Published in final edited form as:
Psychol Rev. 2009 April ; 116(2): 439–453. doi:10.1037/a0015251.

The Importance of Proving the Null

C. R. Gallistel
Rutgers University

Abstract

Null hypotheses are simple, precise, and theoretically important. Conventional statistical analysis cannot support them; Bayesian analysis can. The challenge in a Bayesian analysis is to formulate a suitably vague alternative, because the vaguer the alternative is (the more it spreads out the unit mass of prior probability), the more the null is favored. A general solution is a sensitivity analysis: Compute the odds for or against the null as a function of the limit(s) on the vagueness of the alternative. If the odds on the null approach 1 from above as the hypothesized maximum size of the possible effect approaches 0, then the data favor the null over any vaguer alternative to it. The simple computations and the intuitive graphic representation of the analysis are illustrated by the analysis of diverse examples from the current literature. They pose 3 common experimental questions: (a) Are 2 means the same? (b) Is performance at chance? (c) Are factors additive?

Keywords

Bayesian hypothesis testing; Pavlovian conditioning; statistical learning; attention; motor planning

Experimentalists very often wish to conclude that their data imply a lack of effect under some conditions or that two different conditions produce the same effect. Almost without exception, they use conventional null hypothesis significance testing (NHST) to buttress these conclusions, a mode of analysis that is antithetical to their purpose. In a recent critique of NHST, Wagenmakers (2007, p. 795) reminded us:

> A *p* value indicates the evidence against the null hypothesis. It is not possible to observe the data and corroborate the null hypothesis; one can only fail to reject it. Hence, the null hypothesis exists only in a state of suspended disbelief. The APA Task Force on Statistical Inference underscored this point by issuing the warning "Never use the unfortunate expression 'accept the null hypothesis'" (Wilkinson, 1999, p. 599).

Experimentalists are generally aware of these officially sanctioned strictures, and yet they persist in doing conventional analyses and reporting "insignificant" *p* values in oblique support of conclusions to the effect that performance was at chance in a given condition or that performance in one experimental condition did not differ from performance in a control or comparison condition. They may not use the "unfortunate expression," but what they intend is precisely that. The challenge that theoreticians and experimentalists alike need to confront is to bring the null hypothesis out of the conceptual closet, where it has been put by

Correspondence concerning this article should be addressed to C. R. Gallistel, Department of Psychology and Rutgers Center for Cognitive Science, 152 Frelinghuysen Rd., Piscataway, NJ 08854. galliste@ruccs.rutgers.edu.

Supplemental materials: http://dx.doi.org/10.1037/a0015251.supp

[6]One often sees this written with the *H* and **D** and μ and **D** interchanged within the argument of *L*( ). I write it this way to emphasize that, in each case, the likelihood attaches to the symbol on the left of the "given" symbol, whereas the symbol on the right is what is given or assumed.

an inappropriate statistical methodology. The null hypothesis is often the theoretically most elegant and interesting hypothesis, as witness the invariance laws in physics (see Rouder, Speckman, Sun, Morey, & Iverson, in press, for many further examples). It is almost always the more precise hypothesis. We should proudly champion it when the data support it.

The following are three illustrations of the importance of proving the null. They are drawn from among a great many examples in the recent literature because of the diversity of the research topics encompassed and because the researchers were willing to share their raw data with me. Their generosity enabled me to apply the to-be-described, computationally straightforward, normative analytic strategy to marshal support for their null hypotheses—in lieu of the non-normative, antithetical NHSTs that had been published. Together they emphasize the broad importance for psychological theory of empirically supported null hypotheses and the intuitive and graphically obvious way in which Bayesian analysis translates data into odds for or against them.

Do two means differ?

The first null hypothesis is that eightfold variation in the number of trials in a learning experience of fixed duration has no effect on the progress of learning (Gottlieb, 2008). It is hard to imagine a more counterintuitive hypothesis. The assumption that the trial is the basic unit of a learning experience, from which it follows that the progress of learning must depend on the number of such units, is deeply embedded in almost all formalized models of the associative learning process (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981), as well as in almost all neural net and connectionist models. Thus, the null hypothesis that the progress of learning is *not* a function of the number of trials has far-reaching theoretical significance.

Is performance at chance?

The second example of a theoretically consequential null hypothesis is that statistical learning does not occur when a stream of items is seen but not attended to (Turk-Browne, Jungé, & Scholl, 2005). If performance on items not attended to is truly at chance, then lack of attention does not simply attenuate the learning of statistical properties; it closes a gate, denying the sequence access to the mechanisms that extract statistical properties. The question of whether the learning of the statistical properties of an unattended icon stream is truly at chance speaks to the question of whether attention is a matter of selection (gating) or of reducing the allocation of capacity-limited processing resources. This, too, is a null hypothesis whose empirical confirmation has substantial theoretical import.

Are factors additive?

A third theoretically consequential null hypothesis arises whenever the question of additivity is posed. If factors combine additively, then they are most likely operating independently; that is, the mechanism that determines the contribution of one of the factors gets no effective input from the other (Roberts & Sternberg, 1993; Sternberg, 1998). Rosenbaum, Halloran, and Cohen (2006) investigated the effects of target height and target width on grasp height when subjects reached out to grasp an everyday object that was to be moved to a target location. They found additivity, but they remark, “Of course, obtaining evidence for additivity means that we failed to reject the null hypothesis that the contribution of target width would be the same at all target heights. We cannot say, however, that the null hypothesis would definitely fail to be rejected with a much larger group of participants” (Rosenbaum et al., 2006, pp. 920–921). The hypothesis that there is no interaction is a null hypothesis: It asserts that the effects of variation in one factor are the same at all levels of the other factor. Conventional analysis can only argue against this hypothesis, not in favor of it, despite the fact that it is an elegant and simple hypothesis, with strong theoretical

implications. No matter how large one makes the groups, conventional analysis can never provide empirical support for the hypothesis that the factors are additive.

Essentials of the Bayesian Analysis

A Bayesian analysis considers the null hypothesis and at least one plausible alternative to it in the light of the data. Instead of computing a *p* value, as in NHST, it computes the odds favoring one or the other hypothesis. The $\log_{10}$ of the odds is called the *weight of the evidence* (Good, 1960).

The two hypotheses make differing predictions about where the data should fall. These differing predictions take the form of *prior probability distributions.* A prior probability distribution specifies our uncertainty about where the data should fall, given a hypothesis—and, often also, some analytic considerations.

The data give us information about the values of θ, the natural parameters of the *source distribution*, from which we assume that we have drawn the data. If, for example, we assume that we are drawing from a normal distribution—if that is our *statistical model*—then the data in our sample give us information about its mean and standard deviation. When we assume that the source distribution is normal, then $\theta = \langle \mu, \sigma \rangle$ are the parameters whose values must be specified to define one particular normal distribution. The data and the assumed form of the source distribution (normal, Bernoulli, exponential, etc.) together determine the *likelihood function, L*(θ|**D**).[1] It gives the relative likelihoods of different parameter values (in the current example, different values for the mean and standard deviation) in the light of the data, **D**.

The likelihood function is the source distribution function run backward. When a distribution function is run forward, the values of its natural parameters (here, μ and σ) are given and fixed. Graphically, the abscissa represents possible data values. As we vary the value on the abscissa, we read off the corresponding probability or probability density from the ordinate. When we run the distribution function backward, the data are given and fixed; they are the parameters of the likelihood function. Graphically, the abscissa now represents possible values of a distribution parameter, like the mean or standard deviation. If, for example, it is the mean, then as we move along the abscissa, we vary the assumed value of the mean. At each location, we read off the corresponding likelihood. The likelihood of a location for the source distribution is the joint probability of the data, **D**, when the mean is at that location. Given data, some values for the mean are more likely than others and likewise for the standard deviation. The larger the data sample is—that is, the more fixed parameters there are in the likelihood function—the more narrowly the likelihood function constrains likely values for the parameters of the source distribution.

If we assume that there are no prior empirical or analytic considerations that constrain the possible values of a distribution's parameters before we look at the data (if, that is, we assume an uninformative prior), then *L*(θ|**D**) specifies our uncertainty about the values of those parameters after we have looked at the data. However, very often, there are prior considerations that do constrain possible values of the parameters. If, for example, our data are response counts or trials to acquisition, the mean of an assumed-to-be normal source distribution cannot be less than zero. If our data are proportions correct, then *p*, the true probability of a correct response, the natural parameter of a Bernoulli distribution, cannot be greater than one nor less than zero. If the data come from a two-alternative forced-choice

[1]This is often written p(**D**|θ). This is confusing because it treats θ as given and **D** as variable and because it suggests that the function thus symbolized is a probability distribution, in which case, it should integrate to one, which it does not.

paradigm, then *p* cannot be less than .5 (chance). These are purely analytic prior constraints on possible values for the parameters of the source distribution.

Our thinking about *plausible* (as contrasted with analytically possible) values for the parameters of the source distribution is influenced by our hypotheses or beliefs. For example, if we believe that the progress of learning depends on the number of trials a subject has experienced, then we believe that subjects who have experienced many trials have, on average, progressed further than subjects who have experienced only a few. If we have an empirical estimate of the average progress in a population of subjects experiencing only a few trials, then our hypothesis or belief implies that the average progress in a population experiencing many more trials will be greater. Belief- or hypothesis-based considerations shift the probabilities in one direction or another, making some values more probable and others less so, whereas analytic considerations may make some values for the parameters of the source distribution impossible, forcing their prior probabilities to zero. A belief or hypothesis, *H*, together with analytic considerations, determines a *prior probability distribution*, $\pi(\theta|H)$. This distribution represents our uncertainty about plausible values for the parameters of a source distribution before we have looked at data drawn from that distribution.

Bayes's rule tells us to multiply the likelihood function point by point with the *prior distribution* to obtain the *posterior likelihood function* $L(\theta|\mathbf{D}, H)$:

$$L(\theta \left| \mathbf{D}, H \right.) = L(\theta \left| \mathbf{D} \right.)\pi(\theta \left| H \right.).$$

This function specifies the relative likelihood of different values for θ after taking into account both the data, on the one hand, which act through the likelihood function, and, on the other hand, the analytic considerations and a hypothesis or belief, which together act through the prior probability distribution. The graph of $L(\theta|\mathbf{D}, H)$ versus θ looks like a probability distribution, rising above the abscissa and then falling back to it, but it is not because it does not integrate to one. Indeed, its integral is usually nowhere near one.

The marginal or integrated likelihood of an hypothesis, $L(H|\mathbf{D})$, is the integral of the posterior likelihood function with respect to the parameter vector:

$$L(H|\mathbf{D}) = \int L(\theta|\mathbf{D})\pi(\theta|H)d\theta.$$

In other words, it is the area under the graph of $L(\theta|\mathbf{D}, H)$.

The *Bayes factor* is the ratio of the marginal likelihoods of two contrasted hypotheses, H_i and H_j: $BF_{i \text{ vs } j} = L(H_i|\mathbf{D})/L(H_j|\mathbf{D})$, that is, the ratio of the two areas. It tells us the relative extent to which the contrasted prior probability distributions correspond to the likelihood function. When the correspondence or lack thereof is graphically visualized, the Bayes factor puts a number on what we see. The graph of the two prior distributions and the likelihood function of the data they predict also shows the relative vagueness of the two hypotheses and the price that the alternative to the null pays for its greater vagueness. We are tempted to be vague to avoid the embarrassment of being wrong, but vague hypotheses can only be vaguely right. Bayesian analysis counteracts the temptation to be vague (cf. Myung & Pitt, 1997).

