Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests

Ruud Wetzels[1], Dora Matzke[1], Michael D. Lee[2], Jeffrey N. Rouder[3], Geoffrey J. Iverson[2], and Eric-Jan Wagenmakers[1]

1. University of Amsterdam
2. University of California, Irvine
3. University of Missouri

Abstract

Statistical inference in psychology has traditionally relied heavily on significance testing, with practitioners adopting a convention that an effect is established if the p value under the null is sufficiently small. This approach to drawing conclusions from data, however, has been widely criticized, and two types of remedies have been advocated. The first proposal is to supplement p values with additional, complementary measures of evidence such as effect sizes. The second is to replace frequentist inference with Bayesian measures of evidence such as the Bayes factor. We provide a practical comparison of p values, effect sizes, and Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology. Our comparison yields two main results: First, although p values and Bayes factors almost always agree about what hypothesis is better supported by the data, the measures often disagree about the strength of this support; for 70% of the data sets for which the p value falls between 0.01 and 0.05, the Bayes factor indicates that the evidence is only anecdotal. Second, effect sizes can provide additional evidence to p values and Bayes factors. We argue that the Bayesian approach can naturally incorporate inferences about both the presence of effects, as well as their magnitude, and so can encompass all of the types of evidence found in our comparisons in a unified and principled way.

Keywords: Hypothesis testing, t test, p value, effect size, Bayes factor

Introduction

Experimental psychologists use statistical procedures to convince themselves and their peers that the effect of interest is real, reliable, replicable, and hence worthy of academic attention. A suitable example comes from Mussweiler (2006) who assessed whether particular actions can activate a corresponding stereotype. To assess this claim, Mussweiler unobtrusively induced half the participants, the experimental group, to move in a portly manner that is stereotypic for the overweight, while the other half, the control group, made no such movements. Next, all participants were given an ambiguous description of a target person and then used a 9-point scale to rate this person on dimensions that correspond to the overweight stereotype (e.g., "unhealthy", "sluggish", "insecure"). To assess whether making the stereotypic motion affected the rating of the ambiguous target person, Mussweiler computed a t statistic ($t(18) = 2.1$), and found that this value corresponded to a low p value ($p < .05$).[1] Following conventional protocol, Mussweiler concluded that the low p value should be taken to provide "initial support for the hypothesis that engaging in stereotypic movements activates the corresponding stereotype" (Mussweiler, 2006, p. 18).

The use of t tests and corresponding p values in this way constitutes a common and widely accepted practice in the psychological literature. It is, however, not the only possible or reasonable approach to measuring evidence and making statistical and scientific inferences. Indeed, the use of t tests and p values has been widely criticized (e.g., Cohen, 1994; Cumming, 2008; Dixon, 2003; Lee & Wagenmakers, 2005; Loftus, 1996; Nickerson, 2000; Wagenmakers, 2007). There are at least two different criticisms, coming from different perspectives, and resulting in different remedies. On the one hand, many have argued that null hypothesis testing inferences should be supplemented with other statistical measures, including especially confidence intervals and effect sizes. Within psychology, this approach to remediation has sometimes been institutionalized, being required by journal editors or recommended by the APA (e.g., American Psychological Association, 2010; Cohen, 1988; Erdfelder, 2010; Wilkinson & the Task Force on Statistical Inference, 1999).

A second, more fundamental criticism comes from Bayesian statistics and holds that that there are basic conceptual and practical problems with frequentist approaches to estimation and inference. Although Bayesian criticism of psychological statistical practice dates back at least to Edwards, Lindman, and Savage (1963), it has become especially prominent and increasingly influential in the last decade (e.g., Kruschke, 2010; Lee, 2008; Myung, Forster, & Browne, 2000; Rouder, Speckman,

[1] The findings suggest that Mussweiler conducted a one-sided t test. In the remainder of this article we conduct two-sided t tests.

This research was supported by a Vidi grant from the Dutch Organization for Scientific Research (NWO). Correspondence concerning this article may be addressed to Ruud Wetzels, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands.

