How Bayesian statistics are needed to determine whether mental states are unconscious

Zoltan Dienes

Sackler Centre for Consciousness Science and School of Psychology

University of Sussex

To appear in:

M. Overgaard (Ed.), Behavioural Methods in Consciousness Research. Oxford: Oxford University Press.

Abstract

Inferring that a mental state is unconscious often rests on affirming a null hypothesis. For example, for perception to be below an objective threshold, discrimination about stimulus properties must be at chance. Similarly, for perception to be below a subjective threshold by the zero correlation criterion, the ability to discriminate one's own accuracy must be at chance. But a non-significant result in itself does not mean there is evidence for the null hypothesis; a non-significant result may just mean that the data are insensitive. Orthodox statistics does not provide a practical solution to this problem. But Bayes Factors provide a simple solution. The solution is vital for progress in the field, as so many claims that mental states are unconscious have relied on non-significant results with no indication of whether that outcome is due purely to data insensitivity.

An important aspect of consciousness research is determining when a mental state (e.g. perception, memory, knowledge, intention) is conscious versus unconscious. Declaring a mental state unconscious often means asserting that some measure of conscious knowledge has a value of zero, or a relationship with a measure of conscious knowledge has a value of zero. That is, declaring a mental state unconscious often depends on asserting a null hypothesis. Conversely, in other situations, asserting that unconscious knowledge does not exist also depends on asserting a null hypothesis.

Researchers have been trained to feel ambivalent about asserting a null hypothesis (e.g. Gigerenzer, 1993). Those feelings are based on the fact that significance testing, as normally conducted, contains no basis for asserting the null hypothesis. While orthodoxy offers two ways of providing a basis (power and confidence intervals) those solutions are often problematic in real scientific contexts (because they crucially depend on specifying a minimal interesting effect size, which is often hard to specify; see the discussion of the principles of 'inference by intervals' in Dienes, submitted). In the absence of a real method for asserting the null hypothesis, researchers freely assert the null hypothesis following a non-significant result for no principled reason (backing down when challenged, or when rhetorically useful). This chapter proposes a simple easy-to-use solution, one that indicates how strong the evidence is for the null versus the alternative hypothesis. Details of using free online software are described, and then concrete examples given in the context of research into unconscious processes. Objective and subjective measures are considered in turn.

1. Do the data support the null hypothesis?

Initially we will consider a series of imaginary examples involving a non-significant result to check our intuitions about what can be concluded. A researcher exposed people to rapidly presented faces. The task was to discriminate which face was presented on each trial. Participants also indicated the clarity of their visual experience on each trial with the Perceptual Awareness Scale (PAS; Ramsøy & Overgaard, 2004). Specifically, participants indicated if the experience for that trial was completely clear (4), almost clear (3), constituted a glimpse of something present (but content could not be specified further) (2), or was non-existent, they had no experience of a stimulus (1). After careful exploration, the researcher found conditions in which participants gave a PAS rating of 2 on each trial (the example is made up). The discrimination ability yielded a mean *d'* of 0.4, t = 2.85, p < .01, with 30 participants. (*d'* is a measure of discrimination ability, giving the estimated internal signal to noise ratio; *d′* is 0 if there is no ability to discriminate, negative if people systematically discriminate incorrectly, and positive if people systematically discriminate correctly.) In sum, there is evidence of a sort of subliminal perception in that people say they don't know what it is they saw, but they can still discriminate what was there.

You would like to know whether subliminal perception occurs when defined by a PAS rating of 1 rather than 2. Simply by changing exposure duration slightly, and keeping everything else the same you find conditions where participants give a PAS rating of 1. So you replicate the original researcher's procedure except for this one change, with the same number of participants (30). For ability to discriminate which face was presented, you obtain a non-significant result, mean *d′* = 0.2 (standard error, SE =0.25) , *t* = 0.80, *p* = .4. What do you conclude about whether people can discriminate stimuli when they give PAS = 1? By how much would these data make you change your confidence in the hypothesis that people can discriminate when PAS = 1? Do you feel you need to collect more data to support any conclusion you draw - or do you have enough for practical purposes? For example, is there enough evidence to assert in a talk, or the discussion section of a paper, that subliminal perception did not occur for PAS =1 for the conditions of the experiment?

Second example. You replicate the original researcher with the same number of participants and with the one change that makes PAS scores 1, as before. But for this example, you obtain mean *d′* = 0.0 (SE = 0.25) , that is, the sample mean is exactly at chance baseline, *t* = 0.00, *p* = 1.0. Now what do you conclude about the existence of subliminal perception when people give PAS = 1? Can people discriminate stimuli when they say they saw nothing at all? How strongly is the matter settled for the conditions of your experiment by this set of data?

Next example. You run the experiment as in the previous examples, but you obtain mean *d′* = -0.20 (SE = 0.25), that is the sample mean goes in the "wrong direction", below chance baseline, *t* = 0.80, *p* = .4. Now what do you conclude about the existence of subliminal perception when people give PAS scores of just 1? How confident are you in null hypothesis versus the theory that there exists subliminal perception for PAS =1?

Final example. You run the experiment as in the previous examples, but with ten times the number of participants (i.e. with 300 instead of 30). You obtain *d′* = 0.03 (SE = .037), *t* = 0.80, *p* = .4. Now what do you conclude about the existence of subliminal perception when people give PAS scores of just 1? How strongly is the matter settled for the conditions of your experiment?

Table 1 summarizes the results for these four hypothetical replications of a significant result under new conditions (PAS= 1 vs 2). There is of course no guarantee that subliminal perception will occur under the new conditions just because it did under the old. (Indeed, you might believe that it should not.) But what evidential value do any of these results have for drawing a firm conclusion? One intuition you may have is that the evidence is stronger in support of the null in the final example than in the first. And, as we will see, this intuition is correct. But notice the p-values in those two cases are the same. So p-values cannot constitute a good measure of evidence for the null hypothesis. We need a better measure. In the next section we consider a better measure, and we apply it to each of these examples.

Table 1 Statistics for four hypothetical tests for an effect

Raw effect size	SE	t	p	Confidence in theory relative to null?
+0.20	0.25	0.8	0.4	?
0.00	0.25	0.0	1.0	?
-0.20	0.25	0.8	0.4	?
+0.03	0.037	0.8	0.4	?

2. The Bayes factor

2.1 The nature of evidence: The Devil and the cat

We will use the principle that evidence supports the theory that most strongly predicted it. To illustrate the principle, I have a box called Zoltan's Box Of Mystery. Inside it is one of two creatures with equal probability. Inside there is either a Tasmanian Devil or else a cat. Tasmanian Devils have one of the strongest bites amongst land mammals, so if you lower your hand down to pet the little fellow, the bite could go through your finger bones like butter. In fact, if you put your hand in the box and a devil is in it, there is a good chance that your hand will be left with only four fingers. The other creature that could be there is a cat instead of a devil. You are much less likely to lose your finger if the cat is there. The cat is sweet, but he does have a vicious streak so there remains some probability that the cat will remove a finger as well. The box is well tested so I can be precise: If the devil is in the box there is a probability of 9/10 of losing a finger when a hand is lowered in it; by contrast, if the cat is in the box there is a probability of only 1/10 of losing a finger.

The box is on the table. We do not know which creature is in it. John puts his hand in the box. When he removes his hand, he has lost a finger. Which hypothesis does the data support more strongly, the devil or the cat?

Which theory most strongly predicted the outcome? The outcome is predicted with probability 9/10 by the devil hypothesis and only 1/10 by the cat hypothesis. So the devil hypothesis is more strongly supported. We can quantify how much more strongly the devil hypothesis is supported over the cat by dividing the 9/10 by the 1/10: The evidence is nine times as strong for the devil over the cat hypothesis. Or we can say the Bayes factor, *B*, for the devil over the cat = 9.

Now imagine that what happened when John put his hand in is that he pulled it out with all five fingers intact. Now which hypothesis is most strongly supported? Now the cat hypothesis predicts this outcome with probability 9/10 and the devil with probability 1/10. So the data support the cat hypothesis nine times more strongly than the devil; or *B* = 9 for the cat over the devil; or, equivalently, *B* = 1/9 for the devil over the cat.

This time a new devil and a new cat have been found, more equal in their character[1]. In fact, for these new creatures, thorough testing shows that a finger is lost 6/10 of the time if the devil is in the box and 4/10 of the time if the cat is. Now if John loses a finger, would you have a strong opinion as to which creature was in the box? The evidence only slightly favours the devil over the cat, by a factor *B* = 6/10 divided by 4/10 = 1.5, that is, not by much at all. The evidence is simply inconclusive.

