Running head: BAYESIAN HYPOTHESIS TESTING

The Need for Bayesian Hypothesis Testing in Psychological Science

Eric-Jan Wagenmakers[1], Josine Verhagen[1], Alexander Ly[1], Dora Matzke[1], Helen Steingroever[1], Jeff N. Rouder[2], Richard Morey[3]

[1] University of Amsterdam

[2] University of Missouri

[3] University of Groningen

Correspondence concerning this article should be addressed to:

Eric-Jan Wagenmakers

University of Amsterdam, Department of Psychology

Weesperplein 4

1018 XA Amsterdam, The Netherlands

E-mail may be sent to EJ.Wagenmakers@gmail.com.

The Need for Bayesian Hypothesis Testing in Psychological Science

Mike is an honest, hard-working graduate student at a respectable psychology department somewhere in the Mid-West. Mike's thesis centers on the unconscious processing of fear-inducing stimuli. Mike is well aware of the recent crisis of confidence in psychology (Pashler & Wagenmakers, 2012), a crisis brought about by a toxic mix of fraud, questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), lack of data sharing (Wicherts, Borsboom, Kats, & Molenaar, 2006), publication bias (Francis, 2013), and a blurred distinction between statistical analyses that are pre-planned and post-hoc (De Groot, 1956/2014; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Undeterred, Mike sets out to conduct his own research according to the highest standards. He immerses himself in the relevant literature and after some thought devises the "Abstract Unconscious Fear Processing" (AUFP) theory, which predicts that due to the way the brain processes certain stimuli, there are abstract patterns of shapes that when processed only unconsciously, will produce a very high fear response. The AUFP theory makes three specific predictions about processing of fear-inducing stimuli. The first prediction is that when these special, abstract shapes are simply shown to participants, they will be only mildly more disliked than similar, but non-AUFP, stimuli. The second prediction is that when the stimuli are shown in a dual-task scenario (where participants are required to perform two tasks simultaneously), AUFP stimuli will produce a moderate fear-related physiological response due to the occasional lapses of conscious attention to the stimuli. The third prediction is that when presented to the participants in a hypnotic state, the physiological response will be very large compared with non-AUFP stimuli.

Mike proceeds to test each of the three predictions in a separate experiment, each

with 25 participants receiving AUFP stimuli and 25 participants receiving non-AUFP stimuli. To counteract hindsight bias and HARKing (Hypothesizing After the Results are Known; De Groot, 1956/2014; Kerr, 1998), Mike first preregisters each experiment on the Open Science Framework (Open Science Collaboration, 2012), detailing in advance his entire analysis plan including criteria for excluding outliers and transformations of dependent variables. Mike then collects the data and conducts the planned statistical analyses. The results show that $p = .04$ in all three experiments; none of the 95% confidence intervals for effect size overlap with zero. Consequently, Mike concludes that in each of the experiments the results are significant, the null hypothesis can be rejected, the effects are present, and the data support Mike's AUFP theory. His peers congratulate Mike on his exemplary academic conduct, and the party to celebrate the significant results lasts well into the night. Mike later manages to publish the findings in *Psychological Science*, earning Open Science badges for "Open Materials", "Open Data", and "Preregistration" along the way.

Mike has done almost everything right, and in many ways his research is a blueprint that all studies in experimental psychology should seek to emulate: no questionable research practices, no confusion between exploratory and confirmatory research, and almost perfect transparency in methodology and data.[1] Nevertheless, as we explain below in detail, Mike's conclusions are based on flimsy evidence. Hence, Mike's findings run the risk of being spurious, polluting the field and setting back research in his field several years. Mike's party, we suggest, was wholly premature.

The goal of this chapter is twofold. Our main goal is to explain why the logic behind p value significance tests is faulty, leading researchers to mistakenly believe that their results are diagnostic when they are not. Our secondary goal is to outline a Bayesian alternative that overcomes the flaws of the p value procedure, and provides researchers with an honest assessment of the evidence against or in favor of the null hypothesis.

The Logic of p Values: Fisher's Disjunction

Almost without exception, psychologists seek to confirm the veracity of their findings using the statistical method of null hypothesis significance testing (NHST). In this method, first proposed by Sir Ronald Aylmer Fisher (1890-1962), one puts forward a null hypothesis that represents the absence of the effect of interest. The inadequacy of this null hypothesis is then considered evidence for the presence of the effect. Hence, the core idea behind NHST is similar to a proof by contradiction: to show that A holds, one hypothesizes the opposite (i.e., not-A), and demonstrates that this situation is impossible (or, in NHST, unlikely).

The inadequacy of the null hypothesis is measured through the infamous p value (Nuzzo, 2014). The p value is the probability of encountering the value of a test statistic at least as extreme as the one that was observed, given that the null hypothesis is true. In other words, the p value captures the extremeness of the data under the null hypothesis. Extreme results –usually, p values smaller than a threshold of .05– are cause to reject the null hypothesis. Indeed, as proposed by Fisher, the p value quantifies "the strength of the evidence against the [null] hypothesis" (Fisher, 1958, p. 80); when $p = .001$ this is more compelling evidence against the null hypothesis than when $p = .049$.[2]

As discussed in Wagenmakers (2007), some authors have given explicit guidelines with respect to the evidential interpretation of the p value. For instance, Burdette and Gehan (1970, p. 9) associated specific ranges of p values with varying levels of evidence (see also Wasserman, 2004, p. 157): When $p > .1$ this yields "little or no real evidence against the null hypothesis"; $.05 < p < .1$ implies "suggestive evidence against the null hypothesis"; $.01 < p < .05$ yields "moderate evidence against the null hypothesis"; and $p < .01$ constitutes "very strong evidence against the null hypothesis".