Sun, Morey, & Iverson, 2009). One standard Bayesian statistic for quantifying the amount of evidence from the data in support of an experimental effect is the *Bayes factor* (Gönen, Johnson, Lu, & Westfall, 2005; Rouder et al., 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). The measure takes the form of an odds ratio: it is the ratio of the evidence that the data provide for the alternative hypothesis relative to that for the null hypothesis (Kass & Raftery, 1995; Lee & Wagenmakers, 2005).

With this background, it seems that psychological statistical practice currently stands at a three-way fork in the road. Staying on the current path means continuing to rely on p values. A modest change is to place greater focus on the additional inferential information provided by effect sizes and confidence intervals. A radical change is struck by moving to Bayesian approaches like Bayes factors. The path that psychological science chooses seems likely to matter. It is not just that there are philosophical differences between the three choices. It is also clear that the three measures of evidence can be mutually inconsistent (e.g., Berger & Sellke, 1987; Rouder et al., 2009; Wagenmakers, 2007; Wagenmakers & Grünwald, 2006; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010).

In this paper, we assess the practical consequences of choosing among inference by p values, effect sizes, and Bayes factors. By practical consequences, we mean the extent to which conclusions of extant studies change according to the inference measure that is used. To assess these practical consequences, we reanalzed 855 t tests reported in articles from the 2007 issues of *Psychonomic Bulletin & Review* (PBR) and *Journal of Experimental Psychology:Learning, Memory and Cognition* (JEP:LMC). For each t test, we compute the p value, the effect size, and the Bayes factor and study the extent to which they provide information that is redundant, complementary, or inconsistent. On the basis of these analyses, we suggest the best direction for measuring statistical evidence from psychological experiments.

Three Measures of Evidence

In this section, we describe how to calculate and interpret the p value, the effect size, and the Bayes factor. For concreteness, we use Mussweiler's study on the effect of action on stereotypes. The mean score of the control group, M_c, was 5.79 on a weight-stereotype scale ($s_c = 0.69, n_c = 10$), and the mean score of the experimental group, M_e, was 6.42 ($s_e = 0.66, n_e = 10$).

The p value

The interpretation of p values is not straightforward, and their use in hypothesis testing is heavily debated (Cohen, 1994; Cortina & Dunlap, 1997; Cumming, 2008; Dixon, 2003; Frick, 1996; Gigerenzer, 1993, 1998; Hagen, 1997; Killeen, 2005, 2006; Kruschke, 2010; Lee & Wagenmakers, 2005; Loftus, 1996; Nickerson, 2000; Schmidt, 1996; Wagenmakers & Grünwald, 2006; Wainer, 1999). The p value is the probability of obtaining a test statistic (in this case the t statistic) at least as extreme as the

one that was observed in the experiment, given that the null hypothesis is true. Fisher (1935) interpreted these p values as evidence against the null hypothesis. The smaller the p value, the more evidence there is against the null hypothesis. Fisher viewed these values as self-explanatory measures of evidence that did not need further guidance. In practice, however, most researchers (and reviewers) adopt a .05 cutoff. Commonly, p values less than .05 constitute evidence for an effect, and those greater than .05 do not. More fine-grained categories are possible, and Wasserman (2004, p. 157) proposes the gradations in Table 1. Note that Table 1 lists various categories of evidence *against* the null hypothesis. A basic limitation of null hypothesis significance testing is that it does not allow a researcher to gather evidence *in favor of* the null (Dennis, Lee, & Kinnell, 2008; Gallistel, 2009; Rouder et al., 2009; Wetzels et al., 2009).

p value			Interpretation
	$<$	0.001	Decisive Evidence Against H_0
0.001	–	0.01	Substantive Evidence Against H_0
0.01	–	0.05	Positive Evidence Against H_0
	$>$	0.05	No Evidence Against H_0

Table 1: Evidence categories for p values, adapted from Wasserman (2004, p. 157).