The steps in a one- or two-sample Bayesian analysis for a possible difference in the mean (the Bayesian equivalent of a two-sample *t* test) are as follows:

1. Assume a statistical model; that is, specify the form of the distribution from which we have drawn the data.
2. Specify the null prior for the expectation of the distribution from which the experimental data were drawn, $\pi(\mu|H_0)$. When the question is whether performance is at chance, the null prior is obtained by purely analytic considerations. When there is a control or comparison condition, the null prior is obtained from the analysis of the control data.
3. Specify the upper limit on a plausible effect size, thereby limiting the vagueness of the alternative to the null.
4. Combine the limit(s) on the possible size of the effect with the uncertainty, if any, about the expectation when there is no effect. The combination of these two uncertainties (these two probability distributions) yields the prior probability distribution that represents the alternative to the null, $\pi(\mu|H_1)$.
5. Compute the likelihood function given the experimental data.
6. Multiply it point by point with the two contrasting prior probability distributions to get the two posterior likelihood functions.
7. Integrate them to get the two marginal likelihoods. These numbers represent the extent to which the prior probability distributions, one for the null and one for the alternative to it, correspond to the likelihood function.
8. The ratio of the marginal likelihoods is the Bayes factor, the odds favoring one hypothesis over the other. Unlike traditional *p* values, these odds may favor the null.
9. If a researcher is not comfortable with the assumption about the limit(s) on the possible size of the effect, he or she can repeat the computation for different limiting values. When the odds favoring the null approach one from above as the upper limit on the possible size of the effect approaches zero, the null is unbeatable. When the odds never favor the alternative by more than a small amount for any assumed upper limit on the possible size of an effect, considerations of precision and parsimony favor the null. The null is rejected in favor of a vaguer alternative only when, for some assumption about the upper limit on the possible size of the effect (for some plausible degree of vagueness), the odds substantially favor the alternative.

These steps are illustrated thrice over in the following examples. The most important part of each analysis is the graphic presentation of the competing hypotheses (a null hypothesis and a vaguer alternative to it) together with the likelihood function for the experimental data. These graphs (see Figures 5, 7, and 10) make Bayesian analysis conceptually transparent. One can literally see why one hypothesis is superior to the other and what, if any modification might make the alternative a better predictor of the data. For Bayesian analysis to supplant the non-normative method of statistical inference that now prevails, researchers must understand it. The graphs are the key to a conceptual understanding of simple Bayesian analysis.

The computational obstacles to simple Bayesian analysis no longer loom. The Bayesian equivalent of the two-sample, equal variance *t* test, complete with the graphic presentation of the two priors and the likelihood function, has been reduced to a point-and-click operation within Microsoft Excel, comparable in its simplicity to the invocation of the “TTEST”

function. So has the function for testing observed percentage correct against a null of "chance performance."

Example 1: Does Reducing the Number of Trials Really Have No Effect?

In a simple classical conditioning protocol, the onset of a motivationally neutral stimulus called the conditioned stimulus (CS; e.g., a tone or a light) predicts the occurrence of a motivationally important stimulus called the unconditioned stimulus (US; e.g., food or shock) after a fixed delay (the CS-US interval; see Figure 1). That the progress of learning in this paradigm depends on the number of trials (CS-US pairings) has generally been taken to be so obvious as to not require experimental demonstration. Gottlieb (2008) was led to seek experimental confirmation in part by a theory (Gallistel & Gibbon, 2000) that predicted that the number of trials (>1) was irrelevant when the duration of training was fixed. More recently, the hypothesis that the number of trials in a training interval of fixed duration should have no effect has been shown to follow from a consideration of the Shannon information that the CS provides about the timing of the next US (Balsam & Gallistel, 2009).

The long established, but little known, quantitative form of the tradeoff between the delay of reinforcement (the CS-US interval) and the average interval between trials (the average US-US interval) may also be taken to predict this (see Figure 2). To my knowledge, this implication has not previously been pointed out, and it did not motivate Gottlieb's (2008) experiments.

In sum, we have, on the one hand, a strongly intuitive, historical and theoretical presumption in favor of the hypothesis that the number of trials is the primary determinant of the progress of learning. On the other hand, some theoretical considerations and some previous experimental results suggest that reducing the number of trials has no effect on the progress of learning over a given duration of training. There is a lot riding on the outcome of an experiment designed to decide between these contrasted hypotheses, but NHST is, in principle, incapable of supporting one of them. This highlights the fact that NHST is a non-normative procedure for weighing the extent to which data favor competing scientific hypotheses (Rouder et al., in press).

Gottlieb (2008) published the results of four experiments testing the null hypothesis that an eightfold variation in the number of trials had no effect on the progress of learning during a protocol with a given duration. In all of them, he found no significant effect. However, because NHSTs are, in principle, incapable of supporting the hypothesis that motivated his experiments, the published analyses give no indication of the strength of the support that his data provide for this startling conclusion. Bayesian analysis shows that all of his results (and the results of two experiments he did not publish) strongly support the null hypothesis.

From a methodological perspective, the best experiment was the fourth, in which he measured conditioned responding trial by trial and used a change-point algorithm, developed by Gallistel, Balsam, and Fairhurst (2004), to determine for each subject the trial on which it first showed a conditioned response.[2] For one group (Few & Sparse), there were four widely spaced trials per session, hence, one trial per quarter session. In a comparison group (the Many & Dense group), there were eight trials for every one in the Few & Sparse group, hence, eight trials per quarter session. Because the durations of the sessions were the same, these trials were eight times more densely packed. For each subject in each group, Gottlieb

[2]The conditioned response in most paradigms appears abruptly (Gallistel et al., 2004; Morris & Bouton, 2006). The widespread belief that it grows gradually stronger as training progresses is an artifact of averaging across subjects (Papachristos & Gallistel, 2006).

(2008) determined the quarter session, *Q*, during which the change-point algorithm said that a consistent differential between their responding during the CS and their responding during an equal interval immediately preceding the CS first appeared (hereafter, *Q*, or quarter sessions to acquisition). For the Few & Sparse group, *Q* is an estimate of the *trial* on which the conditioned response first appeared; for the Many & Dense group, it is an estimate of the eight-trial *block* during which a consistent conditioned response first appeared. The null hypothesis is that the distribution of quarter sessions to acquisition (*Q*) is the same in the two groups.

Figure 3 shows the cumulative distributions. They are intimately interwined, as predicted by the null hypothesis. If the number of trials mattered, the Few & Sparse distribution would lie well to the right of the Many & Dense distribution, because, on average, it would take a subject in the former group more quarter sessions before it began to respond. If the number of trials were all that mattered, the Few & Sparse distribution would be spread out to the right by approximately a factor of eight.

Assuming a Statistical Model

The empirical cumulative distributions in Figures 3 are well fit by cumulative Gaussians, so we assume that, for both groups in both experiments, the data were drawn from normal distributions.[3] The data do not suggest any difference in the standard deviations of the distribution(s) from which the data were drawn, and we have no hypotheses about these standard deviations. Therefore, we simplify the computational problem by assuming that both source distributions have the same standard deviation.[4] The pooled estimate of the standard deviation of the source distribution is $\sigma = 6.3$. So we assume that we are drawing from a single source distribution (the null hypothesis) or two different source distributions (the alternative hypothesis) that is/are Gaussian, with $\sigma = 6.3$.

Computing the Likelihood Functions

The key calculation is the computation of the likelihood functions for the two samples. We obtain a likelihood function by sliding the assumed source distribution along the abscissa, as shown in the top two panels of Figure 4. At each successive location, we read off the likelihoods of the data points (see the arrows projecting up and over from three of them) and take their product. As we move the distribution around (as we vary its mean), the likelihood of each datum changes (see Figure 4, top two panels). At each location, the product of the likelihoods (one for each datum) is the likelihood of that location. The likelihood function (bottom panel of Figure 4) is the plot of these products as a function of the assumed location. The form of the resulting curve need not be the same as the assumed form of the source distribution, because the parameters of the likelihood function are the data. The parameters of a normal distribution are only two, μ and σ, whereas the parameters of the likelihood function for its mean are the data, of which there may be arbitrarily many.

[3]An objection to this statistical model is that it assumes that there can be data with negative values, which is impossible, because *Q* cannot be less than zero. One might want to assume a distribution that is only supported on the positive reals. However, the common continuous distributions with this property have zero probability density for a zero datum (that is, they are not supported at zero). The data from both groups include a value of zero, so we cannot assume a model that makes such a datum impossible. The common discrete distributions, like the binomial and the Poisson, are supported at zero. However, they are one-parameter distributions, with a standard deviation proportional to the square root of the mean. They cannot be made to fit these data, because the standard deviation in these data is much greater than the square root of the mean.

[4]This is the assumption we make when we do an equal-variance two-sample *t* test, which is probably the most common analysis in the experimental literature.

The Null Prior

The null hypothesis is that the source distributions for the two samples are one and the same (i.e., $\mu_{F\&S} = \mu_{M\&D}$). On this hypothesis, when we come to compute the likely location of the source distribution from which the Few & Sparse data were drawn, we have prior information, because we already made six draws from that distribution when we looked at the data from the 6 subjects in the Many & Dense group (the control or comparison group). Before we looked at the control data, we had no information about what the control mean might be (other than that it would have to be positive). Our prior probability distribution on the possible values of the control mean was uniform over the interval in which the control data in fact fall. Under these circumstances (when there is what is called an uninformative prior), the information we have about the probable location of the mean of this supposedly common source distribution is entirely contained in the probability density function that we get by rescaling the likelihood function for the control data (the function in the bottom panel of Figure 4) to make it integrate to one. Thus, the rescaled likelihood function for the Many & Dense group is our null prior, $\pi(\mu_{F\&S}|H_0)$. It predicts where the mean of the Few & Sparse sample should fall, on the hypothesis that this sample is drawn from the same distribution as the Many & Dense sample.

The First Alternate Prior: Only Trials Matter

This experiment is unusual in that there is an obvious quantitatively explicit alternative to the null. In models of learning in which the trial is the unit of experience—which is to say almost every model ever formally specified, including almost all connectionist models—the progress of learning depends monotonically on the number of trials. The point at which a conditioned response appears is a milestone in that progress. If only trials matter, it should take subjects in the Few & Sparse group eight times as many quarter sessions to reach this milestone, because they experience eight times fewer trials in each quarter session. The prior probability distribution for this alternative hypothesis is the null prior widened rightward along the *Q* axis by a factor of eight and scaled down by that same factor, to keep its integral, that is, the total probability mass, at one.

The Second Alternative Prior (Trials Matter Somewhat)

A vaguer hypothesis is that there is an effect of the number of trials on the progress of learning, but it is not necessarily as big as would be expected if the number of trials were all that mattered. According to this hypothesis, the mean of the distribution from which the Few & Sparse data were drawn might be as much as eight times greater than the mean of the distribution from which the Many & Dense data were drawn, but, on the other hand, it might be much less than this. It is hard to say how much less (theory gives very little guidance here), so we might as well allow for the possibility of no effect at all. In doing so, we nest the null, that is, we make it a special case of our alternative hypothesis, the case in which the effect size is zero. The alternative, however, unlike the null, allows for the possibility of a nonzero effect.

The increment prior—In specifying the size (and sign) of the possible effect entertained by a given hypothesis, we specify what I call the increment prior. The question is, “How big an effect do we contemplate?” We see by looking at the likelihood function for the Many & Dense data at the bottom of Figure 4 that the mean of that source distribution might be as high as 15. Eight times that is 120. Suppose we set the upper limit on the possible effect of deleting seven out of every eight trials at an increase of 120 in the mean *Q*. We have already decided that we will allow the smallest possible effect under this alternative hypothesis to be zero. If we assume that any effect within this range is as likely as any other, then we have an (unnormalized) effect size prior that is uniform between 0 (no effect) and 120 (a huge

effect). We note parenthetically that increments can be negative, that is, there can be decremental as well as incremental effects of an experimental manipulation. A "two-tailed" increment prior allows for effects that lie within some specified range on either side of the control mean. In this case, however, no one expects increasing the number of trials to push back the progress of learning.

