatory Research

Abstract

The veracity of substantive research claims hinges on the way experimental data are collected and analyzed. Here we emphasize two uncomfortable facts that threaten the core of our scientific enterprise. First, psychologists generally 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 fine-tuning varies widely across experiments and experimenters but is almost impossible for reviewers and readers to gauge. Second, $p$ values overestimate the evidence against the null hypothesis and disallow any flexibility in data collection. We propose that researchers pre-register 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. All other analyses should be labeled “exploratory”. We also propose that researchers interested in hypothesis tests use Bayes factors rather than $p$ values. Bayes factors allow researchers to monitor the evidence as the data come in, and stop whenever they feel a point has been proven or disproven. We illustrate our proposals with a confirmatory replication attempt of a study on ESP.
You cannot find your starting hypothesis in your final results. It makes the stats go all wonky. – Ben Goldacre, 2009, p. 221, Bad Science.

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 experience cognitive dissonance when the facts do not correspond to the desired state of the world, an unwelcome tension that people will attempt to reduce; moreover, we know that people fall prey to hindsight bias, the tendency to judge an event as more predictable after it has occurred (Christensen–Szalanski & Willham, 1991).

In light of these and other human biases\(^1\) it would be naive to believe that, without special protective measures, the scientific research process is somehow exempt from these systematic imperfections of the mind. When bias influences the research process this means that researchers seek to confirm, not falsify, their main hypothesis (Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). This is especially relevant 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, in press; John, Loewenstein, & Prelec, 2012; Nosek, Spies, & Motyl, in press; Neuroskeptic, in press). 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).

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, this suggests that our discipline is in serious trouble (S. 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 first discuss four popular practices that result in bad science\(^2\); we call these “fairy tale factors”, because each factor increases the probability that a presented finding is fictional and hence non-replicable. Next we propose two radical remedies to ensure scientific integrity and inoculate the research process against the inalienable biases of human reasoning. We end by illustrating the remedies to a replication attempt of an ESP experiment reported by Bem (2011).

\(^1\)For an overview see for instance http://en.wikipedia.org/wiki/List_of_cognitive_biases.

\(^2\)This list is not meant to be exhaustive.
8.1 Bad Science

Science can be bad in many ways. Flawed design, faulty logic, and limited scholarship engender no enthusiasm whatsoever. Here we list four factors that bias the research process and make experimental results appear to be more compelling than they really are (see also Simmons et al., 2011).

Fairy Tale Factor 1: Exploratory Analyses, Confirmatory Conclusions

Our main concern is that almost no psychological research is conducted in a purely confirmatory fashion (e.g., Kerr, 1998; Wagenmakers et al., in press). 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 et al., 2011; for what may happen to psychologists in the afterlife see Neuroskeptic, in press). 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; they can include in their papers only those experiments that produced the desired outcome, even though these experiments were designed as pilot experiments, ready to be 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. In the past decades, the development of statistical software has resulted in a situation where 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” 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, increasing the odds of a 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 like conspiracy theories are never disproved 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, for testing one’s hypothesis, one always needs a fresh data set. This also means that the interpretation of common statistical tests in terms of

---

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

4 The expression is attributed to Ronald Coase. Earlier, Mackay (1852/1932, p. 552) 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!"
type I and type II error rates is valid only if (a) the data were used only once, and (b) the statistical test was not chosen on the basis of relevant characteristics of 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). If a researcher uses double dipping strategies, type I error rates will be inflated considerably, and as a result p values are no longer trustworthy.

As illustrated in Figure 8.1, 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 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.

Fairy Tale Factor 2: Publication Bias or Aversion to the Null

Few researchers like null results. Compared to statistically significant results (i.e., $p < .05$), null results (i.e., $p > .1$) are inherently ambiguous: perhaps the research was carried out poorly, or perhaps the experiment did not have enough power. Also, null results are sometimes uninteresting (e.g., “people are just as creative when they are inside or outside a large box”). When only positive studies are published and the null results are rejected or disappear into the file drawer, the literature does not fairly represent the true state of affairs. There is ample evidence that the file drawer effect in psychology is rather large; for instance, Sterling et al. (1995) found that more than 95% of articles in psychology journals confirm their main hypothesis (see Bones, 2012, for an alternative account).

