The Importance of Proving the Null—and How to Do It

C.R. Gallistel
Rutgers University, New Brunswick

Abstract

Experimental results often support a null hypothesis, as shown by three illustrative examples from the recent literature. Conventional statistical analysis cannot support a null hypothesis, whereas Bayesian analysis can. The challenge in a Bayesian analysis is to formulate a suitably vague alternative to the null. The null is a precise hypothesis, while the alternatives to it are usually vague. Bayesian analysis penalizes vagueness: the vaguer the alternative, the more the null is favored when the observed effect is small. The question is: how vague should the alternative be? A general solution is to compute the odds for or against the null as a function of the upper limit on the vagueness of the alternative. If the odds favoring the null approach 1 from above as the hypothesized size of the effect approaches 0, then the data favor the null over any alternative to it. The simple computation and the highly intuitive graphic representation of the analysis are illustrated by the analysis of the three examples, for each of which the null is the theoretically consequential hypothesis.

BAYESIAN HYPOTHESIS TESTING LEARNING ATTENTION MOTOR PLANNING EFFECT SIZE

It is famously the case that you cannot prove the null with a conventional statistical analysis. This very fact implies that conventional analysis is not a normative procedure for deciding which of two competing hypotheses or descriptive models provides a better account of the data, because the null hypothesis is often not a straw man. In many cases, it is the theoretically more interesting hypothesis. Indeed, it can be the only interesting hypothesis, wildly counter-intuitive, and freighted with far-reaching implications for underlying mechanisms—the kind of hypothesis that pure scientists most value. Fortunately, as others have pointed out (Berger & Sellke, 1987; Edwards, Lindman, & Savage, 1963; Glover & Dixon, 2004; MacKay, 2003, Chap 4; Wagenmakers & Grünwald, 2006), in a Bayesian analysis, which is the normative procedure for deciding among competing hypotheses, the null is not a straw man. It is possible to prove it in the sense of showing that it is more consistent with the data than the other hypotheses that have been or might reasonably be suggested. This is the only sense in which any hypothesis can be "proved" by statistical analysis.

Conventional analysis is rooted in agricultural and medical contexts, where minimizing false claims of treatment efficacy is the goal. In such contexts, treatments that have no effect are of no interest. But in a broader scientific context, hypotheses to the effect that some experimental manipulations have no effect are among the deepest and most important principles of science, as witness the conservation (invariance) laws in physics. The goals of experimental psychologists are often more closely aligned with

those of the experimental physicist than they are with the goals of those who test medical or agricultural treatments. We generally want to know which hypothesis the data favor and how strongly they favor it. There are many cases in which the hypothesis that a given manipulation has no effect are extremely interesting hypotheses, whose confirmation or falsification bear strongly on our conjectures about the nature of underlying mechanisms.

Consider, for example, the hypothesis that the number of learning trials has no effect on the amount of learning produced. Now there is an hypothesis to make the student of learning sit up and take notice! The hypothesis that the more trials, the more learning, is about as close to holy writ as it is possible to come within the experimental study of learning. It is so intuitive, so "obviously" true, that it has almost never been experimentally tested. Remarkably, the only extensive test of the hypothesis of which I am aware showed it to be false: when the duration of the training is held constant, the number of trials has no effect (Gottlieb, 2007, in press). Here, if ever, is a case where it is of quite fundamental importance to assess how strong the evidence for this lack of effect is. One might object that Gottlieb's experiments simply failed to reject the null for want of statistical power. (Indeed, reviewers did make this objection.) The claim that a failure to reject the null is due to lack of sufficient statistical power can always be made, no matter how great the sample. Thus, it amounts in effect to the claim that, for statistical reasons, it is impossible to obtain evidence that increasing the number of trials does not in and of itself produce more learning. If that were really so, then the hypothesis that more trials produce more learning would be a truism, something not susceptible to empirical refutation; hence, not something that belongs in a serious science (Popper, 1959). Fortunately, as we will see, this is not true: Gottlieb's (2007, in press) data provide strong evidence in favor of the null hypothesis; hence, strong evidence against the hypothesis that the number of trials has an effect.

Next, consider the hypothesis that statistical learning, which has been thought to be an automatic process, because it occurs without intent or awareness, does not occur when a stream of items is seen but not attended to, because they are of the wrong color (Turk-Browne, Jungé, & Scholl, 2005). The most interesting hypothesis is that when items are not attended to, nothing is learned about the statistical properties of their sequencing. On this hypothesis, performance on a measure of sensitivity to sequence statistics will be at chance. This prediction is an example of an interesting "point" null, an hypothesis that specifies a single value, in this case $p = 0.5$, as the true value of an experimentally estimated parameter. If performance on items seen but not attended to is truly at chance, then this implies that 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. Put another way, the question whether the learning of the statistical properties of an unattended (but unquestionably seen) icon stream is truly at chance, or merely less than the learning that occurs when the stream is attended to, speaks to the question whether attention is a matter of selection (gating) or of reducing the allocation of capacity-limited processing resources. In the latter case, a diminished but still significant effect might be expected. Turke-Brown, et al. (Turk-Browne, Jungé, & Scholl, 2005), using conventional analysis, reported repeated failures to reject the null (chance performance) when subjects were tested for the implicit recognition of statistical dependencies in the sequence of unattended items. Because they

used conventional methods, they could not marshal statistical support for the hypothesis that performance actually was at chance, although it was the hypothesis they favored.

Finally, consider the question of additivity. Do factors combine additively to determine a measured quantity? If they do, 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. Rosenbaum , Halloran and Cohen (2006) investigated the effects of target height and target width on grasp height when subjects reach out to grasp a target. 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…." – p. 920-921. Nor should they be asked to conjecture what might happen with a much larger group. That is not the question. The question is, Do the data they gathered support the additivity hypothesis, as opposed to reasonable alternatives to it; and, if so, how strongly do they support it? If the odds favoring the null hypothesis over even small deviations from it are large, then running more subjects is a waste of time—unless, of course, a theory comes along that deduces from first principles the prediction of a *very small* effect.

Because support from data is always relative—all that data can do is make some hypotheses more likely than others—an important aspect of marshalling statistical support for the null hypothesis is the posing of reasonable alternatives to it. In this paper, I elaborate strategies for doing this, and stress how instructive the pursuit of these strategies can be. Particularly instructive is the penalty that Bayesian analysis imposes on vagueness. One hopes that many psychologists will appreciate an analytic method that strongly encourages the reduction of vagueness in psychological theorizing by bringing the consideration of possible effect size into the heart of the analysis.

Illustration 1: Does the number of trials matter?

It is a matter of everyday observation, repeatedly confirmed by experiment, that increasing the number of trials in which a neutral stimulus (commonly called the CS, for conditioned stimulus) precedes a motivationally important stimulus (commonly called the US, for unconditioned stimulus) leads to the appearance of a conditioned response to the CS, a response that anticipates the predicted US. However, this everyday observation confounds two different parameters of the training experience: its overall duration and the number of observed predictions (trials). As the number of trials increases, so does the overall duration of the training experience. There has long been evidence that the overall duration is important independent of the number of trials comprised within that duration, although the evidence is not usually summarized in that way. The evidence comes from the well-established trial-spacing effect " (Papini & Brewer, 1994), which is usually summarized by saying that "spaced practice/training is better than massed practice." Spacing trials more widely is known to increase their efficacy. Spacing them more widely increases the duration of the conditioning experience without increasing the number of trials. Therefore, duration itself is an important parameter of a learning experience, independent of how many trials are comprised within a given duration. As with most phenomena in the study of learning, there has been relatively little attention to the quantitative side of this effect. The most extensive compilation of quantitative data on the

effect of spacing trials more widely has startling implications that have never been commented on.

The most widely used Pavlovian (aka classical) conditioning paradigm is pigeon autoshaping. It is exactly Pavlov's procedure in which CS presentation (e.g., the ringing of a bell) repeatedly precedes US presentation (food), leading to the appearance of a conditioned response to the CS. However, it uses pigeons, rather than dogs. It uses the illumination of a round key on a wall of the test chamber as the CS, rather than the ringing of a bell. And the conditioned response that is recorded is the pigeon's pecking of the key, rather than the dog's salivation. The pigeons learn to peck an illuminated key whose illumination predicts food delivery, whether or not their pecking has any effect on that delivery.

Gibbon and Balsam (1981) gathered raw data from laboratories all over the world that use this protocol. Different laboratories in different experiments used different average intervals between the trials and different durations of key-illumination prior to food presentation. In the raw data, though not usually in the published reports, one could discover how many trials there were prior to the appearance of pecking. [The conditioned response in most paradigms appears abruptly (Gallistel, Balsam, & Fairhurst, 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).] Gibbon and Balsam plotted the number of trials to acquisition against the proportion between the average US-US interval (also called the cycle time, $\bar{I}_c$) and the CS-US interval (the delay of reinforcement, *T)* on double logarithmic coordinates (Figure 1)

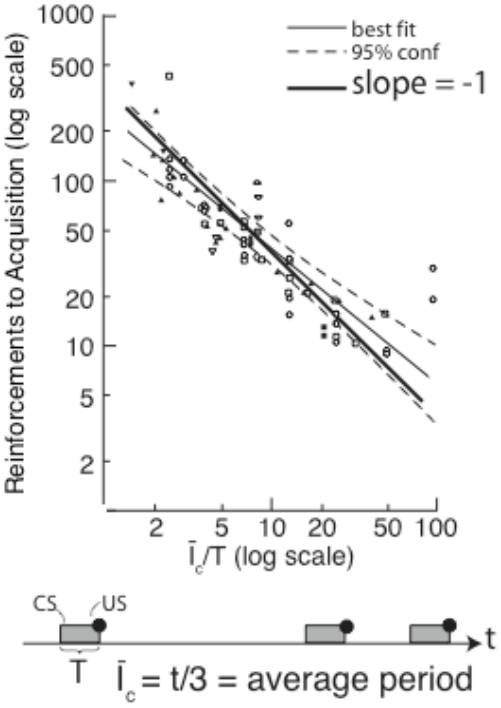

Figure 1. *The number of reinforced trials (CS-US pairings) to acquisition in pigeon autoshaping plotted against the $\bar{I}_c / T$ ratio, on double logarithmic coordinates. (Replotted from Gibbon & Balsam, 1981). $\bar{I}_c$ is the average interval between reinforcements. T is the duration of the warning interval, that is, the CS-US interval, commonly called the delay of reinforcement. As the $\bar{I}_c / T$ ratio grows, the delay becomes relatively smaller and smaller. Thus, the CS becomes a relatively better predictor of imminent reinforcement. Put another way, it becomes more informative.*

As the Figure 1 shows, the number of trials required decreases as the ratio between the cycle period ($\bar{I}_c$) and the CS-US interval (T) increases. The data are consistent with the hypothesis that the true value of the slope of the regression line is -1. If that is its true value, then each doubling of the $\bar{I}_c / T$ ratio halves the number of (reinforced) trials prior to the appearance of a conditioned response to the CS. The results in Figure 1 suggest that the efficacy of a trial in promoting the appearance of a conditioned response is a scalar function of (is strictly proportionate to) the $\bar{I}_c / T$ ratio, which is to say to the relative amount by which the onset of the CS reduces the expected time to the next US. Information theory gives a principled justification for such a dependence, because variation in the amount of Shannon information that a CS provides about a US is proportional to the log of this ratio (Balsam & Gallistel, submitted). For that reason, Balsam and Gallistel call this parameter of a conditioning protocol (the ratio, $\bar{I}_c / T$) the informativeness of the protocol. They propose that associability (defined as the reciprocal of the number of trials to acquisition) is proportional to informativeness.

If it is the informativeness of the CS-US relation, not their temporal pairing, that matters in conditioning experiments, that is, if the relation seen in Figure 1 holds quite generally, and if the slope is truly 1, then the number of trials is not in and of itself a determinant of the progress of learning, because 2 trials will be just as effective as 16 trials when the 2 trials are spread out over the same interval as the 16 trials (Figure 2). Although it was not a consideration of the implications of the regression in Figure 1 that motivated him, the question Gottlieb posed in four different autoshaping experiments with rats or mice as subjects was, Are 8 times fewer trials just as effective at promoting learning if they are spread out over the same total training duration? His experiments can be seen as testing the conclusion that the slope of the regression line in Figure 1 really is -1.

Figure 2. *If the slope of the regression line in Figure 1 is -1 and if it generalizes to species other than the pigeon, then these two protocols will be equally effective in*

promoting the development of a conditioned response to the CS. The lower protocol has only 2 trials, whereas the upper protocol has 16, but the $\bar{I}_c / T$ *ratio in the lower protocol is 8 times greater than in the upper protocol, so each trial should be 8 times more effective because it is 8 times more informative.*

In his first experiment, the subjects were rats. The CS was a 30 s white noise, terminating with the delivery of two food pellets (the US, that is, the reinforcement). The measured response poking into the feeding hopper during the CS (in anticipation of food pellet delivery). There were two training sessions, each lasting approximately an hour and a half duration. For one group of 8 rats, there were 32 trials in each session; thus, 64 in all. For the other group (originally 2 groups, but the minor difference in their protocols had no effect), there were 4 trials in each session; thus, 8 in all. Because the first group had a much richer experience in the training environment (they got 8 times as much food), they showed much greater context conditioning, that is, they poked into the hopper frequently even when the noise was not on. For this reason, the two sessions in which the noise predicted food delivery were followed by two sessions of context extinction, during which nothing happened: no noise and no food. This reduces both groups' excitement about being in the test chambers to a common low level prior to the critical test session in which their response to the noise CS was tested on 6 trials, with no reinforcement (no food). Figure 3A shows the cumulative distributions of the mean numbers of responses to the CS. Although the mean for the group that got 64 trials is slightly higher than the mean for the group that got only 8, the striking thing is how completely intertwined the two distributions are. It is more than a little contorted to summarize this evidence by saying that it does *not* favor the *rejection* of the hypothesis that increasing the number of trials by a factor of 8 had no effect. What the data clearly suggest is that the 8-fold increase had no effect (the null hypothesis). Moreover, that is the interesting hypothesis, the one that Gottlieb sought to test. Bayesian analysis shows that in fact these data support that hypothesis relative to any hypothesis distinguishable from it; the more distinguishable from the null hypothesis an alternative hypothesis is, the more strongly the data favor the null.

