A Bayes factor meta-analysis of Bem's ESP claim

Jeffrey N. Rouder · Richard D. Morey

Published online: 15 May 2011

Abstract In recent years, statisticians and psychologists have provided the critique that *p*-values do not capture the evidence afforded by data and are, consequently, ill suited for analysis in scientific endeavors. The issue is particular salient in the assessment of the recent evidence provided for ESP by Bem (2011) in the mainstream *Journal of Personality and Social Psychology.* Wagenmakers, Wetzels, Borsboom, and van der Maas (*Journal of Personality and Social Psychology, 100*, 426–432, 2011) have provided an alternative Bayes factor assessment of Bem's data, but their assessment was limited to examining each experiment in isolation. We show here that the variant of the Bayes factor employed by Wagenmakers et al. is inappropriate for making assessments across multiple experiments, and cannot be used to gain an accurate assessment of the total evidence in Bem's data. We develop a meta-analytic Bayes factor that describes how researchers should update their prior beliefs about the odds of hypotheses in light of data across several experiments. We find that the evidence that people can feel the future with neutral and erotic stimuli to be slight, with Bayes factors of 3.23 and 1.57, respectively. There is some evidence, however, for the hypothesis that people can feel the future with emotionally valenced nonerotic stimuli, with a Bayes factor of about 40. Although this value is certainly noteworthy, we believe it is orders of magnitude lower than what is required to overcome appropriate skepticism of ESP.

J. N. Rouder (✉)
University of Missouri,
Columbia, MO, USA
e-mail: RouderJ@missouri.edu

R. D. Morey
University of Groningen,
Groningen, The Netherlands

Keywords Statistics · Statistical inference · ESP

Bem (2011) has claimed that people can feel or sense salient events in the future that could not otherwise be anticipated. For example, in his Experiment 2, Bem presented participants with two rather ordinary pictures and asked them to indicate which one would be chosen subsequently by a random number generator. If a participant correctly anticipated the random choice, he or she was rewarded with a brief display of a positively valenced picture. Conversely, if a participant incorrectly anticipated the random choice, he or she was punished with a negatively valenced picture. Bem claimed that people could indeed feel these future reward and punishment events and, consequently, were able to anticipate the random choice at a rate deemed statistically above chance. Bem presented a sequence of similar experiments and results and, on this basis, concluded that people can feel the future. This phenomenon and others like it in which people can show seemingly impossible awareness of events are termed *psi phenomena*, or, more colloquially, *extrasensory perception* (ESP).

If ESP is substantiated, it would be among the most important findings in the history of psychology. The existence of ESP would force us to revise not only our theories of psychology, but also those of biology and physics. In our view, when seemingly implausible claims are made with conventional methods, it provides an ideal moment to reexamine these methods. The conventional approach used by Bem (2011) has two properties. First, as is typical in many empirical investigations, Bem presented a sequence of experiments, each targeting the same basic phenomena from a slightly different angle. Second, Bem employed null-hypothesis significance testing in which *p*-values are reported as evidence and evaluated against a fixed criterion to reach judgments. In previous work, we joined a growing consensus that conventional inference by significance testing overstates the evidence for an effect (see Berger & Sellke, 1987; Edwards, Lindman, & Savage, 1963; Wagenmakers, 2007, among several others), and proposed a Bayes factor replacement for the *t*-test (Rouder, Speckman, Sun, Morey & Iverson, 2009).This Bayes factor quantifies the evidence in data for competing hypotheses from a single experiment or, more precisely, for a single comparison. Unfortunately, while this Bayes factor is appropriate for assessing evidence for a single contrast, it is ill suited for meta-analytically combining evidence across several experiments. Herein, we develop a meta-analytic version of the Bayes factor *t*-test and use it to assess the evidence across Bem's experiments. We find some support for ESP; the probability of the combined data are 40 times more likely under an ESP alternative than under a no-ESP null. This evaluation differs from that of Bem, who, in our opinion, overstated the evidence. It also differs from that of Wagenmakers, Wetzels, Borsboom and van der Maas (2011), who found no support for ESP. Our interpretation of this Bayes factor is that while it is noteworthy, it is insufficient in magnitude to sway the beliefs of an appropriately skeptical reader.