Fairy Tale Factor 3: Optional Stopping

Optional stopping or “sampling to a foregone conclusion” is a popular method of data collection that, within the framework of $p$ value hypothesis testing, is nevertheless tantamount to cheating (e.g., Jennison & Turnbull, 1990; Strube, 2006; Wagenmakers, 2007). The method consists of taking multiple looks at the data as they come in, and stopping data collection whenever the desired result is obtained. The problem is that the standard tests work as advertised only when the number of participants has been determined in advance. Without a correction for multiple looks, the probability of falsely rejecting the null hypothesis is larger than .05. It is again important to note that optional stopping in itself is not a problem, as long as it is openly reported and as long as the relevant statistics are corrected for this strategy.

The problem is perhaps clearest when a researcher tests 20 participants, finds $p = .11$, proceeds to test 10 more participants, and then reports $n = 30$ and $p = .04$, the latter value computed as if 30 participants were scheduled from the start. It is less clear that the
Figure 8.1: 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 pre-register their studies such that data collection and data analyses leave no room whatsoever for exploration; 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.

problem of optional stopping may also exist when a researcher never takes a sneak peak at the data at all. For instance, a researcher may test 30 participants in one sitting, find $p = .04$, and report the result – and this could nevertheless be tantamount to cheating. This is because one may ask the researcher “what would you have done in case the results had not been significant after the initial 30 participants”? One possible answer is “When the $p$ value is higher than .05 but lower than .15 I would have tested 10 participants more”. This answer reveals that the sampling plan did not fix the number of participants in advance, and hence the problem of optional stopping persists.

**Fairy Tale Factor 4: $p$ Values Overestimate the Evidence Against the Null**

The $p$ value is the probability of obtaining a value for a test statistic that is at least as extreme as the one that was actually observed, given that the null hypothesis is true and relevant statistical assumptions are met (these may involve characteristics of the data, like normality and linearity, or characteristics of the research process, such as a the prior specification of a sampling plan). If the $p$ value is low, researchers reject the null hypothesis and assume, explicitly or implicitly, that the alternative hypothesis is much better supported by the data. But this reasoning is fallacious. The observed data (summarized
8. An Agenda for Purely Confirmatory Research

by a test statistic) could be just as unlikely under the alternative hypothesis. Thus, what matters is the relative likelihood of the data under various competing explanations, not the probability of the data under a single explanation (e.g., D. V. Lindley, 1993).

Because the p value is not comparative, it can be shown to overestimate the evidence against the null. In particular, Sellke et al. (2001) considered the diagnosticity of the p value: how much more likely is a particular observed p value under \( H_1 \) than it is under \( H_0 \)? The results are shocking. A p value of \( p = .037 \), for instance, is at best 3 times more likely under \( H_1 \) than under \( H_0 \); a p value of \( p = .01 \) is at best 8 times more likely under \( H_1 \) than under \( H_0 \). Since these values are upper bounds, derived by cherry-picking the single most competitive \( H_1 \), the true diagnosticity is likely to be even lower.

When researchers use a low-diagnostic criterion to detect experimental effects, it should come as no surprise if many of these effects do not replicate.

Together, these and other fairy tale factors create a perfect storm that threatens to unravel the very fabric of our field. This special issue features several papers that propose remedies to right what is wrong, for instance through changes in incentive structures (Nosek et al., in press) and an increased focus on replicability (Bakker et al., in press; Frank & Saxe, in press; Grahe et al., in press). In the next section we stress two radical remedies that hold great promise, not just for the state of the entire field but also for researchers individually.

8.2 Good Science

Science can be good in many ways, but a key characteristic is that the researcher is honest. Unfortunately, a 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 tangible characteristic of good science, namely that it should minimize the impact of the fairy tale factors discussed above.

Solution 1: Separate Exploratory 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 importantly, 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., in press; Nosek et al., in press, and the NeuroSkeptic blog5) By pre-registering the study design and the analysis plan, the first fairy tale factor (i.e., presenting and analyzing exploratory results as if they were confirmatory) is eliminated entirely. The second factor (i.e., aversion to the null) is also avoided, as non-published pre-registered experiments “stick out like a sore thumb”.

To some, pre-registering 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, 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 that the psychological science community set up an online repository comparable to the usual article submission websites such as Manuscript Central.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. 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 pre-registration. Although it is always possible for researchers to cheat, the main advantage of pre-registration is that it removes the effects of confirmation bias and hindsight bias. In addition, researchers

---


6This should not be taken personally: we distrust ourselves as well.