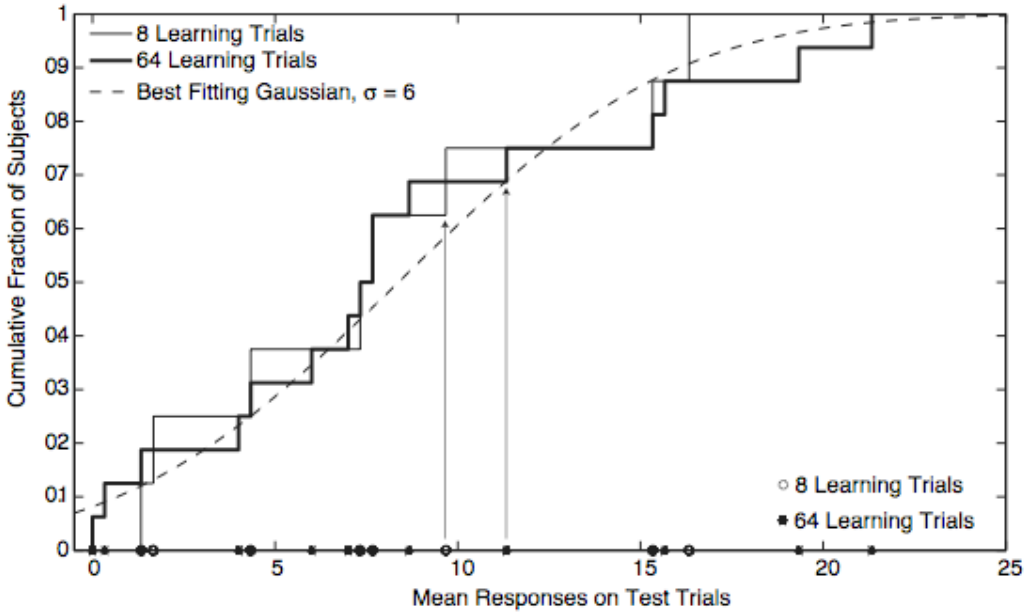

Figure 3. *The cumulative distributions of mean response rates for the two groups. On the abscissa are that data (open circles) from the 8 rats in the group that got 8 trials and the data (asterisks) from the rats that got 64 trials. The cumulative distributions plot the fraction of the data from a given group with a given value or less. Thus, the cumulative plot steps up at each circle or each asterisk in the corresponding scatter plot (see illustrative vertical arrows). It is much easier to compare cumulative distributions than to compare the scatter plots. Because the cumulative distributions are interwoven, it is apparent that the ranges and central tendencies of the response rates are essentially the same in the two. The dashed curve is the maximum likelihood cumulative Gaussian for the pooled data.*

In a Bayesian analysis, one computes the posterior likelihood of the data under each of the competing hypotheses (alternative statistical models for the data). The Bayes Factor, which is the odds favoring one model over the other, is the ratio of the two posterior likelihoods. The posterior likelihoods are the integrals of the posterior likelihood functions. A posterior likelihood function is in essence the correlation (cross product) between a prior probability distribution, which is specified by a proposed model or hypothesis, and the likelihood function, which is specified by the data. The posterior likelihood of the model is the integral of this cross-product, just as the correlation coefficient is the (normalized) sum of the cross products of paired observations. If the data fall in the center of the prior probability distribution, then the posterior likelihood for that model—its likelihood in the light of the data—is (relatively) high. If they fall in one or another tail of the prior distribution, then the posterior likelihood of that model is (relatively) low.

All likelihoods are relative. When the posterior likelihood is low, it is so only in comparison to the posterior likelihood of a model that specifies a prior probability distribution in better accord with the data. The likelihood is then relatively low because an alternative model is better correlated with the data.

In the present case, we are interested in the relative likelihoods of competing models of the data from the 16 rats that got 64 trials. The null hypothesis is that these 16 data are drawn from the same distribution from which we drew the data from the 8 rats that got only 8 trials. Thus, the prior probability distribution specified by the null hypothesis is just the posterior probability distribution for the mean of group that got 8 trials. This posterior probability distribution specifies how uncertain we are about the mean of the distribution from which we drew that sample.

The competing hypothesis says that the data from the 64-trial group were drawn from a distribution with a higher mean. The question is, How much higher? That is where the vagueness comes in. Although it has been an article of faith for more than a century that the strength of the conditioned response increases with an increase in the number of conditioning trials, neither theory nor prior experiment tells us how much of an increase to expect. Thus, we cannot really say what the prior probability distribution for the expected increment in performance should be. This is the essence of "the problem of the prior" (Killeen, 2005). Our vagueness about the prior probability distributions implied by our hypotheses has stood in the way of the wider adoption of Bayesian analysis. For reasons to be explained in a moment, the outcome of a Bayesian analysis depends on how vague we make the alternative prior: the vaguer we make it, the more that the precise null hypothesis will be favored. I propose a generally applicable solution to this problem. The solution is to compute the odds (Bayes Factor) as a function of the vagueness of the alternative. This function shows how close to the null we have to make the alternative hypothesis in order for it to compete with the null in explaining the data. If we have to make it so close to the null that it is essentially indistinguishable from it, then the data favor the null.

The hypothesized source distribution. The source distribution is the distribution from which data have been drawn. We must make an assumption regarding its form. The empirical cumulative distributions are well fit by a cumulative Gaussian (dashed curve in Figure 3), so we assume that we are drawing from normal distributions[1]. The data do not suggest any difference in the standard deviations of the distributions from which both the 8-trial and the 64-trial 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[2], which, we estimate from the data to be very nearly 6. Now, the question is, which hypothesis do the data favor: 1) the null hypothesis, which is that the 64-trial data were drawn from the same distribution as the 8-trial data, because the 56 additional trials had no effect? 2) or the non-null hypothesis, which is that the 64-trial data were drawn from a distribution with a higher mean, because of the effect of the 56 additional trials?

---

[1] An objection to this assumption is that it assumes that there can be data with negative values, which is impossible with mean-rate data. One might want to assume a distribution that is only supported on the positive reals. However, all the common distributions with this property have 0 probability density for a 0 datum. The data from the 16-trial group include a value of 0, so we cannot assume a source distribution that makes such a datum impossible.

[2] This is the assumption we make when we do an equal-variance 2-sample t test, which is probably the most common analysis in the experimental literature

The likelihood function. The likelihood function is the function that we get when we take the hypothesized source distribution and slide it along the x-axis, as shown in Figure 3. At successive locations of the source distribution (the top two panels show two locations), we find the probability densities for all of the data points (see the arrows from all but 2 of the data points in the top two panels). In this context, these probability densities are called likelihoods. The product of the 8 likelihoods (one for each datum) is the likelihood of the data for that location of the hypothesized source distribution. The likelihood function (bottom panel) is the plot of these products as a function of the assumed value of its mean, that is, the assumed location of the distribution along the x axis.

Although the likelihood function looks like a probability distribution and although it was obtained by multiplying together the probability densities for the data as the source distribution was moved from location to location, the likelihood function does not (with rare exceptions) integrate to 1—unlike a probability distribution, which always integrates to 1.

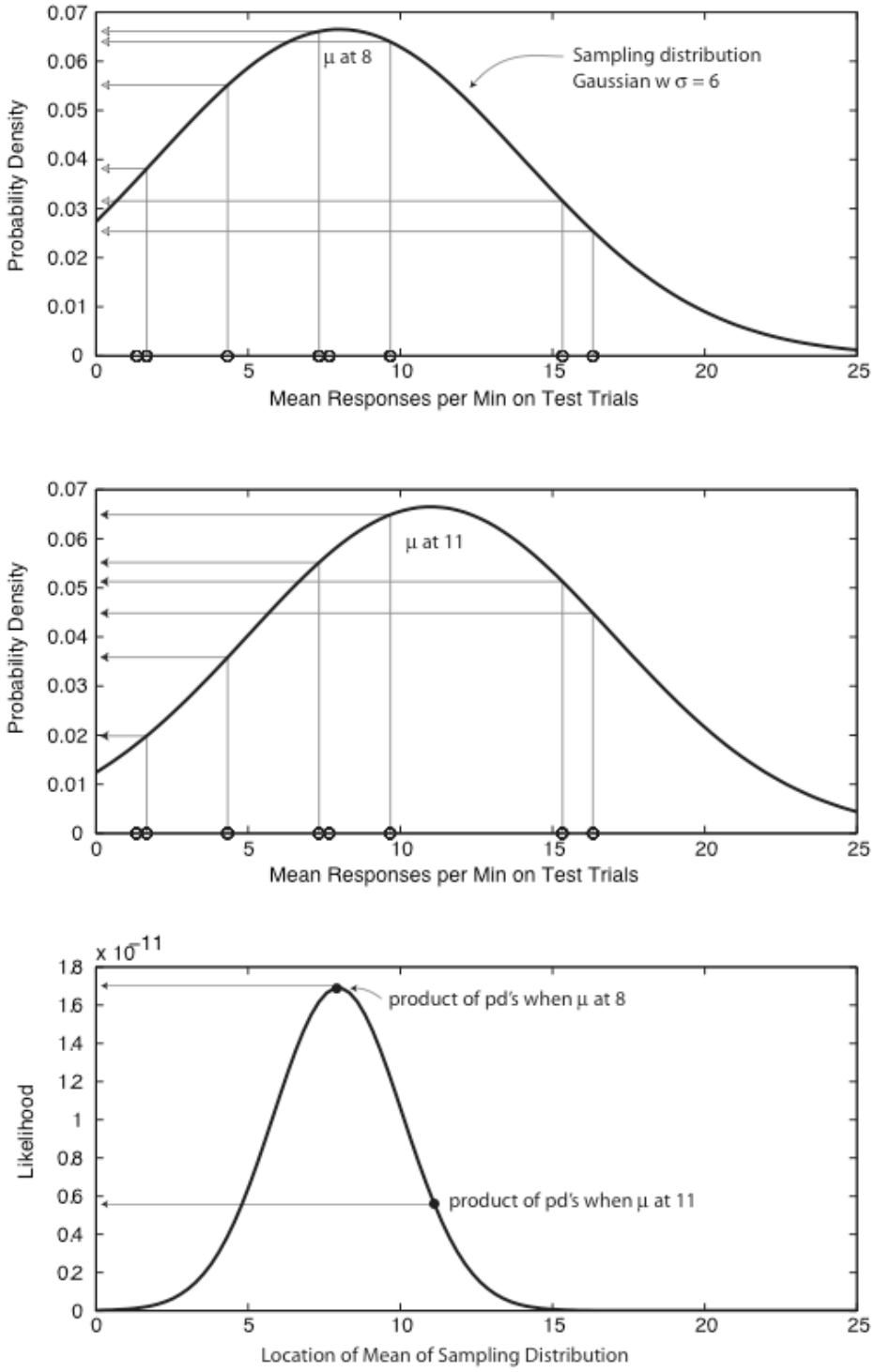

Figure 3. *Computing the likelihood function for the data from the 8-trial group. The source distribution is slid along the x-axis; the top 2 panels show it at two locations. At each location, one reads off the probability densities for the data (arrows for 6 of the 8 data) and multiplies the 8 likelihoods thus obtained to get the likelihood of the data when the source distribution is assumed to be at that location. The likelihood function (bottom panel) is the plot of these products as a function of the (assumed) location of the source*

distribution. Note that the area under the likelihood function is nowhere near 1 (numbers on ordinate of bottom panel are x 10^{-11}).

The posterior likelihood function. The likelihood function *tout court* is determined by the data only, while the *posterior* likelihood function is jointly determined by the likelihood function and the prior distribution. The prior distribution, often called the prior for short, captures what we already know or hypothesize about the likely location of the data before we look at it—on the basis of other data, analytic considerations, and/or theory. However, in the case of the data for subjects getting only 8 trials, we have no theory nor other observations that have implications about where along the (positive) x axis we should expect to find the data. Thus, for this calculation, we have an uninformative or flat prior. Because this prior has the same value everywhere, multiplying it point by point with the likelihood function to obtain the posterior likelihood function has only a meaningless scaling effect. With a flat prior, the posterior likelihood function looks exactly like the likelihood function itself. Only the numbers on the ordinate are different, and that change in scale has no significance to any further computation.

The posterior likelihood function is still a likelihood function, not a distribution function, and so it does not integrate to 1. It specifies the relative likelihood of different possible locations (means) of the source distribution, in the light of the data that went into its computation and the prior information. We see in the bottom panel of Figure 3 that the most likely (maximum likelihood) location of the mean of the source distribution in the light of the data from the 8-trial group is right around 8. Locations at 2 and 15 are much less likely.

The posterior distribution function. The posterior likelihood function can be converted to the posterior distribution function simply by rescaling it to make the area under the curve equal to 1 (Figure 4). Thus, it always looks exactly like the posterior likelihood function. The only differences are that the numbers on the ordinate are now probability densities rather than likelihoods, and, of course, the area under the curve is now 1.

The null prior. The null hypothesis says that the source distribution for the 8-trial group and the 64-trial group are one and the same, that is, they have the same mean. Thus, on this hypothesis, when we come to compute the likely location of the source distribution from which the 64-trial data were drawn, we have relevant prior information, because we have already made 8 draws from that distribution. We have a probabilistic specification of where the mean of the source distribution is, namely, the posterior distribution function obtained from the 8-trial data. Thus, on this hypothesis, the *prior* distribution function for the computation of the posterior likelihood function using the 64-trial data is the *posterior* distribution function obtained using the 8-trial data. This prior is not flat. When we multiply it point by point with the likelihood function for the 64-trial data (see Figure 5, top panel), it affects the shape and height of the resulting posterior likelihood function (Figure 5, bottom panel).

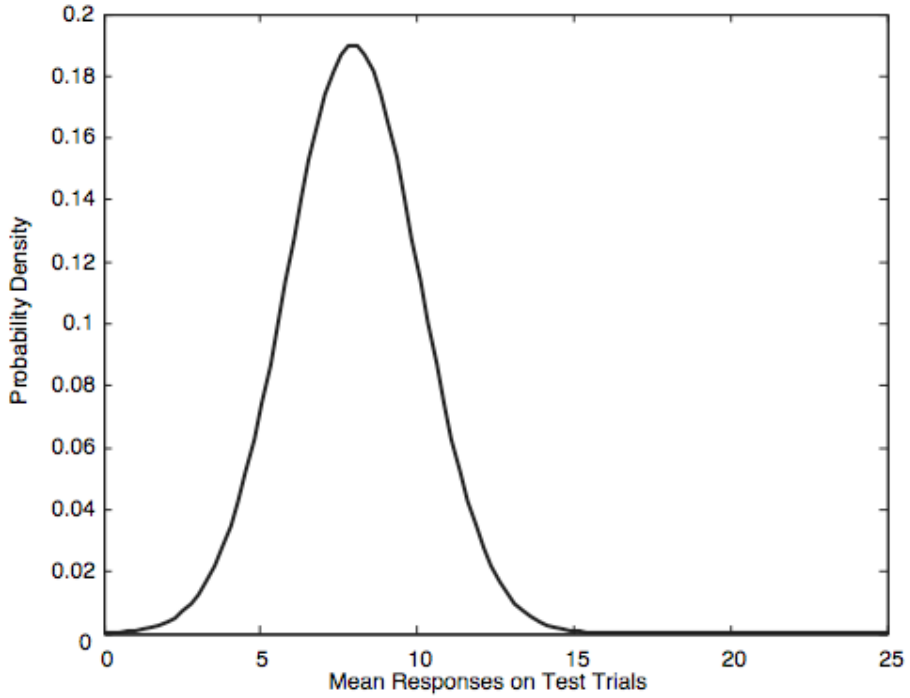

Figure 4. *The posterior probability distribution for the mean of the distribution from which the 8-trial data were drawn. This is also the prior probability distribution for the mean of the 64-trial data on the null hypothesis that those data are drawn from the same distribution as the 8-trial data. The posterior probability distribution quantifies our uncertainty about the mean of the (assumed to be) common source distribution, after we have examined the first 8 draws from it and before we have made the next 16 draws. This graph looks exactly like the one in the bottom panel of Figure 3, because it is that graph after rescaling the ordinate so as to make the area under the curve equal to 1 (because probability distributions always integrate to 1). The rescaling converts the likelihoods to probability densities; hence the difference in labeling of the ordinate between the bottom of Figure 3 and this plot.*

Alternative hypotheses. The alternative to the null hypothesis is that the 56 trials that the 64-trial group got over and above the 8 trials that the 8-trial group got moved the distribution from which the 64-trial data were drawn to the right along the x-axis. In other words, the 64-trial data were sampled from a distribution with a higher mean. A Bayesian analysis requires us to specify just how much of an effect these additional trials may have had. It does not allow us to be noncommittal about the size of the hypothesized effect, because we are going to have to specify a prior probability distribution for the 64-trial data under the alternative (non-null) hypothesis, just as we have already done for the null hypothesis.

