To appear in: *Wiley Interdisciplinary Reviews: Cognitive Science*
Draft of September 17, 2009.

Bayesian data analysis

John K. Kruschke
Indiana University

Bayesian methods have garnered huge interest in cognitive science as an approach to models of cognition and perception. On the other hand, Bayesian methods for data analysis have not yet made much headway in cognitive science against the institutionalized inertia of 20th century null hypothesis significance testing (NHST). Ironically, specific Bayesian models of cognition and perception may not long endure the ravages of empirical verification, but generic Bayesian methods for data analysis will eventually dominate. It is time that Bayesian data analysis became the norm for empirical methods in cognitive science. This article reviews a fatal flaw of NHST and introduces the reader to some benefits of Bayesian data analysis. The article presents illustrative examples of multiple comparisons in Bayesian ANOVA and Bayesian approaches to statistical power.

Keywords: statistics, null hypothesis significance test, multiple comparisons, statistical power, replication probability

This brief article assumes that you, dear reader, are a practitioner of *null hypothesis significance testing*, hereafter abbreviated as NHST. In collecting data, you take care to insulate the data from your intentions. For example, double-blind procedures in clinical trials insulate the data from experimenter intentions. As another example, in field research the observers construct elaborate "duck blinds" to minimize the impact of the observer on the data. After carefully collecting the data, you then go through the ritual invocation of $p < .05$. Did you know that the computation of the p value depends entirely on the covert intentions of the analyst, or the analyst's interpretations of the unknowable intentions of the data collector? This is true despite the emphasis by the data collector to make the data unaffected by his/her intentions, as will be shown below. Moreover, for any set of data, an intention can be found for which p is *not* less than .05.

There is a better way to draw inferences from data. Bayesian data analysis is gaining acceptance in many fields as the best way to conduct data analysis, but many disciplines within cognitive science have been slow to re-tool. This brief article reviews a fundamental problem with NHST, and shows some of the advantages of Bayesian data analysis. Although there have been a number of previous articles that have come to a similar conclusion, the present article emphasizes different points. In particular, this article emphasizes the fatal role of experimenter intention in NHST, and that this fault is inherent in confidence intervals too. The article highlights multiple comparisons of groups as an illustration of the advantages of Bayesian analysis. This article also presents two perspectives on Bayesian interpretation of null effects. Finally, this article describes Bayesian approaches to statistical power, more generally framed as the probability of achieving a research goal.

The author thanks Michael Erickson and Luiz Pessoa for helpful comments on a draft of this article. Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to kruschke@indiana.edu. The author's world wide web page is at http://www.indiana.edu/~kruschke/

The road to NHST is paved with good intentions

Many previous articles have reviewed various problems with NHST (e.g., Wagenmakers, 2007). This article will focus on the crucial and fatal problem: The p value in NHST, upon which we base our inference, is dependent upon the intentions of the experimenter. This dependence exists despite the fact that conscientious researchers deliberately insulate their data collection from their intentions.

To make the issue concrete, consider an example. You have a scintillating hypothesis about the effect of some different treatments on a metric dependent variable. You collect some data (carefully insulated from your hopes about differences between groups) and compute a t statistic for two of the groups. The computer program, that tells you the value of t, also tells you the value of p, which is the probability of getting that t by chance from the null hypothesis. You want the p value to be less than 5%, so that you can reject the null hypothesis and declare that your observed effect is significant.

What's wrong with that procedure? Notice the seemingly innocuous step from t to p. The p value, on which your entire claim to significance rests, is conjured by the computer program with an assumption about your intentions when you ran the experiment. The computer assumes you intended, in advance, to fix the sample sizes in the groups.

In a little more detail, and this is important to understand, the computer figures out the probability that your t value could have occurred from the null hypothesis *if the intended*

experiment were replicated many, many times. The null hypothesis sets the two underlying populations as normal populations with identical means and variances.. If your data happen to have 6 scores per group, then, in every simulated replication of the experiment, the computer randomly samples exactly 6 data values from each underlying population, and computes the t value for that random sample. Usually t is near zero, because the sample comes from a null-hypothesis population in which there is zero difference between groups. By chance, however, sometimes the sample t value will be fairly far above or below zero. The computer does a bizillion simulated replications of the experiment. The top panel of Figure 1 shows a histogram of the bizillion t values. According to the decision policy of NHST, we decide that the null hypothesis is rejectable by an actually observed t_{obs} value if the probability that the null hypothesis generates a value as extreme or more is very small, say $p < .05$. The arrow in Figure 1 marks the critical value t_{crit} at which the probability of getting a t value more extreme is 5%. We reject the null hypothesis if $t_{obs} > t_{crit}$. In this case, when $N = 6$ is fixed for both groups, $t_{crit} = 2.23$. This is the critical value shown in standard textbook t tables, for a two-tailed t-test with 10 degrees of freedom.