The evidence from *p*-values and Bayes factor

There is a well-known asymmetry in significance testing: Researchers can reject the null hypothesis but can never accept it. This asymmetry works against the goals of scientific inquiry, because null hypotheses often correspond to theoretically useful statements of invariance and constraint (Gallistel, 2009; Kass, 1992; Rouder et al., 2009). For Bem's (2011) case, the null hypothesis is the theoretically attractive, reasonable, and highly interpretable constraint that ESP does not exist. In order to fairly assess the evidence for ESP, it is necessary to be able to state the evidence for or against the null provided by the data. Yet, with significance testing, we may only accept ESP and never reject it.

The above point about asymmetry is easy to grasp. Its implications, however, are subtle and consequential, because they extend beyond not being able to state evidence for the null hypothesis; they extend to assessing evidence in the data for the alternative as well. A good starting point is consideration of the distribution of *p*-values under two competing hypotheses (examples are shown in Fig. 1A). If

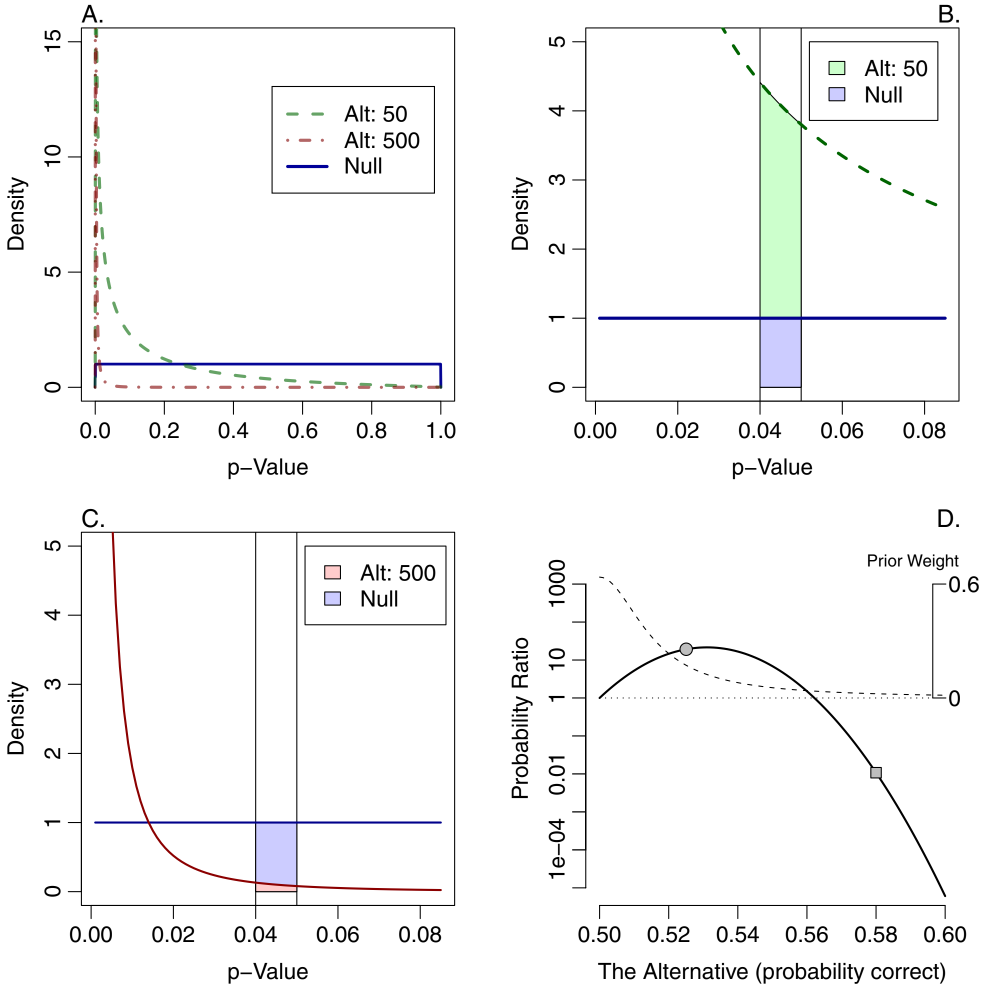

Fig. 1 Significance tests overstate the evidence against the null hypothesis. **A** Distribution of *p*-values for an alternative with effect-size of .2 (dashed and dashed-dotted lines are for sample sizes of 50 and 500, respectively) and for the null (solid lines). **B** Probability of observing a *p*-value between .04 and .05 for the alternative (effect size = .2, N = 50) and for the null. The probability favors the alternative by a ratio of about 4:1. **C** Probability of observing a *p*-value between .04 and .05 for the alternative (effect size = .2, N = 500) and for the null. The probability favors the null by a factor of 10. **D** The solid line is the probability of observing a *t*-value of 2.51 (p = .007) for $N = 100$ under the alternative, relative to that under the null, as a function of the alternative. The circle and square points highlight the ratios that favor the alternative and null, respectively. The dashed line shows the one-tailed prior distribution used throughout

the null hypothesis is false, p-values tend to be small, and they decrease as sample size is increased. The dashed green line shows the distribution of p-values when the underlying effect size is .2 and the sample size is 50; the dashed-dotted red line shows the same when the sample size is increased to 500. The distribution of p-values under the null, however, is quite different. Under the null, all p-values are equally likely (solid blue line in Fig. 1A). Perhaps surprisingly, this distribution holds regardless of sample size; p-values do not increase under the null as sample sizes increase.