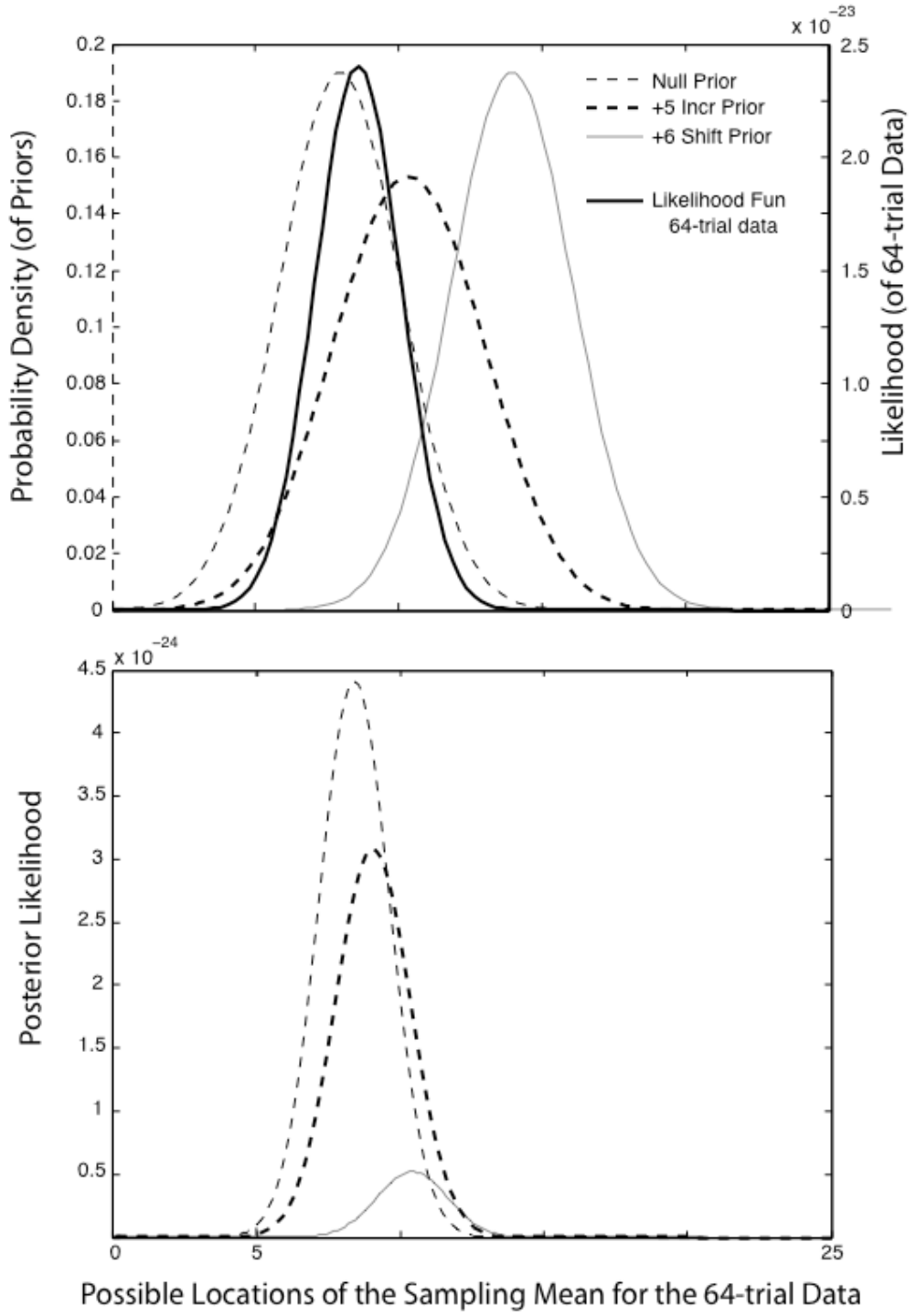

Figure 5. *Top panel plots the prior probability distributions associated with 3 different hypotheses against the left axis and the likelihood function for the 64-trial data against the right axis. The likelihood function is evidently better correlated with the prior probability distribution specified by the null hypothesis than it is with the distributions specified by the other two. Bottom panel plots the posterior likelihood functions, which are the crossproducts of the likelihood function with each of the prior probability*

distributions. The crossproduct with the best-correlated prior probability distribution has the greatest area (biggest integral). The ratio of two integrals is the Bayes factor for the corresponding comparison. Notice that the least plausible prior (+6 Shift) does not shift the peak of the posterior likelihood function very much, even though the peak of the prior is far to the right of the data. What the mismatch between the prior and the likelihood function mostly does is lower the posterior likelihood of the hypothesis associated with the mismatched prior. This illustrates "the power of the data" in Bayesian analysis.

There are various plausible approaches to formulating an alternative prior. One approach is to assume that we have an hypothesis that predicts an effect of a specific size. The prior associated with any such hypothesis is obtained by shifting the null prior rightward by the amount of the predicted effect. The upper panel of Figure 5 plots against the right ordinate the likelihood function determined by the data from the 64-trial group (heavy curve) and, against the left ordinate, the prior probability distributions associated with three different hypotheses. One of these prior distributions is for the null hypothesis (thin, dashed curve). As has already been explained, it is simply the posterior probability distribution for the 8-trial data. The thin solid curve to the right of it is the null prior shifted rightward by 6. This magnitude of rightward shift—this size of "predicted" effect—is not entirely arbitrary. It was chosen for illustration because it is the size of the effect that would just be enough to be "highly significant" ($p < .01$) by a conventional 1-tailed t-test. The problem is that although (almost) everyone believes that there will, of course, be an effect of the additional trials, no one has an hypothesis that makes a prediction of the size of that effect. One solution, a solution that ignores the vagueness of the alternative hypothesis, but is nonetheless instructive, is to compute the odds for or against the null as a function of the size of this shift, that is, as a function of the size of the effect that we might predict if we had hypotheses that actually made quantitative predictions. Figure 6 shows the results of this computation.

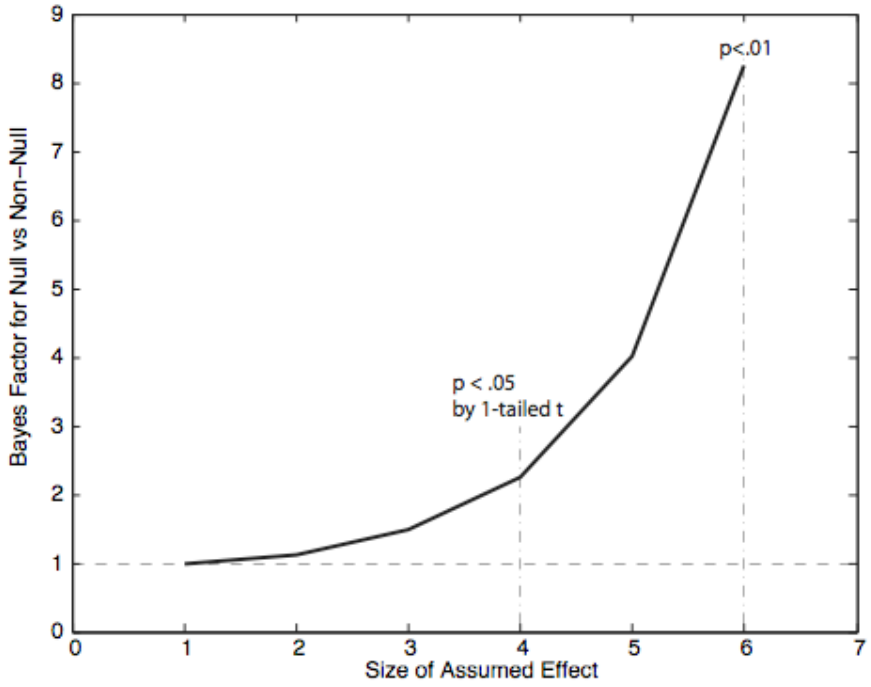

Figure 6. *The Bayes Factor (odds for or against the null) between the null and an hypothesis that predicts an effect of a specific size, as a function of the size predicted. The horizontal dashed line at 1 indicates even odds. The vertical dash-dot lines indicate the size of the effects that would be just sufficient to be "significant" by a conventional 1-tailed t test, using the indicated alpha values. The bigger the predicted effect, the worse such an hypothesis fares when pitted against the null.*

What Figure 6 shows is that, given the data, no such hypothesis could beat the null hypothesis. The only way that an alternative hypothesis that predicted a specific effect size could compete with the null is if it predicted an effect so small as to make the prediction all but indistinguishable from the null. This, of course, reflects the fact that the data from the 64-trial group fall right on top of the data from the 8-trial group (Figure 3).

Another thing to note is that if the difference between the mean of the 8-trial group and the mean of the 64-trial group were equal to 4, which is to say if the difference were just significant at the .05 level by a 1-tailed t-test, the null hypothesis would only be about 2.3 times less likely than an hypothesis that predicted an effect of exactly the size observed! The null hypothesis would fare even better when pitted against a vaguer hypothesis, which predicted an effect that big or bigger. This emphasizes that observing an effect that is just significant at the .05 level does not imply that the hypothesis that the treatment had an effect is 20 times more likely than the null hypothesis. It does not even imply that such an hypothesis is more likely than the null hypothesis! If the alternative hypothesis (often the experimenter's favorite hypothesis) is sufficiently vague about the size of the effect that it predicts, then the null hypothesis will be more likely than that hypothesis even though the data show a (just) significant effect. That is why hypotheses that merely predict the direction of an effect but say nothing about its size should be held to stricter account than they usually are. Vague hypotheses should suffer when pitted

against precise hypotheses. In a conventional analysis, they do not; in a Bayesian analysis they do. The vaguer they are, the more they suffer, for reasons that will soon become clear.

This example is representative of the many situations in which we believe there should be an effect, but we have only a hazy idea of how big it should be. In thinking about this kind of alternative, we must consider how we can reasonably delimit the statistical effects of our quantitative vagueness. We must consider the limits on the plausible size of an effect, because the broader we make those limits (the vaguer we make our alternative hypothesis), the more poorly it will compete with the null hypothesis. Intuitively, a vague hypothesis can only be vaguely right.

The increment prior. In thinking about size of a possible effect, we might begin by noting that the highest datum was only somewhat greater than 20 and that the mean of the 8-trial group was very nearly 8. Thus, it is unlikely that the difference between the 8-trial mean and the 64-trial mean is greater than 20 – 8 = 12. That makes 12 a conservative upper bound on the conceivable size of an observed difference between the two means. It is hard to say what a plausible lower bound on this increment might be, so let us be maximally generous to the non-null hypothesis and assume that the lower bound is 0. Let us consider an increment prior that is uniform between 0 and 12. Of course, should the increment in fact prove to be 0 that would make this "alternative" hypothesis indistinguishable from the null hypothesis. Technically, the null hypothesis is nested within this alternative; it is a special case, the case when the increment is 0.

The non-null prior. The increment prior is not the prior distribution function that we need in order to compute the posterior likelihood function for the 64-trial data under the non-null hypothesis, because it does not take account of our uncertainty about where the 0-end of the increment should be placed. Conceptually, we want to place it at the true value of the mean of the source distribution from which the 8-trial data were drawn. But we do not know where exactly that is. What we know about its probable location is completely expressed by the posterior probability distribution obtained from the 8-trial data (see Figure 4). To get an appropriate prior for computing the posterior likelihood of the 64-trial data under the non-null hypothesis we need to convolve the increment prior with this posterior probability distribution.

The convolution operation places the 0-end of the increment prior at successive locations along the x 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 8-trial 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. So we won't give that possibility much weight." Then, it moves it step by step to the right, leaving behind after each step successive copies of the increment prior, each weighted (scaled) by the corresponding probability density of the null prior. Thus, when the left (0) end of the increment prior is placed at the peak of the posterior probability distribution from the 8-trial data, the convolution says, "Here is a highly probable true starting 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, when the increment prior is assumed to be uniform between 0 and 5, is shown in Figure 5 top panel, heavy dashed curve). This result is our alternative to the null prior.

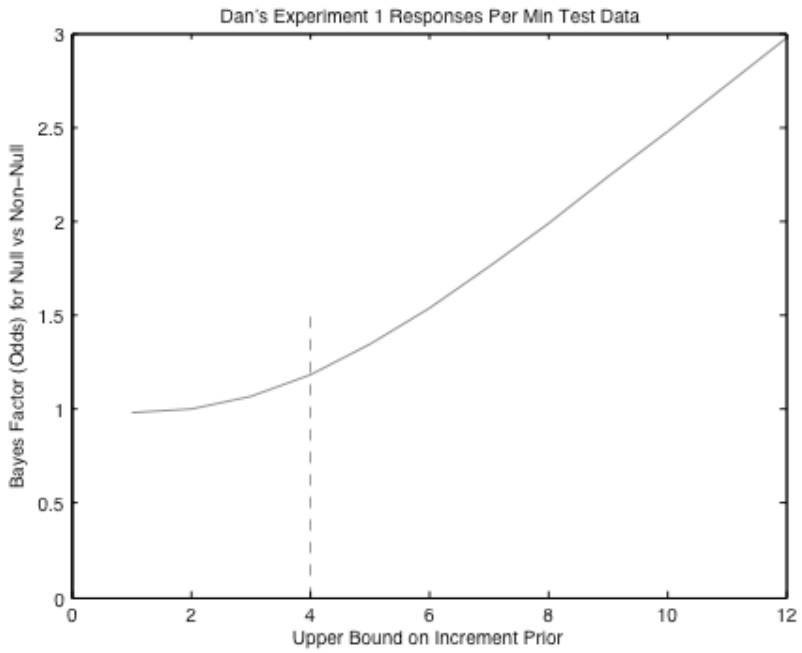

Figure 7. *Bayes Factor as a function of the upper bound on the increment prior. The vertical dashed line indicates the increment that would just be significant by a 1-tailed t test. Notice that in order to bring the odds to 50-50, the maximum possible increment has to be assumed to be so small that the alternative hypothesis has been essentially reduced to the null hypothesis.*

Figure 7 plots the odds in favor of the null hypothesis as a function of the upper bound on the alternative hypothesis (the increment prior). What it shows is that the likelihood of the null hypothesis in comparison to an alternative non-null hypothesis depends on the vagueness of the alternative. If the alternative allows for any effect between 0 and 12, then the null is three times more likely. If the alterative for some reason stipulates that although there is an effect, it is so small that the difference between the means will be at most only just significant (no greater than 4), then the null hypothesis is still slightly more likely (Bayes Factor =1.2). The only way one can get to even odds is to reduce the upper bound on the possible increment to essentially 0, making the alternative indistinguishable from the null. This emphasizes that the alternative models do as well as they do against the null only because they include it as a special case. They do worse than the null, despite including it, because they are vaguer. Their vagueness places some of the prior probability well away from the likelihood function.

As was clear from the reviews that Gottlieb received when he submitted his work, the learning community is not going to accede readily to the contention that the number of trials contained within a training protocol of a specified duration has no effect on the amount of learning that it will produce. Nor should they. This is after all a startling claim.

What other explanation can there be for the fact that the 16 rats that got 64 trials showed no more conditioned behavior than the group that got only 8? One explanation is that there is a ceiling effect. Both the 8-trial protocol and the 64-trial protocol produced the maximum observable amount of conditioned responding. As mentioned above, the widespread belief that conditioned responding gets progressively stronger as the number of training trials increases rests, in many cases, on an averaging artifact. In individual subjects, the learning curve is often a step function (Gallistel, Balsam, & Fairhurst, 2004). Different subjects step after different numbers of trials and the size of the step, the amount of conditioned responding they show also varies greatly (Papachristos & Gallistel, 2006). Averaging across these steps of different sizes at different locations produces the gradually asymptoting curve seen in many textbooks. In the Gottlieb experiment, perhaps it was the case that every subject in both groups had already stepped.