For the data from Mussweiler, we compute a p value based on the t test. The t test is designed to test if a difference between two means is significant. First, we calculate the t statistic:

$$t = \frac{M_e - M_c}{\sqrt{s^2_{pooled}\left(\frac{1}{n_e} + \frac{1}{n_c}\right)}} = \frac{6.42 - 5.79}{\sqrt{0.46\left(\frac{1}{10} + \frac{1}{10}\right)}} = 2.09,$$

where M_c and M_e are the means of both groups, n_c and n_e are the sample sizes, and s^2_{pooled} is the common population variance:

$$s^2_{pooled} = \frac{(n_e - 1)s^2_e + (n_c - 1)s^2_c}{n_e + n_c - 2}.$$

Next, the t statistic with $n_e + n_c - 2 = 18$ degrees of freedom results in a p value slightly larger than 0.05 (≈ 0.051). For our concrete example, Table 1 leads to the conclusion that the p value is on the cusp between "no evidence" and "positive evidence against H_0".

there is no evidence against the null hypothesis that both groups have an equal mean.

The Effect Size

Effect sizes quantify the magnitude of an effect, serving as a measure of how much the results deviate from the null hypothesis (Cohen, 1988; Thompson, 2002; Richard, Bond, & Stokes-Zoota, 2003; Rosenthal, 1990; Rosenthal & Rubin, 1982). We denote effect size by δ, for the data from Mussweiler, it is calculated as follows:

$$\delta = \frac{M_e - M_c}{s_{pooled}} = \frac{6.42 - 5.79}{0.68} = 0.93.$$

Effect sizes are often interpreted in terms of the categories introduced by Cohen (1988), as listed in Table 2, ranging from "small" to "very large". For our concrete example, $\delta = 0.93$, and we conclude that this effect is large to very large. Interestingly, the p value was on the cusp between the categories "no evidence" and "positive evidence against H_0" whereas the effect size indicates the effect to be strong.

Effect size			Interpretation
	$<$	0.2	Small Effect Size
0.2	–	0.5	Small to Medium Effect Size
0.5	–	0.8	Medium to Large Effect Size
	$>$	0.8	Large to Very Large Effect Size

Table 2: Evidence categories for effect sizes as proposed by Cohen (1988).

The Bayes Factor

In Bayesian statistics, uncertainty (or degree of belief) is quantified by probability distributions over parameters. This makes the Bayesian approach fundamentally different from the frequentist approach, which relies on sampling distributions of data (Berger & Delampady, 1987; Berger & Wolpert, 1988; Lindley, 1972; Jaynes, 2003).

Within the Bayesian framework, one may quantify the evidence for one hypothesis relative to another. The Bayes factor is the most commonly used (although certainly not the only possible) Bayesian measure for doing so (Jeffreys, 1961; Kass & Raftery, 1995). The Bayes factor is the probability of the data under one hypothesis relative to the other. When a hypothesis is a simple point, such as the null, then the probability of the data under this hypothesis is simply the likelihood evaluated at that point. When a hypothesis consists of a range of points, such as all positive effect sizes, then the probability of the data under this hypothesis is the weighted average of the likelihood across that range. This averaging automatically controls for the complexity of different models, as has been emphasized in Bayesian literature in psychology (e.g., Pitt, Myung, & Zhang, 2002).

We take as the null that a parameter α is restricted to 0 (i.e., $H_0 : \alpha = 0$), and take as the alternative that α is not zero (i.e., $H_A : \alpha \neq 0$). In this case, the Bayes

factor given data D is simply the ratio

$$BF_{A0} = \frac{p\left(D \mid H_A\right)}{p\left(D \mid H_0\right)} = \frac{\int p\left(D \mid H_A, \alpha\right) p\left(\alpha \mid H_A\right) \, \mathrm{d}\alpha}{p\left(D \mid H_0\right)},$$

where the integral in the denominator takes the average evidence over all values of α, weighted by the prior probability of those values $p\left(\alpha \mid H_A\right)$ under the alternative hypothesis.