In computing p, the computer assumes that you did not intend to collect data for some time period and then stop; you did not intend to collect more or less data based on an analysis of the early results; you did not intend to have any lost data replaced by additional collection. Moreover, you did not intend to run any other conditions ever again, or compare your data with any other conditions. If you had any of these other intentions, or if the analyst believes you had any of these other intentions, the p value can change dramatically.

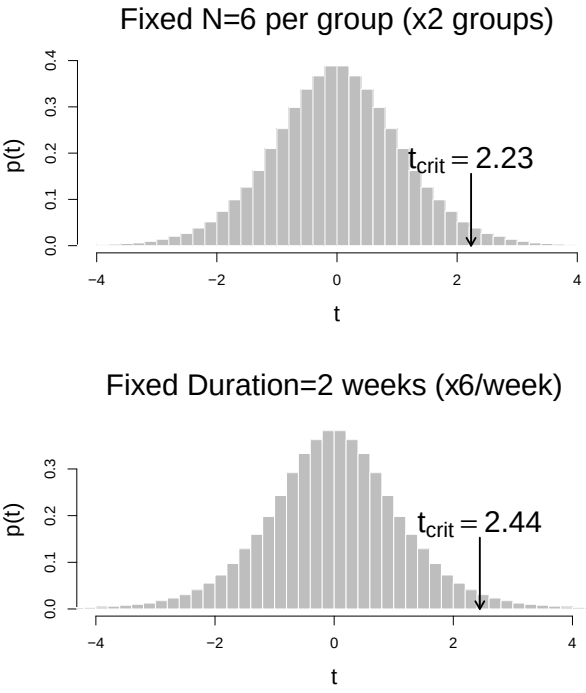

Figure 1. Sampling distribution of t for two groups when the null hypothesis is true. Upper: When the intention is to fix $N = 6$ for both groups, regardless of how long that takes. Lower: When the intention is to fix the duration of data collection at 2 weeks, when the mean rate is $N = 6$ per week.

The intention to collect data until the end of the week

In most of my research, I have only a rough sample size in mind, and I collect data for a period of time until that rough sample size is achieved. For example, I will post session times for which volunteers can sign up, and the posted times span, say, a two week period. I expect some typical rate of subject recruitment during that span of time, hoping to get a sample size in the desired range.

It is easy to generate a sampling distribution for t under these intentions. Specifically, suppose that the mean rate of subject sign-ups is 6 per week, with the actual number randomly generated by a simple Poisson process, as is often used to model the arrival of customers in a queue (e.g., Sadiku & Tofighi, 1999). The Poisson distribution generates an integer between zero and infinity, with a mean, in this case, of 12 for a two-week duration. Each subject is randomly assigned to one of the groups by a flip of a fair coin. Thus, in some random replications we will happen to get 6 subjects in each group, but in other replications we will get, say, 5 subjects in one group and 8 in the other. (On those rare occasions when this process allocates fewer than 2 subjects to a group, the number of subjects is promoted to 2.) For every simulated replication, the t value is computed. The resulting distribution of t values is shown in the lower panel of Figure 1. The value of t, at which only 5% of the distribution is more extreme, is $t_{crit} = 2.44$.

In summary, if the intention was to collect 6 subjects per group, then the null hypothesis predicts t values distributed according to the upper panel of Figure 1, but if the intention was to collect data for two weeks, with a mean rate of 6 subjects per week, then the null hypothesis predicts t values distributed according to the lower panel of Figure 1.