We have constructed three situations, the first in which B showed the evidence supported one hypothesis more strongly than the other, the second the other way round, and in the third, the evidence did not strongly indicate anything either way. In general, the Bayes factor, *B*, indicates how much more probable the data are on one theory (say H1, the alternative hypothesis, your pet theory) rather than on another theory (say, H0, the null hypothesis); thus, it measures the amount of evidence data provide for H1 compared to H0 (e.g. Berger & Delampady, 1987; Dienes, 2011, submitted; Gallistel, 2009; Goodman, 1999; Jeffreys, 1939/1961; Kass & Wasserman, 1996; Kruschke, 2011; Lee & Wagenmakers, 2005; Rouder et al., 2009) . *B* can indicate whether (i) there is strong evidence for H1 over H0; or (ii) whether there is strong evidence for H0 over H1; or (iii) whether the data are insensitive and do not discriminate H1 and H0. In effect, p-values only make a two-way distinction, they contrast (i) with either (ii) or (iii), but in no way discriminate (ii) from (iii). A p-value

[1] For an account of the relation between the devil and the cat, see Bulgakov (1997).

of .1 or .9 has nothing to say over whether there is substantial evidence for the null or whether the data are insensitive. The discrimination between (ii) and (iii) is just what has been missing from statistical practice. Bayes factors make the required discrimination, they plug the hole in orthodoxy.

Jeffreys (1939/1961) suggested conventions for deciding whether evidence was substantial or not. If *B* is greater than 3, then there is substantial evidence for H1 over H0 (for the way round the Dienes, 2008, calculator is programmed); if *B* is less than 1/3 there is substantial evidence for H0 over H1; and if B is between 1/3 and 3 the data are insensitive, nothing follows from the data (other than more needs to be collected). The conventions are not arbitrary. If a significant result at the 5% level is obtained and the obtained effect size is about that expected, then B is likely to be about 3 (Dienes, submitted). So *B* > 3 corresponds to the standard of evidence we are accustomed to as scientists in rejecting the null hypothesis, as we will see in the examples below. (Though there is in fact no necessary one-to-one relation between p-values and *B*; Lindley, 1957.) By symmetry, we get a standard of evidence for accepting the null: *B* < 1/3.

Evidence supports the theory that most strongly predicted it. Thus, in determining the strength of evidence for H1 versus H0 the predictions of each must be specified. This is easy for the null hypothesis. For example, the null hypothesis may say that the population mean difference is exactly zero. But what does H1 predict? A major task we tackle below is just how to answer this question for the case of establishing the conscious status of perception or knowledge. Our goal will be to consider for H1 the range of possible population values (is there a minimum or maximum plausible value?) and whether some values are more likely than others. If the question strikes you as fiddly and irksome just remember: You can't tell if evidence supports a theory if you don't know what it predicts.

One reaction is to ask if we can postulate a default H1? That is, could we specify predictions in a way suitable for many situations psychologists might come across, so that it could be used generally, and is hence "objective"? Rouder et al (2009) provided such a default Bayes factor calculator (see http://pcl.missouri.edu/bayesfactor) (cf also the Bayesian Information Criterion, BIC, which approximates a vague default Bayes factor, Wagenmakers, 2007). To cover the range of situations psychologists are interested in, the Rouder calculator assumes that according to H1, the effect could be in either direction, and the standardized effect size (Cohen's d) could be up to 4 or 5, but 6 or more is very unlikely. Ultimately, this is just a particular set of predictions, a particular model, and it may or may not be relevant to a particular scientific problem. Thus, the calculator allows modifications of these predictions. In fact, Rouder et al (submitted) argued elegantly that there is no "free lunch" in statistical inference: We have to do the work in specifying predictions of the alternative to allow sensible statistical inference at all. Inference should always be sensitive to precise scientific context; thus, Gelman and Rubin (1995) argue inference should in the end go beyond the statistics. Here we will attempt to make sure that the statistics themselves address the scientific context as much as possible.

2.2 Representing the alternative hypothesis

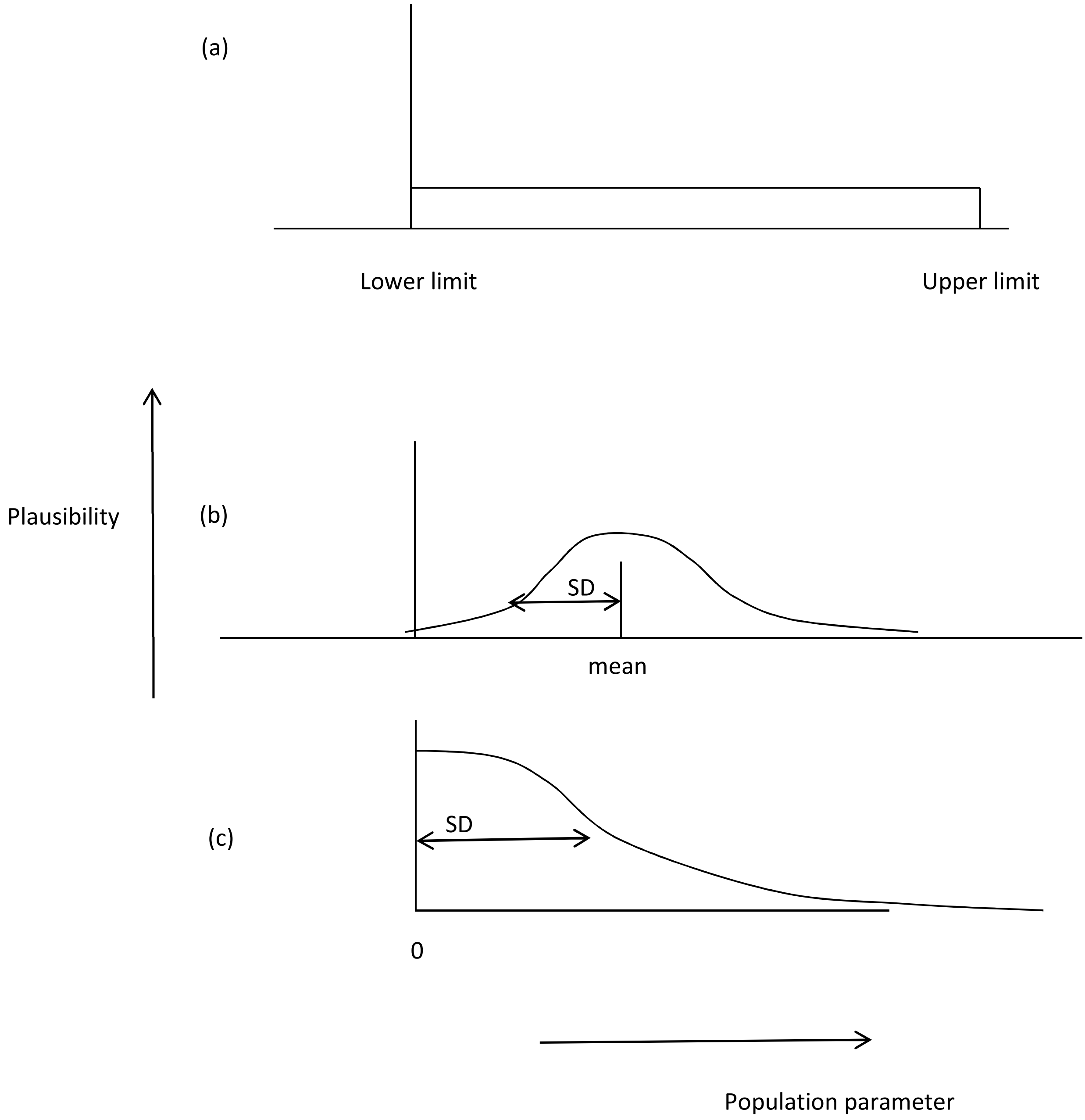
The Dienes (2008) Bayes factor calculator gives three options for specifying the predictions of H1: The plot of plausibility against different possible population values could be a) a uniform; b) a normal; and c) a half-normal (see Figure 1). A uniform indicates that all population effect sizes in certain range (from minimum to maximum) are equally plausible, and anything outside that range is ruled out. A normal indicates that one effect size is most plausible, and smaller or larger values are increasingly unlikely. A half-normal is constructed from a normal which had a mean of zero, but now all of the left-hand side of the distribution has been removed. So the half-normal indicates that values close to zero are most likely, the bigger the size of the effect in the positive direction, the less likely it is, and all negative effects are ruled out (see Figure 1). These distribution shapes capture the predictions of many theories in psychology to a sufficient accuracy. (It turns out that in many cases the exact shape does not change substantial conclusions, as we will shortly see: This is crucial in establishing that the predictions of a theory are specified only to the accuracy that they deserve.)