An alternative—but formally equivalent—conceptualization of the Bayes factor is as a measure of the change from prior model odds to posterior model odds, brought about by the observed data. This change is often interpreted as the *weight of evidence* (Good, 1983; Good, 1985), and is given by

$$BF_{A0} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{\text{odds } (H_0 \text{ vs. } H_A \mid D)}{\text{odds } (H_0 \text{ vs. } H_A)}.$$

Under either conceptualization, the Bayes factor has an appealing and direct interpretation as an odds ratio. For example, $BF_{A0} = 2$ implies that the data are twice as likely to have occurred under H_A than under H_0. One way to characterize these ratios is by the labeling proposed by Jeffreys (1961), presented in Table 3.

Bayes factor			Interpretation
	$>$	100	Decisive evidence for H_A
30	–	100	Very Strong evidence for H_A
10	–	30	Strong evidence for H_A
3	–	10	Substantial evidence for H_A
1	–	3	Anecdotal evidence for H_A
	1		No evidence
1/3	–	1	Anecdotal evidence for H_0
1/10	–	1/3	Substantial evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very Strong evidence for H_0
	$<$	1/100	Decisive evidence for H_0

Table 3: Evidence categories for the Bayes factor BF_{A0} (Jeffreys, 1961). We replaced the label "worth no more than a bare mention" with "anecdotal". Note that, in contrast to p values, the Bayes factor can quantify evidence in favor of the null hypothesis.

In general, calculating Bayes factors is more difficult than calculating p values and effect sizes. However, psychologists can now turn to easy-to-use webpages to calculate the Bayes factor for many common experimental situations or use software such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000; Wetzels et al., 2009;

Wetzels, Lee, & Wagenmakers, in press).[2] In this paper, we use the Bayes factor calculation described in Rouder et al. (2009). Rouder et al.'s development is suitable for one- and two-sample designs, and the only necessary input is the t value and sample size.

In our concrete example, the resulting Bayes factor for $t = 2.09$ and a sample size of 20 observations is $BF_{A0} = 1.56$. Accordingly, the data are 1.56 times more likely to have occurred under the alternative hypothesis than under the null hypothesis. This Bayes factor falls into the category "anecdotal". In other words, this Bayes factor indicates that although the alternative hypothesis is slightly favored, we do not have sufficiently strong evidence from the data to reject or accept either hypothesis.

Comparing p values, Effect Sizes and Bayes Factors

For our concrete example, the three measures of evidence are not in agreement. The p value the p value was on the cusp between the categories "no evidence" and "positive evidence against H_0", the effect size indicates a large to very large effect size, and the Bayes factor indicates that the data support the null hypothesis almost as much as they support the alternative hypothesis. If this example is not an isolated one, and the measures differ in many psychological applications, then it is important to understand the nature of those differences.

To address this question, we studied all of the comparisons evaluated by a t test in the Year 2007 volumes of *Psychonomic Bulletin & Review* (PBR) and *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP:LMC). This sample was comprised of 855 comparisons from 252 articles. These articles covered 2394 journal pages, and addressed many topics that are important in modern experimental psychology. Our sample suggests, on average, that an article published in PBR and JEP:LMC contains about 3.4 t tests, which amounts to one t test every 2.8 pages. We describe the empirical relation between the three measures of evidence, starting with the relation between effect sizes and p values.

Comparing Effect Sizes and p values

The relationship between the obtained p values and effect sizes is shown as a scatter plot in Figure 1. Each point corresponds to one of the 855 comparisons. Different panels are introduced to distinguish the different evidence categories, as given in Tables 1 and 2. Comparisons (points) are colored by effect size, ranging from white for small effects to red for large ones.

Figure 1 suggests that p values and effect sizes capture roughly the same information in the data. Large effect sizes tend to correspond to low p values, and small effect sizes tend to correspond to large p values.The two measures, however, are far

[2] A webpage for computing a Bayes factor online is `http://pcl.missouri.edu/bayesfactor` and a webpage to download a tutorial and a flexible R/WinBUGS function to calculate the Bayes factor can be found at `www.ruudwetzels.com`.