In his fourth experiment, Gottlieb addressed this possibility. He measured conditioned responding trial by trial and used a change-point algorithm developed by (Gallistel, Balsam, & Fairhurst, 2004) to determine for each subject the trial on which it first showed a conditioned response (that is, the location of the step in the step-like learning curves that he in fact observed). The subjects with numerous trials per session (the Many&Dense group) got 8 trials for every one that a subject in the group that got only a few widely spaced trials in each session (the Few&Sparse group). Thus, for the Few&Sparse group, Gottlieb determined the *trial* at which they stepped, while for the Mny&Dense group, he determined the 8-trial *block* at which they stepped. If their getting 8 times more trials over the same span of time in which subjects in the other group got only one trial did not advance the progress of their learning, then the distribution of acquisition *trials* in the Few&Sparse group should be the same as the distribution of acquisition *blocks* in the Many&Dense group. Figure 8 shows the cumulative distributions for these two groups and for a third group (the Few&Dense group) that got only 4 trials per session, like the sparse group, but at the same density (trial spacing) as the group that got many dense trials per session. For this third group, the session duration was shorter by a factor of 8.

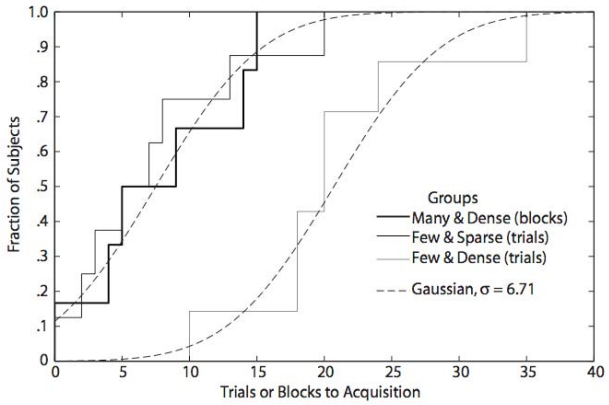

Figure 8. *Cumulative distributions of trials or blocks to acquisition for the three groups in Gottlieb's Experiment 4. The Many&Dense group got 32 trials per session. For that*

group, what is plotted is the 8-trial block during which a subject began to make a conditioned response. Thus, the number of trials such a subject had experienced is (approximately) 8 times the value plotted. The Few&Sparse group got only 4 trials per session (8 times fewer) but they were spaced 8 times farther apart. The Few&Dense group got only 4 trials per session, spaced no more widely than in the first group; hence their sessions were 8 times shorter. For both Few groups, what is plotted is the trial at which they first showed conditioned responding. The dashed curves are a cumulative Gaussians with a standard deviation of 6.71, which is the pooled maximum unbiased estimate of the standard deviation of the distribution from which the data are drawn (that is, assuming that the distributions have a common variance, but not necessarily a common mean).

The distributions for the Few&Sparse group and the Many&Dense group fall on top of one another, as predicted by the hypothesis that removing 7 of the trials in each block of 8 makes the single remaining trial eight times more effective, thereby nullifying the effect of trial deletion on the progress of learning. By contrast, the distribution of trials to acquisition in the Few&Dense group clearly lies well to the right. It took more trials to get those subjects to respond, implying that dense trials are less effective than spaced trials.

Again, we wish to assess the strength of the support that these results provide for the null hypothesis relative to various plausible non-null hypotheses. The null hypothesis is that the distribution of *blocks* to acquisition in the Many&Dense group is the same as the distribution of *trials* to acquisition in the Few&Sparse group. As before, the prior distribution for this null hypothesis is the posterior probability distribution for the mean of the Many&Dense group. This distribution is the thin solid curve in Figure 9. Hereafter, I call it the null distribution. It closely matches the likelihood function for the Few&Sparse data (heavy solid curve in Figure 9), so we see graphically that it is an almost unbeatable hypothesis.

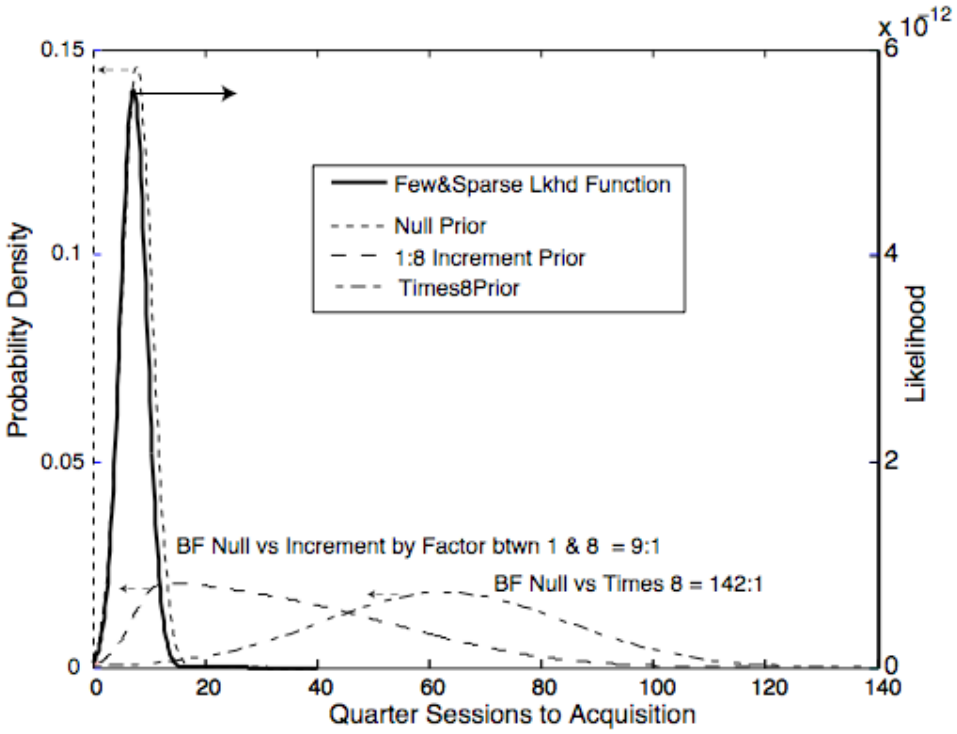

Figure 9. *The likelihood function for the mean of the Few&Sparse group (heavy curve, plotted against right axis) and the three prior probability functions corresponding to three different hypotheses: 1) the null hypothesis (that trials to acquisition in the Few&Sparse group are drawn from the same distribution as blocks to acquisition in the Many&Dense group); 2) the Times-8 hypothesis which is that only trials matter, in which case the trials to acquisition in the Few&Sparse group will be drawn from a distribution with an 8-fold greater mean; 3) the hedged hypothesis that the mean of the distribution from which trials-to-acquisition in this group is drawn is greater than the mean of the other group by a factor somewhere between 1 and 8. The three prior probability distributions are plotted against the left axis. They all, of course, integrate to 1.*

There are at least three plausible alternative hypotheses. The first is our counterintuitive null hypothesis, which is that the fact that there are 8 trials in the Mny&Dense group for every one trial in the Few&Sparse group will have no effect on the number of quarter sessions to acquisition. In that case, the prior probability distribution for the mean trials to acquisition in the Few&Sparse group is the posterior probability distribution for the mean of the quarter-sessions-to-acquisition in the Mny&Dense group. This prior is the finely dashed curve in Figure 9. It very closely matches the likelihood function for the Few&Sparse data, so we see immediately that it is an unbeatable hypothesis.

The second hypotheis is the intuitive one that only the number of trials determines the progress of conditioning. On that hypothesis, the mean of the distribution of trials to acquisition for the Few&Sparse group should be 8 times greater than the mean of the distribution of blocks to acquisition in the Many&Dense group. The prior probability function for this Times-8 hypothesis is the null distribution scaled by a factor

of 8 and reduced in height by the same factor so that it still integrates to 1. This distribution is the curve plotted with alternating dash lengths in Figure 9. It puts very little probability under the Few&Sparse likelihood function, so we see graphically that it is a poor hypothesis relative to the null hypothesis. Indeed, the Bayes Factor for this comparison is more than 142:1. Thus, the null hypothesis (that differences in the number of trials are irrelevant when the durations of conditioning are fixed and trials are not clustered) is overwhelmingly more likely than the hypothesis that the progress of conditioning depends only on the number of trials. Also, if only the number of trials mattered, then the Few&Sparse data are drawn from the same distribution as the Few&Dense data. When this hypothesis is contrasted with our null hypothesis that the Few&Dense data are drawn from the same distribution as the quarter-sessions-to-acquisition data from the Many&Dense group, the odds favor the latter hypothesis by 1,000:1.

Unlike the naïve reader of psychology texts, the specialist would be aware of the "trial spacing effect" the well-established fact that spacing trials more widely reduces the number of trials required for the appearance of the conditioned response. However, the study of animal learning has never been a quantitatively oriented field, so most specialists would not have any convictions about how strong the trial-spacing effect might be in this case. (The null hypothesis is that the effect is so strong that it fully cancels the effects of deleting trials when we keep the duration of training constant.) Thus, the third hypothesis, which is vague in a way that is characteristic of predictions in this sub-field, as well as in many other subfields of psychology, is that the mean of the trials-to-acquisition distribution for the Few&Sparse group should fall somewhere between 1 and 8 times the mean of the distribution of blocks to acquisition in the Many&Dense group. This vaguer hypothesis contains both of the other hypotheses (the suprising Null hypothesis and the intuitive Times8 hypothesis) as limiting cases.

The prior probability distribution for this alternative hypothesis is obtained, as before, by convolving the null prior with an increment prior, but now we do this in the logarithmic domain so as to make multiplication (scaling up by some factor) an additive operation. When we plot the posterior probability distribution for the Many&Dense data (aka the Null distribution) above a $\log_2$ axis, we can represent the distribution of possible multiplicative increment factors (our increment prior) as uniform over the $\log_2$ interval from 0 to 3. The interval 0 to 3 on a log base 2 axis corresponds to the interval from 1 to 8 on the linear axis. This covers the range of scale factors by which the location of the Few&Sparse mean might be expected exceed than the Many&Dense mean on our vague alternative hypothesis.

In making the increment prior uniform over the interval from 0 to 3 on the log base 2 axis, we place equal amounts of probability mass on increments of between 1 and 2 times the null as on increments between 2 and 4 times and increments between 4 and 8 times. Thus, the probability mass per unit of linear increment in the hypothetical scale factor decreases with increasing increment factors. Put another way, we assume that an increase by a smaller factor is relatively more likely. Indeed, we place 1/3 of the entire probability mass between 1 and 2, which is to say, close to what the null hypothesis predicts. We could make the increment prior exponentially increasing over the interval between 0 and 3. That would place equal amounts of probability mass between equal

linear increments in the hypothesized multiplication factor, moving the probability mass toward higher factors. However, we see by simple inspection (of Figure 9) that the likelihood function is not in fact consistent with any multiplication factor substantially greater than one. Therefore, we use the first (uniform) increment prior in this illustration. The resulting distribution (the convolution) when plotted back onto a linear axis is the curve with medium dashes in Figure 9.

The peak probability under this vague hypothesis is close to the likelihood function. However, the vagueness of the hypothesis spreads the prior probability out over a broad range, taking most of it out from under the likelihood function. Although it includes the null hypothesis as a special case, and although by making our increment prior uniform on the logarithmic axis we moved probability mass toward the likelihood function, this alternative hypothesis is nonetheless considerably less likely than the null hypothesis, as is evident from the graph. The Bayes Factor for the comparison is more than 9:1 in favor of the null. Thus, the data provide good evidence in favor of the null hypothesis relative to a plausible but vague alternative.

What makes this case particularly interesting is that elementary theory places an upper limit on the vagueness. There could plausibly be an 8-fold effect. That is what the widespread and conventional assumption that only the trials matter predicts. But there is no reason to expect an effect any bigger than that and fairly good reason to expect a smaller effect. What is not expected is that there should be no effect. That is, however, what the data in fact show. A mode of statistical analysis in which it is not possible to assess the strength of the evidence for such a theoretically consequential conclusion should not be the preferred mode of statistical data analysis (cf Glover & Dixon, 2004; Jaynes, 2003).

Finally, we may ask how well the different hypotheses explain the results from all three groups. None of them does well at predicting the data from the Few&Dense group. The hypothesis that it is the informativeness of trials rather than their number—the hypothesis that predicts the identity of the Many&Dense and Few&Sparse means—predicts that mean trials to acquisition in the Few&Dense group should exceed mean trials to acquisition in the Few&Sparse group by a factor of 8 (because the trials being 8 times more dense are 8 times less effective). The observed factor is 2.9. The hypothesis that only trials matter predicts that the mean of the Few&Dense group (scored in trials) should exceed the mean of the Many&Dense group (scored in blocks of 8 trials) by a factor of 8, while the observe factor is 2.6. Thus, both hypotheses are clearly off the mark.

The likelihood function for the Few&Dense data is very nearly the same width as that for the Few&Sparse data (see Figure 9), but shifted rightward by a factor of somewhat less than 3, placing its peak just beyond 20. Can a vaguer hypothesis predict this? And can we find some motivation for this prediction? There is some evidence for a “trials-per-session” effect such that grouping trials into shorter but more numerous sessions leads to fewer trials to acquisition (get citation from Stathis). The effect is not as well established as the trial-spacing effect, but it might nonetheless be cited as motivating a vague ex post facto hypothesis to the effect that learning rate, as measured by trials to acquisition, should be somewhat faster in the Few&Dense group than in the Many&Dense group. If we set similar limits on the quantitative vagueness of this

hypothesis—an increase in learning rate by a factor of somewhere between 1 and 8—then we get again the 1:8 increment prior in Figure 9 (the medium-dashes curve). This vague prior puts more probability under the Few&Dense likelihood function than either of the others.

What we want, however, is a measure of how far off the mark the informativeness hypothesis really is, statistically speaking. We also want a measure of how well it fares relative to the kind of multifaceted and vague hypotheses that are commonly encountered in explaining the results from psychological experiments. I have in mind an hypothesis that says that, because of both the trial-spacing effect and the trials-per session effect, one should expect the means of both Few groups to be greater than the mean of the Many&Dense group by a factor of no more than 8 and no less than 1. We want to contrast the explanatory adequacy of this multifaceted, quantitatively vague hypothesis with that of the precise informativeness hypothesis, which says that the Few&Sparse and Many&Dense means are one and the same, while the Few&Dense mean exceeds them by a factor of 8. We can achieve both objectives by looking at the relation between the joint likelihood function for the data from the two Few groups and the joint prior probability distributions for the competing explanations (Figure 10).

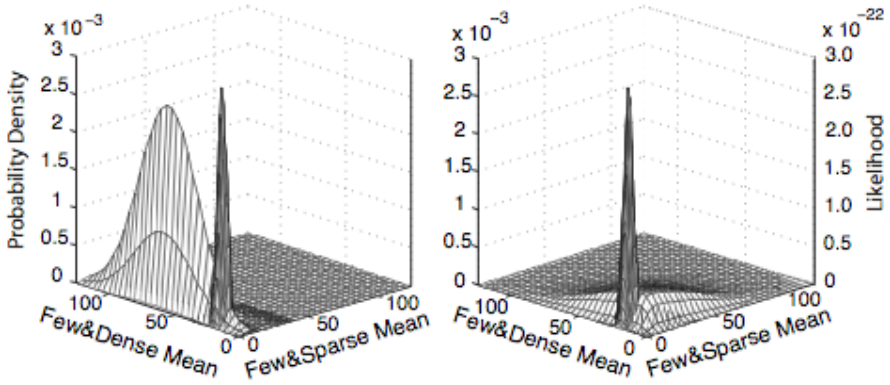

Figure 10. *Joint prior probability distributions (the spikes, plotted against left axes) and the joint likelihood distribution for the Few groups (plotted against right axes). On the left is the joint prior probability distribution for informativeness hypothesis. On the right is the joint prior probability distribution for the vague 2-free-parameters hypothesis.*

The joint likelihood function for the data from the Few groups represents our uncertainty about where the true means for these two groups are. The joint prior probability distributions represent the extent to which we can predict where they should be given the data from the Many&Dense group and our contrasting hypotheses. The peak of the joint likelihood function lies much closer to the peak prior probability on the vague model (right plot in Figure 10) than it does to the peak prior probability on the precise model. However, the vague model with its two free parameters spreads the prior probability out over a much greater area than does the precise model, which has no free parameters. The result is that the vague model places somewhat less total probability under the joint likelihood function than does the precise hypothesis. The Bayes Factor marginally favors the precise model (1.4:1).