Convolving the increment prior with the null prior—The increment prior does not by itself give us an alternative prior probability distribution for the mean of the distribution from which we drew the experimental data, because the zero-increment must be placed at the mean of the distribution from which the control data were drawn, and we do not know where exactly that mean is. Our uncertainty about its location is specified by the null prior. To get a prior for computing the marginal likelihood of the experimental data under this hypothesis, we convolve the increment prior with the null prior. The convolution operation places the zero of the increment prior at successive locations along the abscissa (here, the *Q* axis), scales the increment prior at each location by the probability density of the null prior at that location, then sums point by point all the scaled copies of the increment prior. Intuitively, this operation says, "Well it [the true location of the mean from which the control data were drawn] could be here," placing the rectangular increment distribution well toward the left end of the *x* axis, "but that is very improbable, as we see from the null prior. So we won't give that possibility much weight." Then it moves the zero anchor step by small step to the right, leaving behind after each step successive copies of the increment prior, each copy weighted (scaled) by the corresponding probability density of the null prior. Thus, when the zero is at the peak of the null prior distribution, the convolution says, "Here is a highly probable true anchor point for the increment, so we will give this copy of the increment prior correspondingly greater weight." When it has finished leaving scaled copies of the increment prior at successive small steps along the *x* axis, it goes back and sums point-by-point over all of the copies. The result of convolving the null prior with our increment prior and normalizing to make the resulting distribution integrate to one is the dashed flat-topped probability density distribution in Figure 5.

Computing the marginal likelihoods—Figure 5 plots the functions that enter into the computation of the marginal likelihoods. Plotted against the left axis (probability density), we have three prior probability distributions, representing three alternative hypotheses: the null hypothesis, the 8× hypothesis, and the somewhere-in-between hypothesis. Plotted against the right axis (likelihood), we have the likelihood function for the experimental data. It represents our uncertainty about where the mean of the distribution from which the Few & Sparse data were drawn lies after we have looked at those data.[5] The question is, "Which hypothesis—which prior probability distribution—does a better job of predicting the likelihood function?" The answer is obvious: the null hypothesis does a much better job; it puts the unit mass of prior probability right under the likelihood function, where the experimental data are. The other two hypotheses spread this same mass of prior probability out to the right, away from the likelihood function. They put most of the prior probability where the experimental data are not.

As in computing a correlation, one obtains the marginal likelihood of the experimental data under a given hypothesis by taking point-by-point cross products between the prior probability distribution representing that hypothesis and the likelihood function and summing them. Formally,

[5]Technically, it is the posterior distribution that represents this uncertainty. However, because we are assuming an "improper" prior that is uniform (flat) in the region where the likelihood function is nonzero, the posterior distribution is simply the rescaled likelihood function. The rescaling is of no consequence, because the common scale factor in the numerator and denominator cancels out of the Bayes factor.

$$L(H|\mathbf{D}) = \int L(\mu|\mathbf{D})\pi(\mu|H)d\mu,^{6}$$

where $L(\mu|\mathbf{D})$ is the likelihood function (the likelihood of various values of the source mean, μ, given the data vector, **D**; see, for example, the heavy curve in Figure 5), $\pi(\mu|H)$ is the prior probability distribution associated with a given hypothesis, *H* (see, for example, one of the dashed curves in Figure 5), and $L(H|\mathbf{D})$ is the marginal likelihood of the hypothesis. Note that $L(H|\mathbf{D})$ is simply a number, whereas $L(\mu|\mathbf{D})$ and $\pi(\mu|H)$ and their product are all functions (of the variable μ).

The Bayes factors for the different contrasts is the ratio of the marginal likelihoods. Given what we see in Figure 5, it is no surprise that the Bayes factor for the null versus the 8× hypothesis is 216:1 in favor of the null; for the null versus somewhere-in-between hypothesis, it is 32:1 in favor of the null.

Bayesians dislike the promulgation of "critical values" (alpha levels), which is such a prominent feature of NHST. Telling the research community what odds are required to decide in favor of a hypothesis is a bit like telling the betting community what odds are required to make a safe bet. The odds are what they are. The data favor the hypothesis that has the larger side of the odds ratio. Those in need of advice on how to use odds to make "safe" bets on scientific hypotheses can consult Kass and Raftery (1995, p. 777). Following Jeffreys (1961), they suggest that odds less than 3:1 are weak, odds between 3 and 10 are "substantial," odds between 10 and 100 are "strong," and odds greater than 100 are "decisive." In terms of the absolute value of the weight of the evidence, that is, $\log_{10}(BF)$, a weight less than 0.5 is modest to negligible, a weight in the range from 0.5 to 1 is substantial, a weight greater than 1 is heavy, and a weight greater than 2 is crushing.

Varying the vagueness—The most basic difference between Bayesian analysis and a conventional NHST is that the Bayesian analysis brings an alternative hypothesis into the analysis, whereas NHST never considers alternatives to the null. Alternatives to the null are almost always relatively vague ("some effect"). The outcome of a Bayesian analysis depends on how vague. The greater the range of possible effect admitted into the alternative to the null, the more the analysis favors the null. The question of how vague to make the alternative is the essence of what Killeen (2005) has called "the problem of the prior." There is a general solution: Compute the Bayes factor as a function of the vagueness, that is, as a function of the limits on the increment prior. Figure 6 plots the Bayes factor as a function of the upper limit on the increment prior (keeping the lower limit at zero). What it shows is that the null is literally unbeatable. No matter how small one makes the assumed maximum possible effect of deleting seven out of eight trials, the odds still favor the hypothesis that the deletion has no effect. Varying the vagueness by varying the width of the increment prior is an instance of the more general practice of testing the sensitivity or robustness of a Bayesian analysis (Berger et al., 1994).

Autoscaling the vagueness—Another approach to delimiting the vagueness of an alternative to the null is to let the data tell us the plausible range of possible effects. Approaches to simple Bayesian analysis more formally rigorous than that elaborated here do this by placing a prior on the normalized effect size (see, for example, Rouder et al., in press). If we assume, for example, that possible effect sizes, δ, are themselves normally distributed about zero with, for example, $\sigma_\delta = 1$, then effect sizes greater than three are taken to be a priori very unlikely. A normalized effect size is the possible difference between the source means scaled by the reciprocal of the pooled standard deviation of the samples. In other words, the data determine the scale factor. In the approach here elaborated, a rough

and ready autoscaling of the vagueness is achieved simply by letting the width of the uniform increment prior be equal to the span of the data. The rationale is that the true difference between the means of two source distributions is very unlikely to be greater than the span of the pooled data. When we autoscale the vagueness, we ask, in essence, "Assuming that the difference between the two source means is no greater than the span of our pooled data, which is a better model: one that assumes the two samples come from a single source, or one that assumes they come from sources with different means?"

Rough and ready as this approach to autoscaling the vagueness is, it gives reasonable results. For the possible difference between the Many& Dense and Few & Sparse populations, it gives odds of 2.2 in favor of the null. The unit information prior on normalized effect sizes gives odds of 2.1; the JZS prior—the Cauchy distribution on δ and $p(\sigma^2) = 1/\sigma^2$ on the variance—gives odds of 2.8. (See Rouder et al., in press, for an explanation of these priors and for a website that computes the odds.)

Autoscaling the vagueness allows investigators to follow a convention in posing the alternative to the null. It delivers a single number characterizing the strength of the evidence for or against the null when contrasted with a conventionally vague alternative to it. Thus, it very nearly automates the process of weighing the evidence provided by the data.

Consistency?—This way of structuring simple Bayesian inference is somewhat unorthodox. The question of whether it is consistent arises. Given two samples, A and B, it is often debatable which should be regarded as the control and which the experimental group. Under many circumstances, the hypothesis that the mean of Source Distribution B is greater than the mean of Source Distribution A by some amount and the hypothesis that the mean of Source Distribution A is less than the mean of Source Distribution B by that same amount are interchangeable. When this is true, consistency requires that we get the same Bayes factor regardless of how we frame the question.

Repeated testing with simulated data (see Supplemental Materials) shows that in practice, the computation just described is consistent: It gives the same result regardless of which sample is treated as the control and which the experimental, and it does so even when the sample sizes are grossly different.

Consistency is an issue only when analytic considerations do not preclude symmetrical increment priors. The Gottlieb data are an example in which symmetrical increment priors are not possible, because *Q*, the number of quarter sessions to acquisition, can be arbitrarily large but it cannot be less than zero. Before we set out to test hypotheses about differences or the lack thereof in the means of Gottlieb's two source distributions, we have already fixed both standard deviations at 6.3. A priori, the mean of whatever we take to be the control group can be any value greater than zero, which means that it could be, for example, .01. If we were to frame the comparison by taking the Few & Sparse group as the control and asking whether the mean of the Many & Dense was less than this, then we would know a priori that, depending on what the mean of the Few & Sparse group turned out to be, our maximum possible effect size might be arbitrarily small. Indeed, in the event, a look back at Figure 3 shows that the maximum possible zero-ward effect size would be only slightly bigger than one, whereas one theory clearly predicts an effect size much bigger than that. Our measure (*Q*) has an (analytic) floor but no (analytic) ceiling. Therefore, we must frame the comparison so that the direction of the hypothesized difference is upward. Another way of making this same point is to say that both groups have an informative, purely analytic prior that gives zero probability to all values of *Q* less than zero.

Example 2: Is Performance (With Unattended Triplets) Truly at Chance?

In the experiments by Turk-Browne et al. (2005), subjects saw a stream of 24 different icons (shapes), some red, some green. The icons of a given color were grouped into triplets, which were always presented in a fixed within-triplet sequence during the familiarization phase. The appearance of the first member of a triplet infallibly predicted that the next icon *of that color* would be the second member of that triplet and that the next after that (of that color) would be the third. Icons from the two different color categories were, however, randomly interleaved, so the next one or two or three icons that appeared in the mixed stream might be from the other color category. To induce the subjects to attend to icons of one color and not the other, the experimenters occasionally repeated the third item of a triplet in the to-be-monitored color. Subjects were told to monitor that color stream and to press a key whenever they observed two identical icons of that color occur in sequence. In the familiarization phase, subjects saw streams in which each of the four triplets within each of the two colors was repeated 24 times, with randomization of the triplet orders within each color and random interleaving of the icons from the two color streams. On the 64 trials of a test phase, subjects saw two triplets of icons, all black. One was a triplet that had reappeared 24 times in the familiarization phase, albeit in color (and with other icons of the other color interleaved). The other was an unfamiliar triplet consisting of three different icons they had seen but not in any of the triplet orders they had seen. On half the test trials, the familiar triplet was from the attended color stream, whereas on the other half, it was from the unattended stream. Subjects were asked to press one of two keys, indicating which of the two triplets seemed more familiar.

In four versions of this experiment, which varied the exposure duration and whether the color at test was or was not the color during familiarization, Turk-Browne et al. (2005) consistently found that the mean proportion correct (out of 32) on the attended triplets was significantly above chance, whereas the mean proportion correct on the unattended triplets was not. Again, intuitively, their data support the null hypothesis that nothing was learned about the sequential dependencies among icons of the unattended color. However, conventional analysis allows us to say only that we cannot reject the hypothesis that performance is at chance for the unattended items. This is an oddly contorted way of describing the data, given that all four mean proportions were at or slightly below chance. Moreover, the theoretical interest is not in whether we can reject the null hypothesis but, rather, in whether we can accept it. That is the conclusion for which Turk-Browne et al. understandably argue, even though their use of conventional statistical analysis did not allow them to marshal statistical support for it. Bayesian analysis supports their conclusion, except for one subject, out of the 34 tested across the four experiments.