Now we will consider the examples with the PAS scale. Previous research has found that as the PAS scale increases, so does discrimination accuracy (e.g., Atas, Vermeiren & Cleeremans, in press; Ramsøy & Overgaard, 2004). Thus, whatever accuracy as might occur for PAS = 1 will be less than that for PAS = 2. We have an estimate of the accuracy for PAS = 2 from the previous researcher: *d'* = 0.4. Thus, we can use this as an upper limit of a uniform. What should the lower limit be? This is harder to say. Technically, any value above zero would indicate some degree of subliminal perception for PAS = 1. Maybe some degrees of subliminal perception would be so tiny though, that they are uninteresting? It turns out that while answering this question is crucial for using power or confidence intervals to draw inferences about the meaning of a non-significant result, the lower limit is typically not influential for the conclusions that follow from a Bayes factor. Thus, we can use 0 as the effective lower limit. We might have intuited a lower limit of say *d'* = .05. Different people may give different precise numbers, but we can see what results this lower limit gives us as well.

Taking the first example, with mean *d'* = .20, SE = 0.25, go to the Dienes (2008) Bayes factor calculator (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayes_factor.swf). Enter

0.20 as the sample mean and .25 as the standard error. Say that the plausibility of different population values given the theory (p(population value|theory) is uniform. Enter "0" as the lower bound and ".4" as the upper bound (thereby defining the interval [0, 0.4]), and click "Go!". The result is *B* = 1.24. That is, this non-significant result does not count against the theory of subliminal perception at all. One should not reduce one's confidence, even by one iota, in the theory of subliminal perception for PAS = 1. Now enter the uniform [0.05, 0.4]. This yields *B* = 1.27, barely changed. We do not have to worry about the minimum, entering "0" is good enough and perhaps even truer to our interests (in which case power and confidence intervals become entirely useless, as using them to draw inferences about theories depends on specifying a minimum).

Figure 1 Representing the alternative hypothesis



(a) A uniform distribution with all population parameter values from the lower to the upper limit equally plausible. Here the lower limit is zero, a typical but not required value.

(b) A normal distribution, with population parameter values close to the mean being more plausible than others. The SD also needs to be specified; a default of mean/2 is often useful.

(c) A half-normal distribution. Values close to 0 are most plausible; a useful default for the SD is a typical estimated effect size. Population values less than 0 are ruled out.

Does it matter that we used a uniform distribution? To keep a maximum of 0.4, we could use a normal with a mean of 0.2. By two standard deviations out, the normal distribution comes fairly close to zero. Thus if we used SD = 0.1, there are plausible population values between 0 and 0.4, and the plausibility beyond those limits is small. So now tell the calculator you do not want a uniform. Enter a mean of 0.2 for the normal, a SD of 0.1, and indicate the number of tails as "2" (just meaning it is not a half-normal, the distribution extends in both directions). Now we get *B* = 1.28, again virtually unchanged. Alternatively we could use the half-normal. Given the same principle that a normal comes down close to zero by two standard deviations out, we set the SD of the half-normal equal to 0.2 (set the mean to 0 and the tails to 1; these last two numbers are always the setting for a half-normal). Again we have effectively specified plausible population values between 0 and 0.4. Now we get *B* = 1.23. The upshot is that so long as we used a maximum of about .40 and a minimum near 0, we could shift the distribution around from flat, to pushed up against zero, to peaked in the middle, and it did not affect conclusions. It is precisely this property of Bayes factors that allows confidence in their conclusions. (Note that the robustness of the conclusion to different possible specifications of the theory is not guaranteed; that is a matter to check and confirm. Where there is robustness, the conclusions are to that extent meaningful. Where the conclusion depends on equally plausible parametric specifications, more participants can be run until the conclusion is robust.)

The Bayes factor can be sensitive to the effective maximum specified for the alternative. But even if we used a uniform [0, 0.8] instead of [0, 0.4], B would be 0.84 instead 0f 1.24. Even in this case, the qualitative conclusion is the same: The data are insensitive. Crucially, the maximum of 0.4 has not been arbitrarily intuited; it was based on the established theory that discrimination increases with PAS score and on the estimate of 0.4 for PAS = 2 obtained from data. This illustrates how a maximum can be simply specified in a non-arbitrary way[2]. There are further examples below

[2] One could rightly argue that setting the upper limit as 0.4 does not take into account uncertainty in that estimate. Thus, the upper limit should be increased to reflect that uncertainty; for example, one could use the upper limit of the 95% confidence (or credibility) interval of that estimate. In the original study, *t* = 2.85, so the standard error = mean difference / *t* = 0.4/2.85 = 0.14 d' units. So the confidence interval on the raw effect size for the original study was [0.12, 0.68]. So for B for our first example we could use a uniform to specify the alternative of [0. 0.68] instead of [0, 0.4], which gives a *B* of 0.97, again indicating the same qualitative conclusion, namely data insensitivity. In practice, in my papers I have just used the estimate of the maximum from data and not taken into account uncertainty in that estimate (see Dienes, submitted, for a set of examples). One reason is simplicity in specifying what is being done. The second is that in the examples I used *B* to interpret non-significant results. The higher the maximum of a uniform, the more *B* will support the null. Thus, by using the estimated maximum the outcome errs slightly towards indicating data insensitivity rather than support for the null. This is the cautious way to proceed. It also simplifies the situation where it is hard to specify the actual uncertainty in the estimate, given, for example, a change in context. In many of the

illustrating how this can be done in different situations relevant to consciousness research. As the simplest way of specifying the alternative in this case is the uniform [0, 0.4], we will continue to use this specification for the remaining examples illustrated in Table 1.

In the second example, mean *d'* = 0 for PAS = 1, SE = 0.25, as in the previous example. Now we get *B* = 0.70. That is, the data are insensitive. Just because the sample mean is zero, it does not indicate in itself that one's confidence in the null hypothesis should be substantially increased. If the standard error is large enough, a sample mean difference of around zero is quite possible even when the alternative is true. There is nothing magic about a sample mean difference of zero.

In the third example, mean *d'* = -0.20. The mean is entered as "-.20", negative because it goes in the wrong direction, according to the theory. Now we get *B* = 0.44. That is, while the evidence is favouring the null more than before, it is still not substantial. Just because the sample means go in the wrong direction, it does not mean one has substantial evidence against one's theory and in favour of the null. Again, a sample effect size of zero is not a magic line whose crossing entails that inferential conclusions change their qualitative character.

One should not think that non-significant results are always insensitive. Sensitivity depends on the standard error. In the final PAS example, mean *d'* = 0.03, SE = 0.037. The standard error is considerably smaller than the previous examples (and crucially, it is small relative to the maximum, 0.4). Now we get *B* = 0.25, substantial evidence for the null hypothesis. Table 2 shows the pattern for all our examples. Note that the p-value for the final example is the same as the first (and third), and even less than the second example. Yet the final example provides stronger evidence for the null hypothesis than any of these other examples. P-values do not measure evidence for the null hypothesis.

Of all the examples we have considered, it is only in the final one that we have a reason for asserting, in a results or discussion section, that subliminal perception does not occur for PAS = 1, for the conditions of the experiment[3]. In the previous examples we could have asserted "there was no

examples we will consider, we infer unconscious knowledge from evidence for the null hypothesis. To conclude that unconscious knowledge exists, a simple, cautious, yet practical approach seems appropriate.

[3] Inference by intervals provides an interpretation of these examples largely consistent with one based on Bayes factors, though with interesting differences (see Overgaard, Lindeløv, Svejstrup, Døssing, Hvid, Kauffmann, & Mouridsen, 2013, for a useful application of inference by intervals to subliminal perception). For illustration we will use 95% confidence (or credibility) intervals, but the same logic applies to e.g. 90% confidence or credibility intervals. The upper limit of the 95% confidence or credibility interval for the first example is .2+ 2*.25 = 0.7 d' units. This is larger than the d' of 0.4 for the case of PAS =2, so the data must be declared insensitive, as it was in the text. In the second example the upper limit is 0.5 and the same conclusion follows. Notice to declare data insensitive by inference by intervals does not always require specifying a minimum, so long as one can say the interval includes values definitely above the minimum. In the third

significant subliminal perception" because that is just a fact about our sample and not a claim about the population (it leaves open that subliminal perception may have actually occurred), and it deserves and requires no theoretical explanation. But we could not have asserted "there was no subliminal perception" because that is a claim about the underlying state of affairs. It would be easy to slip erroneously between the two claims, making not a pedantic error but a fundamental scientific mistake.