The logic behind significance testing is a form of argument by contradiction. If observed data (or data more extreme) are improbable under the null, then the null is contradicted, and presumably, there is some alternative under which the data are more probable. It is reasonable to ask, then, about the factor by which the observed data are more probable under some alternative than under the null. This factor serves as a measure of evidence for the alternative, relative to the null. Suppose that a data set with sample size of 50 yields a p-value in the interval between .04 and .05. Figure 1b shows the distributions of p-values for the null and the alternative (effect size = .2) around this interval, and the probabilities are the shaded areas under the curve. The probability of observing a p-value under the null and alternative is .01 and .04, respectively. Therefore, the alternative fares four times better than the null. Although such a ratio constitutes evidence for the alternative, it is not as substantial as might be inferred by such a small p-value.

Figure 1c shows a similar plot for the null and alternative (effect size = .2) for a large sample size of 500. For this effect size and sample size, very small p-values are the norm. In fact, a p-value between .04 and .05 is about 10 times more likely under the null than under the alternative. In fact, a p-value at any one point—say .05—constitutes increasing evidence for the null in the large sample size limit. This paradoxical behavior of significance testing in which researchers reject the null even though the evidence overwhelmingly favors it is known as *Lindley's paradox* (Lindley, 1957) and is a primary critique of inference by p-values in the statistical literature.

We can examine the evidence from Bem's (2011) data for various alternatives, relative to the null. In Experiment 1, for example, participants needed to anticipate which of two erotic pictures they would be shown. The average performance across 100 naive subjects was .531, and this level was significantly different from the at-chance baseline of .5, $t(99) = 2.51$, $p = .007$. Figure 1d shows the evidence for various alternatives. The probability ratios on the y-axis are the probability of the observed p-value under a specific alternative, relative to that under the null. Not surprisingly, these ratios vary greatly with the choice of alternative. Alternatives that are very near the null of .5—say, .525—are preferred over the null (filled circle in Fig. 1D). Alternatives further from .5—say .58 (filled square)—are definitely not preferred over the null. Note that even though the null is rejected at $p = .007$, there is only a small range of alternatives where the probability ratio exceeds 10, and for no alternative does it exceed 25, much less 100 (as might naïvely be inferred from the p-value). We see that the null may be rejected by p-values even when the evidence for every specific point alternative is more modest.