The only place in experimental psychology where Bayesian analysis has made inroads is in the fitting of competing regression functions to data (see Lee, 2008, for a recent example). In that literature, it is well understood that averaging across subjects before fitting a model often leads to odds favoring models that are not the models that best describe the data from individual subjects, a fact whose importance has been stressed by Estes for more than 50 years (Estes, 1956; Estes & Maddox, 2005). It is seldom noted that the same consideration applies to simple hypothesis testing: If there are theoretically consequential differences between subjects, the group mean is meaningless. (No one has 2.85 cone types.) Therefore, before computing group statistics, I asked whether the data did or did not support the assumption of subject homogeneity. (The details of this analysis may be found in the Supplemental Materials.)

In all but the fourth experiment, when subjects were tested with unattended triplets, the odds either strongly favored the hypothesis that subjects had a common underlying *p* value or

they only weakly favored the heterogeneous *p* hypothesis. Moreover, the estimate of the hypothesized common *p* was very close to .5 in all three cases (.49, .49, and .50, respectively). This analysis already favored the hypothesis that when subjects have not attended to a stream, they have not detected the sequential dependencies in it. In the fourth experiment, the odds favored the heterogeneous-*p* hypothesis, but this was largely due to one subject, who got 27 out of 32 correct. One might conjecture that this subject deduced the design of the experiment and estimated dependencies by conscious procedures unconnected with the (unconscious) statistical analysis procedure under investigation. An informal debriefing of this subject tended to confirm this conjecture (Brian Scholl, personal communication, March 14, 2008).

By contrast, for the attended color, the odds generally favored the hypothesis that different subjects ended up with different underlying probabilities of correctly detecting a familiar triplet. Simple inspection reveals that some subjects had a very high probability, whereas others were at chance. The only experiment in which this was not true was the first one, where the exposure duration was short, and all of the subjects had a rather low probability of detecting the familiar triplets. In that experiment, the odds favored the hypothesis that there was a homogeneous cross-subject *p* value for the attended color stream.

A more sophisticated approach would use a hierarchical model that computes a representation of the uncertainty about the true value of each subject's probability parameter and the uncertainty about the distribution of those values in the population of subjects. Methods for doing this are well developed, but the additional complexity would be antithetical to my purpose, which is to show the power and simplicity of elementary Bayesian analysis. Here, these preliminary results were used only to decide when to pool data across subjects and when not to do so. For the unattended color in Experiments 1a, 1b, and 2a, and for the attended color in Experiment 1a, pooling the 8 subjects × 32 trials gave 256 trials (in each case), and summing the number of correct responses across subjects gave the *K*, the number of "successes." The source distribution is the binomial distribution, with $N = 256$ and K = number of successes (total correct responses). The likelihood function is the probability of getting those *K* successes in 256 trials as a function of the assumed value of the parameter, *p*, which is the underlying probability of a success. The likelihood functions for three of the four data sets are the thick curves in the left panels in Figure 7 (plotted against the right axes, as usual).

The marginal likelihood of the data, given a prior probability distribution, is the sum of the probability mass at different points weighted by the corresponding values of the likelihood function. The null hypothesis (chance responding) puts the entire prior probability mass at .5, so the marginal likelihood of the null is simply the value of the likelihood function at .5. This null prior is merely suggested by the heavy dashed vertical lines at .5 in the panels on the left in Figure 7. These lines are infinite in height (technically, they are Dirac delta functions), but, like all probability distributions, they integrate to one, that is, the total probability mass is one, as always.

The alternative hypothesis that the underlying probability of a correct response is greater than chance spreads the unit probability mass evenly over the interval from .5 to 1. The height of the resulting rectangular priors is 2 (because $.5 \times 2 = 1$). In the top two panels on the left of Figure 7, one sees immediately why the chance hypothesis is more likely in the light of these data. The heavy dashed vertical lines at .5 intersect the likelihood functions very near their peaks, so the entire probability mass gets nearly the maximum possible weight from the likelihood functions. By contrast, the rectangular alternative priors spread much of the probability mass well out to the right, where it gets negligible weight from the likelihood function.

Because the peaks of the likelihood functions in the top two left panels are at or to the left of the peak of the null prior, any rightward spread of the probability mass lowers the marginal likelihood. Any hypothesis that spreads any probability mass above chance is less likely than the null hypothesis, which puts it all at chance. This is confirmed by the panels on the right, which plot the odds favoring the null, as a function of how vague we are about how far above chance the probability of success might possibly be. As one brings the upper limit on the possible probability down to the lower limit at chance, the odds favoring the null approach one from above. The graphs for the unattended color in Experiment 1b are identical to those in Figure 7A. Thus, for three of the four experiments, the data from all subjects combined strongly favor the null hypothesis that responding to items with the unattended color was at chance. The odds on the null are 17, 17, and 13 to 1. Moreover, in all three cases, the null is unbeatable.

On the other hand, when given the data on number correct from items of the attended color in Experiment 1a, the null cannot compete with *any* vaguer alternative, as shown by the plots in Figure 7C. The likelihood function for these data has fallen almost to 0 at .5, so it gives very little weight to the probability mass from the null hypothesis. The odds are 8 to 1 against the null, even when the alternative spreads the probability mass over the entire above-chance interval. When the upper limit on the possible is reduced to .6, the odds grow to 28 to 1 against the null.

Given the other four data sets (the data from the attended color in Experiments 1b, 2a, and 2b and from the unattended color in Experiment 2b), we must analyze them subject by subject, because the above-mentioned preliminary analysis showed that it was unlikely that the subjects had a common probability of success in these conditions. The analysis of a single subject's data is the same as with the pooled data, only it is done with $N = 32$ (the number of trials for a single subject) and K = however many correct responses that subject made.

We see in Figure 8 that on the test with unattended triplets, the data from the great majority of subjects favor the null hypothesis, not only in Experiments 1a, 1b, and 2a but also in 2b. How strongly they favor it depends on the hypothesis against which it is pitted. When pitted against the hypothesis that p is anywhere in the above-chance interval (black bars), the odds usually (but not always) favor the null more strongly than when it is pitted against the more precise hypothesis that p is only slightly greater than .5. However, the odds consistently favor the null, even in the latter comparison.

In 34 subjects, there is only one clear exception: For the third subject from the top in Figure 8, both rightward-projecting bars go off scale; the odds against the null for this subject were more than 1,000:1 when it was pitted against the vaguer hypothesis and almost 70:1 when it was pitted against the less-vague alternative. The odds are lower in the latter case because the percentage correct was substantially greater than allowed by this hypothesis, which puts all the probability mass substantially *below* the proportion of successes observed in this subject. This was the experiment in which we could decisively reject the hypothesis that all the subjects had the same probability of success when tested on unattended triplets. Clearly, the rare subject can either learn sequential dependence in unattended sequences or can respond on the basis of conscious assessments not made by most subjects and not based on unconscious statistical learning. In fact, this subject did substantially better on the unattended triplets than on the attended triplets, suggesting that he/she tried consciously to detect the dependencies in both color streams.

The picture from the tests with attended triplets is mixed. A considerable majority of subjects do learn some or even all of the sequential dependencies in attended triplets. How

many they learn, however, varies dramatically from subject to subject; some subjects seem not to learn any (see the leftward projecting bars on the "Attended" side of Figure 8). Because the methods of Turk-Browne et al. (2005) are similar to those used by other investigators, one suspects that the same conclusion would emerge if the data from other reports of statistical learning were analyzed on a subject-by-subject basis.

Example 3: Are Factors Additive? (Is There No Interaction?)

Two factors have an additive effect on some measure if the size of the effect of a given change in one factor is the same regardless of the level (value) of the other factor. Thus, "proving" additivity requires proving a null hypothesis of the form "the magnitudes of these differences are the same."

The application to this quite fundamental theoretical issue of the already-explained methods for supporting null hypotheses through Bayesian analysis is illustrated by the analysis of data from a recent experiment by Rosenbaum et al. (2006).[7] They investigated the effects of target height and target diameter on height at which subjects grabbed an everyday implement (a plunger) to move it back and forth between a "home" location and a target location. The target location was a ring on a shelf. There was also a ring at the home location. There were two dimensions of variation: the height of the shelf and the diameter of the rings (hence the precision with which the plunger bulb had to be guided into or out of them). The experiment showed that (some) subjects took both factors into account when they grasped the plunger's handle: the higher the shelf, the lower their grasp, and the smaller the diameter of the target ring, the lower their grasp. A repeated-measures analysis of variance (ANOVA) gave significant main effects for both factors and an insignificant interaction, but this latter is, of course, not what is wanted; what is wanted is a measure of the strength of the evidence *for* additivity.

There were five different shelf heights (five different levels of the shelf factor) and four variations on the two rings, the ring from which the plunger was removed and the ring into which it was placed. In one condition (the easy-easy, henceforth EE) condition, both rings were wide. In the other three (easy-hard [EH], hard-easy [HE], and hard-hard [HH]), one or both were narrow. Subjects moved the plunger between the home and target locations and then back again twice, giving four grasps within each of the $5 \times 2 = 10$ combinations of factor levels. The main effect of ring diameter was seen only in the EE condition: When either or both of the rings was/were narrow, the average subject grasped relatively lower (closer to the plunger's bulb, which had to be removed from one ring and steered into the other), but when both were wide, the average subject grasped higher. The question of additivity bears critically on the model of the complexity of the movement planning process: Interactive models are more complex.

Although the published repeated-measures ANOVA showed a significant effect of the ring diameter, my preliminary Bayesian subject-by-subject analysis (described in the Supplemental Materials) showed that there was an effect in some subjects but not in others. It does not make sense to include in an analysis for additivity subjects whose behavior is unaffected by the ring-diameter factor. The question only arises when varying one factor has a clear effect within at least some levels of the other factor. The larger the effect is, the more powerful the analysis for additivity becomes, because there is more room within which to detect differences in the size of the effect at different levels of the other factor. To that end, the data from Subject 7, the subject who showed the biggest effect of the variation in ring diameter, were selected for analysis.

[7]David Rosenbaum generously supplied the raw data for 10 of the subjects.

The ANOVA test for additivity does not test for the systematicity of the effects. Any pattern of departures from a constant size of effect as one shifts from level to level of the shelf-height factor, if large enough, yields a significant interaction. Arbitrary, subject-specific patterns of departure are of little scientific interest. What we would like to know is, in those subjects for whom the ring-diameter is taken into account in their movement planning, does the effect of ring diameter tend to converge toward zero as the shelf gets higher? If so, then the assessment of shelf height enters into the subject's determination of how much of an adjustment to make in response to a smaller ring diameter—an interaction. If not, then the effect of this factor on movement planning is independent of the subject's assessment of shelf height.

The simplest monotonically converging pattern is linear convergence. If we pose as an alternative to the null hypothesis the hypothesis that the effect of a difference in ring diameter converges linearly toward zero as the shelf height increases, then our alternative only has one additional free parameter, namely the slope of the EE data. On the null hypothesis (no convergence), this slope is the same as the slope of the base data (the data from the EH, HE, and HH conditions). On the convergence hypothesis, this slope is somewhat greater, so that as shelf height increases, the effect of a difference in ring diameter decreases linearly.