This example of testing a model with two free parameters (the magnitudes of the trial-spacing and trials-per-session effects) against a model with no free parameters illustrates the important point that models with more free parameters are not necessarily better models than more constrained models, even when the looser model does a better job of accommodating to the data. It is the freeness of the parameters (the quantitative vagueness about their values) that is the problem, not their number. The precise model in Figure 10 also has two parameters, but their values are not free. They are pre-specified and cannot be adjusted to accommodate the data. This point about free parameters is often made by statisticians, but one has the feeling that it is seldom understood or taken to heart by experimentalists and theorists, because models with an abundance of free parameters are so popular. SOP and its extensions are an example in the subfield of animal learning (Brandon, Vogel, & Wagner, 2002; Wagner, 1981). Connectionist models are another ubiquitous example. They swim in a sea of free parameters. The essential point to appreciate about free parameters is that they increase the dimensionality of the prior probability distribution. The increase in dimensionality radically reduces the prior probability mass in the vicinity of any actual data. Increases in dimensionality have much greater diluting effects than do expansions in the range of uncertainty regarding possible values of a single parameter. The truism that probability distributions always integrate to 1 is profoundly important because it fixes the total probability mass. The more our hypotheses spread that mass out, the less of it will be found in the vicinity of any actual data. That is why vague hypotheses can only be vaguely right.

Illustration 2: Is the percent correct for the unattended stream 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. Thus, 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 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 third item of a triplet in the to-be-monitored color was occasionally repeated. 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 a 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, while 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, they consistently found that the mean proportion correct (out of 32) on the attended triplets was significantly above chance, while 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, but conventional statistical analysis does not allow us to assess the strength of that support. All it allows us to say—and all the authors in fact said—is 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 means 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 that Turk-Browne et al understandably argue for, even though their use of conventional statistical analysis does not allow them to marshal statistical support for it. As I now show, Bayesian analysis supports their conclusion—except for one subject (out of the 34 tested across the four experiments).

Digression on Group Averaging. In the psychophysical tradition (and elsewhere in some parts of natural science), the (usually unstated) assumption prevails that each subject is representative of all subjects. In this tradition, more than one subject is used in order to make sure that the experiment is replicable from one subject to the next. Statistical analysis is generally carried out subject by subject. There is very little group averaging—particularly when there are pronounced differences between subjects. (The color vision psychophysicist thinks that knowing that males have on average 2.08 different cone types is not a useful statistic, because it is an average across those males who have 3 types (the great majority), those who have only 2 (about 8% of males), and the tiny minority that have only 1. No one has 2.08.) When 3-6 subjects have been run in a typical psychophysical experiment, each giving basically the same results, the experiment is regarded as having been repeatedly replicated, and ready for publication. Students trained in other parts of psychology are often astonished to learn that in psychophysics one can publish an experiment with only 3 subjects, because they have learned to regard the number of subjects as equivalent to the n in a statistical analysis, and they know that with an n that small, one has very little statistical power.

In most other parts of experimental psychology, group averaging is the norm and there is little or no statistical analysis of the data from individual subjects. Oddly, however, this practice also commonly rests on the (unstated) assumption that subjects are homogeneous. On this assumption estimates of a parameter (e.g., the probability of correctly identifying a familiar triplet) obtained from several different subjects are tacitly assumed to be stochastically differing manifestations of the same underlying parameter, whose value is the same, or roughly so, in every subject. This is presumably the logic that justified Turk-Browne et al's following the (perfectly conventional!) practice of using a t-tests to determine whether the mean proportions correct were or were not significantly above chance. In social psychology, group means are arguably meaningful and of interest when applied to socially recognized groups, but in those parts of psychology concerned with the functioning of the individual mind, it is not clear that group means are of any interest, unless they can be assumed to apply to the individuals in the groups. If half the subjects got 32 out of 32 correct on the attended triplets and the other half got only 16 out of 32 correct, the fact that the group mean proportion is .75 correct and significantly

above chance is of little interest. It is not representative of any subject, and since the group of subjects was a more or less randomly constituted group of no social significance, their collective mean is presumably not of any scientific interest.

The upshot of this digression is that before computing group statistics, I ask whether the data do or do not support the assumption of homogeneity. This turns out to be of some interest in and of itself.

Analyzing for Subject Homogeneity. Looking over the raw data, which Turk-Browne et al graciously sent, there are strikingly different proportions for some subjects in the latter experiments. For example, in their fourth experiment (Experiment 2b), one subject got 12/32 correct when tested on familiar triplets, while another subject got 32/32 correct on the same test. If any us were grading a true-false quiz, we would not assume that the two subject had the same underlying probability of answering correctly. While most subjects in their fourth experiment had a proportion correct of around .5 when tested with unattended triplets, one subject got 27/32 = 84% correct. The chances of doing that well or better when just guessing are about 1 in 100,000, so, here, too, there is reason to doubt the assumption of subject homogeneity.

In testing for subject homogeneity, one hypothesis is that the proportions from the different subjects in a group on one of the tests (with either attended or unattended triplets) are stochastic variations generated by a common underlying probability of correct recognition. The contrasting hypothesis is that each subject has a different probability of correctly recognizing a triplet on a given one of the two tests. For the first hypothesis, there is only one free parameter—the common *p*. For the second hypothesis, there are as many free parameters as there are subjects in a group, because we estimate from each subject's proportion correct what the underlying probability of a correct response from that subject is. In other words, the second model tailors its "predictions" subject by subject. Therefore, it is always going to be the more likely model if we fail to take account of the large number of free parameters that it has, hence, its vagueness. If we did a full Bayesian analysis with an 8- or 10- dimensional prior probability space (depending on whether there were 8 or 10 subjects in the group), the spreading of the unitary probability mass over the high-dimensionality prior probability space would automatically penalize the second model for its extravagant use of free parameters. Computing integrals in a space of high dimensionality is, however, a bit tricky and hard to visualize. There is a short-cut, called the Schwarz Criterion, that does not require us to specify and integrate over a prior probability space (Kass & Raferty, 1995). We can compute the likelihood of the data under the maximum likelihood values for the parameters and correct in a principled way for the inherently greater likelihood of the multi-parameter model:

$$SC = \log(\text{Bayes Factor}) = MLL_1 - MLL_2 - .5\,(d_1 - d_2)\log(n). \qquad (1)$$

MLL_1 is the maximum log likelihood of the data under Model 1, that is, the sum of the logs of the likelihoods when the source distribution posited by Model 1 has been positioned so as to make the data maximally likely; MLL_2 is the maximum log likelihood of the data under Model 2, and d_1 and d_2 are the respective degrees of freedom, that is, the number of parameters estimated from the data in each model (1 and *n*, respectively, where *n* is the number of subjects). The third term on the right hand side of this

Expression (1) corrects for the difference in the number of free parameters in the two models. Thus, for example, in a computation with a group of 8 subjects,

$$-.5\ (d_1 - d_2)\log(n) = -.5(1 - 8)\log(8) = 3.161,$$

which means that in order for the odds to favor the multi-parameter model ($Model_2$), it must make the data 10^3.161 = about 1500 times more likely than the one-parameter model. That is an approximation of the extent of the "unfair" advantage that the multi-parameter model gains from its extravagant use of free parameters.

When we run this comparison for the four groups of subjects (a different group in each of the four experiments) and the two tests, we get the Bayes Factors (that is, odds, or relative likelihoods) given in Table 1. The pattern is somewhat complex. In all but the fourth experiment, the odds when subjects are tested with unattended triplets either strongly favor the hypothesis that subjects have a common underlying p value or, in the case of Experiment 2a, they only weakly favor the heterogeneous *p* hypothesis. Moreover, the the estimate of hypothesized common *p* is very close to .5 in all three cases (.49, .49, and .50, respectively). This analysis already favors the hypothesis that when subjects have not attended to a stream (even though they saw it), they have not detected the sequential dependencies in it, so they all have an underlying *p* value of .5. In the fourth experiment, the odds favor the heterogeneous-*p* hypothesis, but this is largely due to the one subject already mentioned, who got 27/32 correct. This subject's performance differs strikingly from that seen with the unattended color in most of the other subjects, both in this experiment and the other experiments. 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, 3/14/08)

On the other hand, when subjects are tested on the attended color, the odds generally favor the hypothesis that they end up with different underlying probabilities of correctly detecting a familiar triplet. Simple inspection reveals that some subjects have a very high probability, while others are at chance. The only experiment in which this is not true is the first one. In that experiment, the exposure duration was short, and all of the subjects had a rather low probability of detecting the familiar triplets, even in the test with the attended triplets. In this experiment, the odds favor the hypothesis that there was a homogeneous *p* value across subjects for the attended color stream.

In assessing statistical support for the non-null hypothesis in the tests on attended items, we do not want to compare group averages. The average from a randomly constituted group of subjects has no scientific significance if there is reason to believe that there are pronounced individual differences on the parameter of interest. In that case, the tests with different subjects are not replications. The only thing that makes scientific sense is to estimate the distribution of values for the parameter of interest. On the other hand, in assessing statistical support for the null hypothesis in the tests with unattended items, it makes sense to treat the results from different subjects as replications of the same experiment, because we have reason to believe that the subjects share a common underlying value of the parameter being estimated.

Table 1. *Bayes Factors (Odds) Favoring the Individual-p Hypothesis or the Common-p Hypothesis in the Turk-Browne et al (2004) Experiments*

Experiment (*n*)	Attended Individ p: Common_p	UnAttended Individ p: Common_p
Experiment 1a (8)	1 : 68	1 : 891
Experiment1b(8)	256 : 1	2.45 : 1
Experiment2a(8)	astronomical: 1	1 : 15
Experiment2b(10)	astronomical : 1	25,000: 1

The source distribution for these data is the Bernoulli (the binary coin-flip distribution) with parameter *p*, which is the probability of "success" (or "heads"). For analytic reasons, we have an informative prior, because we believe that the underlying probability of correct identification cannot be less than .5 (chance). Therefore, all of the prior probability must be between .5 and 1 (inclusive). There are at least three hypotheses we might want to compare (Figure 11): 1) the null hypothesis, which is that the subject is responding at chance on a given test; 2) the "greater-than-chance" hypothesis, which is that the subject's *p* is somewhere in the interval between .5 and 1; 3) the "only-slightly-greater-than-chance" which is that the subject's *p* is above chance but not by much. The prior probability distribution for the first hypothesis places the entire probability mass at .5. The prior probability distribution for the second hypothesis is uniform on the non-chance interval (.5 – 1).

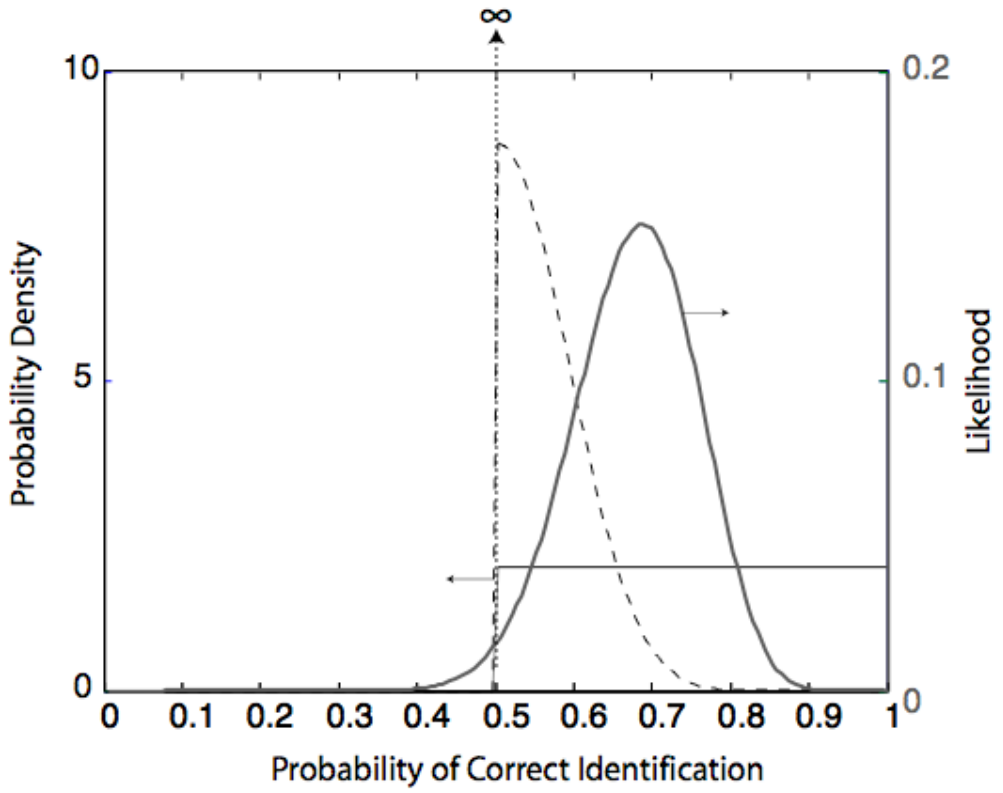

Figure 11. *Prior probability distributions for three hypotheses (plotted against left axis) and the likelihood function in the case where the subject got 22 out of 32 correct (plotted against right axis). . The finely dashed null prior puts all of the probability at .5. It is the Dirac delta function at .5, that is, it is 0 everywhere except at .5, where it is infinite in such a way that its integral goes to 1. The flat prior spreads the probability mass evenly over the greater-than-chance interval. The coarsely dashed prior puts 75% of the probability between .5 and .6*

To get a prior distribution for the third hypothesis (above chance, but only slightly), we take the beta distribution for 16 successes and 16 failures and fold it over on itself at .5, so that all of the probability is above .5 (coarsely dashed curve in Figure 11). The beta distribution gives our uncertainty about the true value of *p* after a given number of draws from the Bernoulli distribution—in this case, 32 draws, of which exactly half are assumed to be successes. In the absence of the analytic constraint that *p* >= .5, this <16,16> version of the beta distribution would be symmetrical about .5. When folded over so as to put all of the probability above .5, it puts 75% of the prior probability below .6 and almost 50% between .50 and .55. Like the uniform prior, it includes the null hypothesis as a special case. Indeed, that value (.5) is the maximally probable value of *p* on this hypothesis! This may seem odd for an hypothesis that is meant to contrast with the null hypothesis. The problem in constructing this prior is that we want most of the probability to be *near* .5 but not *at* it. Any attempt to specify some value that is "only slightly" greater than .5 as the lower limit on permissible probability is compromised by the essential arbitrariness of the choice of a particular value for this lower limit. Moreover, and more importantly, there is no need to wrestle with this paradox. The null hypothesis puts the entire probability mass exactly at .5, whereas, with the folded beta

distribution, the closer we come to .5, the less probability mass lies between where we are and .5. In the limit, as we approach arbitrarily close to .5 from above it, the probability mass becomes arbitrarily small, despite the fact that the probability density is increasing to its maximum. Thus, the folded-beta distribution accomplishes what we want: it places negligible probability mass at .5 even though it attains its maximum at .5.