The probability ratio in Fig. 1D may be denoted by B and expressed as follows:

$$B = \frac{Pr(\text{Data}|H_1)}{Pr(\text{Data}|H_0)},$$

where H_0 is the null and H_1 is that the alternative is that true performance is a specific value—for example, .52. In Bayesian statistics, probability ratios B are called *Bayes factors*, and they are well-calibrated measures of evidence from the data for one hypothesis relative to another. One drawback of the preceding formulation, however, is that the alternative is a single point hypothesis. In Bayesian statistics, it is possible and desirable to consider composite hypotheses in which parameters range over many possible values. To consider composite hypotheses, the analyst specifies how each single value should be weighted. Figure 1d shows such weights (dashed lines), and for this alternative hypothesis, small effects are weighted more than large ones. The distribution of weights over parameters is called the *prior distribution*. When an alternative H_1 consists of a weighted range of parameter values, the probability of the data is

$$Pr(\text{Data}|H_1) = \int Pr(\text{Data}|\theta)f(\theta)d\theta,$$

where θ are the parameters and f is the prior distribution on these parameters. The probability of the data given the hypothesis is the expected or weighted averaged probability across the possible parameter values. The Bayes factor for a composite versus a point null is

$$B = \frac{Pr(\text{data}|H_1)}{Pr(\text{data}|H_0)} = \frac{\int Pr(\text{data}|\theta)f(\theta)d\theta}{Pr(\text{data}|\theta = \theta_0)},$$

where θ_0 is the value of θ under the null, or .5 for Fig. 1D. Figure 1d also shows an example of a prior over the parameter (dashed line), and for this prior, the Bayes factor evidence for the observed p-value is 3.23; that is, the observed level of performance is about 3 times more probable under the alternative than under the null.

To compute Bayes factors, researchers must choose the prior distribution f. Fortunately, there is ample guidance in the literature about how to do so for the linear models, including the t-test (Gönen, Johnson, Lu, & Westfall, 2005;

Liang, Paulo, Molina, Clyde, & Berger, 2008; Zellner, 1986; Zellner & Siow, 1980). We advocate a prior that serves as a generic default broadly applicable for scientific use. This prior was proposed by Jeffreys (1961), was developed for linear models by Zellner and Siow, among several others, and was termed the *JZS prior* by Bayarri and Garcia-Donato (2007). The JZS prior, along with the resulting *JZS Bayes factor*, are presented in the Appendix. The JZS Bayes factor has a number of advantages: It makes intuitive sense, it has beneficial theoretical properties,[1] it is not dependent on the measurement scale of the dependent variable, and it can be conveniently computed.[2] Further details are provided in Rouder et al. (2009).

The Bayes factor measure of evidence is the probability ratio of data given hypotheses. A related quantity of interest is the probability ratio of hypotheses given data, called the *posterior odds*. The posterior odds describe the analyst's degree of belief in the hypotheses after observing the data. The following equation describes the relationship between posterior odds and the Bayes factor:

$$\frac{Pr(H_1|\text{data})}{Pr(H_0|\text{data})} = B \times \frac{Pr(H_1)}{Pr(H_0)},$$

where the terms $\frac{Pr(H_1|\text{data})}{Pr(H_0|\text{data})}$ and $\frac{Pr(H_1)}{Pr(H_0)}$ are posterior and prior odds, respectively. The prior odds describe the beliefs about the relative plausibility of hypotheses before the data are observed, and the Bayes factor describes how the evidence from the data should affect beliefs. For example, suppose the evidence from a set of ESP experiments yielded a Bayes factor of 40 in favor of ESP. Consider a skeptical reader with prior odds of a 1,000,000:1 against ESP. In this case, the reader should revise their beliefs by a factor of 40, to 25,000:1 against ESP. Likewise, a reader that has prior odds favoring ESP should multiply these odds by 40 in light of the data to reach an even more favorable posterior odds. Bayes factors are logically independent of prior odds and, consequently, are ideal for scientific communication (Jeffreys, 1961). We recommend that researchers report Bayes factors and that readers use the context of prior knowledge, such as knowledge about physical laws or plausible mechanisms, to set prior odds in interpreting these Bayes factors.