The logic that underlies the p value as a measure of evidence is based on what is known as *Fisher's disjunction*. According to Fisher, a low p value indicates either that an

exceptionally rare event has occurred or that the null hypothesis is false. The next section shows that this logic is not as compelling as it appears at first glance.

The Illogic of p Values

Despite their dominance in scientific practice, p values have been criticized on many counts (for reviews see Berger & Wolpert, 1988; Nickerson, 2000; Wagenmakers, 2007). Here we focus on an inherent weakness of p values: the fact that they depend only on what is expected under the null hypothesis $\mathcal{H}_0$ — what is expected under an alternative hypothesis $\mathcal{H}_1$ is simply not taken into consideration. As we will see below, this omission disqualifies the p value as a measure of evidence.

To the best of our knowledge, this general critique was first put forward by Gosset, the inventor of the t test, who wrote Egon Pearson in 1926 and argued that "...an observed discrepancy between a sample mean and a hypothesized population mean 'doesn't in itself necessarily prove that the sample was not drawn randomly from the population even if the chance is very small, say .00001: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05...you will be very much more inclined to consider that the original hypothesis is not true' (Gosset[1926], quoted in Pearson, 1938)" (Royall, 1997, p. 68).

This critique was echoed by Berkson (1938, p. 531): "My view is that *there is never any valid reason for rejection of the null hypothesis except on the willingness to embrace an alternative one.* No matter how rare an experience is under a null hypothesis, this does not warrant logically, and in practice we do not allow it, to reject the null hypothesis if, for any reasons, no alternative hypothesis is credible." (italics in orginal).

To appreciate the logical validity of the Gosset-Berkson critique, it is important to recognize that Fisher's disjunction is similar to the *modus tollens* argument in deductive reasoning. In abstract form, this syllogistic argument proceeds as follows:

(Premise) If A, then B;

(Premise) not B;

(Conclusion) not A.

A specific example is the following:

(Premise) If Mark has been hanged, then he is dead;

(Premise) Mark is alive;

(Conclusion) Mark has not been hanged.

Fisher's disjunction is of the same form, and as cast below, it is logically valid:

(Premise) If $\mathcal{H}_0$, then not y;

(Premise) y;

(Conclusion) not $\mathcal{H}_0$.

Henceforth, we will use y to denote the observed data; in the NHST syllogism above, one summarizes y by the p value, integrating over more extreme outcomes that have not been observed. For the discussion in this chapter, the distinction is irrelevant (but see Berger & Wolpert, 1988 for scenarios where the distinction is relevant).

For deductive reasoning then, Fisher's disjunction is a valid case of *modus tollens*. However, statistical inference is probabilistic, and therefore Fisher's disjunction is really of the following form:

(Premise) If $\mathcal{H}_0$, then y very unlikely;

(Premise) y;

(Conclusion) $\mathcal{H}_0$ very unlikely.

But this probabilistic version of *modus tollens*, however, is not logically valid. To see this, consider the following non-sequiturs; first, an example suggested by Pollard and Richardson (1987):

(Premise) If Tracy is an American then it is very unlikely that she is a US congresswoman;

(Premise) Tracy is a US congresswoman;

(Conclusion) It is very likely that Tracy is not an American.

Of course, the conclusion should be that Tracy is an American – if she were not it would be impossible for her to be a US congresswoman. Another example is inspired by Beck–Bornholdt and Dubben (1996):

(Premise) If an individual is a man, he is unlikely to be the Pope;

(Premise) Francis is the Pope;

(Conclusion) Francis is probably not a man.

One final example:

(Premise) If John does not have ESP, then he probably will not make money at the casino tonight;

(Premise) John made money at the casino tonight;

(Conclusion) John probably has ESP.

The fact that the typical reasoning from Fisher's disjunction is logically invalid is well-known (e.g., Beck–Bornholdt & Dubben, 1996; Cohen, 1994; Cortina & Dunlap, 1997; Falk & Greenbaum, 1995; Falk, 1998; Pollard & Richardson, 1987; Krämer & Gigerenzer, 2005; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2014; Schneider, 2014; but see Edwards, 1996; Hagen, 1997, 1998; for a review see Nickerson, 2000). Surely there must be a way of reasoning in situations of uncertainty that *is* logically valid. In the next section, we present a generalization of propositional logic that can be used for just this purpose.

Generalizing Logic: The Bayesian Perspective

Consider observed data y, a null hypothesis $\mathcal{H}_0$, and an alternative hypothesis $\mathcal{H}_1$. The first two premises in the *modus tollens* NHST argument state that $p(y \mid \mathcal{H}_0)$ is low. What we would like is a method of using the premises we have to make a statement about the plausibility of the hypothesis, given the data. If the plausibility is sufficiently low, we can reject $\mathcal{H}_0$. The central question is: what are the laws of plausibility?