When we compute the Bayes Factors for the two comparisons (Null vs Anywhere-Above .5 and Null vs Only-Slightly-Above .5) for the tests with Unattended and with Attended triplets for each of the 34 subjects in the 4 experiments, we get the results shown in Figure 12.

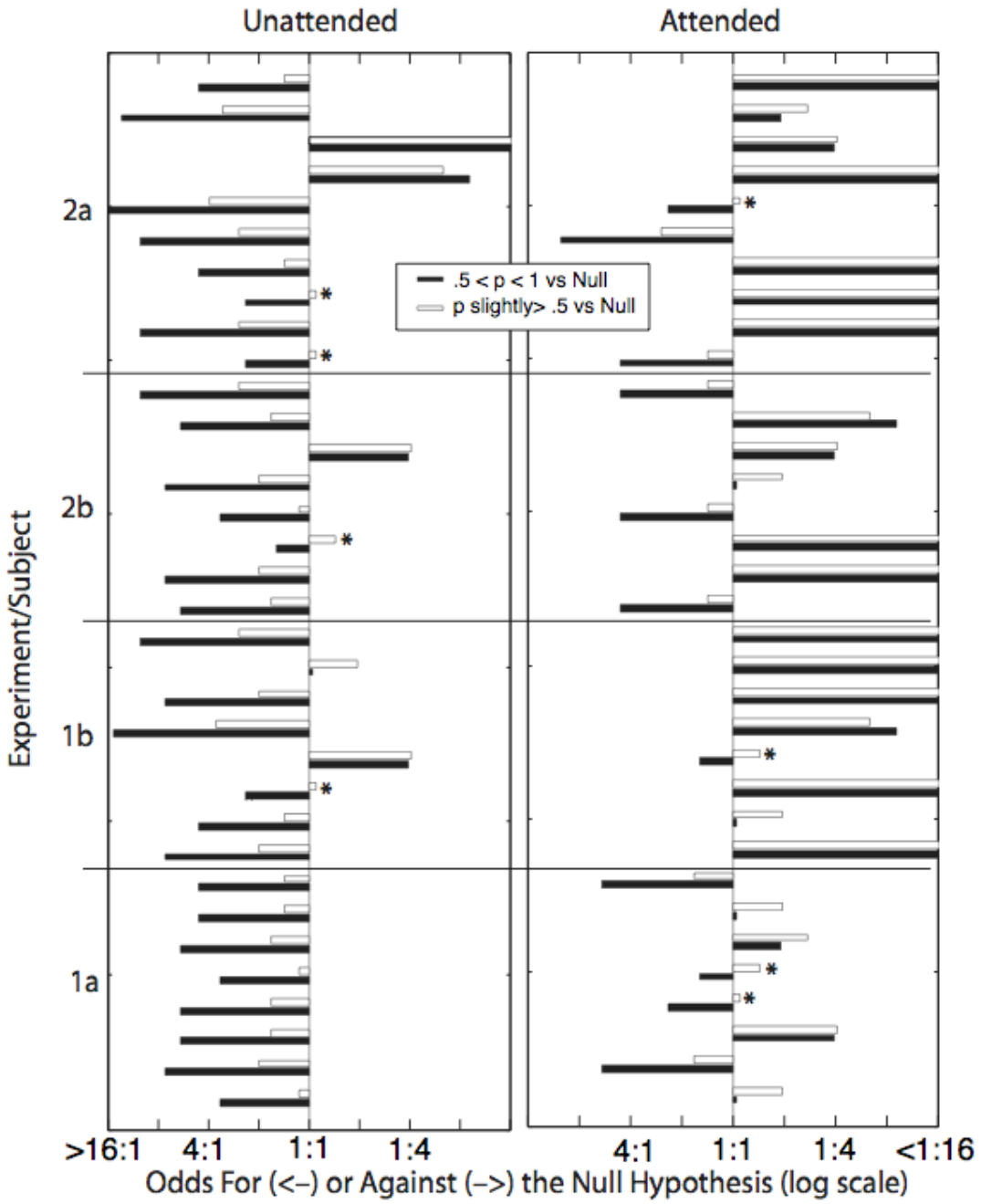

Figure 12. *Odds for (leftward projecting bars) or against (rightward projecting bars) the null hypothesis (chance), for each of 34 subjects in four experiments with tests of unattended and attended triplets. For each subject and each test, the null is pitted against two different alternatives: i) the value of the underlying probability of correct identification is anywhere on the interval from .5 to 1 (black bars); ii) the value of the underlying probability is only slightly greater than chance (white bars). 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 more precise hypothesis (p slightly above chance).*

We see immediately that on the test with unattended triplets, the data from the great majority of subjects favor the null hypothesis. How strongly they favor it depends on which hypothesis it is pitted against. When pitted against the vaguer hypothesis, that the underlying *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 the underlying *p* is only slightly greater than .5. However, the odds consistently favor the null even in the latter comparison.

That the odds mildly favor the alternative to the null in some subjects is to be expected for purely stochastic reasons, even if the true value of the underlying *p* for every subject is .5. That is why we conducted the initial test to weigh the evidence for and against the hypothesis that all the subjects had the same underlying *p*. For the test with unattended triplets, that hypothesis was either favored or acceptable in the first three experiments (1a, 1b, and 2a). Therefore, in those experiments, we may regard the test on each subject as a replication of the experiment. In that case, we can combine the data across subjects by multiplying the Bayes Factors (odd ratios) to get an overall Bayes Factor, which is the odds in favor of the Null in the light of all the replications. When we do this, we get, for Experiment 1a, overall odds of more than 300,000:1 in favor of the Null when it is pitted against the vaguer alternative and about 25:1 when it is pitted against the more precise (but otherwise weaker) alternative. In Experiment 1b, we get overall odds of almost 21,000:1 in favor of the Null in the first instance and almost 6:1 in the second instance. In Experiment 2a, we get odds of almost 25,000:1 in the first instance and almost 5.7:1 in the second instance. Thus, it matters dramatically what the alternative to the null is. However, even for an alternative that puts 75% of the prior probability between .5 and .6, the overall odds nonetheless still strongly favor the Null. If we multiply the overall Bayes Factors across the three experiments, we get odds of better than 800:1 in favor of the Null when pitted against this “only a bit above chance” hypothesis.

When we come to the fourth experiment, it is clear that at least one subject, and possibly two, achieved substantially above chance performance on the test with unattended triplets. For the third subject from the top in Figure 12, both rightward-projecting bars go off scale; the actual 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 more precise hypothesis. The odds are lower in the latter case because the percent correct was substantially greater than readily allowed by the more precise hypothesis, which puts almost 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 underlying *p* when tested on unattended triplets. It is clear that 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.

All considered, it seems safe to conclude that under most conditions, the great majority of subjects do not learn anything about the sequential dependencies in

unattended triplets. That was the principal point of the Turk-Browne et al paper. This analysis marshals strong statistical support for it, support that could not be marshaled with the conventional analyses they used.

The picture from the tests with attended triplets is more 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. Because the methods of Turk-Browne 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 in this more psychophysical way.

The analysis of the data from Experiment 1a test with attended triplets illustrates the consequences of quantitative vagueness. For this experiment, our first analysis favored the hypothesis that subjects had a common underlying *p* even when tested with attended triplets. Thus, for this experiment, we can combine the Bayes Factors for different subjects to get an overall Bayes Factor for the test with attended triplets. When we combine the Bayes Factors for the Null vs the "Anywhere-Above-.5" hypothesis, the overall Bayes Factor favors the Null by better than 15:1, despite the fact that the mean proportion correct is greater than .5 by a 1-tailed *t* test. By contrast, when we combine the Bayes Factors for the Null vs the "Only-Slightly-Above .5" hypothesis, the overall Bayes Factor favors the latter by more than 24:1. This strikingly illustrates the penalty for quantitative vagueness about the size of a "predicted" effect: When one has an hypothesis that a given treatment has an effect and when a t test enables one to reject the null hypothesis at the .05 level, it does not follow that one's hypothesis is more likely than the null hypothesis! It depends on how vague one's hypothesis is about the size of the effect it predicts. If it is sufficiently vague, then the null hypothesis may be more likely than one's hypothesis, despite the "significance" of the experimental result. Rejection of the null cannot be equated with acceptance of just any alternative to it! A very great superiority of Bayesian analysis over conventional analysis is that it requires one to consider what if anything one's hypothesis—or other information, or analytic considerations—have to say about the possible size of a predicted effect. Consideration of effect size is inescapable in Bayesian analysis, whereas it is, at best, an after thought in conventional analysis.

Illustration 3. Proving additivity

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 two differences are the same, or, equivalently, there is no interaction." The question whether the effects of different factors are additive arises often, and it is of considerable theoretical importance (Roberts & Sternberg, 1993; Sternberg, 1998). Thus, once again we see that experimental psychology is ill served by conventional hypothesis testing statistics in which null hypotheses are straw men for which the method of analysis cannot in principle marshal support.

To show the application of the already explained methods for supporting null hypotheses through Bayesian analysis, I analyze data from a recent experiment by Rosenbaum , Halloran and Cohen (2006). They investigated the effects of target height and target diameter on height at which subjects grabbed an everyday implement (a plunger) in order 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 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 ANOVA gave significant main effects for both factors and an insignificant interaction. The claim for additivity rested on the lack of significance of the interaction, which, as the authors noted, is not in fact a support for such a claim, because an ANOVA cannot in principle yield quantitative support for any null hypothesis, only support for alternatives to them. An ANOVA does not measure the weight of the evidence in support of the claim that there is no interaction; it merely tests how likely it is that the observed interaction (there always is one) will have arisen simply from the variability in the data.

Like the other experimentalists on whose data I have relied, David Rosenbaum generously supplied the raw data for 10 of the subjects. There were 5 different shelf heights (5 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 (EH, HE and HH), one or both were narrow. Subjects moved the plunger between home location and then back again twice, giving four grasps within each of the 5 x 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 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 our model of the complexity of the movement planning process: interactive models are inherently more complex.

The analysis of these data requires us to again ponder whether when we average across subjects we get something that is representative of *any* subject—let alone of all subjects. As a psychophysicist, I am committed to answering this question whenever possible *before* averaging across subjects, because, if subjects differ strikingly, then it is unclear that any scientific purpose is served by averaging across them. Thus, I began by asking whether the effect of ring diameter was of comparable magnitude in all subjects. To this end, I computed the mean and standard deviation of the 12 grasps that a subject made for any given shelf height (that is, level of the first factor) in the EH, HE, and HH conditions, which I will call the base precision conditions, because the results in these conditions are the base from which the grasp height in the EE condition departs. Figure 13 plots the 4 grasps made in the EE condition as deviations from the corresponding base mean, at each of the five levels of shelf height, for each of the 10 subjects.

It is evident in Figure 13 that some subjects (e.g., Subject 7) were sensitive to the precision factor (how tight the rings were), while others were not. Subject 6 is an

example of the latter: at every shelf level, all four of this subject's EE grasp heights fell within +/- one standard deviation of the heights of the grasps made in the other three precision conditions. This is clear evidence for the null hypothesis that the precision factor had no effect on the behavior of this subject, but a conventional analysis cannot establish that conclusion. A subject-by-subject Bayesian analysis can, however, compute Bayes Factors (odds) both for and against the null hypothesis that ring diameter had no effect. We combine the 5 x 12 = 60 baseline grasps into a single pool by expressing all grasps as deviations from their mean for that shelf level. Similarly, we combine the 5 x 4 =20 EE grasps into a pool by expressing them as deviations from the EE mean at their shelf level. The standard deviations (variabilities about the means) appear to be unaffected by shelf height, so we fix the standard deviation of our Gaussian model for the source distribution at the standard deviation of the 60 mean-normalized baseline grasps. Using this standard deviation (and a uniform prior), we compute the likelihood function for the baseline mean.

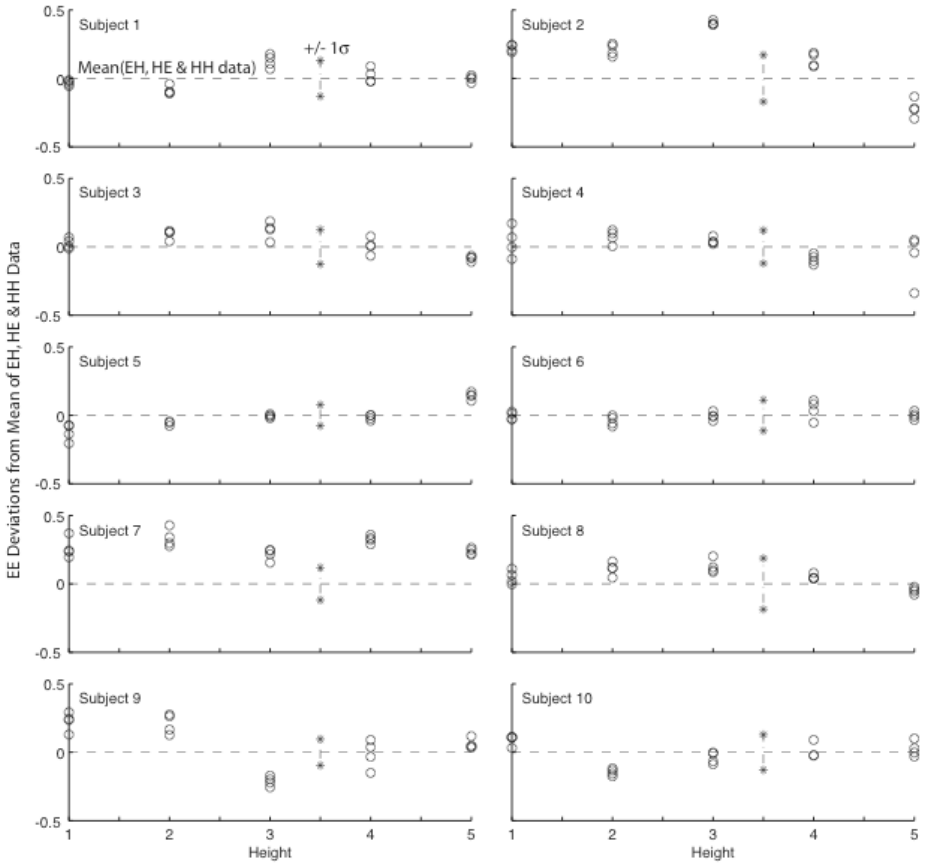

Figure 13. *For each of 10 subjects, the heights of the four EE grasps are plotted as deviations from the mean of the heights of the 12 grasps made under the other three precision conditions (EH, HE and HH) with the shelf at that height. The dashed line at 0 represents these baselines. The asterisks in the middle of each plot indicate +/- one standard deviation about the mean.*

Because we are working with mean-normalized data, the likelihood function for the mean of the normalized data is symmetric about a peak at 0. Because we implicitly assume a uniform prior, the posterior probability distribution for the baseline mean on this mean-normalized axis is simply the likelihood function normalized so that it integrates to 1. This distribution specifies our uncertainty about what the true value of the mean of the source distribution from which these baseline data were drawn really is. After looking at Figure 13, where we note that the deviations of EE data from the baseline mean are always less than .5, and because we conjecture that the relaxation of demands on precision placement will, if anything, lead subjects to grasp higher on the handle, we adopt an increment prior that is uniform on the interval between 0 and .5. In other words, we hypothesize that the effect of a wider ring, with the consequent relaxation of demands on precision placement of the bulb, may increase grasp height by at most half the length of the handle. (Remember, the vaguer we are about how big the effect may be, the more poorly the hypothesis that the ring diameter has an effect will fare when pitted against the hypothesis that it has no effect.) When we convolve this uniform increment prior with the posterior probability distribution of the baseline mean in order to take into account our uncertainty about where exactly the baseline is, we get the dashed line prior seen in the two panels of Figure 14.