Wagenmakers et al.'s (2011) analysis of ESP

Table 1 shows the 10 contrasts originally reported by Bem (2011) and reanalyzed by Wagenmakers et al. (2011). Wagenamakers et al. computed *two-tailed JZS Bayes factors*, and some contrasts yielded modest support for the no-ESP null, while others yielded modest support for the ESP alternative. On balance, according to Wagenmakers et al., there is little systematic evidence for ESP. We have added a third column as a validity check, and it provides the direction of the effect. In several of Bem's experiments, one could be reasonably sure that if ESP held, the effect should be in one direction and not the other. For example, in Bem's Experiment 1, discussed previously, participants were instructed to indicate the curtain behind which there was an erotic picture, and, if ESP held, their performance should be greater rather than worse than chance. If there were no ESP, we would expect the observed performance to be slightly below chance for some experiments and slightly above chance for others. Table 1 shows that the direction of all 10 were in the direction hypothesized by Bem. This concordance serves as evidence for ESP that is not captured by Wagenmakers et al.'s analysis. In fact, the Bayes factor of getting all 10 contrasts to be in the same direction is about 100:1 in favor of ESP.[3] This inconsistency motivates our development of a meta-analytic Bayes factor.

The meta-analysis problem

Meta analysis seems like it should be a strong point of the Bayes factor. If one has several replicate experiments, it seems reasonable that the posterior odds from the first can serve as the prior for the second, and so on. Under this framework, the combined evidence across all the replicate experiments is simply the product of the Bayes factors. This intuition that the meta-analytic Bayes factor is the product of individual Bayes factors is not correct, and Table 2 provides an example of how it fails. The first four rows show the results of four replicate experiments, each of sample size 100. The data are independently and identically normally distributed observations with a mean of .2 and a variance of 1.0. Hence, the true effect size is .2, and the observed effect sizes in the replicate experiments vary reasonably around this true value. The corresponding Bayes factors for the replicate experiments are shown, and these indicate that the evidence in each experiment is marginal, with one sample favoring the alternative and the other three favoring the null. The product of these Bayes factors is also

[1] The theoretical properties of the JZS Bayes factor are as follows. First, the Bayes factor is always finite for finite data. Second, the Bayes factor is consistent; as sample size is increased, B grows to infinity if the null is false and shrinks to zero if it is true. This consistency may be contrasted with p-values, which do not converge in the limit when the null is true (see Fig. 1). Finally, for any sample size, the Bayes factor grows to infinity as t grows to infinity.

[2] Web applets to compute Bayes factors for paired and grouped t-tests may be found at pcl.missouri.edu/bayesfactor.

[3] We assume that the direction of each experiment is distributed as a Bernoulli trial. Under the no-ESP null, the probability parameter p = .5; under the ESP alternative, p is distributed as a uniform between 0 and 1 (see Wagenmakers, 2007, for details).

Table 1 Wagenmakers et al. (2011) assessment of Bem's evidence

Experiment	Bayes Factor (Alt/Null)	Direction Predicted
1	1.64	Yes
2	1.05	Yes
3	1.82	Yes
4	.58	Yes
5	.88	Yes
6	.32	Yes
6	.30	Yes
7	.13	Yes
8	.47	Yes
9	5.9	Yes

Note. According to Bem (2011), the direction of each contrast supported ESP

shown (B = .092), and it indicates that the null is preferred with evidence slightly larger than 10:1. The row labeled "Data pooled" shows the results of pooling the data, rather than multiplying the Bayes factor. In this case, where the data are drawn from a common distribution, pooling is valid and preferred. The resulting Bayes factor is 54:1 in favor of an effect. Hence, multiplying JZS Bayes factors is not a valid meta-analytic approach.

This seeming contradiction comes about because JZS Bayes factors respect the resolution of data (Rouder et al., 2009). When the sample size is small, small effects may be considered evidence for the null because the null is the more parsimonious description given the resolution provided by the data. As the sample size grows, however, the resolution provided for the data is finer, and small effects are more concordant with the alternative. An appropriate analogy may be a criminal court trial in which each of several witnesses provides only partial information as to the guilt of a defendant who has committed a crime. If the jury is forced to assess the odds after hearing the testimony of any single witness, these odds may all favor innocence, since no one witness may be compelling enough in isolation to provide evidence for guilt. However, if the jury considers the totality of all testimonies, the weight will assuredly shift toward guilt.