Figure 9 portrays the computation graphically. The means of the three base conditions (EH, HE, and HH) are plotted as asterisks on the base plane of the graph. The axes of this base plane are shelf height and mean grasp location. The means of the EE condition are plotted on the base plane as open circles. Also plotted on this base plane are the maximally likely regression lines, computed in the usual least-squares way. (The line that minimizes the squared residuals is the maximally likely regression line on the assumption that the residuals are drawn from a normal source distribution with a constant sigma, a sigma that is the same at all locations along the line.) The two slopes are approximately equal, so the evidence for convergence is, at best, not strong. A conventional analysis compares the slope estimates using a *t* test. The *t* value does not approach significance, from which the conventional analysis concludes that we cannot reject the hypothesis that the two slopes are the same. But that is not an assessment of how strong the evidence is that they *are* the same. To get that, we compute the marginal likelihood function for the slope of the EE line using two different prior probability distributions, one based on the null hypothesis, which is that the slope of the EE line is the same as the slope of the base line, the other based on the hypothesis that the slope of the EE line is “somewhat” greater than the slope of the base line. As always, we have to put an upper limit on how much greater we think that slope might be. It does not seem reasonable to suppose that the regression lines would actually cross. If they did, then for some shelf height, the grasp height for a precise placement would actually be higher than the grasp height for an imprecise placement of the plunger bulb. Therefore the slope increment that makes the line through the EE data intersect (converge on) the line through the base data at the highest shelf seems a reasonable upper limit.

The analysis proceeds as usual (see details in the Supplemental Materials): (a) Compute the likelihood function for the slope of the baseline data and normalize it to get the prior probability distribution for the null hypothesis (parallel regression lines). (b) Convolve this distribution with the increment prior (uniform between 0 and −.15) to get the prior probability distribution for the alternate hypothesis (converging regression lines). (c) Compute the likelihood function for the slope of the EE data. At this point one can make the usual graph (see Figure 10), which tells us all we really need to know about which hypothesis the data favor and how strongly. (d) Compute the marginal likelihoods and take their ratio to get the Bayes factor.

The Bayes factor gives odds of 2:1 in favor of the null. These are not strong odds, and we may feel that had we not set the width of the increment prior as wide as we did—had we been less vague about what a "somewhat" greater slope might be—they would be even weaker. Figure 11 plots the Bayes factor as a function of (negative) limit on the increment prior used in obtaining the convergence prior. Figure 11 shows that there is no way to improve on the null hypothesis. Thus, the data favor the additivity hypothesis (no interaction) over the interaction hypothesis, but not strongly. Strengthening the evidence should focus on identifying those subjects who take both factors into account and running more trials with them.

Discussion

Proving the null is centrally important to the development of theory in psychology, cognitive science, and neuroscience. Claims of no effect or no difference in effect very frequently enter into the argument for the theoretical conclusions drawn from experimental results. Often, the claims of no effect are as important to the argument as are claims that under other conditions (or with other variables), there is an effect. A claim that there was no effect or no difference was the central focus in the examples here discussed. A theoretically important claim of no effect is even more often seen in articles that contrast one experimental manipulation, which is said to have no effect, with another, which is said to have a "significant" effect. Almost always, NHST analyses are used to support both conclusions, but such an analysis is in principle incapable of supporting the first one. Moreover, as has often been pointed out (see Rouder et al., in press, for examples and citations), the support offered for the second one—in the form of a "significant" *p*—may be misleading, because *p* values may seriously misrepresent the strength of the evidence against the null. To take an example from Rouder et al. (in press), it is possible to find an effect on the mean reaction time that is significant at the $p = .02$ level when a comparison of the null to a well-motivated alternative gives odds of 3,800:1 *in favor* of the null. Because of the theoretical importance of null hypotheses, experimentalists should adopt methods of statistical analysis that measure the strength of the evidence *for* them as well as *against* them. As one bonus for adopting the Bayesian mode, we are forced to consider—and to limit plausibly—the vagueness of the alternatives that we pose, thereby counteracting the temptation to be maximally vague to minimize the risk of being wrong. As a second bonus, we get valid measures of the strength of the evidence in favor of these alternatives. The *p* values from NHST are not such measures, although they are often treated as if they were.

One conceptual problem and one practical problem have stood in the way of the wider use of the normative (that is, Bayesian) form of statistical inference. The conceptual problem is the problem of the prior on the hypothesis posed as the alternative to the null: How vague should it be? The general solution to the conceptual problem is a sensitivity analysis: Compute the odds for and against the null as a function of the limit(s) on the increment prior, that is, as a function of the assumed maximum possible size of an effect (see Figures 6, 7 [right panels], and 11). The null is rejected only when this function has a minimum substantially below the odds reversal line (see Figure 7C2). As a bonus, the location of such a minimum is an estimate of the effect size. If the function approaches the reversal line from above as the maximum assumed effect size approaches zero, then the null is unbeatable. Because it is more precise and assumes less (equivalently, is a less complex model), it should be accepted. Alternatively, one can use the scale of the data to autoscale the vagueness, either by putting a prior on the normalized effect size (Rouder et al., in press) or by setting the range of the increment prior equal to the span of the data.

Another aspect of the conceptual problem has been a lack of understanding of the basic idea behind Bayesian analysis. Graphic presentation of the competing prior probability density

functions, together with the likelihood function of the experimental data, renders simple Bayesian analysis graphically transparent (see Figures 5, 7 [left panels], and 10). Looking at the graphs, one sees that the question is simple: Which prior probability density function better matches the likelihood function? The approximate answer is usually graphically obvious. The Bayes factor puts a number on it.

The practical problem has been the difficulty of the computations. Recently developed software greatly reduces this difficulty. Rouder et al. (in press) give a website that computes the one- and two-sample Bayes factors using two different priors on normalized effect sizes. The online Supplemental Materials for this article give Matlab code that reduces the above-described computations of the Bayes factors for the two-sample (normal source) comparison and the comparison with chance (Bernoulli source) to a call to one of two custom Matlab functions. Each function returns the weight and the Bayes factor and generates the graph of the competing priors together with the likelihood function. The graph makes it visually apparent what the Bayes factor represents, namely, the extent to which the prior distributions posited by the competing hypotheses match the likelihood function of the experimental data. Non-Matlab users may access these computations online at http://ruccs.rutgers.edu/faculty/GnG/gallistel.html or download them in the form of Microsoft Excel plug-ins that run on Windows machines.

Acknowledgments

I thank Peter Killeen, Michael Lee, Geoffrey Loftus, and Eric-Jan Wagenmakers for helpful comments and suggestions.

References

Balsam P, Gallistel CR. Temporal maps and informativeness in associative learning. Trends in Neurosciences 2009;32(2):73–78. [PubMed: 19136158]

Berger J, Moreno E, Pericchi L, Bayarri M, Bernardo J, Cano J, et al. An overview of robust Bayesian analysis. TEST 1994;3(1):5–124.

Estes WK. The problem of inference from curves based on group data. Psychological Bulletin 1956;53:134–140. [PubMed: 13297917]

Estes WK, Maddox WT. Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. Psychonomic Bulletin & Review 2005;12(3):403–409. [PubMed: 16235625]

Gallistel CR, Balsam PD, Fairhurst S. The learning curve: Implications of a quantitative analysis. Proceedings of the National Academy of Sciences of the United States of America 2004;101(36): 13124–13131. [PubMed: 15331782]

Gallistel CR, Gibbon J. Time, rate, and conditioning. Psychological Review 2000;107:289–344. [PubMed: 10789198]

Gibbon, J.; Balsam, P. Spreading associations in time. In: Locurto, CM.; Terrace, HS.; Gibbon, J., editors. Autoshaping and conditioning theory. New York: Academic Press; 1981. p. 219-253.

Good IJ. Weight of the evidence, corroboration, explanatory power, information and the utility of experiments. Journal of the Royal Statistical Society, Series B 1960;22:311–322.

Gottlieb DA. Is the number of trials a primary determinant of conditioned responding? Journal of Experimental Psychology: Animal Behavior Processes 2008;34:185–201. [PubMed: 18426303]

Jeffreys, H. Theory of probability. 3rd. Oxford, England: Oxford University Press; 1961.

Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association 1995;90:773–795.

Killeen PR. Replicability, confidence, and priors. Psychological Science 2005;16(12):1009–1013. [PubMed: 16313669]

Lee MD. Three case studies in the Bayesian analysis of cognitive models. Psychonomic Bulletin & Review 2008;15(1):1–15. [PubMed: 18605474]

Mackintosh NJ. A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review 1975;82:276–298.

Morris RW, Bouton ME. Effect of unconditioned stimulus magnitude on the emergence of conditioned responding. Journal of Experimental Psychology: Animal Behavior Processes 2006;32:371–385. [PubMed: 17044740]

Myung IJ, Pitt MA. Applying Occam's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin & Review 1997;4(1):79–95.

Papachristos EB, Gallistel CR. Autoshaped head poking in the mouse: A quantitative analysis of the learning curve. Journal of the Experimental Analysis of Behavior 2006;85:293–308. [PubMed: 16776053]

Pearce JM, Hall G. A model for Pavlovian learning: Variation in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review 1980;87:532–552. [PubMed: 7443916]

Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH.; Prokasy, WF., editors. Classical conditioning II. New York: Appleton-Century-Crofts; 1972. p. 64-99.

Roberts, S.; Sternberg, S. The meaning of additive reaction-time effects: Tests of three alternatives. In: Meyer, DE.; Kornblum, S., editors. Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience. Vol. 14. Cambridge, MA: MIT Press; 1993. p. 611-653.

Rosenbaum DA, Halloran ES, Cohen RG. Grasping movement plans. Psychonomic Bulletin & Review 2006;13(5):918–922. [PubMed: 17328395]

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson GJ. Bayesian *t*-tests for accepting and rejecting the null hypothesis. Pychonomic Bulletin & Review. in press.

Sternberg, S. Discovering mental processing stages: The method of additive factors. In: Scarborough, D.; Sternberg, S., editors. An invitation to cognitive science Volume 4: Methods, models and conceptual issues. Cambridge, MA: MIT Press; 1998. p. 703-863.

Turk-Browne NB, Jungé JA, Scholl BJ. The automaticity of visual statistical learning. Journal of Experimental Psychology: General 2005;134:552–564. [PubMed: 16316291]

Wagenmakers EJ. A practical solution to the pervasive problem of *p* values. Psychological Bulletin & Review 2007;14:779–804.

Wagner, AR. SOP: A model of automatic memory processing in animal behavior. In: Spear, NE.; Miller, RR., editors. Information processing in animals: Memory mechanisms. Hillsdale, NJ: Erlbaum; 1981. p. 5-47.

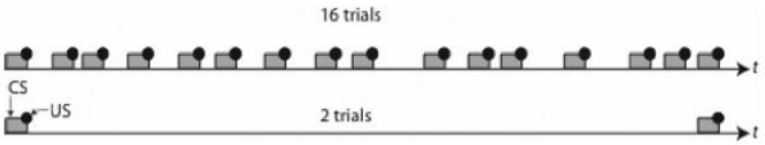

Figure 1.
Schematic representation of (a portion of) two different classical conditioning protocols of equal duration but with an eightfold difference in the number of trials (CS-US pairings) in a given amount of time (*t*). Gottlieb (2008) tested the hypothesis that these two protocols have equivalent effects on the progress of learning.