Cox (1946) showed that given three simple axioms –including one that requires the laws of plausibility to be generalizations of propositional logic –the laws of plausibility are precisely the laws of *probability*. Our target for inference is $p(\mathcal{H}_0 \mid y)$, which represents the plausibility of $\mathcal{H}_0$ given the observed data. Assume one is reluctant to reject $\mathcal{H}_0$ when it has considerable plausibility, that is, when $p(\mathcal{H}_0 \mid y)$ is relatively high. Since the laws of plausibility are the laws of probability, we know that

$$p(\mathcal{H}_0 \mid y) = \frac{p(y \mid \mathcal{H}_0)p(\mathcal{H}_0)}{p(y \mid \mathcal{H}_0)p(\mathcal{H}_0) + p(y \mid \mathcal{H}_1)p(\mathcal{H}_1)}, \tag{1}$$

by Bayes' theorem, which forms the foundation for Bayesian statistics.

As expected, when y is an impossibility under $\mathcal{H}_0$, Equation 1 reproduces the result from deterministic syllogistic reasoning: when $p(y \mid \mathcal{H}_0)$ equals zero then so will $p(\mathcal{H}_0 \mid y)$. However, when y is merely improbable rather than impossible, the a posteriori plausibility of $\mathcal{H}_0$ depends crucially on (1) the prior plausibility of $\mathcal{H}_0$ (cf. the ESP example above); and (2) $p(y \mid \mathcal{H}_1)$, that is, the unlikeliness of the data under the alternative hypothesis (cf. the US congress example above). In the words of Sellke, Bayarri, and Berger (2001, p. 64-65): "The clear message is that knowing that the data are 'rare' under $\mathcal{H}_0$ is of little use unless one determines whether or not they are also 'rare' under $\mathcal{H}_1$."

At this point, those invested in NHST may interject that the syllogistic counter-examples are far-fetched, that science does not necessarily have to use logical rules for inference, and that –from a practical point of view– the negative consequences of using

p values are overstated. The next section intends to demonstrate with a concrete example that such counterarguments fall flat: the drawbacks of p values are real and noticeable even in standard, run-of-the-mill statistical paradigms.

A Concrete Example: Results from AUFP Re-examined

The practical ramifications of p value logic are apparent from Mike's AUFP research discussed in the first paragraphs of this chapter. Recall that Mike tested 25 participants with AUFP stimuli and 25 participants with non-AUFP stimuli. In each of the experiments, the dependent measure was assumed to be approximately normally distributed, and therefore the adequacy of the null hypothesis $\mathcal{H}_0 : \delta = 0$ (i.e., AUFP and non-AUFP stimuli do not differ on the dependent measure) was assessed using a two-tailed, unpaired t test. In each experiment, the result was $t(48) = 2.11$, $p = .04$. The 95% confidence interval for δ ranges from .03 to 1.16 and does not overlap with zero.

The statistical outcomes of each experiment are displayed in the three right-hand panels of Figure 1. In each panel, the solid line indicates the t distribution that is expected under $\mathcal{H}_0$, and the gray vertical line indicates the test statistic that was observed in the experiment. For all three experiments, the observed test statistic is in the 98th percentile and can therefore be considered relatively extreme, given that $\mathcal{H}_0$ holds. Hence, it appears that in all three experiments, the data provide ample justification to reject $\mathcal{H}_0$, a line of reasoning that pervades current-day statistical reasoning in all empirical disciplines including psychology.

However, consider what happens when we add, for each experiment, the expectations based on a plausible alternative hypothesis $\mathcal{H}_1$, the hypothesis that the p value ignores. The top two panels of Figure 1 feature an alternative hypothesis for Experiment 1 (i.e., the test that AUFP stimuli are liked somewhat less than non-AUFP stimuli when simply shown). This alternative hypothesis is characterized by a relatively

small effect size: $\mathcal{H}_1 : \delta = .15$. In the top right panel, the dotted line shows the expectation for the test statistic under this alternative hypothesis. The top left panel illustrates what this means in terms of the population difference between participants viewing AUFP stimuli and those viewing non-AUFP stimuli. It is immediately apparent that, even if AUFP stimuli are more disliked than non-AUFP stimuli, the predicted differences are relatively small. Hence, the observed p value is not diagnostic; the top right panel of Figure 1 shows that the observed data y are almost as likely to have occurred under $\mathcal{H}_0$ as under $\mathcal{H}_1$. The *likelihood ratio* (i.e., the ratio of the ordinates of the two distributions at the point of the observed test statistic) is only 2.56.