Fortunately, a meta-analytic extension of the JZS Bayes factor is tractable and convenient. One of the key properties of the JZS priors is that the full influence of the data is captured by the t-statistic. Under the JZS priors, we may think of t-statistic as a single piece of datum and the parameter of interest as the *effect size*, δ. Under the null, the effect size is constrained to zero; Under the alternative, it follows a fat-tailed distribution. The resulting Bayes factor is

$$B = \frac{Pr(t|H_1)}{Pr(t|H_0)} = \frac{\int Pr(t|\delta)f(\delta)d\delta}{Pr(t|\delta = 0)},$$

where expressions for the probabilities and prior f is provided in the Appendix. The generalization to M independent experiments, each with t-values $t_1, t_2, \ldots, t_M$ is given by

$$B = \frac{\int \prod_{m=1}^{M} Pr(t_m|\delta)f(\delta)d\delta}{\prod_{i=1}^{M} Pr(t_m|\delta = 0)}, \qquad (1)$$

where $\prod$ indicates the product of terms. The key property of this meta-analytic approach is that the true effect size is assumed to be constant across each experiment. Although the meta-analytic Bayes factor assumes common true effect size across experiments, it does not assume a common variance. Hence, it is applicable to experiments where the unit of measure may vary, such as those that span accuracy and response time effects. A script in **R** program for computing this Bayes factor may be obtained from the authors.

It is reasonable to wonder whether the constant effect size model underlying the meta-analytic Bayes factor is warranted. We chose this approach because it is tractable when researchers have access to the test statistics, rather than the raw data. Alternative models that posit variation in effect size across experiments are possible (Utts, Norris, Suess, & Johnson, 2010), although analysis may require access to the raw data. These variable effect-size alternatives are certainly more complex than the constant effect-size model, and if the true effects are about the same size, it may be at a competitive disadvantage. Whereas Bem (2011) reports near constant effect sizes across the experiments, we believe that the constant effect size model is a convenient and appropriate alternative to the null model.

To illustrate this meta-analytic Bayes factor, we applied it to the four replicate experiments in Table 2. The value is about 49:1 in favor an effect, which is quite close to the value of 54 from pooling the data. The reason these values differ slightly is that the meta-analytic Bayes factor posits a separate variance (σ^2) for each experiment, while the JZS Bayes factor on pooled data assumes a common, single variance.

Table 2 JZS Bayes factor across four replicate experiments

	N	$\hat{\delta}$	t	B
Experiment 1	100	.18	2.16	0.75
Experiment 2	100	.12	1.25	0.17
Experiment 3	100	.29	2.80	3.29
Experiment 4	100	.14	1.44	0.22
Data pooled	400	.18	3.83	54.1
Product of Bayes factors				0.092
Meta-analytic Bayes factor				49

The evidence in Bem's (2011) data

Bem (2011) provided 10 contrasts from nine separate experiments to support the claim of ESP. The contrasts chosen, however, strike us as too opportunistic. For example, in his Experiments 8 and 9, Bem found a positive result for ESP with neutral stimuli and entered the corresponding *t*-value into his final tally in his Table 7 . In Experiment 1, Bem found a positive result for ESP with erotic stimuli (accuracy of .53 vs. .50 baseline) but a null result for neutral stimuli or emotionally evocative stimuli (accuracy of .49 vs. .50 baseline). Bem entered only this positive result with erotic stimuli into his final tally. In our view, tallying the positive results without the null result is not justified. One way of improving the assessment is to evaluate the evidence for neutral, emotionally evocative, and erotic stimuli separately, such that conflicting results can be contrasted. The corresponding *t*-values, sample sizes, and resulting meta-analytic Bayes factors for these three classes of stimuli are shown in Table 3.

We have not included results from Bem's (2011) Experiments 5, 6, and 7 in our meta-analysis because we are unconvinced that these are interpretable. These three experiments are retroactive mere-exposure effect experiments in which the influence of future events purportedly affects the current preference for items. The main difficulty in interpreting these experiments is forming an expectation about the direction of an effect, and this difficulty has consequential ramifications. In the vast majority of conventional mere-exposure effect studies, participants prefer previously viewed stimuli (Bornstein, 1989). Bem observed this pattern for negative stimuli, but the opposite pattern, novelty preference, for positive stimuli. Bem claimed that this crossover was anticipated by the findings of Dijksterhuis and Smith (2002), who documented that participants habituate to emotional stimuli. Accordingly previously encountered negative stimuli are judged less negative and previously encountered positive stimuli are judged less positive. We, however, remain unconvinced that Dijksterhuis and Smith's emotional habituation is applicable here because of methodological differences. Dijksterhuis and Smith, for example, used six subliminal presentations to achieve habituation, and it is unclear if habituation will follow from a single presentation. What is sorely missing is the analogous conventional mere-exposure experiment with the same negative, positive, neutral, and erotic stimuli to firmly establish expectations. In fact, Bem took this approach with his retroactive priming experiments (Bem's Experiments 3 and 4), and the inclusion of conventional priming studies to establish firm expectations greatly increases the interpretability of those results. Without these control experiments to establish the direction of mere exposure effects with emotional and evocative stimuli, the most judicious course is to exclude Experiments 5, 6, and 7 from analysis.