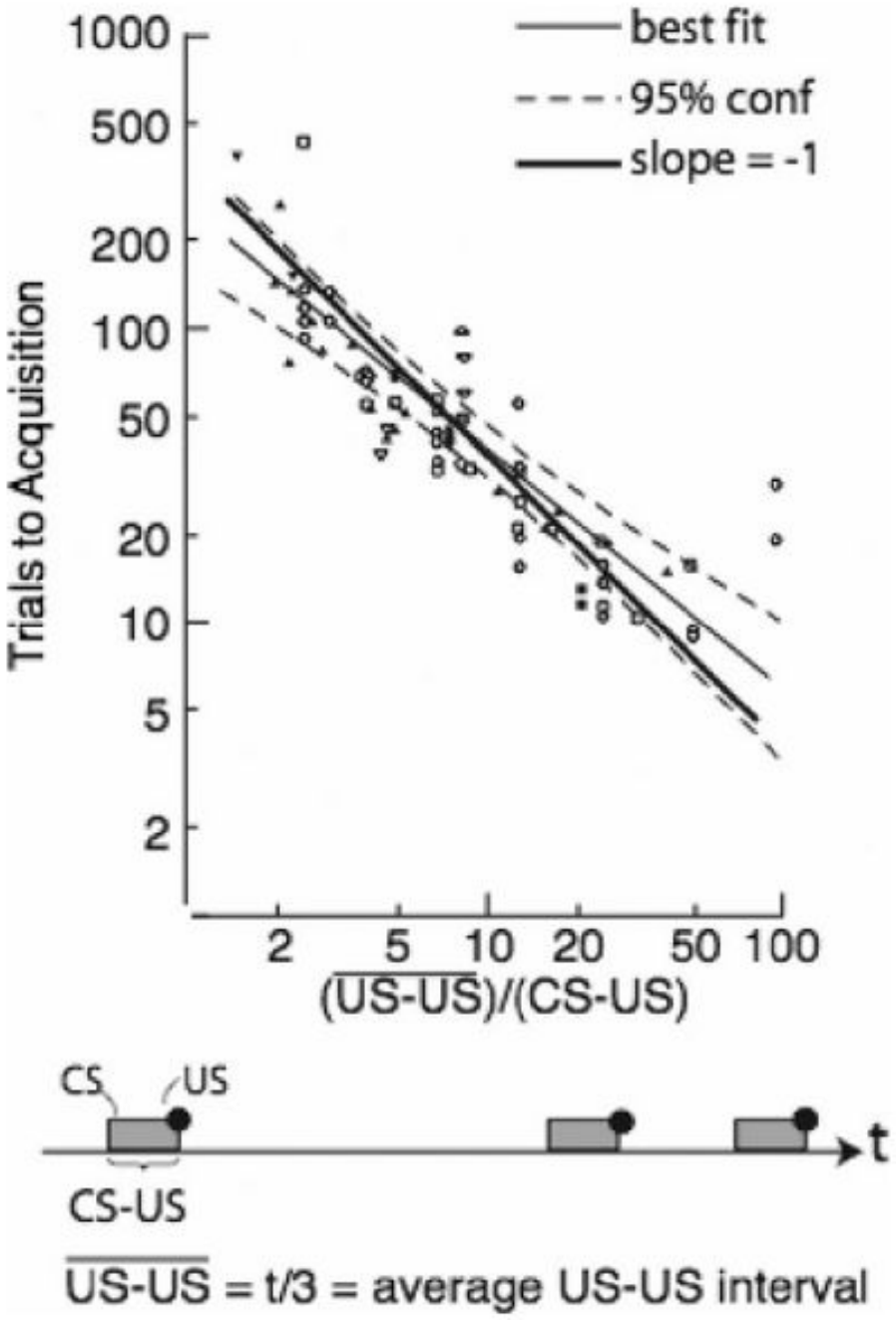

Figure 2.

The number of reinforced trials (CS-US pairings) to acquisition in a standard classical conditioning protocol with pigeon subjects, plotted against the ratio of the US-US and CS-US intervals, on double logarithmic coordinates. (Replotted from Gibbon & Balsam, 1981.) US-US is the average interval between USs (a.k.a. reinforcements). CS-US is the duration of the warning interval, commonly called the delay of reinforcement. As the (US-US)/(CS-US) ratio grows, this delay becomes relatively small, making it a relatively better predictor of imminent reinforcement. That is, the CS becomes more informative (Balsam & Gallistel, 2009). The slope of the regression (light solid line) does not differ significantly from −1 (heavy solid line). If it truly is −1 (itself a null hypothesis), then when trials are deleted, the

increase in the informativeness of the CS precisely compensates for the decrease in the number of trials, in which case the number of trials is not itself important. What is important is the informativeness of the trials. t = time.

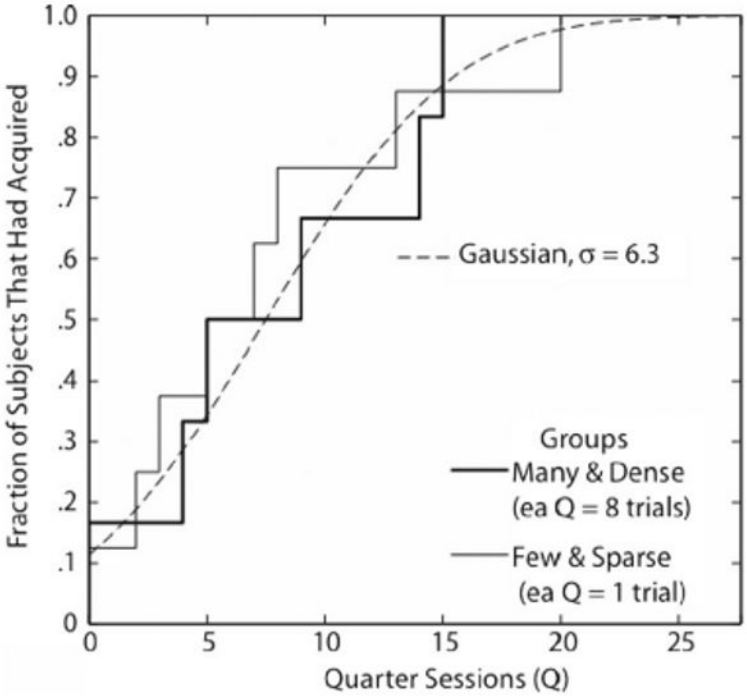

Figure 3.

Cumulative distributions of quarter sessions to acquisition for two groups in Gottlieb's (2008) Experiment 4. These empirical cumulative distributions step up at the locus of each datum. Thus, the number of steps indicates the N ($N_{M\&D} = 6$; $N_{F\&S} = 8$), and the location along the Q axis of any one step is the Q for one subject. The dashed curve is a cumulative Gaussian with a standard deviation of 6.3, which is the pooled maximum unbiased estimate of the standard deviation of the distribution from which the data are drawn (assuming that the source distributions have a common variance, but not necessarily a common mean).

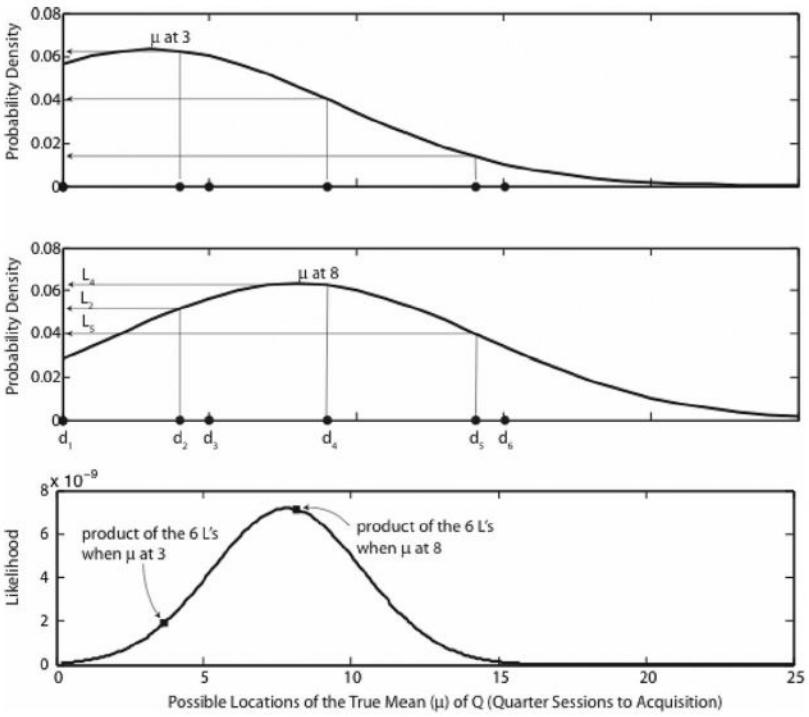

Figure 4.

Computing a likelihood function. The assumed source distribution (the statistical model) is slid along the abscissa (here the μ axis), on which are plotted the known data (solid dots, labeled d_1–d_6 in middle panel, which are the data from the Many & Dense group). The top two panels show it at two locations. At each location, the likelihoods are read off (arrows projecting up from 3 of the 6 data points and over to the corresponding likelihoods). The product of the likelihoods is the likelihood of that location (that value of μ), given the data. The likelihood function (bottom panel) is the plot of the likelihood as a function of the possible locations of the source distribution, possible values of μ_Q. Note that the area under the likelihood function is nowhere near one (numbers on ordinate of bottom panel are $\times 10^{-9}$).

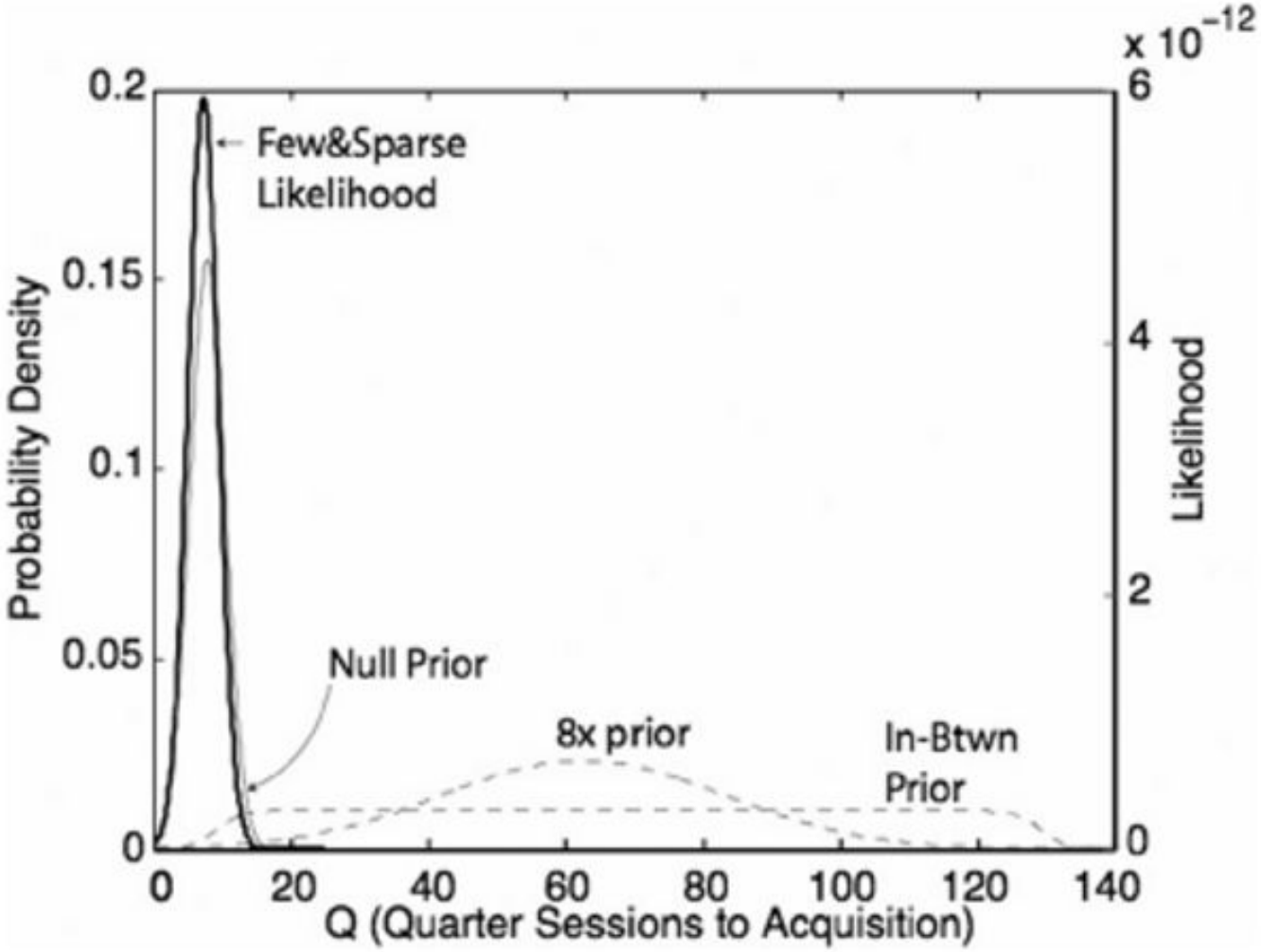

Figure 5.