The two middle panels feature an alternative hypothesis for Experiment 2 (i.e., the test that AUFP causes moderate physiological responses in dual-task scenarios) that is a little more extreme: $\mathcal{H}_1 : \delta = .60$. The middle left panel illustrates what this means in terms of the population difference between participants viewing AUFP stimuli and those viewing non-AUFP stimuli. The middle right panel shows that the observed data are now clearly more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$; the likelihood ratio is 8.61. Note that under $\mathcal{H}_1 : \delta = .60$ the expectation is at its peak for the observed test statistic. Under any other alternative hypothesis, the peak expectation shifts away from the observed test statistic. Consequently, considered across all possible alternative hypotheses $\mathcal{H}_1 : \delta = x$, the maximum likelihood ratio is achieved for $\mathcal{H}_1 : \delta = .60$. In other words, suppose a researcher reports a likelihood ratio and is motivated to present the null hypothesis in the least favorable light. The researcher cheats and cherry-picks the alternative hypothesis that maximizes the likelihood ratio; the alternative hypothesis of choice is $\mathcal{H}_1 : \delta = .60$, where the expectation peaks at the observed test statistic and the likelihood ratio equals 8.61.

The bottom panels feature an alternative hypothesis for Experiment 3 (i.e., the test that AUFP causes large physiological responses when participants are in a hypnotic state)

that is relatively extreme: $\mathcal{H}_1 : \delta = 2.0$. The bottom left panel illustrates what this means in terms of the population difference between participants viewing AUFP stimuli and those viewing non-AUFP stimuli. Surprisingly perhaps, the bottom right panel shows that the observed data are now more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, even though $p = .04$. How can this be? As indicated by the solid curve, the null hypothesis $\mathcal{H}_0 : \delta = 0$ predicts t values that are relatively small; as indicated by the dashed curve, the alternative hypothesis $\mathcal{H}_1 : \delta = 2.0$ predicts t values that are relatively high. The observed t value (indicated by the gray line) falls somewhere in between these two expectations, but is more consistent with $\mathcal{H}_0$ than it is with $\mathcal{H}_1$. In other words, the observed data are somewhat rare under the null hypothesis (as indicated by $p = .04$), but they are more rare under the alternative hypothesis $\mathcal{H}_1 : \delta = 2.0$. This difference in rarity is quantified by a likelihood ratio that is $13,867$ in favor of $\mathcal{H}_0$. This result illustrates the phenomenon that "(...) the more powerful the test, the more a just significant result favors the null hypothesis." (Pratt, 1961, p. 166).

This trio of p values highlights the importance of the alternative hypothesis; the evidence is weak in all but the second experiment shown in the middle panel of Figure 1. For the top and bottom panels, the data do not provide compelling evidence for AUFP; hence, *Psychological Science* should not have accepted Mike's paper, and the party celebrating the results was uncalled for. This should be shocking: in all three experiments, $p = .04$, the confidence intervals do not overlap with zero, and yet it is wholly premature to reject the null hypothesis, for at least two out of the three experiments.

This is so important, so vital, that we repeat it here. All three of Mike's experiments yielded a significant result, $p < .05$, yet for only one of them did the statistical evidence actually support his claim that the null hypothesis should be rejected (albeit not as strongly as the p value may suggest). This occurs because the data may be extreme under $\mathcal{H}_0$, but they are not likely under $\mathcal{H}_1$ either, and it is the balance between

the two that provides the evidence. As noted by Edwards (1965, p. 402): "The trouble is that in classical statistics the alternative hypothesis is essentially undefined, and so provides no standard by means of which to judge the congruence between datum and null hypothesis; hence the arbitrariness of the .05, .01, and .001 levels, and their lack of agreement with less arbitrary measures of congruence. A man from Mars, asked whether or not your suit fits you, would have trouble answering. He could notice the discrepancies between its measurements and yours, and might answer no; he could notice that you did not trip over it, and might answer yes. But give him two suits and ask him which fits you better, and his task starts to make sense, though it still has its difficulties."

The paradox is visualized in Figure 2: the referee is Fisherian, and, considering the abysmal state of boxer H_0, declares his opponent $\mathcal{H}_a$ the winner. To the audience, however, it is clear that boxer $\mathcal{H}_a$ does not look too healthy either, and a decision based only on the state of boxer H_0 is irrational, premature, and potentially misleading.

The Bayesian Remedy

Implicit in the above discussion is that a more appropriate measure of evidence is given by the likelihood ratio, that is, the relative plausibility of the observed data y occurring under $\mathcal{H}_1$ versus $\mathcal{H}_0$: $p(y \mid \mathcal{H}_1)/p(y \mid \mathcal{H}_0)$ (Royall, 1997). Unfortunately, it happens rarely that we know $\mathcal{H}_1$ exactly (e.g., $\delta = .25$ or $\delta = .30$). However, we might know $\mathcal{H}_1$ approximately – and when we are Bayesian, our uncertainty about the true value of δ can be formalized using a probability distribution. This way we can define an alternative hypothesis not by a single, specific effect size, but rather by a collection of different effect sizes, weighted by their plausibility.

After assigning effect size a distribution, we wish to compute the overall evidence for $\mathcal{H}_0 : \delta = 0$ versus the "composite" alternative hypothesis $\mathcal{H}_1 : \delta \sim f(\cdot)$. This can be accomplished by averaging the likelihood ratios over the distribution that has been

assigned to effect size under $\mathcal{H}_1$ (e.g., Lee & Wagenmakers, 2013, Chapter 7). This average likelihood, better known as the *Bayes factor* (Jeffreys, 1961), quantifies the extent to which the data are more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$.