Table 2 Statistics for four hypothetical tests for an effect, whose plausible range of effect sizes can be specified with a uniform [0, 0.4].

Raw effect size	**SE**	**t**	**p**	**B**
+0.20	0.25	0.8	0.4	1.24, data insensitive
0.00	0.25	0.0	1.0	0.70, data insensitive
-0.20	0.25	0.8	0.4	0.44, data insensitive
+0.03	0.037	0.8	0.4	0.25, substantial evidence for null

Now that the use of the Dienes (2008) Bayes calculator has been illustrated, some remarks about its use. The easiest way to use it in a t-test situation is to run the t-test first. The calculator asks for a "mean" which is the mean difference, M, tested by the t-test. It also asks for a "standard error" (call this SE) which is the standard error of the difference. As t = M/SE, SE = M/t. Thus, as you know M and you know t, you can easily find the required standard error, no matter what design (within-subjects, between-subjects, one sample, mixed). The calculator assumes that the population distribution is normally distributed, as, for example, a t-test does. The calculator also assumes that

example, the upper limit is 0.3, less than the 0.4 of the PAS=2 experiment. In the final example the upper limit is .10. Now judgment is needed about a minimum. This is the first problem with inference by intervals: asserting a null hypothesis does require specifying a minimum in a non-arbitrary way. The second problem is that even when a minimum is specified, intervals typically require more data to reach sensitive conclusions than Bayes factors: In this case the interval includes a minimum of .05, so the null cannot be asserted, while a Bayes factor (using a uniform [.05, .40]) indicates there is evidence for the null. See Dienes (submitted) for fuller discussion. In sum, the rough agreement between the methods is reassuring, and inference by intervals can often be useful and quickly used to indicate data insensitivity. Where a rough typical or a maximum effect size can be specified, Bayes factors are easier to use than intervals. Where the minimum is the aspect of the alternative hypothesis easiest to specify, and where the minimum is the most important aspect of the alternative hypothesis, inference by intervals may be more useful than a Bayes factor.

the population variance is known, which in many applications it will not be. If the degrees of freedom, df, are greater than 30, then the assumption can be ignored. If df < 30, a correction should be applied to the size of the standard error. Specifically, increase SE by a factor (1 + 20/df^2) (see Dienes, submitted). For example, if df = 10, the correction factor is (1 + 20/100) = 1.2. Thus, if the standard error was 0.25, you would actually enter into the calculator 1.2*0.25 = 0.3 as the standard error. The calculator can be used in many ANOVA, regression, correlation and Chi squared situations (see Dienes, submitted, for how). See Appendix 2 for a discussion of the use of Rouder's Bayes factor calculator for binomial situations.

As is now obvious, there is no such thing as the THE Bayes factor for a given set of data. A Bayes factor compares two models; for example a model of H1 against H0. How we model H1 can vary according to theory and context. To make this explicit, when a uniform is used, *B* could be notated B_U; when a half-normal is used, B_H; and when a normal B_N. (And when the full default Rouder calculator is used, B_{JZS}.)[4] Further, $B_{U[0,3]}$, for example, could specify that the uniform was the interval [0, 3]; $B_{N(10,5)}$ could specify that the normal had a mean of 10 and a SD of 5; and $B_{H(0, 5)}$ could specify that the half-normal used an SD of 5. Appendix 1 shows the results using different specifications of H1 for all the examples to follow. (See also Verhagen and Wagenmakers, in press, for a different Bayes factor calculator that takes as its theory that the current study is an exact replication of a previous one, so H1 can be set as predicting the effect size previously obtained, with an uncertainty defined by the previous standard error in the estimate.)

With background and technicalities out of the way (see Dienes, 2008, 2011, and submitted, for more discussion), we now consider the application of Bayes to different situations in which the conscious status of knowledge and perception are established.

3. The use of Bayes when using objective measures

According to objective measures, knowledge is unconscious when priming shows knowledge but direct classification of the relevant distinction is at chance. Asserting that performance is precisely at chance requires Bayes factors; it cannot be done with orthodox approaches unless a theoretically relevant minimum can also be stated. But it is not clear how such a minimum could be decided in order to declare knowledge unconscious. Thus, using objective measures to assert knowledge is unconscious, according to the definition just given, requires Bayes factors. We will first consider the case of implicit learning, and then subliminal perception.

[4] Thanks to Wolf Vanpaemel for suggesting both this notation and the table in the Appendix

3.1 Objective measures in implicit learning.

A common paradigm for exploring implicit learning is the serial reaction time (SRT) task (Nissen & Bullimer, 1987). People indicate which of, say, four possibilities occurred on a given trial by pressing one of four buttons (for example, they indicate which of four locations a stimulus appeared in that trial). From the participant's point of view this is all there is to the task: It is a complex reaction time task. Unbeknownst to participants the sequence of events is structured. It can be shown that people learn the structure because they are faster on structured rather than unstructured trials. The question is, is this knowledge of the structure as shown in reaction times conscious or unconscious? One common method for determining the conscious status of the knowledge is to give participants a recognition test afterwards. The logic is that if people are at chance on recognizing the structure then the knowledge must be unconscious. To employ this logic, a null hypothesis must be asserted.

Shang, Fu, Dienes, Shao, and Fu (2013) used an SRT task followed by a recognition task. The SRT task involved over 1000 trials, where 90% of the trials followed a sequence and 10% violated the sequence in some way. Reaction times showed that people had acquired knowledge of structure (people were faster for trials that followed rather than violated the sequence, $p < .001$). Was that knowledge conscious? For one set of conditions, people were non-significantly different from chance on a subsequent recognition task ($p = 0.30$). A common reaction would be to look at the p-values and declare the objective threshold satisfied. But the $p = 0.30$ in itself does not mean people were at chance in recognizing, and thus it does not mean that the knowledge was unconscious.

We need a handle on an expected level of recognition, if knowledge had been conscious. Shang et al (2013)used a sequence of elements defined by triplets. That is, just knowing the preceding element did not allow one to predict the next element. But knowing the two preceding elements allowed one to predict the next one with certainty. There were in total 12 such triplets that could be learned. The recognition test consisted of indicating whether or not a triplet was old (explicitly defined as the one occurring 90% of the time). Thus the question arises, how many triplets did people learn in the reaction time phase of the experiment? Each triplet had been paired with an infrequent violation triplet with the same first two elements but a different final element. Thus for each triplet Shang et al could determine if there was evidence for RT saving. Let us say in one condition there was significant learning of each of five triplets. If on the recognition task, people expressed all this knowledge they would get those five correct. The remaining seven triplets did not contribute detectably to the RT effect; participants are thus expected to get half of those right in recognition (i.e. 7/2 = 3.5 correct). Thus, in total participants could be expected to get 5 + 3.5 correct or 8.5/12 = 71%, if they expressed all knowledge in the recognition test. However, people are unlikely to express all their knowledge on every trial (e.g. Shanks & Berry, 2012), so it would be more cautious to consider the 71% as a maximum possible recognition performance, rather than the performance we expect (i.e. recognition performance may be significantly lower than the 71%, but this could happen even if the knowledge were conscious). Thus, we could represent the hypothesis that the knowledge was conscious, and thus expressible in recognition, as a uniform from chance (50%) to 71%. In fact, the Dienes (2008) calculator assumes the null hypothesis is always 0. Thus we need to consider our scores in terms of how much above chance they are. Scored in this way, the minimum of the uniform is 0 (i.e. 0% above chance baseline) and the maximum of the uniform is 21 (i.e. 21% above chance baseline).

Say recognition performance was 52% (SE = 6%) (so $t(40) = 0.33$, $p = 0.74$). Is this evidence for people being at chance on the recognition test? Enter “2” (i.e. 2% above a baseline of 50%) as the mean in the Dienes (2008) calculator, enter “6” as the standard error. Indicate the alternative is a uniform and enter the limits [0, 21]. The result is $B_{U[0,21]} = 0.48$, which is not substantial evidence for the null hypothesis that recognition was at chance[5]. Thus these data would not legitimate concluding that there was unconscious knowledge. But now let us say that recognition performance was 52% as before, but with a standard error of 2% (so $t(40) = 1$, $p = 0.32$). Then $B_{U[0,21]} = 0.33$, substantial evidence for the null and hence for unconscious knowledge. (Note that in this case the higher p-value is associated with less substantial evidence for the null hypothesis. Moral: p-values do not indicate evidence for the null hypothesis.)