7The Open Science Framework may provide such a service, see http://openscienceframework.org/.
who cheat with respect to pre-registration 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.

**Solution 2: Bayesian Hypothesis Tests**

Fairy tale factor four (i.e., \( p \) values overestimate the evidence against the null) can be eliminated by calculating a comparative measure of evidence. One such measure, the one we focus on here, is the Bayes factor. The Bayes factor \( BF_{01} \) quantifies the evidence that the data provide for \( H_0 \) vis-a-vis \( H_1 \) (Hoijtink et al., 2008; Jeffreys, 1961; Kass & Raftery, 1995; Masson, 2011; Rouder et al., 2009; Wagenmakers et al., 2010; Wetzels et al., 2009). 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 \( H_0 \). It is important to realize that the Bayes factor quantifies the evidence brought about by the data and does not in any way depend on the prior probabilities that are assigned to \( H_0 \) and \( H_1 \).

Thus, the complication with the Bayes factor does not reside in the prior probabilities that are assigned to the hypotheses. Instead, the main complication lies in the requirement to fully specify \( H_1 \). In particular, we need to state what effect sizes can be expected should \( H_1 \) be true; in Bayesian terminology, we need to assign effect size a prior distribution. This is a choice that should be made judiciously, and much work in Bayesian statistics has concerned this crucial problem (Jeffreys, 1961; Liang et al., 2008; Rouder et al., 2009; Zellner & Siow, 1980). One method, the one we prefer, is to use a default specification process based on general principles (i.e., an “objective Bayesian hypothesis test”). Another method is to use substantive knowledge and specify prior distributions for effect size on the inference problem at hand. Of course, such subjective specification should be carried out before the data are analyzed in order to prevent the data analyst from falling prey to hindsight bias and assigning effect size a prior distribution that matches the observed data too closely.

An additional bonus of using the Bayes factor is that it eliminates fairy tale factor three (i.e., optional stopping). As noted in the classic article by Edwards et al. (1963, p. 193), “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.” (see also Kerridge, 1963). This means that researchers who use the Bayes factor should feel entirely uninhibited to continue data collection in case the initial results are not sufficiently compelling. Likewise, when early results are compelling, researchers who use the Bayes factor can just stop data collection and report the result without feeling any pressure to continue collecting more data. Space constraints prevent us from discussing the other advantages of using Bayes factors (e.g., the ability to quantify evidence in favor of \( H_0 \), the fact that discovery of the truth is guaranteed as the number of data points increases).

---

8 One method is to use unit-information priors, that is, priors that contain as much information as a single observation.

9 Denny Borsboom wishes to state that, unlike the first author, he is not an evangelical Bayesian.
8.3 Example: Precognitive Detection of Erotic Stimuli?

In 2011, Dr. 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). 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 are more pronounced for extraverts, and that for certain erotic pictures women show psi but men do not.

In order 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. 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 them in a different random order. The idea is that individual differences in psi –if these exist– lead to a positive correlation between performance on 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 analyses. First we outlined our main analysis tool, the Bayes factor t-test:

“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). We will compute the Bayes factor for $H_0$ vs. $H_1$ ($BF_{01}$) and for $H_0$ vs. $H_2$ ($BF_{02}$).”

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 an online appendix. Here we only present the results from our sixth hypothesis. Figure 8.2 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 et al. (2011) the data are 6.2 times more likely under $H_0$ than under $H_1$. Note that 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 Figure 8.2 and decided that for the present purposes the results were sufficiently compelling. Note how the Bayes factor can be used to quantify evidence in favor of the null hypothesis.

---

12Available from the first author’s webpage or directly from https://dl.dropbox.com/u/1018886/Appendix_PoPS_WagenmakersEtAl.pdf.
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 had shown something unexpected and interesting, or that we could have forgotten to include an important analysis in our pre-registration document. It is also possible that reviewers of this paper will ask 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 the inferential statistics may therefore to some extent be wonky.

Pre-registration of our study was sub-optimal. The key document was posted on the first author’s website and a purpose-made blog, and therefore the file would have been easy to alter, remove, or ignore. 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 pre-registration will quickly realize how different this procedure is from what is now standard practice. 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.