What remains is to choose a distribution for effect size under $\mathcal{H}_1$. This choice can be guided by general desiderata such as scale invariance (i.e., the prior should result in the same Bayes factor regardless of the unit of measurement) and model consistency (i.e., the prior should give rise to a Bayes factor that asymptotically converges upon the true model). Based on these and other desiderata, outlined in Bayarri, Berger, Forte, and García-Donato (2012), an attractive prior for effect size is a Cauchy distribution[3] with scale 1. Of course, other choices are possible: a standard normal distribution, a Cauchy distribution with smaller width, etc. Each choice corresponds to a different assumption about the alternative hypothesis; consequently, each choice yields a different measures of evidence, something that is already apparent from Figure 1. Researchers may check the robustness of their conclusions by examining a range of prior distributions (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). As an example, consider again Mike's data. Table 1 shows the Bayes factors for different prior distributions on effect size.

For Mike's data, the Cauchy(0,$r = 1$) prior yields $\text{BF}_{10} = 1.45$, indicating that the data are about equally likely under $\mathcal{H}_1$ and $\mathcal{H}_0$. A similar conclusion (i.e., $\text{BF}_{10} = 1.83$) follows when we halve the scale of the Cauchy distribution. A standard normal distribution for effect size yields $\text{BF}_{10} = 2.02$. These different choices underscore the robustness of the general conclusion: the data are not very informative. To explore the upper limits of the evidence we can use "oracle priors", distributions on effect size that are informed by the data themselves. Specifically, an oracle prior is constructed by peaking at the data and tinkering with the shape of the prior distribution until the results provide the maximum possible support in favor of the alternative hypothesis. When it comes to the assessment of evidence, data-based tinkering of the prior distribution amounts to

nothing less than statistical cheating. Nevertheless, oracle priors serve a function because they provide an upper bound on the evidence in favor of the alternative hypothesis – the true level of evidence is necessarily less impressive than that obtained by cheating. In particular, the "oracle width prior" cherry-picks the width of a normal distribution to make the evidence in favor of $\mathcal{H}_1$ appear as strong as possible. This unrealistic prior yields $\text{BF}_{10} = 2.51$ – despite cherry-picking the prior width, this evidence is still relatively weak. An absolute upper bound on the evidence can be obtained by using a distribution that is centered as a point on the most likely value (Edwards, Lindman, & Savage, 1963); this "oracle point prior" yields $\text{BF}_{10} = 8.61$, the same as the likelihood ratio from the middle panel of Figure 1.

Other, non-standard prior choices are possible as well. In particular, one may use "non-local" priors that are centered away from zero. Such priors can be selected according to formal rules (Johnson, 2013), constructed from the outcome of previous experiments (Verhagen & Wagenmakers, in press), or be based on subjective considerations (Dienes, 2008). A discussion of such priors would take us too far afield.

In sum, Mike's data are ambiguous – only for the oracle point prior is the Bayes factor higher than 3, and in all other cases the evidence is anecdotal or "not worth more than a bare mention" (Jeffreys, 1961, Appendix B). It is important to stress that, even though different specifications of $\mathcal{H}_1$ lead to different answers, these answers are generally much closer to each other than to the answer one obtains when the existence of $\mathcal{H}_1$ entirely ignored. As argued by Berger and Delampady (1987, p. 330): "(...) formal use of P-values should be abandoned. Almost anything will give a better indication of the evidence provided by the data against $\mathcal{H}_0$."

An in-depth discussion of Bayesian hypothesis testing is beyond the scope of this chapter, but relevant details can be found in Rouder, Speckman, Sun, Morey, and Iverson (2009), Rouder, Morey, Speckman, and Province (2012), Rouder and Morey (2012),

Wetzels and Wagenmakers (2012), Wetzels et al. (2011).

Concluding Comments

By means of several examples, we have tried to demonstrate that the current method for measuring empirical "success" is dangerously lenient. By ignoring the alternative hypothesis, researchers routinely overestimate the evidence against the null hypothesis. An additional factor, one we could not discuss for reasons of brevity, is the a priori plausibility of $\mathcal{H}_0$ versus $\mathcal{H}_1$. It matters whether $\mathcal{H}_1$ is "plants grow better when people water them regularly" or "plants grow better when people pray for them regularly". Equation 1 shows that the same demonstration we gave here regarding the impact of the alternative distribution could have been given regarding prior plausibility.

In the Bayesian framework, the relative prior plausibility of two models is given by the prior model odds, $p(\mathcal{H}_1)/p(\mathcal{H}_0)$. The prior model odds reflect a researcher's skepticism, and they can be used to quantify Carl Sagan's dictum "extraordinary claims require extraordinary evidence".[4] Specifically, one starts with prior model odds $p(\mathcal{H}_1)/p(\mathcal{H}_0)$; these are then updated by means of the Bayes factor $p(y \mid \mathcal{H}_1)/p(y \mid \mathcal{H}_0)$ to yield posterior model odds $p(\mathcal{H}_1 \mid y)/p(\mathcal{H}_0 \mid y)$, which represent the relative plausibility of two models after seeing the data y. The final belief state, therefore, is a compromise between prior skepticism and evidence provided by the data. Hence, implausible claims require more evidence from the data to reach an acceptable level of belief.