3.2 Objective threshold in subliminal perception.

Armstrong and Dienes (2014) rapidly presented a low contrast words on each trial and then asked participants to indicate which of two words had just been displayed. The choice was followed

[5] Using a “default Bayes factor”, i.e. one that does not take into account the scientific context, would make obtaining evidence for unconscious knowledge too easy, because defaults necessarily represent alternative hypotheses as vague. (The vaguer a theory, the harder it is to obtain evidence for it.) For example, using the Rouder calculator (http://pcl.missouri.edu/bf-one-sample), gives $B = 0.13$, substantial evidence for the null hypothesis, and hence substantial evidence for unconscious knowledge. (The Rouder calculator is actually scaled in terms of the null relative to the alternative, so it reports 7.79 in favour of the null; take 1/7.79 = 0.13 to get it scaled the same way round as the Dienes calculator.) But in this example, we *can* specify the alternative according to the actual scientific demands of the situation, so a default alternative hypothesis is not appropriate. We can estimate the level of recognition we are trying to detect.

by a confidence rating on a 50-100% scale, where 50% indicated that the participant expected to get 50% of such answers correct because they were literally guessing. If the participant was confident in their answer (i.e. confidence above 50%), the time between the onset of the stimulus and the onset of a back-mask (i.e. the stimulus onset asynchrony, SOA) was reduced, until the participant used 50% five times in a row. This was an attempt to reach a stable subjective threshold (i.e. the point where people believe they are performing at chance). With the SOA's obtained by this method, people were actually correct in indicating which word was presented on 51% of trials (standard error = 0.8%), $t(29) = 1.49$, $p = .15$ (experiment 1). So had the objective threshold been reached (i.e. were people actually at chance) in addition to the subjective threshold?

Armstrong and Dienes (2013; experiment 3) had used the same threshold setting and masking procedures, for slightly different stimuli; they obtained an objective performance of 55%, significantly different from chance. Thus, we can us the 2013 performance as a rough estimate of the sort of performance we could expect in the 2014 study. Remember we need to consider all scores as deviations from the chance baseline; thus the 55% becomes 5% above chance. Figure 1 indicates that for a typical estimated effect size, one can use a half-normal with a standard deviation equal to that estimate (i.e. 5% in this case). Thus, in the Dienes (2008) calculator enter "1" as the mean, and "0.8" as the standard error. Indicate that the alternative will not be represented as a uniform. Boxes for entering the parameters of a normal then appear. Enter "0" for the mean and "1" for the tails (both numbers being the specifications for a half-normal generally). Then enter "5" as the standard deviation. Click "Go!". We obtain $B_{H(0,5)} = 0.60$, indicating the evidence is insensitive. We cannot conclude that the objective threshold has been reached (nor that it has not). The matter could be settled by collecting more data. (In fact, Armstrong and Dienes, 2013, collapsed across three experiments, and with this larger data set had sufficient evidence to show that objective performance was above chance.)

Armstrong and Dienes (2014) were lucky they had another study that provided a rough estimated effect size. What if one did not have a very similar example to draw on in order to specify the alternative? We will now consider another general procedure for considering subliminal perception at the objective threshold.

In general, in a subliminal perception experiment using objective measures, one obtains a level of priming (call it P) in milliseconds or whatever units the priming is measured in. Let us say there was 20 ms of priming. This is significant, $p < .05$. In addition, on a forced choice identification or classification task indicating what was shown, people are non-significantly different from chance, say 51%, $p > .05$. The standard syllogism is to now conclude there was subliminal perception. But this is invalid. The non-significant result in no way indicates that people were at chance on classification. But we cannot go further until we can answer the following scientific question: What level of classification could we expect for 20 ms priming if it had been based on conscious perception? Without further data one cannot say. So run another group of people with stimuli that are difficult to view but nonetheless give a level of conscious experience. As shown in Figure 2 (a), regress classification against priming score for just the data where people are clearly conscious. Now we can estimate for a given level of priming, say P, what level of classification could be expected (call this level E), if perception had been conscious. If P falls within the body of data, we can derive E without extrapolating. But one might not be comfortable using the regression line to obtain E of P falls outside of the data. In fact, one might have an estimate of mean priming and classification for conscious cases, but not access to all the data for obtaining a regression line.

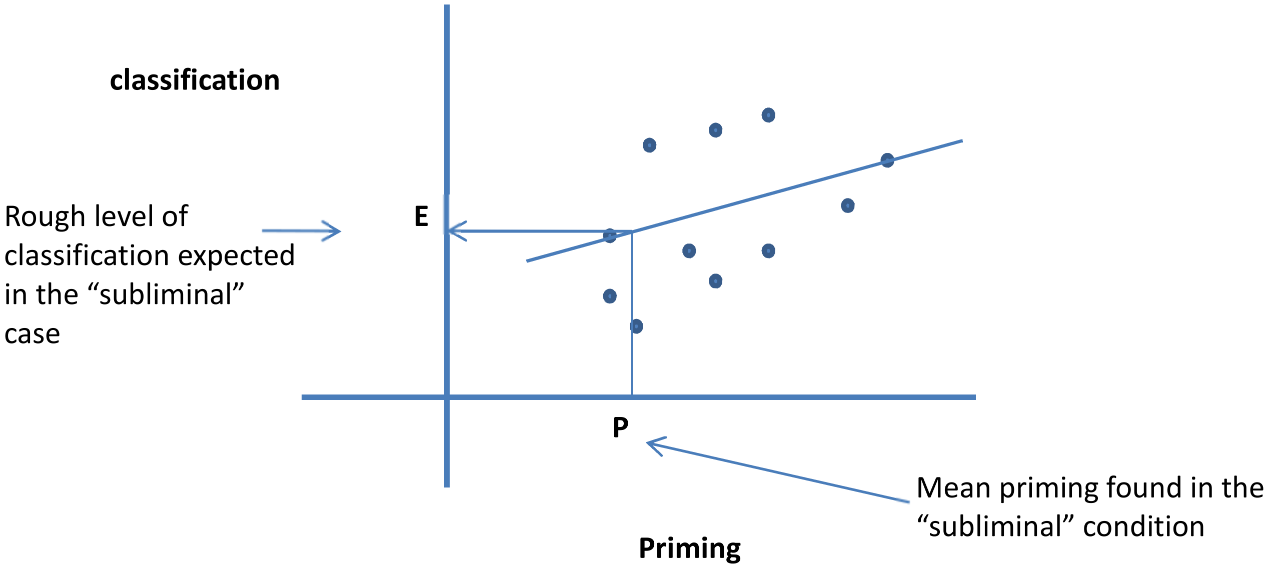
Figure 2 (b) shows a possible solution. The one data point is the mean classification and priming for data where perception was clearly conscious. Draw a line from that point to the origin (0, 0) , i.e. (chance baseline, chance baseline). The assumption is that if perception is conscious it can express itself on either measure (consistent with the global workspace hypothesis); thus, when one measure is at chance, so will the other measure be. The assumption will be false in detail because of noise in measuring or expressing priming; such noise will flatten the regression line. But that only means that the line through the origin will give a smaller E than if one fitted a least squares regression line to the actual data. And a smaller E will make it harder to get discriminating evidence either way. Thus the solution in 2(b) is slightly cautious, while remaining simple. And that is just what we want.

We can work a hypothetical example. As before, say the level of priming is 20 ms, and classification is 51%, $t(40) = 0.5$, $p = 0.62$. By itself, nothing follows from the last result. So a group is run with a longer SOA, under which conditions people say they saw relevant information. Classification is 70% and priming is 40ms. What level of classification do we roughly expect for the potentially subliminal case?