The likelihood function for the mean of the one-trial-per-quarter-session group (heavy curve) and the three prior probability functions corresponding to three different hypotheses: (a) the null hypothesis, which is that the quarter-sessions-to-acquisition data from this group were drawn from the same distribution as the data from the eight-trials/quarter sessions group; (b) the 8× hypothesis, which is that only trials matter, in which case the data from the one-trial group were drawn from a distribution eight times wider; (c) the vague hypothesis that the effect of reducing the number of trials per quarter session from eight to one has an effect somewhere within the range delimited by the null and the 8× hypotheses. The prior probability distributions are plotted against the left axis. They all, of course, integrate to one. The likelihood function is plotted against the right axis; it does not integrate to one.

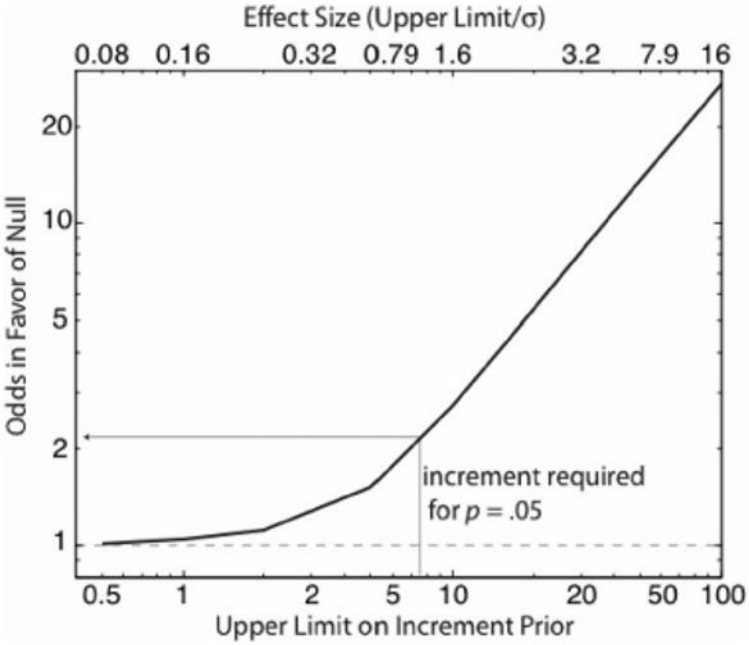

Figure 6.

The odds in favor of the null as a function of the assumed upper limit on the possible size of the effect (double logarithmic coordinates). The dashed line at 1 is where the odds ratio reverses (from favoring the null to favoring the vaguer alternative). Because the plot approaches this reversal point from above as the limit on vagueness goes to 0, the null is unbeatable by any alternative that posits some effect, no matter how small. The thin arrow shows the difference in the sample means that would be just significant at the .05 level. The odds are better than 2:1 against the hypothesis that the effect of deleting seven out of every eight trials is so small that it would be *at most* just detectable by a conventional null hypothesis significance test with samples this size.

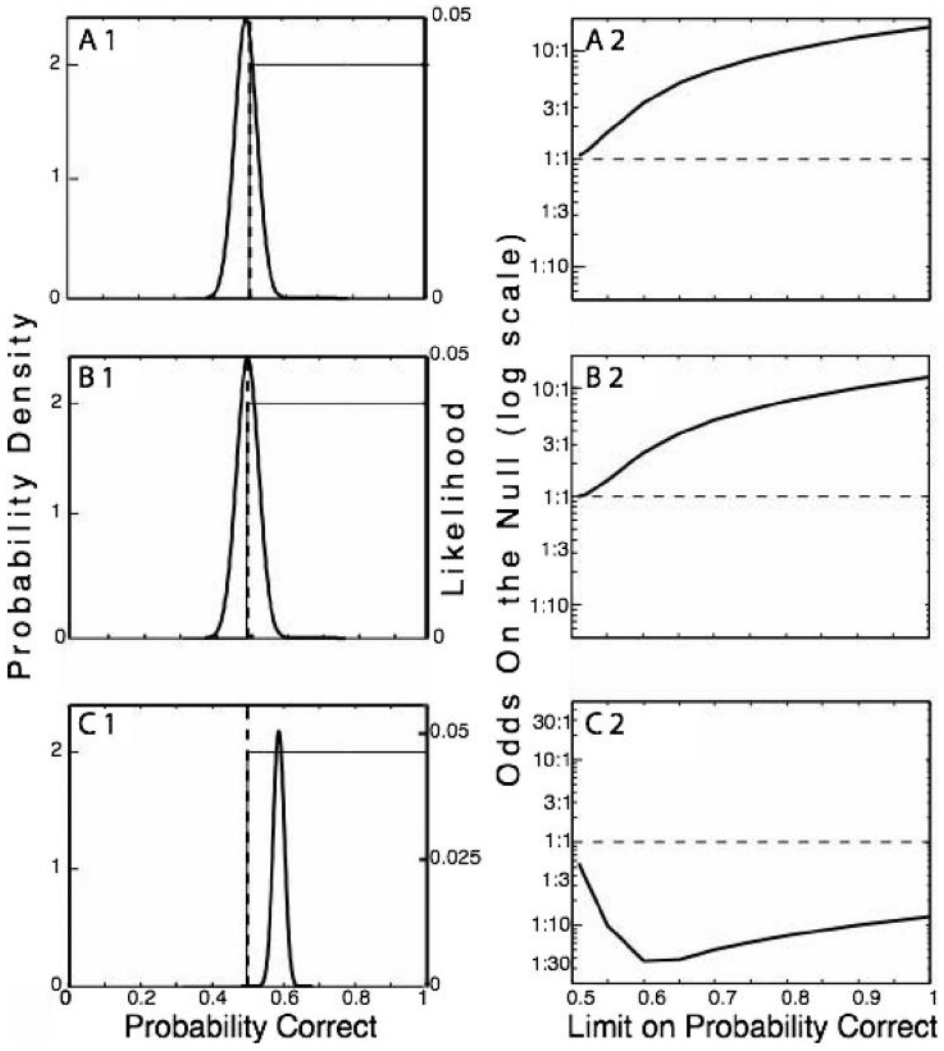

Figure 7.

Left: Likelihood functions (heavy curves, plotted against right axes) and competing prior probability functions (plotted against left axes) for chance (heavy vertical dashed lines at *p* = .5) and for greater than chance (light rectangles with height = 2, with left edge at .5 and right edge at 1). Right: Odds on the null as a function of the upper limit on the possible probability of a correct. A. Turk-Browne, Jungé, and Scholl's (2005) Experiment 1a, unattended color. B. Turk-Browne et al.'s Experiment 2a, unattended color. C. Turk-Browne et al.'s Experiment 1a, attended color.

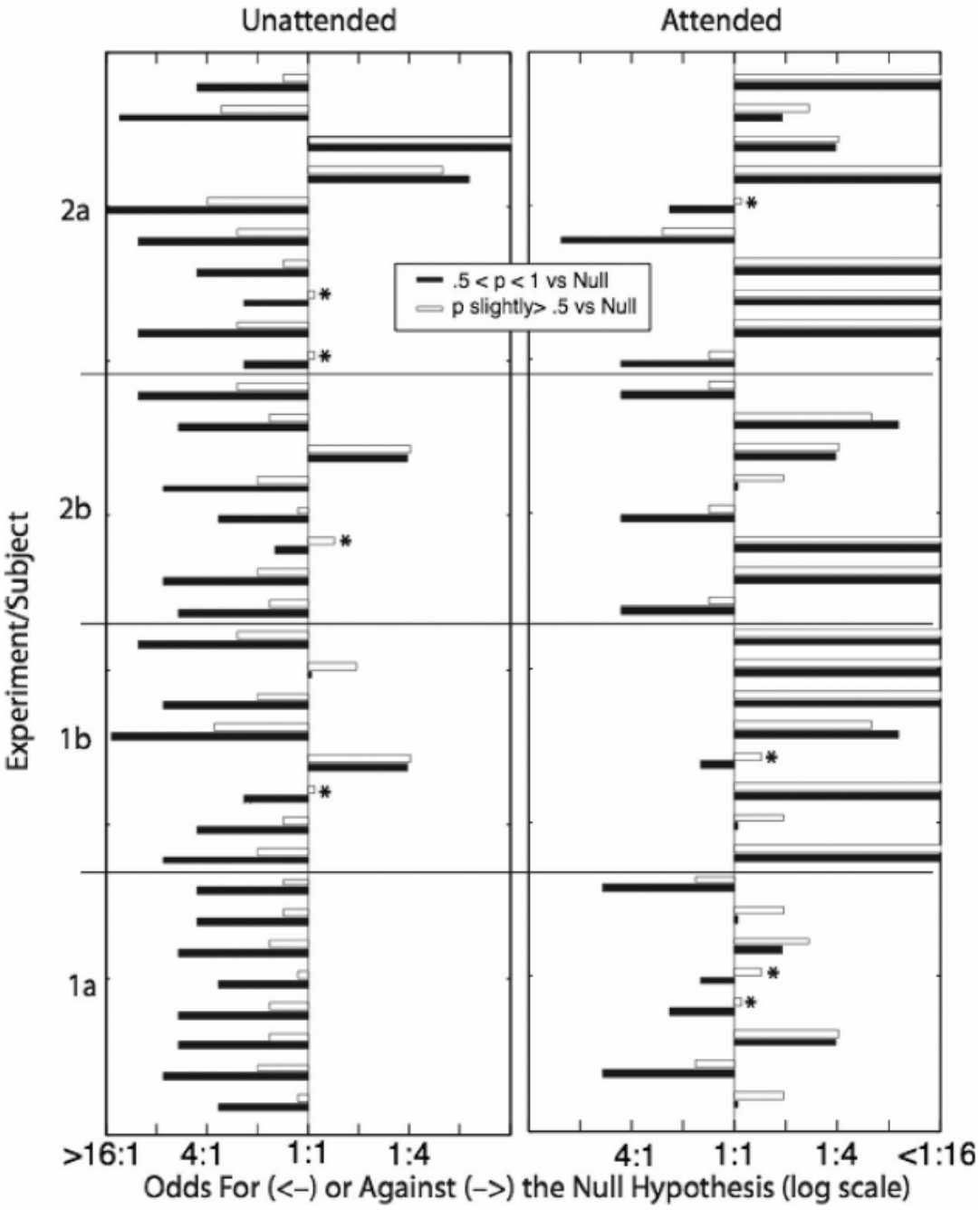

Figure 8.

Odds for (leftward projecting bars) or against (rightward projecting bars) the null hypothesis (chance), for each of 34 subjects in four experiments with tests of unattended and attended triplets. For each subject and each test, the null is pitted against two different alternatives: the probability of correct identification lies anywhere on the interval from .5 to 1 (black bars), or it lies between .5 and .65 (i.e., it is at most only slightly greater than chance). Asterisks mark instances in which the odds favor the null when pitted against the vaguer hypothesis but are against the null when it is pitted against the less vague alternative.