Exactly how to quantify initial skepticism is a subjective endeavour, one that most researchers engage in only implicitly. One exception is Lykken (1968), who probed clinicians' opinion about the hypothesis that people with eating disorders are relatively prone to unconsciously believe in the "cloacal theory of birth" (i.e., oral impregnation and anal parturition).[5] Of course, outside academia the quantification of prior beliefs is quite popular, in particular where it concerns betting on outcomes of sports competitions and

election results (Silver, 2012). But the assessment of initial skepticism can be useful even when it defies exact quantification. For instance, when recent experimental work initially suggested that neutrinos can travel faster than the speed of light, Drew Baden –chairman of the physics department at the University of Maryland– compared its plausibility to that of finding a flying carpet. It is difficult to quantify exactly how likely one is to find a flying carpet these days, but it is clear that this initial skepticism is sufficiently large to warrant attention. Similar considerations hold for the existence of extra-sensory perception (Wagenmakers et al., 2011) and the effectiveness of alternative medicine compared to placebo.

A classical statistician may object that we do not know about prior plausibility, or about how to specify a reasonable alternative hypothesis, and that these uncomfortable concepts are therefore best swept under the rug. We believe the classical statistician is wrong on both counts: in most cases, it is possible to say something about prior plausibility and alternative hypotheses –or at least conduct a sensitivity analyses to explore the impact of model assumptions on inference– and it is misleading to ignore key concepts that matter.

But if we assume with the classical statistician that it is possible that a researcher truly has no information on which to build prior expectations, the implications are staggering. This would mean that the researcher has absolutely no predictions about the phenomenon under study. Any data –regardless of how outlandish– would be equally expected by this researcher. An effect size of $1,000,000$ would be equally as surprising as an effect size of 0.5. Raising all observations to the 10th power would yield an equally plausible data set as the one observed. We cannot think of any phenomenon about which so little is known. If such a phenomenon did exist, surely one should not test *any* hypothesis about it, because the meaning of such hypotheses would be questionable. The conditions under which a hypothesis test would be meaningful presuppose the ability to

construct predictions, and hence a reasonable alternative.

In sum, the current crisis of confidence was brought about not only by questionable research practices and related mischief; below the radar, a contributing factor has been the p value statistical analyses that are routinely conducted and generally considered "safe". The logic that underlies p values, however, is fundamentally flawed as it only considers what can be expected under the null hypothesis. To obtain a valid measure of evidence, psychologists have no choice but to turn to methods that are based on a concrete specification of the alternative hypothesis: this may feel uncomfortable at first, but it is the price that needs to be paid for inference that is reliable, honest, and fair.

Glossary

Alternative hypothesis. The alternative hypothesis ($\mathcal{H}_1$ or $\mathcal{H}_a$) refers to the proposition that the effect of interest is present. In classical statistics, this hypothesis is either ignored, or specified as a single point (e.g., $\delta = .25$); in Bayesian statistics, the alternative hypothesis is often composite, covering a range of plausible values (i.e., $\delta \sim N(0, 1)$).

Bayes factor. The Bayes factor (BF_{10}) is an average likelihood ratio that quantifies the extent to which the data change the prior model odds to the posterior model odds. When $\text{BF}_{10} = 10$, the observed data are 10 times more likely under the alternative hypothesis $\mathcal{H}_1$ than under the null hypothesis $\mathcal{H}_0$; when $\text{BF}_{10} = 1/5$ the data are 5 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$.

Fisher's disjunction. According to Fisher, a low p value indicates either that an exceptionally rare event has occurred or that the null hypothesis is false.

Likelihood ratio. The likelihood ratio quantifies the relative plausibility of the observed data y under a specific alternative hypothesis $\mathcal{H}_1$ versus the null hypothesis $\mathcal{H}_0$: $p(y \mid \mathcal{H}_1)/p(y \mid \mathcal{H}_0)$. The likelihood ratio assumes that the alternative hypothesis is

specified by a single point; when the alternative hypothesis is composite, the likelihood ratio turns into a Bayes factor.

Modus tollens. In deductive reasoning, the modus tollens –also known as denying the consequent– is a logically valid syllogistic argument of the following form: "If P, then Q (first premise). Not Q (second premise). Therefore, not P (conclusion)."

Null hypothesis. The null hypothesis ($\mathcal{H}_0$) refers to the proposition that the effect of interest is absent.

p value. In null hypothesis significance testing, the p value is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true and the data were generated according to a known sampling plan.

References

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*, 1550–1577.

Beck–Bornholdt, H.-R., & Dubben, H.-H. (1996). Is the Pope an alien? *Nature*, *381*, 730.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*, 1–32.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.).* Hayward (CA): Institute of Mathematical Statistics.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526–536.

Burdette, W. J., & Gehan, E. A. (1970). *Planning and analysis of clinical studies.* Springfield (IL): Charles C. Thomas.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, *59*, 121–126.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161–172.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *The American Journal of Physics*, *14*, 1–13.

De Groot, A. D. (1956/2014). The meaning of "significance" for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom,

Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica*, *148*, 188–194.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* New York: Palgrave MacMillan.

Edwards, A. W. F. (1996). Is the Pope an alien? *Nature*, *382*, 202.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, *63*, 400–402.