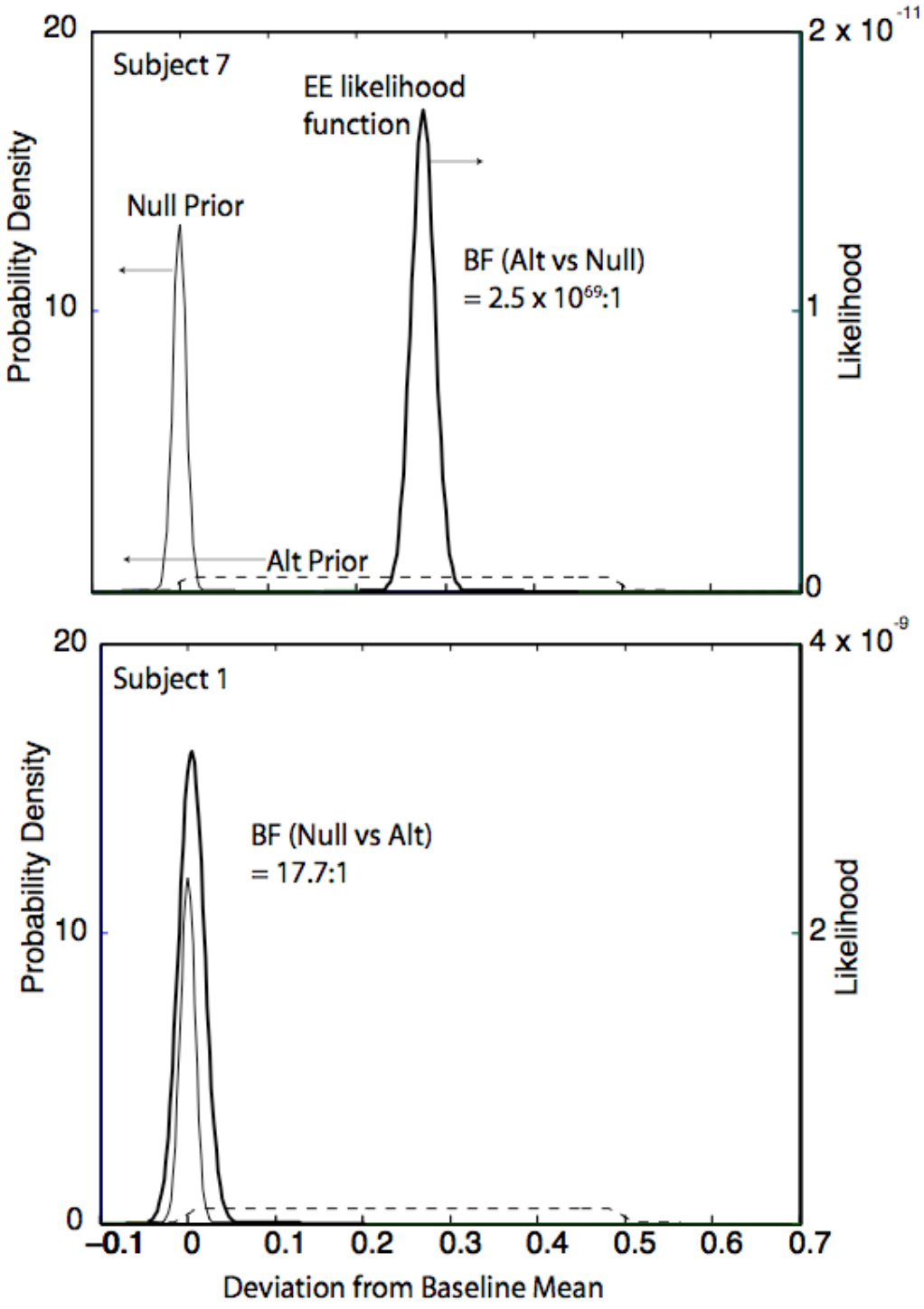

Figure 14. *In Subject 7, the likelihood function for the EE data is far from the null prior, so the Bayes Factor is astronomically in favor of the alternate hypothesis, namely that the ring diameter did affect grasp height in this subject. In Subject 1, by contrast, the likelihood function for the EE data falls on top of the null prior, which is much more precise in its predictions than the alternative hypothesis, so the Bayes Factor favors the null hypothesis by almost 18:1.*

Figure 14 is representative of what emerges in general from this analysis: for four subjects, the odds favor the hypothesis that the diameter of the rings had an effect (by odds of “astro”, “astro”, 2.1 and 900:1), while for six subjects, they favor the null hypothesis (by odds of 18, 1.7, 26, 67, 28 and 36:1). The odds ratios both ways are generally large, meaning that the evidence is decisive for most subjects. Thus, in this experiment, too, averaging across subjects would not seem to make scientific sense, although I would emphasize that in doing so Rosenbaum et al followed standard procedure (as evidenced by the fact that Turk-Browne did the same). The practice of averaging across subjects is ubiquitous, but the rationale for it is rarely discussed, let alone questioned.

In particular, it does not make sense to include in the analysis for additivity subjects whose behavior is unaffected the ring-diameter factor. The question of additivity 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 of the ring-diameter factor at different levels of the shelf-height factor. Thus, I now move to an analysis for additivity in Subject 7, the subject who showed the biggest effect of the variation in ring diameter, as may be seen in Figure 13.

The ANOVA test for additivity does not constrain the question to what one might call a test for systematic, that is, monotonic changes in the size of the effect of ring diameter with changes in shelf height. 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, will yield a significant interaction. In Bayesian terms, the implicit alternative model has as many additional free parameters as there are levels of the shelf-height factor. But arbitrary, subject-specific patterns of departure are presumably of little or no scientific interest. What we would like to know is whether in those subjects for which the ring-diameter is taken into account in their movement planning, does the effect of ring diameter tend to converge toward 0 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 0 as the shelf height increases, then our alternative only has one additional free parameter (rather than five), namely the slope of the EE data. On the null hypothesis (no convergence), the slope of the EE data is the same as the slope of the base data (the data from the EH, HE and HH conditions; whereas, on the convergence hypothesis, the slope of the EE data is somewhat greater, so that as shelf height increases, the effect of a difference in ring diameter decreases linearly.

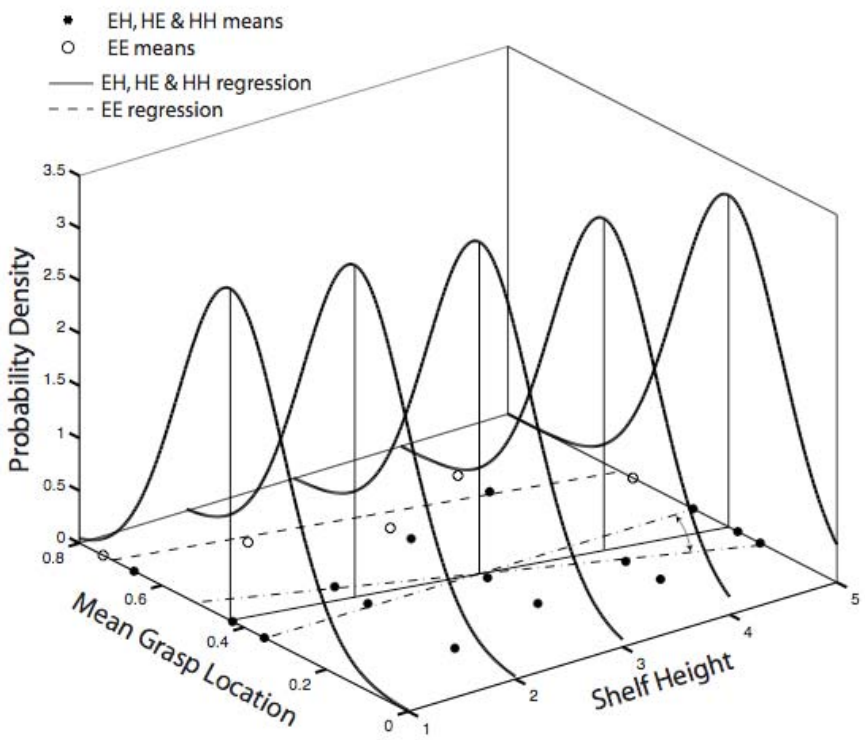

Figure 15. *The heavy solid curves are copies of a common source distribution centered on a regression line, which, like all regression lines, is constrained to pass through the centroid of the data (at Level 3). The positions of the copies at Levels 1, 2 4 and 5 relative to the underlying data, hence, also the likelihoods of those data, depend on the slope of the regression line. The maximally likely regression lines for the base and 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 those data as a function of that slope, we get the likelihood function for the slope.*

Figure 15 portrays the computational problem 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 what we want. What we want is an assessment of how strong the evidence is that they *are* the same. That is what the conventional analysis is in principle incapable of giving us. To get that, we need to compute the posterior 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 in a Bayesian analysis, we are going to have to get quantitative about the "somewhat." We are going to have to put an upper limit on how much greater we think that slope might be, because that upper limit will determine how well the second hypothesis fares when pitted against the null.

Figure 15 shows how we get the likelihood function for the slope of a regression line. We estimate the maximally likely standard deviation for the (assumed-to-be-normal) common source distribution, using the residuals about the maximally likely line of regression. We use the regression line to position copies of that source distribution over the underlying data at each level of Shelf Height (heavy curves extending upward from the base plane). The likelihood of any one datum for any one assumed slope (e.g., the likelihood of any one asterisk on the base plane) is read off the overlying copy of the source distribution. As we rotate the regression line about the centroid of the data, the copies of the source distribution at Levels 1, 2 4 and 5 slide over the underlying data, changing the likelihoods of those data. As always, the overall likelihood of the data (of, for example, the 3 x 5 =15 base means plotted as asterisks) at any one slope is the product of the individual likelihoods. The overall likelihood, as a function of the assumed slope, is the likelihood function for the slope.

If we assume a uniform prior in computing the posterior likelihood function for the slope of the base data, then rescaling the likelihood function so as to make it integrate to 1 gives us the posterior probability distribution for the slope of the regression line through the base data. This is our null prior for the computation of the posterior likelihood function for the slope of the line through the EE data (Figure 16). It represents our uncertainty about what the base slope really is. The null hypothesis asserts that whatever the base slope really is, the slope of the EE line is the same.

The alternative to the null hypothesis is that the slope of the regression line through the EE data is somewhat greater than the slope of the line through the base data, so that it converges on the latter line as the shelf height is increased. The greater the difference in the two slopes, the more rapid the convergence. The question is, How much greater a slope is it reasonable to assume? We want an increment prior, which specifies the range over which it is reasonable to suppose that the two slopes might differ. As always, we fix the lower end of that range at 0, making the null hypothesis a special case of the alternate hypothesis. 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. That slope increment is slightly less than -.15, so we adopt that as the upper limit on our uniform increment prior (a rectangle, whose sides are

located at -.15 and 0 on the slope axis). Because we do not know what the slope of the regression line through the base data really is, we convolve the increment prior with the probability distribution for that slope in order to get the prior probability distribution for the alternative (convergence) hypothesis shown in Figure 16.

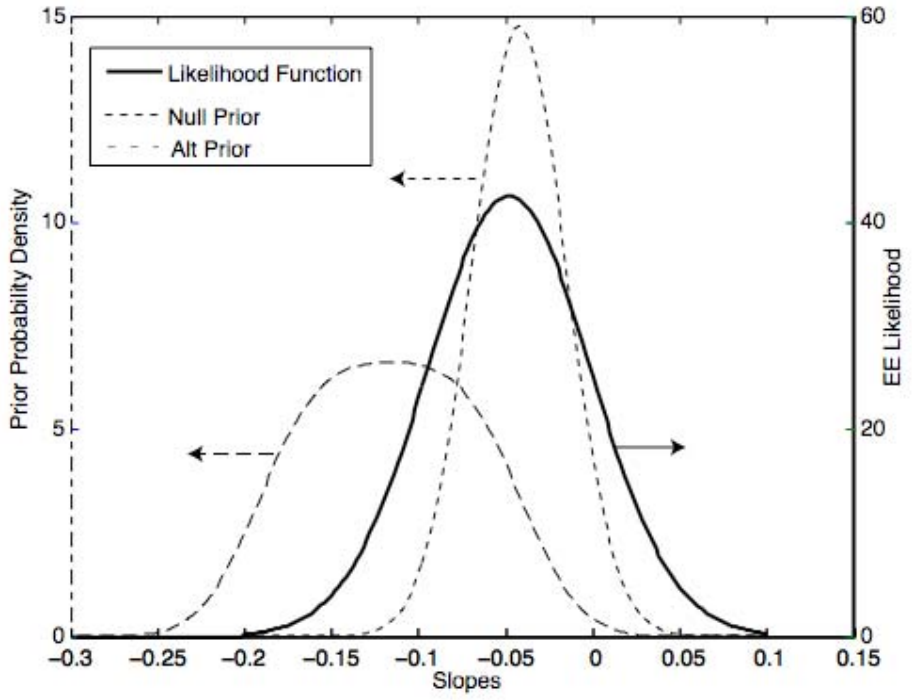

Figure 16. *The null prior and convergence prior for the slope of the regression line through Subject 7's EE data (plotted against left axis), together with the likelihood function for the slope of those data (plotted against right axis). The null prior is the posterior probability distribution for the slope of the base data. The convergence prior is the convolution of that distribution with an increment prior uniform on the interval from –.15 to 0.*

We can see in Figure 16 that null prior places more prior probability mass under the EE likelihood function than does the Convergence prior. Indeed, 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. This uncertainty about just how vague we should be is common and more or less unavoidable. It is an obstacle to the use of Bayesian analysis, because it is somehow silly to get into an argument with yourself or others about how much vagueness is "fair." What I have here proposed is a generally applicable solution to this problem, which is to examine the Bayes Factor as a function of how vague we allow ourselves to be. Figure 17 plots the Bayes Factor as a function of (negative) limit on the increment prior used in obtaining the convergence prior. As this limit is reduced to zero, the convergence prior becomes indistinguishable from the null prior.

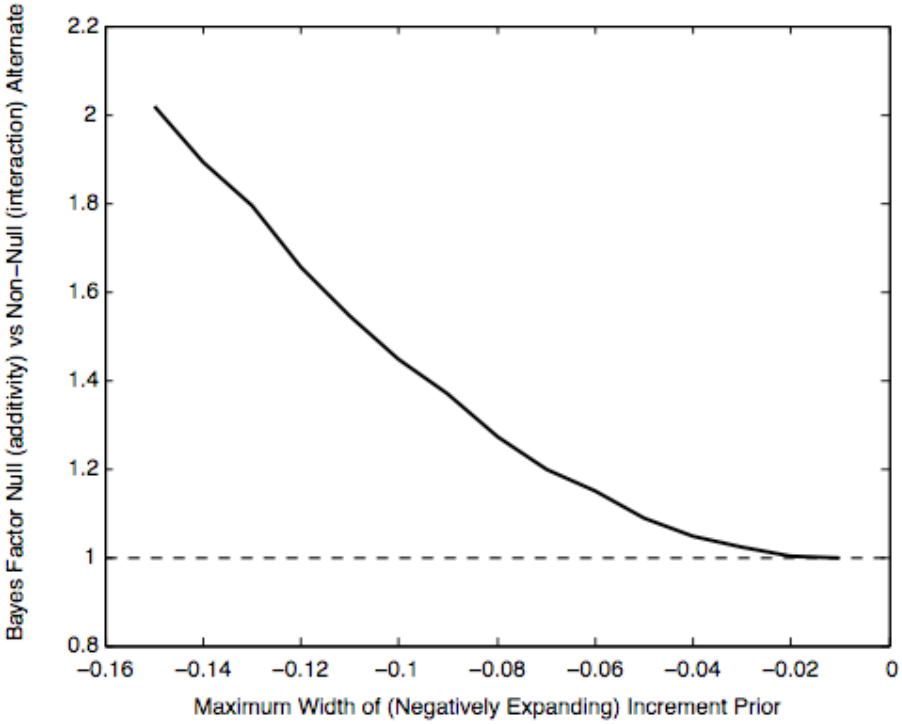

Figure 17. *The Bayes Factor for the Null-vs-Convergence comparison as a function of the assumed upper limit on the rate of convergence (upper limit on how much more negative the EE slope is than the base slope). The dashed line at 1 represents equal odds. The odds converge on this line. For any non-negligible width of the increment prior, the odds favor the null hypothesis, the hypothesis of additivity.*

Figure 17 shows that there is no way to improve on the null hypothesis; the only way to make the hypothesis that the EE slope is somewhat greater than the baseline slope equal in likelihood to the null hypothesis (that the slopes are the same) is to make the assumed difference in slope negligible. Thus, although the evidence for the null hypothesis is not as strong as we might wish; it is nonetheless clear that the data favor the additivity hypothesis (no interaction) over the interaction hypothesis. It is not simply that the data do not allow us to reject the additivity hypothesis (that is, the hypothesis of no interaction); they favor it.