Suppose we are handed some data from two groups, with 6 values per group. We compute t and find that $t = 2.35$. Do we reject the null hypothesis? According to NHST we can only answer that question when we ascertain the intention of the experimenter. We ask the research assistant who collected the data. The assistant says, "I just collected data for two weeks. It's my job. I happened to get 6 subjects in each group." We ask the graduate student who oversaw the assistant. The student says, "I knew we needed 6 subjects per group, so I told the assistant to run for two weeks. We usually get about 6 subjects per week." We ask the lab director, who says, "I told my graduate student to collect 6 subjects per group." Therefore, for the lab director, $t = 2.35$ rejects the null hypothesis, but for the research assistant who actually collected the data, $t = 2.35$ fails to reject the null hypothesis.

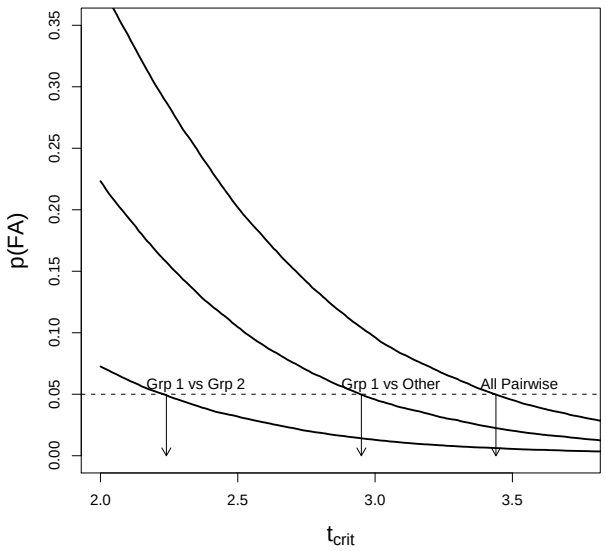

Figure 2. Probability that any of the t values among the comparisons exceeds the critical value, when the null hypothesis is true, and $N = 6$ is fixed for all five groups. The arrow labeled "Grp 1 vs Grp 2" is at $t = 2.23$ as shown in the top panel of Figure 1.

The intention to examine data thoroughly

Essentially all introductory NHST textbooks describe only fixed-N t-distributions. Although the books implicitly acknowledge the dependence of critical t values on this assumption of fixed sample size, they rarely illustrate how much the critical t depends on the assumption, nor how strange it is for the decision to depend on the assumption.

The only application in which standard textbooks regularly acknowledge the role of intent is *multiple comparisons*. When there are multiple groups, the analyst has the option of many different comparisons of different groups and combinations of groups. With every comparison, there is new opportunity to commit a false alarm, i.e., to have a t value that is spuriously or accidentally larger than the critical t value, even though there is no actual difference between groups. For example, consider a situation in which there are four groups, with intentionally fixed $N = 6$ per group. Suppose the analyst computes a t value for comparing group 1 and group 2. If the underlying populations are not actually different, then there is a 5% chance that the sample t will exceed $t_{crit} = 2.23$, i.e., there is a 5% false alarm rate. Suppose the analyst also computes the t value for comparing group 3 and group 4. Again there is a 5% chance that the sample t will exceed $t_{crit} = 2.23$ if the null hypothesis is true. If the analyst conducts *both* tests, however, there is a greater chance of committing a false alarm, because a false alarm can arise if *either* test happens to spuriously exceed the critical value.

Figure 2 shows how the probability of false alarm depends on the intended set of comparisons and the candidate critical t value. In all cases, there are five groups with fixed $N = 6$ per group. The lower-left curve shows the case where only a single comparison is intended. As the candidate critical t value gets larger, the false alarm rate gets smaller. The curve crosses $p(FA) = .05$ when $t_{crit} = 2.23$, which corresponds to the upper panel of Figure 1. The middle curve of Figure 2 shows the case in which the intent is to compare the first group with each of the other four groups. Here the curve crosses $p(FA) = .05$ when $t_{crit} = 2.95$. The rightmost curve shows the case in which the intent is to compare each group with every one of the others (i.e., 10 pairwise comparisons). In this case, the curve crosses $p(FA) = .05$ when $t_{crit} = 3.44$.

Now, suppose we actually run the experiment. We randomly assign 30 people to the 5 groups, 6 people per group. The first group gets the placebo, and the other four groups get the corresponding four drugs. We are careful to make this a double-blind experiment: Neither the subjects nor experimenters know who is getting which treatment. Moreover, no one knows whether any other person is even in the experiment. We collect the data. Our first question is to compare the placebo and the first drug, i.e., group 1 versus group 2. We compute the t statistic for the data from the two groups and find that $t = 2.95$. Do we decide that the two treatments had significantly different effects?