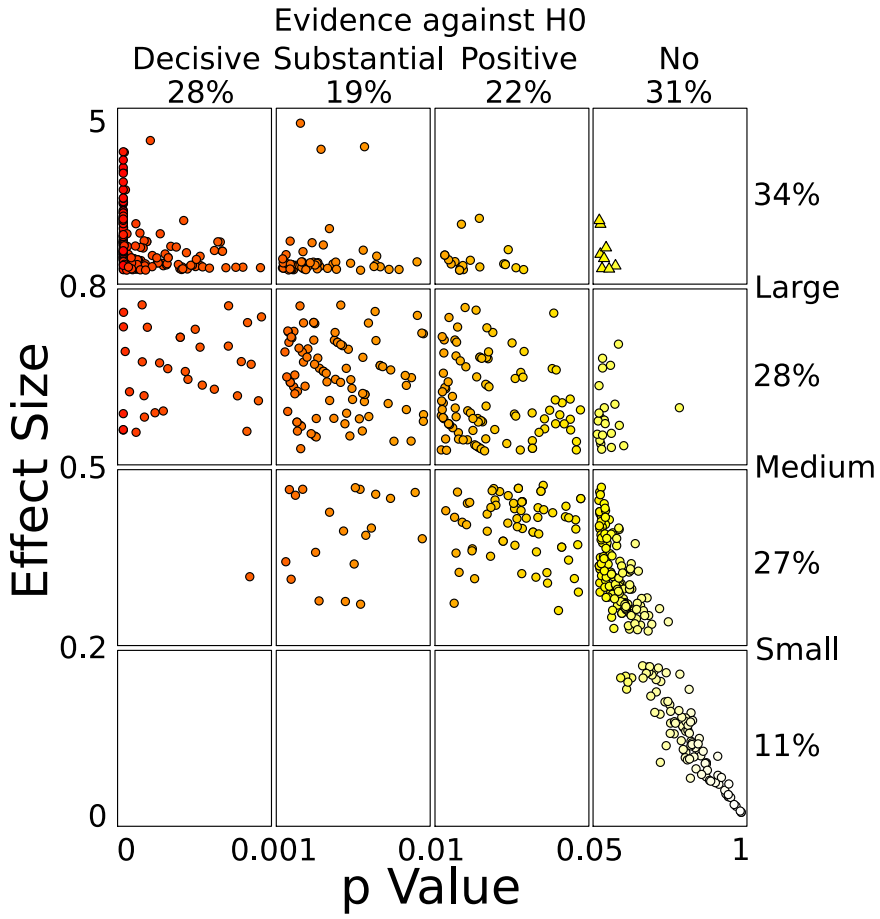

Figure 1. The relationship between effect size and p values. Points denote comparisons (855 in total), and are colored according to effect size, ranging from red for large effects to white for small effects. Points denoted by circle indicate relative consistency between the effect size and p value, while those denoted by triangles indicate gross inconsistency.

from identical. For instance, a p value of 0.01 can correspond to effect sizes ranging from about 0.2 to 1, and an effect size near 0.5 can correspond to p values ranging from about 0.001 to 0.05. The triangular points in the top-right panel of Figure 1 highlight gross inconsistencies. These 8 studies have a large effect size, above 0.8, but their p values do not indicate evidence against the null hypothesis. A closer examination revealed that these studies had p values very close to 0.05, and were comprised of small sample sizes.