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

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist*, *53*, 798–799.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard : The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75–98.

Fisher, R. A. (1958). *Statistical methods for research workers (13th ed.).* New York: Hafner.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24.

Hagen, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist*, *53*, 801–803.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, *57*, 171–182.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: 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.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 19313–19317.

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

Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science*, *20*, 223–230.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course.* Cambridge University Press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*, 150–152.

Open Science Collaboration, T. (2012). An open, large–scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making type I errors. *Psychological Bulletin*, *102*, 159–163.

Pratt, J. W. (1961). Review of Lehmann, E. L., testing statistical hypotheses. *Journal of the American Statistical Association*, *56*, 163–167.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2014). The $p < .05$ rule and the hidden costs of the free lunch in inference. *Manuscript submitted for publication.*

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.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman & Hall.

Schneider, J. W. (2014). Null hypothesis significance tests: A mix–up of two different theories, the basis for widespread confusion and numerous misinterpretations. *Manuscript submitted for publication.*

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

Silver, N. (2012). *The signal and the noise: The art and science of prediction.* London: Allen Lane.

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.

Verhagen, A. J., & Wagenmakers, E.-J. (in press). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General.*

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.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 627–633.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference.* New York: Springer.

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., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

Author Note

We thank the editors for their constructive comments on an earlier draft. This work was supported by an ERC grant from the European Research Council. Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: EJ.Wagenmakers@gmail.com.

Footnotes

[1]Of course, Mike should have tested more participants. We chose the present numbers because it made Figure 1 more appealing graphically; however, our arguments and examples work for both small and large samples sizes.

[2]A competing statistical paradigm was proposed by Neyman and Pearson. For details on the confusion between the two paradigms see Berger (2003), Christensen (2005), Hubbard and Bayarri (2003). Here we focus on the paradigm proposed by Fisher because it is more closely connected to the everyday practice of experimental psychologists.

[3]The Cauchy distribution is a t distribution with one degree of freedom. Compared to the normal distribution, the Cauchy distribution has fatter tails.

[4]Earlier such statements are due to David Hume and Pierre-Simon Laplace.

[5]The clinicians did not buy it: the prior probability for the hypothesis ranged from 10^{-6} to 0.13, and the median was 0.01.

Prior	BF_{10}	BF_{01}
Cauchy(0, $r = 1$)	1.45	0.69
Cauchy(0, $r = .5$)	1.84	0.54
Normal(0,1)	2.03	0.49
Oracle width prior	2.52	0.40
Oracle point prior	8.61	0.12

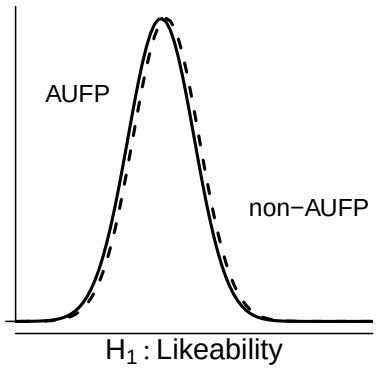
Table 1

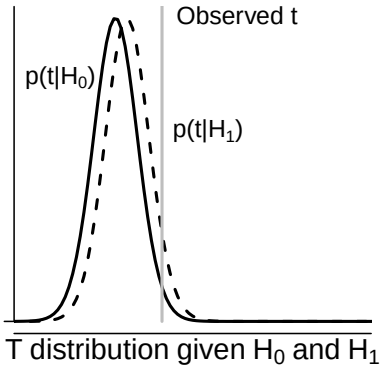
Bayes factors for different priors. $BF_{01} = 1/BF_{10}$.

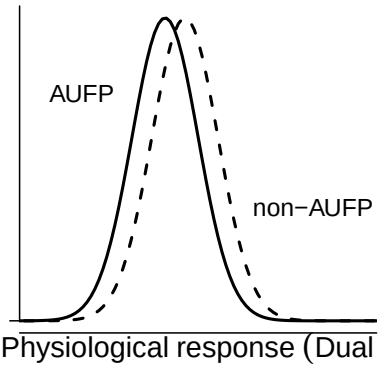
Figure Captions

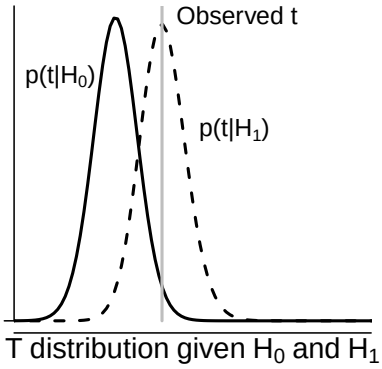
Figure 1. A trio of p values, showing that the diagnosticity of a significant result hinges on the specification of the alternative hypothesis. Top panels: a significant result that is ambiguous; middle panels: a significant result that is moderately informative; bottom panels: a significant result that is evidence in favor of the null hypothesis. The left column shows the population distribution under $\mathcal{H}_1$, and the right column shows the two relevant sampling distributions (i.e., one under $\mathcal{H}_0$, the other under $\mathcal{H}_1$) of the test statistic for the difference between 25 participants viewing AUFP stimuli and 25 participants viewing non-AUFP stimuli.