The answer, bizarrely, depends on the intentions of the person we ask. Suppose, for instance, that we handed the data from the first two groups to a research assistant, who is asked to test for a difference between groups. The assistant runs a t-test and finds $t = 2.95$, declaring it to be *highly significant* because it greatly exceeds the critical value of 2.23 for a two-group t-test. Suppose, on the other hand, that we handed the data from all five groups to a different research assistant, who is asked to compare the first group against each of the other four. This assistant runs a t-test of group 1 versus group 2 and finds $t = 2.95$, declaring it to be *marginally significant* because it just squeezes past the critical value of 2.95 for these four planned comparisons. Suppose, on yet another hand, that we handed the data from all five groups to a different research assistant, who is told to conduct all pairwise comparisons post-hoc because we have no strong hypotheses about which treatments will have beneficial or detrimental or neutral effects. This assistant runs a t-test of group 1 versus group 2 and finds $t = 2.95$, declaring it to be *not significant* because it fails to exceed the critical value of 3.43 that is used for post-hoc pairwise comparisons. Notice that regardless of which assistant analyzed the data, the t-value for the two groups stayed the same because the data of the two groups stayed the same. Indeed, the *data* were completely uninfluenced by the intentions of the analyst. So why should the *interpretation* of the data be influenced by the intentions of the analyst? It shouldn't.

Confidence intervals: Only as confident as intentions

Some practitioners of NHST give the impression that a lot of problems would be solved if researchers would report a *confidence interval* and not only a p value (e.g., Thompson,

2002). A confidence interval does convey more information than a point estimate and a p value, but rarely acknowledged is the fact that *confidence intervals depend on the experimenter's intentions in the same way as p values.*

A confidence interval is merely the range of hypothetical parameter values we would not reject if we replicated the *intended* experiment many times.[1] For any candidate value of a parameter, we generate a sampling distribution from simulated replications of the intended experiment, and determine whether or not our actually observed t falls in the extreme 5% of the sampling distribution (e.g. Cumming & Finch, 2001; Young & Lewis, 1997). Examples of sampling distributions were shown in Figure 2, for the case when the candidate value of the parameter is zero. The confidence interval considers all possible parameter values, not only zero. A candidate value of the parameter is not in the confidence interval if the observed t value falls in the extreme 5% of the sampling distribution. Because sampling distributions depend on the experimenter's intentions, confidence intervals also depend on the experimenter's intentions.

Readers familiar with how to compute the confidence interval for a sample mean may recall that the interval's limits are given by the actual sample mean plus or minus the critical t value times the sample standard error. If the experimenter intended to stop when $N = 6$, then the critical t value is 2.23, as shown in the upper panel of Figure 2, and therefore the width of the confidence interval is 2.23 times twice the standard error. But, if the experimenter intended to stop at the end of two weeks, then the critical t value is 2.44, as shown in the lower panel of Figure 2, and the width of the confidence interval is 2.44 times twice the standard error. In other words, the confidence interval is wider, even though the data have not changed, merely because the experimenter intended to stop after two weeks instead of when $N = 12$.

The dependence of the confidence interval on the intentions of the experimenter is rarely if ever acknowledged by computer programs for single t tests. Computer programs do reveal the dependency, however, when showing the results of multiple comparisons. The confidence intervals will indicate that they were determined by one or another "correction" for multiple comparisons.

Good intentions make any result insignificant

It's trivial to make any observed difference between groups non-significant. Just imagine a large set of other groups that should be compared with the two initial groups, and earnestly intend to compare the two groups with all the other groups, once you eventually collect the data. Poof! The false alarm rate sky rockets and any observed difference between the first two groups becomes insignificant. The analogous result holds for confidence intervals: The confidence interval becomes huge, merely by intending to compare the first two groups with lots of other groups.

Why persist with NHST?

The NHST hostage might attempt to defend his or her warden: "There is only a limited range of realistic intentions, and the p value does not change much within that range." This argument does not work. As was shown above, p values *do* change a lot with realistic changes of intentions. More importantly, data interpretation should not depend on intentions at all. To argue that the dependence is okay as long as it is not big is like saying it is okay for the butcher to put his finger on the scale because most butchers cheat roughly the same amount. The point is that it shouldn't happen at all.