The analysis makes clear how to strengthen the data. One would not want simply to add data from more subjects to the ANOVA. We know from the data already in hand that, at least under these experimental conditions, half or more of the subjects show no effect of the ring-diameter variable. Data from them is useless in testing for additivity. One needs to test subjects that show a big effect of this ring diameter, and one may want to modify the conditions so as to make the effect as big as possible (using, for example, really tight ring with a deep indentation). Moreover, most of the power in a test for convergence comes from the data at the extremes of shelf height. Therefore, in testing for additivity, one would want to eliminate trials at intermediate shelf heights and run instead more trials at the two extremes.

Discussion

A number of theoretical issues of broad scope in experimental psychology have been addressed. First among them, is the point with which we began, namely, that the null hypothesis is often the most interesting and theoretical consequential hypothesis. I have illustrated this with three examples. The examples are of sufficient substance and are drawn from a sufficiently broad spectrum of research to make this important point and establish the generality of its application.

The hypothesis that when the duration of training is fixed, the number of trials within that duration has no effect on the progress of learning, if is true, must result in a profound revision of our understanding of what goes on in the brain in "elementary" learning paradigms. (In connectionist models of all kinds, the progress of learning is a function of the number of trials; no account is taken of trial spacing.)

Similarly, the hypothesis that there is no statistical learning when a stream is not attended to bears strongly on viable models of the process of statistical learning and on models of attention, as well.

Finally, the hypothesis that different factors enter into the planning process independently and therefore have additive effects on movement parameters is of central importance to models of motor planning.

These are all null hypotheses. None of them is a straw man. The field of psychology has a substantial stake in statistical methods that assess the weight of the evidence in favor of hypotheses like these.

The second point is that conventional statistical analysis is incapable of assessing the strength of the evidence *for* these hypotheses. It treats all null hypotheses as straw men; they can aspire only to rejection, never to acceptance. This tells us that conventional statistical analysis is not a normative theory for inference in the face of uncertainty, because no normative theory would make it impossible for probabilistic considerations to weigh in favor of the most interesting and theoretically consequential hypotheses.

The third point is that Bayesian analysis is a normative model of probabilistic inference; indeed, Bayesian's would argue that it is *the* normative model. In Bayesian analysis, the null is on an equal footing with other hypotheses; the analysis may or may not support it.

The fourth point is that in Bayesian analysis, vagueness about the size of the effect predicted by a model weighs against it when it is pitted against less vague models. Null hypotheses are precise; they predict the size of the effect, namely 0. If the function of a model is to tell us what to expect, then this penalty on vagueness is a good thing, because a vague model does not really tell us what to expect. It retrodicts—after the fact, when we have seen the data and can estimate from the data the size of the effect—but it does not predict with any precision. Thus, a vague model is less useful than a precise model whose predictions are close to what is actually observed, even when they are not spot on. With a vague model, you cannot tell how far off the mark the model is once one has seen some data, because it never was clear where it said the data would fall. That is why a vague model can only be vaguely right. If the purpose of a model is only to give us a way of talking about the results after we have seen them, then vague models are to be

preferred, but if the purpose is to anticipate the results of future experiments, then precise models are what we want.

In this connection, it is important to appreciate that the null hypothesis may be a more likely model than a vague alternative even when the data allow us to reject the null at the conventional alpha of $p < .05$. As the analysis of data from Experiment 1 in the Turk-Browne and Scholl series shows, this is true not just for preposterously vague alternatives to the null. In that experiment, for the attended triplets, odds were 15:1 in favor of the null (chance performance) when it was pitted against what is implicitly taken to be the alternative in a conventional analysis, namely that performance is "above chance." 'Above chance' presumably means anywhere in the interval above .5. The data from the attended triplets in the Turk-Browne's Experiment 1 favored the null relative to this hypothesis, even though performance was significantly above chance by conventional analysis. The data favored the alternative to the null only when we reformulated the alternative to be "slightly above chance." This gain in the precision of our conclusions is, I suggest, greatly to be prized.

It must not be thought that Bayesian analysis stacks the deck in favor of the null. The null is supported only when the data fall close to where it predicts they will. When the data fall elsewhere, then a vague hypothesis that puts some non-negligible prior probability under the likelihood function becomes astronomically more likely than a precise hypothesis that puts no prior probability where it is needed—see, for example, the top panel of Figure 14. An advantage of a precise hypotheses is that it can also be precisely wrong; it is unequivocally refutable (Popper, 1959),.

In short, the good news about Bayesian analysis is that it penalizes vagueness. The bad news is that it forces us to specify our vagueness. We cannot be vague about how vague our hypothesis is. This is hard. There are no firm guidelines. The lack of guidelines governing how vague a prior should be can lead to unedifying arguments about how vague it "ought" be in order to be "fair". This, I claim, is the essence of the "problem of the prior," which is often cited as the reason not to use Bayesian analysis (e.g., Killeen, 2005). I have here proposed a general solution to this problem, which is to fix the lower limit on our vagueness at 0, thus including the null as a special, limiting case, and then to examine the strength of the evidence for or against the null as a function of the upper limit on the possible size of an effect. This leads to truly meaningful characterizations of effect size. It can always be done, and it always promises less vagueness in our conclusions. If it shows that the odds against the alternative and in favor of the null approach one (even odds) only as the upper limit on the size of the possible effect approaches 0, that tells us that the null is the best model. It is simple, precise, and no alternative to it is more likely in the light of the data.

Another point that I hope became clearer from the application of full Bayesian analysis to these diverse examples is that, when we stop to think, other information together with analytic considerations almost always limits the plausible size of any difference between two or more means. We often know more than we take into account when we analyze our data by conventional means. For example, analytic considerations tell us that mean trials to acquisition cannot be less than 0 nor more than the maximum number of trials recorded in our data. Analytic considerations tell us that the true probability of correct choice in the 2-alternative forced choice paradigm, such as Turk-

Browne used, cannot be less than .5 nor more than 1. Plausibility considerations tell us that the slope of the EE regression in Figure 15 is unlikely to be so much steeper than the slope of the base regression as to make the lines actually cross. These simple, obvious considerations establish limits on the vagueness of an alternative to the null. Bayesian analysis brings these considerations into the analysis. It discourages us from implicitly admitting into the realm of statistical possibility things that we know a priori either cannot be true or are exceedingly unlikely to be true.

In short, there is a broadly applicable recipe for the conduct of Bayesian analysis in place of the ubiquitous t-test and ANOVA, which test for "significant" differences between means. The recipe starts with the elaboration of a prior probability distribution of the mean of the experimental group, assuming the null hypothesis. If there is a control or comparison or baseline group, this distribution is usually the likelihood function for the mean of the control or baseline data, scaled so as to integrate to 1. The computation of a likelihood function is conceptually straightforward (see Figure 3) and easily carried out by numerical, not analytic means, in R or Matlab or Excel (see Supplementary Materials). If the reference is a chance level of performance, there is nothing to compute; the null prior is the Dirac delta function at the chance level, which is a formal way of saying that all of the prior probability mass is at .5.

The second step is to elaborate an increment prior (or increment/decrement prior). This specifies in probabilistic terms the range of plausible differences between the control or baseline mean and the experimental mean, under some hypothesis about the effect of the experimental manipulation. This is the conceptually challenging part of the analysis; it forces us to think about what is and is not plausible, about what the analytic constraints are, and so on. The elaboration of the increment prior should focus first on the limits on the possible magnitude of the effect, rather than on the form of the distribution of possible effect sizes. In most cases, the lower limit should be set at 0, thereby including the null as a special case (nesting the null). This leaves only the upper limit to be specified. If locating that upper or outer limit(s) is an issue—and it often is—the issue can be resolved by computing the analysis as a function of the upper limit (or, in the case of a bidirectional increment/decrement prior, as a function of the outer limits). This step brings the consideration of possible effect sizes into the heart of the analysis.

The form of the increment/decrement prior distribution within its limits will often be uniform (flat). Then, the distribution is simply a rectangle, which gets higher as it gets narrower, because its area must always be 1. If the experimental effects are plausibly scalar (multiplicative) rather than additive, as in Gottlieb's trials-to-acquisition experiment, then the null prior and the increment prior, should be elaborated over a logarithmic axis, rendering the postulated increments additive.

The third step is to convolve the increment prior distribution with the null prior distribution to obtain the prior distribution for the alternative to the null. This step takes into account the uncertainty about where the control or baseline mean really is. The resulting prior probability distribution for the location of the experimental mean reflects both this uncertainty and the uncertainty about the size (and possibly the sign, that is, direction) of the experimental effect.

If the reference or null condition is chance performance, convolution is not necessary because there is no uncertainty about what the chance level is; it derives from analytic considerations. In that case, the increment prior is the prior probability distribution for an alternative (non-null) model of the data. An attractive alternative to a uniform increment prior in this case is obtained from a beta distribution with its mode at the chance value[3], which is truncated at the mode and renormalized (rescaled so as to integrate to 1). This puts all of the prior probability above the chance level, in a distribution that falls off gradually, rather than abruptly (see, for example, the dashed distribution in Figure 11). As the parameters (A and B) of the truncated beta distribution are scaled up, the distribution becomes more and more closely confined to the region just above chance. Thus, scaling up the parameters of a truncated beta distribution achieves the same effect as lowering the upper limit on a uniform increment prior.

The fourth step is to compute the likelihood function for the experimental data and plot it over the same axis as the competing prior probability distribution functions (Figures 5, 9, 10, 11 and 14). An advantage of Bayesian analysis is the extent to which it can be rendered graphically transparent. When you have plotted the priors for different models of the data and the likelihood function over the same axis, you have before you in immediately apprehensible form all you need to know. The computation of the Bayes Factors is a formality. It always confirms what you can see. Moreover, the likelihood function for the experimental data is better than any confidence interval; it shows the full shape and extent of your uncertainty about where the central tendency of the experimental data is.

Like all Bayesian expositions, the present exposition contains many terms unfamiliar to those trained in the conventional approach: likelihood functions, posterior likelihood functions, posterior distributions, prior distributions, source distributions, increment priors, and so on. The unfamiliar terminology tends to obscure the essential simplicity of the approach: You have hypotheses that predict, however vaguely, what you might expect to observe. These vague predictions are represented by competing prior distributions. You compute the likelihood function, which represents your uncertainty about what the true value of the mean of the population is. Roughly speaking, this is what is represented by the confidence interval about a sample mean. You compare the likelihood function to the competing prior distributions. The hypothesis whose prior distribution better matches the likelihood function is the hypothesis favored by the data. The essential simplicity of the approach is seen in Figures 5, 9, 10, 11 and 14. The supplementary material shows that only a few lines of code are required for the computation of a likelihood function or a prior distribution function.

A final point of broad theoretical relevance that emerged in two of the analyses is that we need to think about the assumptions we implicitly make when we follow the common practice of averaging across subjects in conventional statistical analyses. Sometimes, we have no choice. In learning experiments, for example, it is not possible to measure trials to acquisition more than once in the same subject. Thus, we have no way of estimating our uncertainty about what the "true" value of the trials to acquisition

[3] The mode of the beta distribution is determined by A/(A+B), where A & B are the parameters of the distribution.

variable is in individual subjects. All we can say is whether the distribution of such values across a group of subjects is or is not affected by an experimental manipulation, such as an 8-fold reduction in the number of trials. Often, however, we make multiple measurements on individual subjects. We do so whenever we estimate proportions, for example. In such cases, we can ask whether one model applies to some subjects and a different model to others. In the case of statistical learning, for example, judging from the Turk-Browne data, some subjects do learn and some do not, even when they are attending. Conversely, when they are not attending, most subjects do not learn, but an occasional subject either does learn or successfully evades the measures taken to induce inattention and reliance on unconscious assessments of statistical dependencies. Similarly, in the motor planning experiment, it appears that some subjects do take the precision of the required bulb placement into account and some do not. In considering the question of additivity, there is no point in working with data from those who do not. It is true in both the statistical learning case and the motor planning case that the distribution of underlying values in the subject population is shifted by the manipulation. But is the central tendency of the distribution of much interest, once we know that the distribution includes a substantial fraction of subjects to which the no-effect model applies? Is it not then of more interest to estimate the fraction of subjects to which one model applies and the fraction to which a different model applies, rather than to estimate the mean of a distribution that includes both kinds of subjects?

References

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

Brandon, S., Vogel, E., & Wagner, A. (2002). Stimulus representation in SOP:I. Theoretical rationalization and some implications. *Behavioural Processes (SQAB '02 special issue), ?*(?), ?

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

Gallistel, C. R., Balsam, P. D., & Fairhurst, S. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences, 101*(36), 13124-13131.

Gibbon, J., & Balsam, P. (1981). Spreading associations in time. In C. M. Locurto, H. S. Terrace & J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 219-253). New York: Academic.

Glover, S., & Dixon, P. (2004). Likelihood Ratios: A Simple and Flexible Statistic for Empirical Psychologists. *Psychonomic Bulletin & Review, 11*(5), 791-807.

Gottlieb, D. A. (2007, in press). Is the number of trials a primary determinant of conditioned responding? *Journal of Experimental Psychology: Animal Behavior Processes*.

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

Kass, R. E., & Raferty, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773-795.

Killeen, P. R. (2005). Replicability, Confidence, and Priors. *Psychological Science, 16*(12), 1009-1013.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. New York: Cambridge University Press.

Morris, R. W., & Bouton, M. E. (2006). Effect of Unconditioned Stimulus Magnitude on the Emergence of Conditioned Responding. *Journal of Experimental Psychology: Animal Behavior Processes 32*(4), 371-385.

Papachristos, E. B., & Gallistel, C. R. (2006). Autoshaped Head Poking in the Mouse: A Quantitative Analysis of the Learning Curve. *Journal of the Experimental Analysis of Behavior, 85*, 293-308.

Papini, M. R., & Brewer, M. (1994). Response competition and the trial-spacing effect in autoshaping with rats. *Learning & Motivation, 25*(2), 201-215.

Popper, K. R. (1959). *The logic of scientific discovery*. New York,: Basic Books.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), *Attention and*

performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (Vol. 14, pp. 611-653). Cambridge, MA: MIT Press.

Rosenbaum, D. A., Halloran, E. S., & Cohen, R. G. (2006). Grasping movement plans. *Psychonomic Bulletin & Review, 13*(5), 918-922.

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

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552-564.

Wagenmakers, E. J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science, 17*(7), 641-.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: memory mechanisms* (pp. 5-47). Hillsdale, NJ: Lawrence Erlbaum.