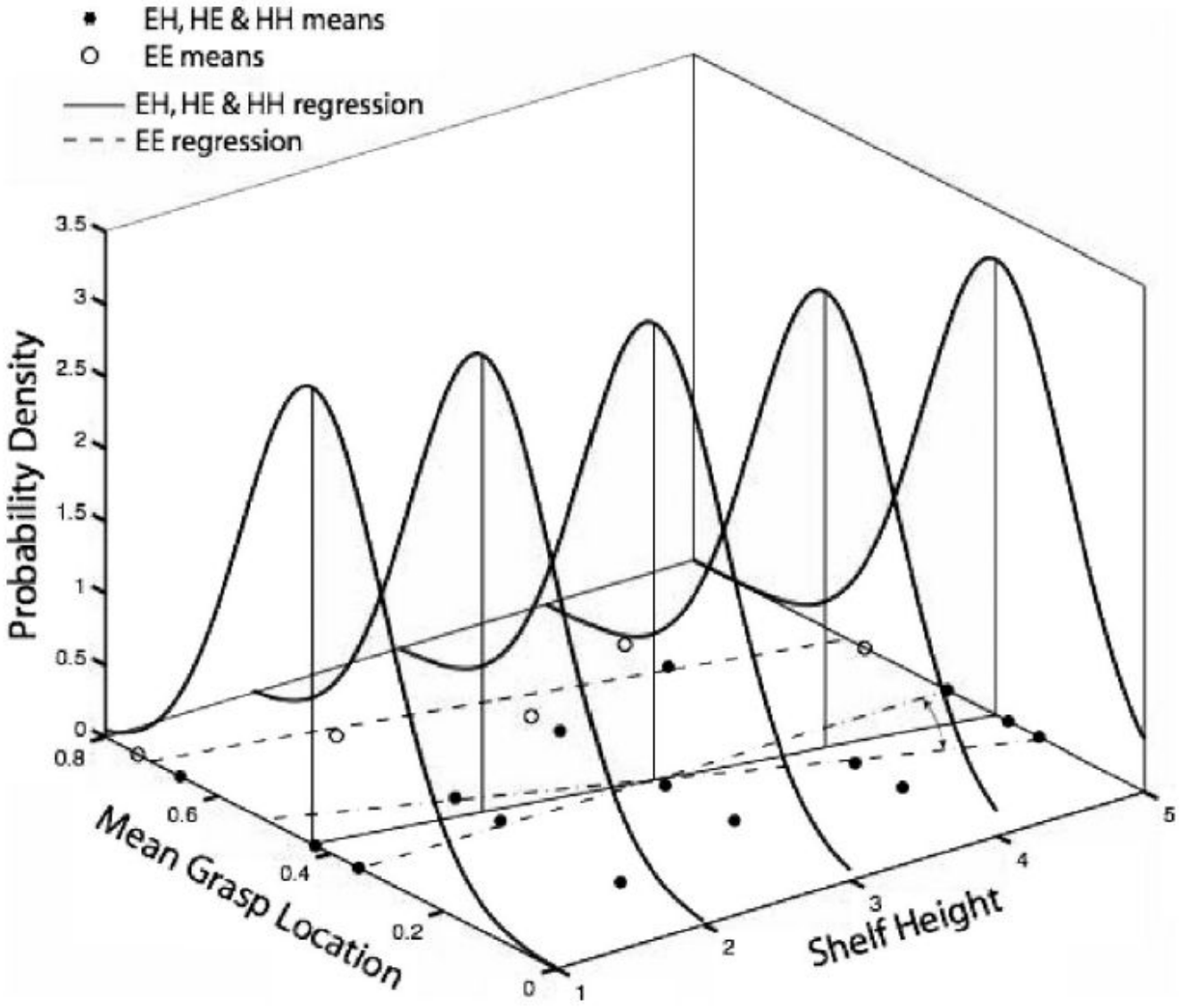

Figure 9.

The heavy solid curves are copies of a common source distribution. Each copy is centered on the regression line, which, like all regression lines, is constrained to pass through the centroid of the data (at Level 3). The positions of the copies at Levels 1, 2, 4, and 5, relative to the underlying data (hence, also the likelihoods), depend on the slope of the regression line. The maximally likely regression lines for the base and easy-easy (EE) data are also plotted on the base plane, along with the data (solid line and dashed line). When we vary the slope of the regression line through the base data (curved arrows and dashed-dot lines) and compute the likelihood of the data as a function of that slope, we get the likelihood function for the slope. HE = hard-easy; HH = hard-hard.

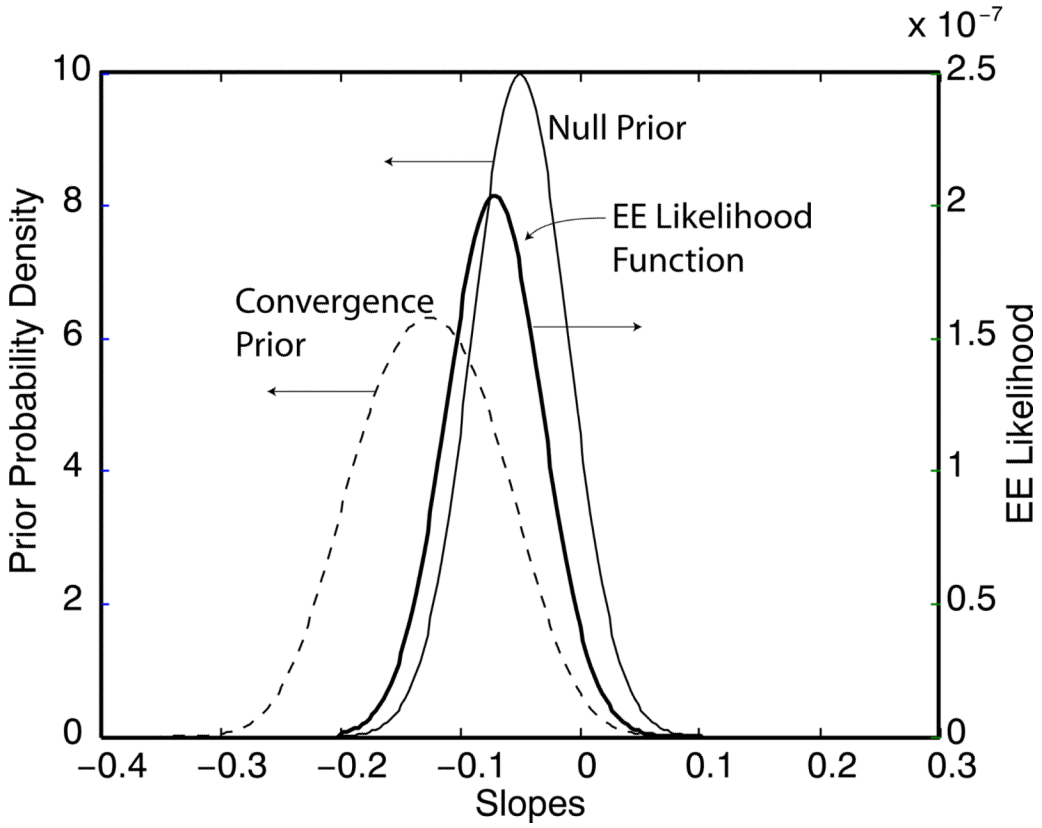

Figure 10.
The null prior and convergence prior for the slope of the regression line through Subject 7's easy-easy (EE) data (plotted against left axis), together with the likelihood function for the slope of those data (plotted against right axis).

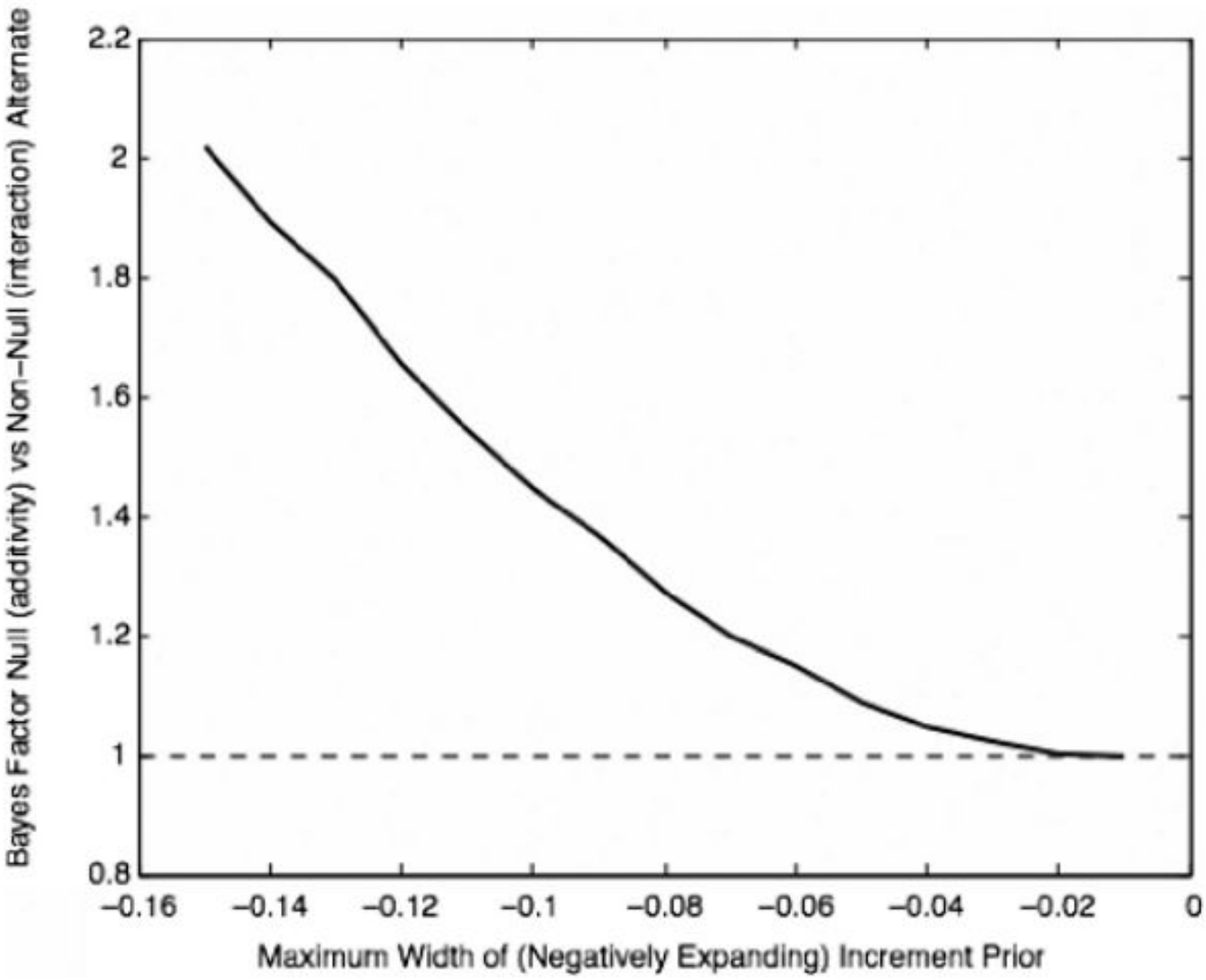

Figure 11.
The Bayes factor for the null versus convergence comparison as a function of the assumed upper limit on the rate of convergence (upper limit on how much more negative the easy-easy slope is than the base slope). The dashed line at 1 represents equal odds. The odds converge on this line. For any non-negligible width of the increment prior, the odds favor the null hypothesis (additivity).