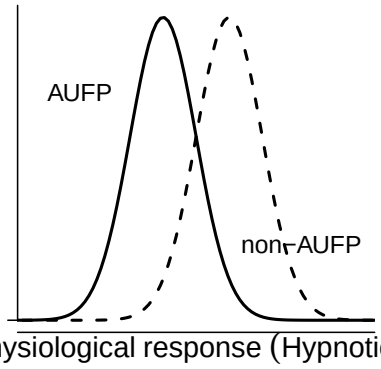
Figure 2. A boxing analogy of the p value. By considering only the state of boxer $\mathcal{H}_0$, the Fisherian referee makes an irrational decision. Figure downloaded from Flickr, courtesy of Dirk-Jan Hoek.

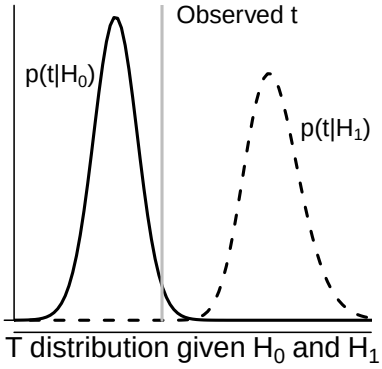

AUFP
non−AUFP
H1 : Likeability

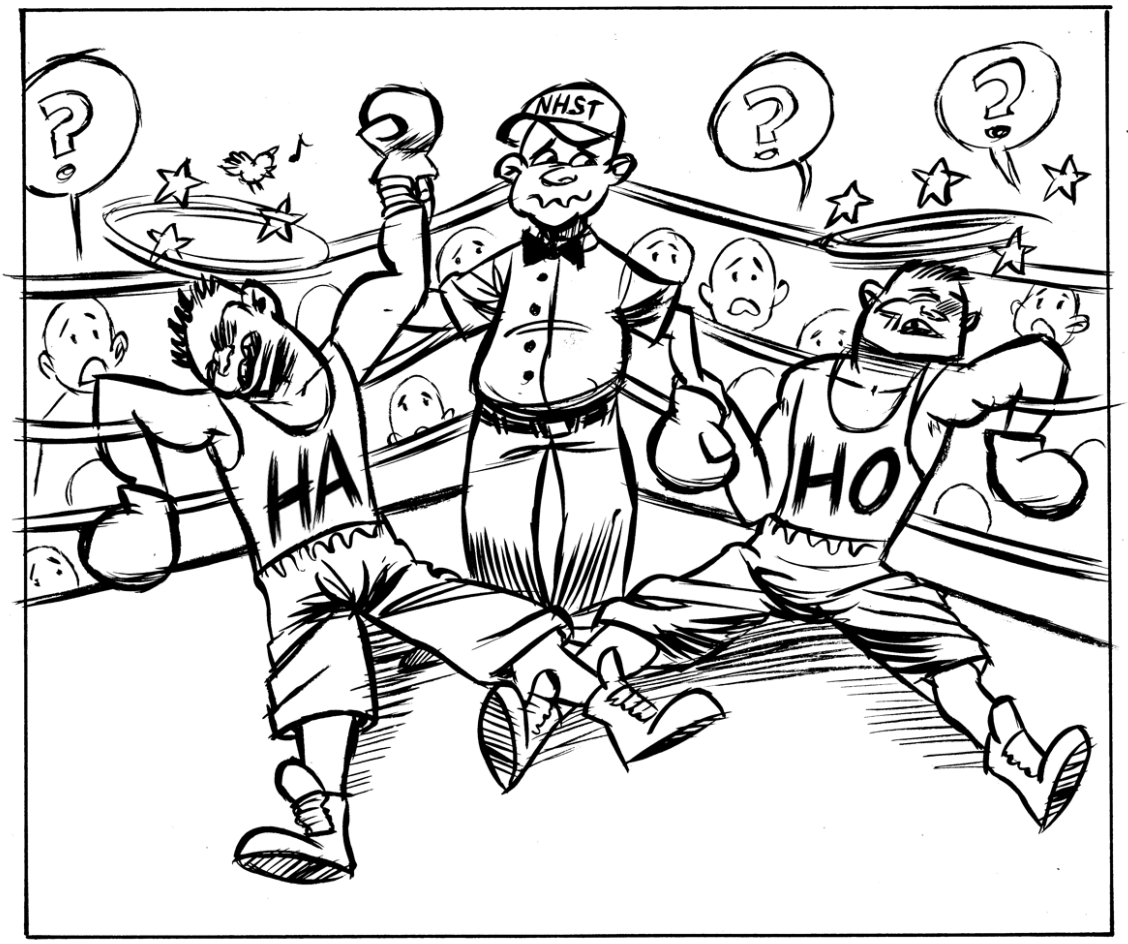

Observed t
p(t|H0)
p(t|H1)
T distribution given H0 and H1


AUFP
non−AUFP
H1 : Physiological response (Dual task)


Observed t
p(t|H0)
p(t|H1)
T distribution given H0 and H1


AUFP
non−AUFP
H1 : Physiological response (Hypnotic state)


Observed t
p(t|H0)
p(t|H1)
T distribution given H0 and H1


?
NHST
?
?
HA
HO