Comparing Effect Sizes and Bayes Factors

The relationship between the obtained Bayes factors and effect sizes is shown in Figure 2. Points (comparisons) are colored according to their Bayes factor, ranging

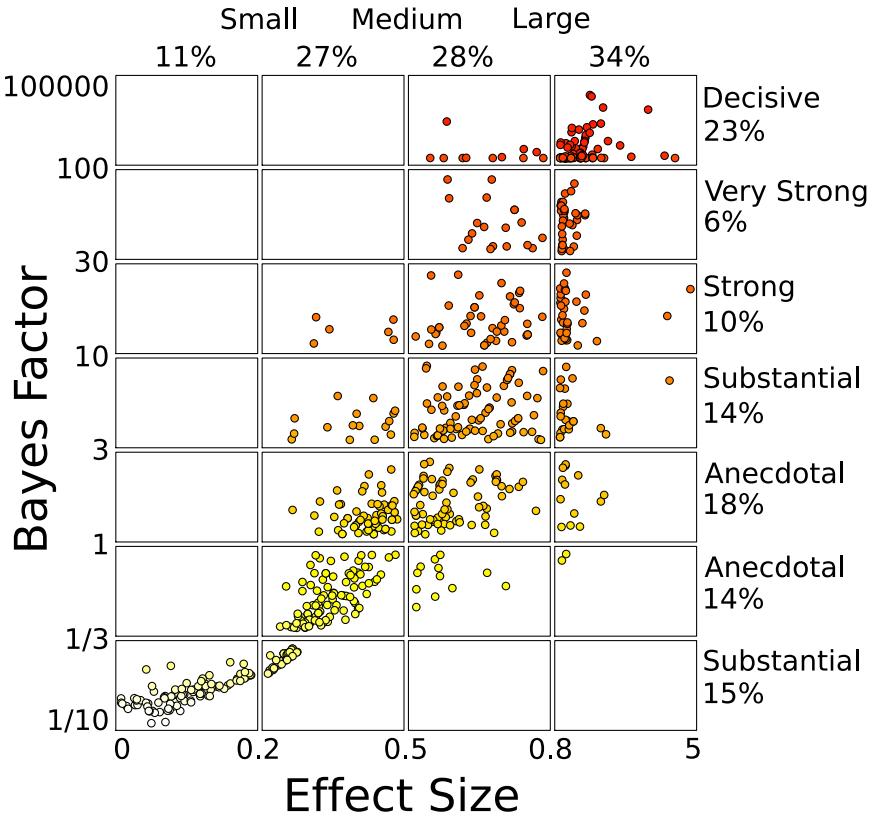

Figure 2. The relationship between Bayes factor and effect size. Points denote comparisons (855 in total), and are colored according to the Bayes factor, ranging from red for decisive evidence in favor of H_A, to white for substantial evidence in favor of H_0.

from red for decisive evidence in favor of the alternative hypothesis to white for decisive evidence in favor of the null hypothesis.

Much as with the comparison of p values with effect sizes, it seems clear that Bayes factors and effect sizes are in general, but not exact, in agreement with one another —there are no striking inconsistencies. No study with an effect size greater than 0.8 coincides with a Bayes factor below 1/3, nor does a study with very low effect size below 0.2 coincide with a Bayes factor above 3. The two measures, however, are not identical. They differ in the assessment of strength of evidence. Effect sizes above 0.8 range all the way from anecdotal to decisive evidence in terms of the Bayes factor. Also note that small to medium effect sizes (i.e., those between 0.2 and 0.5) can correspond to Bayes factor evidence in favor of either the alternative or the null hypothesis.

This last observation highlights that Bayes factors may quantify support for the null hypothesis. Figure 2 shows that about one-third of all studies produced evidence in favor of the null hypothesis. In about half of these studies favoring the null, the