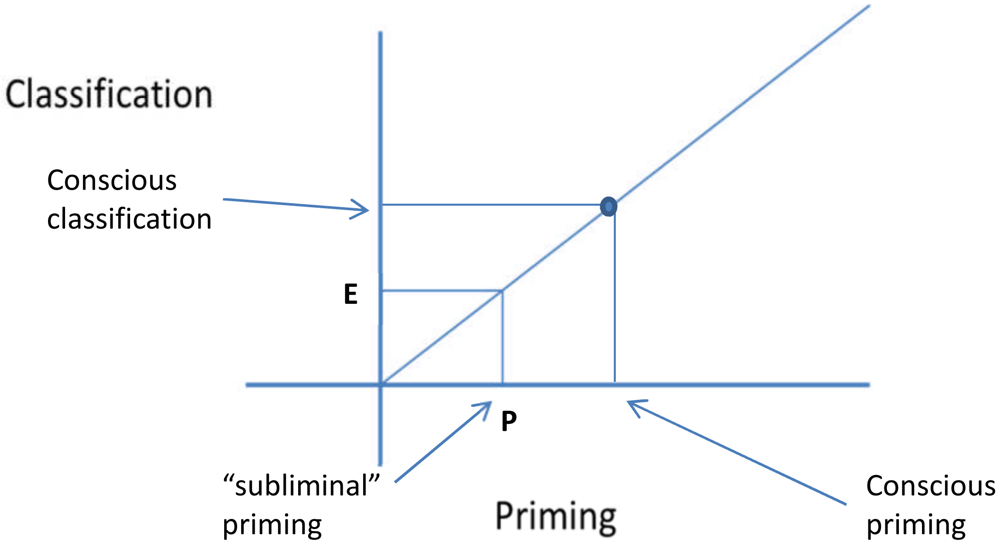
The potentially subliminal condition showed exactly half the priming as the conscious condition. Assuming a proportional relation between priming and classification, as shown in Figure 2(b), the expected level of classification is also halved for the potentially subliminal case. 70% is 20% above baseline; thus a halving of it gives E = 10% above baseline. How should the alternative hypothesis be represented? The simplest method is to use the suggestion in Figure 1(c): use E as the standard deviation of a half-normal. Thus, enter “1” as the mean, “2” as the standard error. Indicate the alternative is not uniform and give its standard deviation as “10”. This gives $B_{H(0,10)} = 0.30$. There is subliminal perception! One could also argue that the expected value E really is the most expected value; thus we could use the suggestion in Figure 1(b) and use a normal with mean “10” and standard deviation half this value (enter “5”)[6]; this gives *B* = .10. In this case, the methods agree qualitatively, so the difference does not matter.

[6] If one had used the method in Figure 2(a), one could use a normal with a more precise standard error, because the prediction from a regression equation has a calculable standard error in prediction. Let SSe = SSclassification $(1 - r^2)$, SE = SSe/(N-2), where SSclassification is the sums of squares for classification scores and *r* is the correlation between classification and priming, then SE in prediction = SE*sqrt(1 + 1/N + (P – mean priming)^2/SSpriming). Represent the alternative as a normal with mean E and standard deviation equal to the SE in prediction.

Figure 2 Predicting a level of classification performance



a) Plot of classification against priming for just those cases where people indicated they saw the stimulus, so the seeing was conscious. The level of priming found in the potentially subliminal condition (X) falls amongst the spread of data for the conscious cases.



b) Same plot but when the level of priming for putative unconscious conditions falls outside of the data for clearly conscious seeing, or else only means are known for the conscious case. The line is drawn to the origin (chance level, chance level).

One general moral to draw is that often interpreting a non-significant result involves solving a scientific problem of what effect size could be expected - a problem that cannot be solved by general statistical methods but, for example, by collecting more data and using it in theoretically relevant ways (in this case, data on how well people do when they consciously see to some degree). Figure 2(b) assumes a relation of close to proportionality between the two measures; in any given case this could be disputed: It is a scientific matter to settle.

For a different approach to assessing perception below an objective threshold, using Bayesian hierarchical models, see Morey, Rouder and Speckman (2008) and Rouder, Morey, Speckman and Pratte (2007).

4 The use of Bayes when using confidence (and other Type II measures of metacognition)

According to subjective measures, one criterion of knowledge being unconscious is if people say they are guessing (or e.g. have no relevant visual experience), and yet they classify above baseline accuracy (the guessing criterion; Dienes, 2008b). In this case, unconscious knowledge is indicated by a significant result; thus, the danger is that unconscious knowledge is declared not to exist simply because of a non-significant result. We considered an example in section 1. According to subjective measures, another criterion of knowledge being unconscious is if confidence is unrelated to accuracy (the zero correlation criterion; Dienes, 2008b). In this case, unconscious knowledge is indicated by evidence for the null hypothesis of no relation between confidence and accuracy. Thus, the danger is that unconscious knowledge is declared just because the data are insensitive.

4.1 Guessing criterion

In terms of the guessing criterion, Guo, Li, Yang, and Dienes (2013) investigated people's ability to associate word forms with semantics. In one condition, Guo et al found that when people said they were guessing, the classification performance was 44% (SE = 5%) where chance baseline was 50%, $t(15) = 1.37$, $p = .19$. Can one conclude there was no unconscious knowledge by the guessing criterion? Not yet. Chen, Guo, Tang, Zhu, Yang, and Dienes (2011) used a very similar paradigm for exploring the learning of form-meaning connections and found the guessing criterion was satisfied with 55% classification accuracy, i.e. a reliable 5% above baseline. Thus, Guo et al modelled the alternative hypothesis with a half-normal with a standard deviation of 5, i.e. using Chen et al as an estimate of the scale of effect size that might be expected if there were an effect. As degrees of freedom were below 30, a correction of $(1 + 20/15^2) = 1.09$ needs to be applied to the standard error; i.e. the standard error to be entered is 1.09*5% = 5.4%. Entering "-6" as the mean (negative because it goes in the opposite direction to theory), “5.4” as the standard error, indicating not a uniform, entering “0” as the mean of the normal, “5” as the standard deviation and “1” for tails, gives $B_{H(0,5)}$ = 0.44. The evidence is insensitive. Guo et al concluded that no claim can be made about whether or not there was unconscious knowledge, according to the guessing criterion.

What if we did not have other examples of reliable guessing criterion knowledge with very similar paradigms? If one had a full range of confidence ratings one could use the information from these data. For example given a continuous 50-100 confidence scale, an intercept at confidence=50% of a regression of accuracy against confidence, using all data *except* for confidence = 50%, could provide an estimated performance for confidence = 50%. That estimate could be used as a standard deviation of a half-normal. However, the intercept might be small, zero or negative if it has a large standard error. And we may be conducting a Bayes factor precisely because the intercept is non-significant and may thus have a large standard error. Thus, the upper limit of the confidence interval on the intercept could be used as the maximum of a uniform, for testing accuracy for when people say they are guessing. (To find the confidence interval, regress accuracy against confidence where you have rescaled confidence so that 0 = guessing, i.e. subtracted 50% from all confidence ratings. Your stats package should give you the intercept, I, and its standard error, SE. I + 2*SE is close enough to the upper limit of the confidence interval of the value of the intercept.) Note this technique is only valid if the accuracy data for confidence=guessing is not used in the regression; otherwise we are double counting the same data, once to make a prediction and then again to test the prediction, which is strictly invalid (Jaynes, 2003). (We can use other aspects of the same data to help make predictions about a mean, but we cannot use the very mean we are testing to predict itself!) The suggested regression technique assumes the theory that performance when judgment knowledge is conscious allows inferences about performance when judgment knowledge is unconscious. (In artificial grammar learning the theory is often true, especially when most structural knowledge is unconscious, Dienes, 2012; but for transfer between domains in artificial grammar learning it is not true, Scott and Dienes, 2010.)

If a binary confidence scale had been used, e.g. “purely guessing” vs “confident to some degree”, classification accuracy for the higher confidence could be used as a maximum for predicted accuracy when guessing. That is, the alternative hypothesis could be represented as a uniform from 0 to a maximum provided by the estimate of performance when people have confidence.

Appendix 2 illustrates using a Bayes factor for binomial data, where a single case achieves a certain proportion of trials correct when they claim they are guessing.

4.2 The zero correlation criterion.

The relation between confidence and accuracy can be expressed in a number of ways. For example, the relation may be expressed in terms of signal detection theory as Type II *d'* (Tunney & Shanks, 2003) or meta-*d'* (Maniscalco & Lau, 2012); or by the slope of accuracy regressed on confidence. We consider each in turn.

In general, for distinguishing conscious versus unconscious states, the interest is in whether having any confidence at all is associated with improved performance compared to believing one is purely guessing. That is, the relevant distinction in confidence is between 'purely guessing' and everything else. The fact that there is no relation between classification accuracy and 'being confident to some degree' versus 'being confident to a larger degree' does not mean that knowledge is unconscious (Dienes, 2004). Thus, for analysis, for determining a zero correlation criterion, confidence should typically be made binary, specifically as 'purely guessing' vs 'any degree of confidence'. Binary ratings could be made from the start (e.g. using 'no loss gambling' methods; Dienes & Seth, 2010), or a more continuous confidence scale could be collapsed.