Table 3 reveals that there is relatively little support for the claim that people can feel the future with erotic or neutral events. The Bayes factor does offer some support for a retroactive effect of emotionally valenced, nonerotic stimuli: The evidence for an effect provided by Experiments 2, 3, and 4 outweighs the evidence against an effect provided by Experiment 1. In Experiment 2, participants were rewarded with brief presentations of positive pictures and punished with brief presentations of negative ones when they anticipated or failed to anticipate, respectively, the future state of a random-number generator. In Experiments 3 and 4, participants identified an emotionally valenced target stimulus more quickly when a subsequently presented prime matched the valence of the target.

General discussion

The publication of Bem's (2011) report on ESP provides an ideal opportunity to discuss how evidence should be assessed and reported in experimental studies. We argue here that inference by *p*-values not only precludes stating evidence for

Table 3 Bayes factor for three feeling—The-future hypotheses

Stimuli	Included Experiments				Bayes Factor
Erotic stimuli					3.23
Bem'sexperiment	1				
Sample size	100				
t-value	2.51				
Negative or positive stimuli					38.7
Bem'sexperiment	1	2	3	4	
Sample size	100	150	97	99	
t-value	0.15	2.39	2.42	2.43	
Neutral stimuli					1.57
Bem'sexperiment	1	8	9		
Sample size	100	100	50		
t-value	0.15	1.92	2.96		

theoretically useful null hypotheses, but also overstates the evidence against them. A suitable alternative is the Bayes factor—the relative probability of observing the data under two competing hypotheses. To use the Bayes factor, it is necessary to specify a prior against which evidence is calibrated. We recommend the JZS prior as a suitable generic default because the resulting Bayes factor is invariant to changes in measurement scale and has beneficial theoretical properties (see note 1). One of the drawbacks of our previous development (Rouder et al, 2009) was that it did not provide a means of combining data across multiple experiments, making meta-analysis difficult. Herein, we extend JZS default Bayesian *t*-test to multiple experiments and use this new development to analyze the data in Bem. Our Bayes factor analyses of Bem's data, which Bem offered as evidence of ESP, show that the data support more modest claims. The data yield no substantial support for ESP effects of erotic or neutral stimuli. For emotionally valenced nonerotic stimuli, however, we found a Bayes factor of about 40, and this is the factor by which readers should increase their odds.

We caution readers against interpreting this Bayes factor as the posterior odds that ESP is true. On the contrary, posterior odds should reflect the context provided by prior odds, as discussed previously. In the present case, there are two relevant sources of context for prior odds: past studies of ESP, and the plausibility of mechanisms underlying ESP. Bem (2011) fallows in a line of parapsychological research that extends from the 1930s. In a recent meta-analyses, Storm, Tressoldi and Di Risio (2010) reported a sizable degree of statistical support for ESP for certain classes of experiments. For example, among the 63 studies that used a four-choice procedure, participants responded correctly on a total of 1,326 out of 4,442 trials, a rate of almost 30% (as compared with a 25% baseline). We worry, however, about the frequency of unreported studies. To us, the more relevant context in setting prior odds is the lack of a plausible mechanism for ESP. ESP seems contradicted by well-substantiated theories in physics and biology. Consequently, it is reasonable to have low prior odds on ESP. In our view, while the evidence provided by Bem is certainly worthy of notice, it should not be sufficient to sway an appropriately skeptical reader. We remain unconvinced of the viability of ESP.