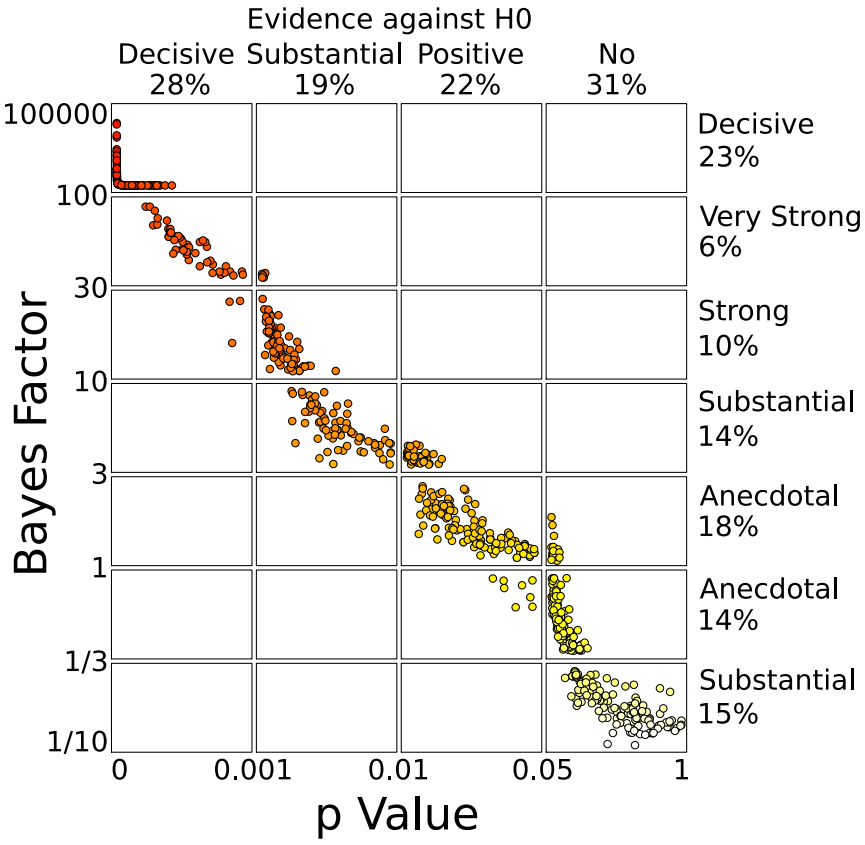

Figure 3. The relationship between Bayes factor and p value. Points denote comparisons (855 in total), and are colored according to the Bayes factor, ranging from red for decisive evidence in favor of H_A, to white for substantial evidence in favor of H_0.

evidence is substantial.

Comparing p values and Bayes Factors

The relationship between the obtained Bayes factors and p values is shown in Figure 3. Interpretative panels are again used, and the coloring again corresponds to Bayes factor values.

It is clear that Bayes factors and p values largely covary with each other. Low Bayes factors correspond to high p values and high Bayes factors correspond to low p values, a relationship that is much more exact than for our previous two comparisons. The main difference between Bayes factors and p values is one of calibration; p values accord more evidence against the null than do Bayes factors. Consider the p values between .01 and .05, values that correspond to "positive evidence" and that usually pass the bar for publishing in academia. According to the Bayes factor, 70% of these comparisons display merely "anecdotal" evidence in favor of the alternative

hypothesis. For p values that are smaller than 0.01, the Bayes factor always implies at least "substantial" evidence in favor of the alternative hypothesis. This suggests that, for the studies we surveyed, a significance level of 0.01 might be more appropriate if the goal is to identify when there is substantial evidence for the alternative hypothesis.

Conclusions

We compared p values, effect sizes and Bayes factors as measures of statistical evidence in empirical psychological research. Our comparison was based on a total of 855 different t statistics from all published articles in two major empirical journals in 2007. In virtually all studies, the three different measures of evidence are broadly consistent. Small p values correspond to large effect sizes and large Bayes factors in favor of the alternative hypothesis. We noted, however, that p values between 0.01 and 0.05 often correspond to what is only anecdotal evidence in favor of the alternative hypothesis; this suggests that—for the studies under consideration here—a p value criterion more conservative than 0.05 is appropriate. Such a criterion will prevent researchers from overestimating the strength of their findings.

So, is there any reason to prefer one measure of evidence over the others? It is easy to make a theoretical case for Bayesian statistical inference in general, based on arguments already well documented in statistics and psychology (e.g., Jaynes, 2003; Lee & Wagenmakers, 2005; Lindley, 1972; Wagenmakers, 2007). Unlike null hypothesis testing, Bayesian inference does not violate basic principles of rational statistical decision-making such as the stopping rule principle or the likelihood principle (Berger & Wolpert, 1988; Berger & Delampady, 1987). In addition, Bayesian inference takes model complexity into account in a rational way. More specifically, the Bayes factor has the attraction of not assigning a special status to the null hypothesis, and so makes it theoretically possible to measure evidence in favor of the null (e.g., Dennis et al., 2008; Kass & Raftery, 1995; Gallistel, 2009; Rouder et al., 2009).