4.2.1 Signal detection measures of confidence accuracy relation

Signal detection theory expresses the relation between discriminative responses (judgments) and the states being discriminated in terms of a *d'*. If there is no relation between judgments and states, *d'* is zero; otherwise it is some positive value if the judgments are accurate to some degree. Type I judgments are about the world (e.g., what stimulus was present, whether at item is old or new, grammatical or non-grammatical), and the resulting *d'* is called a *Type I d'*. Type II *d'* is calculated in the same way as Type I *d'*, but the judgments are guess vs confident and the states discriminated are the accuracy of the Type I judgments (the states being correct versus incorrect). Assuming the confidence judgment is based on essentially the same information as the type I judgment (cf Lau & Rosenthal, 2011), Type II *d'* is very unlikely to be more than the corresponding Type I *d'* for plausible criteria placement (though it can exceed Type I d' in extreme cases; Barrett, Dienes & Seth, in press). Thus, when using Type II *d'* to measure the confidence accuracy relation, we can represent the alternative hypothesis as a uniform up to the Type I *d'* estimate as the maximum. This procedure was used by Armstrong and Dienes (2013, 2014) to argue for subliminal perception using Bayes factors and the zero correlation criterion.

Type II *d'* is in fact not optimal; it is sensitive to both Type I and Type II bias, so Maniscalco and Lau (2012) developed a different signal detection measure of the confidence accuracy relation, meta-*d′*. Meta-*d′* is the Type I *d′* that would obtain, given the participant's Type II performance, if the participant had perfect metacognition[7]. Indeed, meta-*d′* has better properties than Type II *d′* in practice for large numbers of trials, including insensitivity to Type I and II bias (Barrett et al). (Though Type II *d'* may be better than meta-*d'* to use if the number of trials per participant is less than 50, and response bias is small; Sherman, Barrett & Kanai, this volume.) Assuming the confidence judgment is based on essentially the same information as the Type I judgment (cf Lau & Rosenthal, 2011), Type I *d′* is the maximum that meta-*d′* could be. Thus, when using meta-*d′* to represent the confidence accuracy relation, a natural representation of the alternative hypothesis would be a uniform with a maximum estimated by Type I *d′* (just as we did for when using Type II *d′*).

4.2.2 The accuracy confidence slope.

Consider an implicit learning task where people classify sequences as rule-following or not after a training phase. Overall classification performance is 62% and significantly different from a chance baseline of 50%. There is knowledge, but is it unconscious? Let us say the classification accuracy when a person says they are purely guessing is G, and when they say they have some confidence, C. For example, let us say performance was 54% when people said they were guessing and 65% when they had some confidence. We will rescale these numbers so they are the amount of performance above baseline. That is, G = 4%, C = 15%. Also if we represent overall performance ignoring confidence as X, then X = 12%. The accuracy confidence slope is just the difference C – G, i.e. in this case slope = 11%. Let us say it has standard error of 7%, so $t(40) = 11/7 = 1.57$, $p = .12$, non-significant. One might be tempted to conclude all the knowledge was unconscious because the confidence accuracy relation was non-significant. But as you know, such a conclusion would be unfounded. So do the data provide evidence for the knowledge being entirely unconscious or not?

It turns out to be easy to specify a maximum possible slope, given X, and pc, the proportion of confident responses[8]. Namely, the maximum slope = X/pc. So if 70% of responses were associated with some confidence, the maximum slope = 12/0.7 = 17%. Thus, we represent the alternative as the uniform [0, 17]. This gives $B_{U[0,17]} = 2.65$. The evidence is inconclusive. (But, if anything, is more in favour of knowledge being partly conscious rather than completely unconscious.) For actual

[7] See Supplemental Material for Barrett et al (in press) for MATLAB code for calculating meta-*d′*.

[8] X is a weighted average of G and C, with the weights being the proportions of each type of response. That is, X = (1 - pc) * G + pc * C. By definition, our measure of confidence accuracy relation, the slope, is C–G. This will be maximum when all guessing responses are at baseline, i.e. when G = 0. In this case, slope = C–G = C. Also in this case, X = pc * C, with the G term dropping out. Rearranging, C = X/pc. Thus, since maximum slope = C in this case, maximum slope = X/pc.

applications of this method see Armstrong and Dienes (2013, 2014), Guo, Jiang, Wang, Zhu, Tang, Dienes, and Yang (2013), and Li, Guo, Zhu, Yang, and Dienes (2013).

5. Conclusion

Research in many fields, including consciousness science, often involves asserting a null hypothesis. Up to now users of inferential statistics have not typically used any coherent basis for asserting null hypotheses. The result is theoretical claims made without justification. This chapter offers a few suggestions for how we may proceed using Bayes factors, only declaring mental states unconscious when we have substantial evidence for that claim, and also only claiming the absence of unconscious knowledge when we have substantial evidence for that claim. In general we will only assert a hypothesis, null or alternative, when there is substantial evidence for it. And the rest of the time we will, like Socrates, have the wisdom to know that we do not know.

Funding

This research was partly supported by a grant from the Economic and Social Research Council (ESRC) (http://www.esrc.ac.uk/) grant RES-062-23-1975.

References

Armstrong, A. M., & Dienes, Z. (in press). Subliminal Understanding of Active vs. Passive Sentences. *Psychology of Consciousness: Theory, Research, and Practice*,

Armstrong, A. M., & Dienes, Z. (2013). Subliminal Understanding of Negation: Unconscious Control by Subliminal Processing of Word Pairs. *Consciousness & Cognition, 22* (3), 1022-1040.

Atas, A., Vermeiren, A., & Cleeremans, A. (in press). Repeating a strongly masked stimulus increases priming and awareness. *Consciousness & Cognition,*

Barrett, A., Dienes, Z., & Seth, A. (in press). Measures of metacognition in signal detection theoretic models. *Psychological Methods*,

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2* (3), 317-335.

Bulgakov, M. (1997). *The Master and Margarita*. Picador.

Chen, W., Guo, X., Tang, J., Zhu, L., Yang, Z., & Dienes, Z. (2011). Unconscious Structural Knowledge of Form-meaning Connections. *Consciousness & Cognition, 20*, 1751-1760.

Dienes, Z. (2004). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies, 11* (9), 25-45.

Dienes, Z. (2008a*). Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan. Web site: http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm

Dienes, Z. (2008b) Subjective measures of unconscious knowledge. *Progress in Brain Research, 168*, 49 - 64.

Dienes, Z. (2011). Bayesian versus Orthodox statistics: Which side are you on? *Perspectives on Psychological Sciences, 6*(3), 274-290.

Dienes, Z. (2012). Conscious versus unconscious learning of structure. In P. Rebuschat & J. Williams (Eds), *Statistical Learning and Language Acquisition*. Mouton de Gruyter Publishers (pp. 337 - 364).

Dienes, Z. (submitted). Using Bayes to get the most out of non-significant results. Available here: http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/Dienes%20BF%20tutorial.pdf

Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness & Cognition, 19*, 674-681.

Gallistel, C. R. (2009). The Importance of Proving the Null. *Psychological Review, 116* (2), 439–453.

Goodman, S. N. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. Annals of *Internal Medicine, 130* (12), 1005- 1013.

Guo, X., Jiang, S., Wang, H., Zhu, L., Tang, J., Dienes, Z., & Yang, Z. (2013). Unconsciously Learning Task-irrelevant Perceptual Sequences. *Consciousness and Cognition, 22*(1),203–211.

Guo, X., Li, F., Yang, Z., & Dienes, Z. (2013). Bidirectional transfer between metaphorical related domains in Implicit learning of form-meaning connections. *PLoS ONE, 8*(7): e68100. doi:10.1371/journal.pone.0068100

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

Jeffreys, H. (1939/1961). *The theory of probability: First/Third edition*. Oxford, England: Oxford University Press.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91* (435), 1343-1370.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299-312.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373.

Lee, M. D., & Wagenmakkers, E. J. (2005). Bayesian Statistical Inference in Psychology: Comment on Trafimow (2003). *Psychological Review, 112* (3), 662– 668.

Li, F., Guo, X., Zhu, L., Yang, Z., & Dienes, Z. (2013). Implicit Learning of Mappings Between Forms and Metaphorical Meanings. *Consciousness & Cognition, 22*(1), 174-183

Lindley, D.V. (1957). A Statistical Paradox. *Biometrika, 44* (1–2), 187–192.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*, 422– 430.

Morey, R. D., Rouder J. N., & Speckman P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology, 52*, 21-36.