Appendix: Statistical development

The JZS prior and Bayes factor for a single contrast

Model Let $y_1, \ldots, y_N$ be a sequence of N observations. The model of these observations is

$$y_i \overset{iid}{\sim} \mathrm{Normal}\left(\sigma\delta, \sigma^2\right),\ i = 1, \ldots, N,$$

where σ^2 and δ are variance and effect size parameters, respectively.

Hypotheses The null hypothesis is that $\delta = 0$; the alternative is that $\delta \neq 0$.

Priors Priors are needed for parameters σ^2 under the null model, and for parameters σ^2 and δ under the alternative. In the JZS setup, the prior for σ^2 is

$$f(\sigma) = 1/\sigma^2$$

under both hypotheses. The prior for effect size under the alternative hypothesis is

$$\delta \sim \text{Cauchy}.$$

The Cauchy distribution is described in Johnson, Kotz and Balakrishnan (1994).

Bayes factor Rouder et al. (2009) provided the following expression for the corresponding Bayes factor (alternative/null):

$$B = \frac{\int_0^\infty (1+Ng)^{-1/2}\left(1+\frac{t^2}{(1+Ng)(N-1)}\right)^{-N/2}(2\pi)^{-1/2}g^{-3/2}e^{-1/(2g)}dg}{\left(1+\frac{t^2}{N-1}\right)^{-N/2}}$$

where t is the one-sample t statistic $\left(\frac{\bar{y}\sqrt{N}}{S_y}\right)$An applet to compute this expression is provided at pcl.missouri.edu/bayesfactor.

Meta-analytic extension

Model Let $t_1, \ldots, t_M$ and $N_1, \ldots, N_M$ denote a sequence of t-values and sample sizes, respectively, from M experiments. We model these t-values as

$$t_i \overset{iid}{\sim} T\left(N_i - 1, \delta\sqrt{N_i}\right),$$

where T is the noncentral t distribution (Johnson et al., 1994) with N-1 degrees of freedom and noncentrality parameter $\delta\sqrt{N}$.

Hypotheses and priors As before, the null hypothesis is that $\delta = 0$; the alternative is that $\delta \neq 0$. Moreover, the prior on effect size is a Cauchy (for a two-sided alternative) or a half-Cauchy (one-sided alternative). We use the one-sided half-Cauchy in our assessment.

Bayes factor The meta-analytic Bayes factor is

$$B^* = \frac{\int_{\delta=0}^{\delta=\infty} \prod_{j-1}^{M} g\left(t_j, N_j - 1, \delta\sqrt{N_j}\right) f(\delta)d\delta}{\prod_{j-1}^{M} g\left(t_j, N_j - 1, 0\right)}. \tag{2}$$

where g is the probability density function of the noncentral t (Johnson et al., 1994) and f is the probability density function of the Cauchy, or half-Cauchy (Johnson et al., 1994), depending on whether a two-tailed or one-tailed alternative is desired. It may be shown (through algebraic rearrangement) that if M = 1, then $B^* = B$. This property is a consequence of the scale-invariant properties in the JZS prior and does not hold in general for other priors.

References

Bayarri, M. J., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika, 94*, 135–152.

Bem, D. (2011). Feeling the future: Experimental evidence for anamalous retroactive infleces on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425.

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*, 112–122.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin, 106*, 265–289.

Dijksterhuis, A., & Smith, P. K. (2002). Affective habituation: subliminal exposure to extreme stimuli decreases their extremity. *Emotion, 2*, 203–214.

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

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

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

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

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions Vol. 2* (2nd ed.). New York: Wiley.

Kass, R. E. (1992). Bayes factors in practice. *Journal of the Royal Statistical Society, 2*, 551–560.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g*-priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*, 410–423.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192.

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.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136*, 471–485.

Utts, J., Norris, M., Suess, E., & Johnson, W. (2010). The strength of evidence versus the power of belief: Are we all Bayesians? In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics*. Voorburg: International Statistical Institute.

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

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

Zellner, A. (1986). On assessing prior distirbutions and Bayesian regression analysis with *g*-prior distribution. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honour of Bruno de Finetti* (pp. 233–243). Amsterdam: North Holland.

Zellner, A., & Siow, A. 1980. Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp.585–603). University of Valencia.