We believe, however, that our results also provide practical encouragement for using the Bayes factor. There is a sense in which p values and effect sizes answer naturally complementary questions, about, "whether" there is an effect, and "how large" an effect is, respectively. The Bayes factor naturally combines answers to both questions in its assessment of evidence. As we have presented them, Bayes factors larger than 1 favor the alternative hypothesis, and Bayes factors less than 1 favor the null hypothesis. In addition, the magnitude of the Bayes factor quantifies the strength of the evidence, on the familiar and readily interpreted likelihood ratio scale.

Despite the Bayes factor naturally quantifying the strength of an effect in this way, our comparisons show that effect sizes continue to carry some additional information. This seems to occur, looking at individual studies, when small numbers of participants or trials are involved. So, especially for these sorts of studies, there is an argument for reporting both a Bayes factor and an effect size. We note that, from a Bayesian perspective, the effect size can naturally be conceived as a (summary statistic of) the posterior distribution of a parameter representing the effect, under an

uninformative prior distribution. In this sense, a standard Bayesian combination of parameter estimation and model selection could encompass all of the useful measures of evidence we observed.

Our final thought is that reasons for adopting a Bayesian approach now are amplified by the promise of using an extended Bayesian approach in the future. In particular, we think the hierarchical Bayesian approach, which is standard in statistics (e.g. Gelman & Hill, 2007), and is becoming more common in psychology (e.g. Kruschke, 2010; Lee, in press; Rouder & Lu, 2005) could fundamentally change how psychologists identify effects. In a hierarchical Bayesian analysis, multiple studies can be integrated, so that what is inferred about the existence of effects and their magnitude is informed, in a coherent and quantitative way, by a domain of experiments.

Most fundamentally, a complete Bayesian approach would allow for informative priors on effect sizes and prior odds (Vanpaemel, in press), so that inference can use what is already known, and not require researchers to start from scratch to establish an effect each time they conduct an experiment. The accumulation of empirical evidence is a natural conception of scientific progress, and one that can be handled in the hierarchical Bayesian statistical approach. We think it will eventually be adoped as a powerful tool in the psychological sciences. In the meantime, using Bayes factors to evaluate hypotheses, together with inferences about effect sizes, is a natural stepping-stone towards the fully Bayesian goal.

References

American Psychological Association. (2010). Publication manual of the American psychological association (6th ed.). *Washington, DC, American Psychological Association*.

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

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*, 112–139.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

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.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286-300.

Dennis, S., Lee, M., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, *59*, 361–376.

Dixon, P. (2003). The p–value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, *57*, 189–202.

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

Erdfelder, E. (2010). A note on statistical analysis. *Experimental Psychology*, *57*, 1–4.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.

Gallistel, C. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale (NJ): Erlbaum.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two–sample t test. *The American Statistician*, *59*, 252–257.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York: Elsevier.

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

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Killeen, P. R. (2005). An alternative to null–hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, *13*, 549–562.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.

Lee, M. D. (in press). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia (PA): SIAM.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Mussweiler, T. (2006). Doing is for thinking! *Psychological Science*, *17*, 17–21.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology*, *44*.

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

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*, 775–777.

Rosenthal, R., & Rubin, D. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of educational psychology*, *74*, 166–169.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

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.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25–32.

Vanpaemel, W. (in press). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*.

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

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, *17*, 641–642.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. P. P. P. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive psychology*, *60*, 158–189.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212–213.

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

Wetzels, R., Lee, M., & Wagenmakers, E.-J. (in press). Bayesian inference using WBDev: A tutorial for social scientists. *Behavior Research Methods*.

Wetzels, R., Raaijmakers, J., 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.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.