Overgaard, M. , Lindeløv, J., Svejstrup, S., Døssing, M., Hvid, T., Kauffmann, O., & Mouridsen, K. (2013). Is conscious stimulus identification dependent on knowledge of the perceptual modality? Testing the "source misidentification hypothesis". *Frontiers in Psychology*, *4*, 116, doi: 10.3389/fpsyg.2013.00116

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and Subliminal Perception. *Phenomenology and the Cognitive Sciences, 3*(1), 1–23.

Rouder, J.N., Morey, R.D., Speckman, P.L., & Pratte, M.S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review, 14*, 597–605.

Rouder, J. N., Morey R. D., Verhagen J., Province J. M., & Wagenmakers E. - J. (Submitted). The p < .05 Rule and the Hidden Costs of the Free Lunch in Inference.

Scott, R. B., & Dienes, Z. (2010). Knowledge applied to new domains: The unconscious succeeds where the conscious fails*. Consciousness & Cognition, 19*, 391-398.

Shang, J., Fu, Q., Dienes, Z., Shao, C. &, Fu, X. (2013). Negative affect reduces performance in implicit sequence learning. *PLoS ONE 8*(1): e54693. doi:10.1371/journal.pone.0054693.

Shanks, D. R. & Berry, C. J. (2012). Are there multiple memory systems? Tests of models of implicit and explicit memory. *Quarterly Journal of Experimental Psychology, 65*, 1449-1474.

Tunney, R. J., & Shanks, D. R. (2003). Subjective Measures of Awareness and Implicit Cognition. *Memory and Cognition, 31*, 1060–1071.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54*, 491–498.

Verhagen, 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.

Appendix 1 Variations of conclusions with different Bayes factors

Table 1 shows Bayes factors for examples given in the text where the alternative hypothesis, H1, is specified in different ways. Considering different specifications of H1 is important for evaluating the robustness of the conclusions; if the different specifications are different ways of modelling the same theory then the conclusion is robust if the different Bayes factors agree. B_U, B_N, and B_H are all specified so that the lower and upper limits of plausible values are approximately equal; that is, for a lower limit of 0 and an upper limit of L, B_N uses a normal N(L/2, L/4) to model predictions (i.e. a normal distribution with a mean of L/2 and a standard deviation of L/4); and B_H uses a half-normal based on N(0, L/2) to model predictions. B_{JZS} is the default Rouder calculator (http://pcl.missouri.edu/bf-one-sample scaled so that r = 1), and reciprocated so that higher values indicate support for H1 and against H0.

Table 1 Examples from the text

Example	**Mean**	**SE**	$\mathbf{B_U}$	$\mathbf{B_N}$	$\mathbf{B_H}$	$\mathbf{B_{JZS}}$
1 Section 3.1	2	6	$B_{U[0,21]} = 0.48$	$B_{N(10.5,5.25)} = 0.45$	$B_{H(0,10.5)} =0.64$	$B_{JZS} = 0.13$
2 Section 3.1	2	2	$B_{U[0,21]} = 0.33$	$B_{N(10,5)} = 0.20$	$B_{H(0,10)} =0.53$	$B_{JZS} = 0.20$
3 Section 3.2	1	0.8	$B_{U[0,10]} = 0.39$	$B_{N(5,2.5)} = 0.21$	$B_{H(0,5)} =0.60$	$B_{JZS} = 0.28$
4 Section 3.2	1	2	$B_{U[0,20]} = 0.20$	$B_{N(10,5)} = 0.18$	$B_{H(0,10)} =0.30$	$B_{JZS} = 0.14$
5 Section 4.1	-6	5.4	$B_{U[0,10]} = 0.33$	$B_{N(5,2.5)} = 0.30$	$B_{H(0,5)} =0.44$	$B_{JZS} = 0.44$
6 Section 4.2.2	11	7	$B_{U[0,17]} = 2. 65$	$B_{N(8.5,4.25)} = 2.80$	$B_{H(0,8.5)} =2.35$	$B_{JZS} = 0.39$

These examples illustrate how typically B_{JZS} shows stronger support for the null than more context-specific Bayes factors (because B_{JZS} necessarily uses a vague specification of H1, and thus is effectively testing a different theory than the other Bayes factors). B_H tends to give values closer to 1 than the other Bayes factors (i.e. it is more likely to indicate data insensitivity, because it indicates plausible values around the null value). Thus, if the data are shown to sensitively discriminate H1 from H0 using B_H then the conclusion is likely to be robust (e.g. example 4).

B_{JZS} involves a theory about standardized effect size and so depends on the t-value and degrees of freedom; for constant degrees of freedom, the smaller the *t*, the more B_{JZS} supports the null (e.g. examples 1 and 2 above). The other Bayes factors (in these examples) involve theories of raw effect sizes, and hence can show increased support for the null even as *t* increases (examples 1 and 2 above) – because the larger *t* may indicate a smaller standard error (and hence more sensitivity).

Appendix 2 Rouder's Bayes factor for binomially distributed data

The Dienes calculator assumes normally distributed data, so cannot be used for a binomial situation (unless a normal approximation is used). Consider a task consisting of a sequence of binary forced-choice trials (making left or right responses) where the correct answer is on the right a random 50% of the time; consider the number of successes as binomially distributed. Past research suggests that performance on the task should be about 60% when people say they are guessing. Out of 20 trials where an individual participant claims to be guessing, he obtained 12 correct answers, non-significantly different from the chance expected value of 10, $p = .503$ (using http://graphpad.com/quickcalcs/binomial1.cfm). That is, there is no evidence of unconscious knowledge by the guessing criterion. But is there evidence against unconscious knowledge?

The following Rouder calculator can be used for a binomially distributed observation (regardless of whether a normal approximation is appropriate): http://pcl.missouri.edu/bf-binomial. H1 is specified in terms of the parameters '*a*' and '*b*' of a "beta distribution". The mean of the beta distribution is given by $a/(a+b)$; and its variance is $ab/((a+b)^2(a + b + 1))$. Given that past research has found performance about 60%, the mean for the distribution should be 0.6, the proportion expected on H1. What about the variance? We can use the equivalent of the rule given in Figure 1(b); namely, set the SD to be half the distance of the mean from the null value. The mean, 0.6, is 0.1 units from 0.5; thus, we would like an SD of 0.05. We will obtain this by trial and error. If $a = 60$ and $b = 40$, the mean of the beta is 0.6, as required. Variance = 60*40/(100*100*101) = .0024, and thus SD = √.0024 = .05, just as required. (If the variance had been too big, *a* and *b* would be increased to reduce the variance.) Thus, we enter $a = 60$ and $b = 40$ into the boxes for the "prior". This gives a *B* of 0.73 in favour of the null; i.e. 1/0.73 = 1.37 in favour of H1 over H0 (the way round we have been considering in this chapter). That is, the result may be non-significant, but the data do not support the null hypothesis and do not provide evidence against the existence of unconscious knowledge.

Another way of thinking about setting the predictions of H1, is to treat *a* as the number of successes and *b* as the number of failures in a hypothetical past study, upon which we are basing H1. Using an online binomial calculator (e.g. http://graphpad.com/quickcalcs/binomial1.cfm), 60 successes out of 100 trials is almost significantly different from a null value of 0.5 , $p = .057$. Given that a just significant outcome often corresponds to (just) substantial evidence for H1 assuming the mean was about that expected, one way of understanding the rule in Figure 1 (b) is that it represents the expected value as coming from a past experiment that provided just enough evidence for the value to be taken seriously. Alternatively, a just significant difference can be seen as a way of finding how to spread out the plausibility of different population values maximally so that a negligible amount is below baseline. Thus, a heuristic for setting the *a* and *b* values of the beta is: Set the values so that mean (*a*/(*a* + *b*)) is the expected value, and use a binomial calculator to set (*a*+*b*) so that the number of successes, *a*, would be just significantly different from the null value.

For example, consider an expected value of 0.7. Using a binomial calculator (e.g. http://graphpad.com/quickcalcs/binomial1.cfm), if *a* = 7 (*b* = 3), *p* = 0.34 . We need to increase *a*. If *a* = 14, *b* = 6, then *p* = .12. If *a* = 21, *b* = 9, *p* = .04, just significant. So to specify the prior in the Rouder calculator, use *a* = 21, *b*= 9, to represent an expected proportion of 0.7 on H1 (i.e. as the “prior”).

a = *b* = 1 yields a uniform distribution over the interval [0,1]. Such a distribution could be considered a “default” H1 in that all population values are equally probable. For the example in the first paragraph, using *a* = *b* = 1, gives *B* = 2.52 in favour of the null, i.e. 1/2.52 = 0.40 in favour of H1 over H0. However, given information that expected performance is about 0.60, this default is not as relevant as the one used in the first paragraph.