In the movie *Annie Hall*, the final scene has the narrator explaining why people stay in dubious relationships, by analogy to an old joke: A guy goes to a psychiatrist, and says, "Doc, my brother thinks he's a chicken." Psychiatrist responds, "Well, why don't you turn him in?" Guy replies, "I would, but I need the eggs." The narrator concludes that we keep going through it because we need the eggs.

And indeed, NHST lays an egg. The p value, which is supposed to be NHST's main offering, is as fickle as intentions. Confidence intervals are just as groundless. Some NHST procedures do not even provide a confidence interval. For example, a chi-square test of independence yields a p value for the null hypothesis, but no range of believable cell proportions. (A hard-boiled NHST-er could derive sampling distributions for various hypotheses and laboriously approximate a confidence interval, but this is not done in any standard computer packages.) And NHST does not tell us how reliable a result is: The p value tells us little if anything about the probability that we would get a significant result if we repeated the experiment (Miller, 2009). There is a better way: Bayesian data analysis.

Bayesian data analysis

Suppose there is an upcoming election between candidates A and B. You ask a few friends which candidate they prefer, and you thereby get a vague sense that candidate A is preferred. But to get a better prediction of which candidate will win, you would want to poll a larger random sample of the population. Suppose you acquire data from such a poll. What is the effect of the data on your beliefs? The data cause you to shift your beliefs from the uncertain prior beliefs to more certain beliefs informed by the representative sample data.

The Bayesian idea: Data indicate how to reallocate beliefs

The role of data is to reallocate beliefs. Bayesian data analysis is merely the mathematical specification of that reallocation. First, the various candidate beliefs are specified as a space of possibilities. For example, the population's preference for candidate A could be any value from 0% to 100%. Next, the degree of belief in each value is specified as a probability. For example, there might be 90% belief that the population preference is in the range from .52 to .58. (Technically, for infinitesimal intervals, degree of belief is specified

[1] There are various ways of defining a confidence interval. In all definitions, the interval is a random entity that is defined in terms of its behavior across replications of the *intended* experiment.

as probability "density".) The distribution of belief probabilities over candidate parameter values is a mathematical description of our knowledge.

We go into any research situation with prior beliefs, which may be diffuse and uncertain, or which may be heavily concentrated on a narrow range of parameter values. Then we collect data (insulated from the researcher's intentions, of course). Based on the data, we decrease our belief in parameter values that are less consistent with the data, and increase our belief in parameter values that are more consistent with the data. The mathematical formula for the reallocation of beliefs is called *Bayes' rule*. Bayes' rule is just a general form that can be applied to many different specific models and situations, whereby it derives its vast ramifications.

Bayesian reasoning in everyday life is quite intuitive. First, there is the reasoning from Sherlock Holmes: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890, p. 93) This is just a description of Bayesian reallocation of beliefs: Even though the prior belief on a candidate may be small, if the other candidates are eliminated, then the posterior belief on the remaining candidate must be high. The complementary logic is judicial exoneration: If there are several distinct suspects, and one suspect is convincingly shown to have committed the crime, then the other suspects are exonerated. This is again Bayesian reallocation of beliefs: The prior belief in each suspect is moderate, but when one suspect is identified as the culprit and thus belief in that suspect becomes high, then the posterior belief in the other suspects drops.

The question of whether or not Bayesian mathematics captures other aspects of human cognition is of great interest to cognitive scientists (e.g., Anderson, 1990; Chater & Oaksford, 2008; Chater, Tenenbaum, & Yuille, 2006; Kruschke, 2008). Cognitive scientists do not know what models and parameters the mind might be using (or behaving as if it were using). A Bayesian model that adequately mimics human cognition might never be discovered. Data analysts, on the other hand, define the descriptive models that are useful for summarizing data. These descriptive models could be generic domain-independent models such as linear regression, or they could be domain-specific models (Lee, 2008). Given the models, the rational way to allocate beliefs among models and their parameters is via Bayesian analysis. The present article is about Bayesian data analysis, and only indirectly about Bayesian models of mind.

Data interpretation should *depend on prior beliefs*. Some people are leery of the Bayesian requirement to specify a prior belief, because they suspect that the prior belief is vague or idiosyncratic. Fortunately, these fears are unfounded. If prior beliefs are vague, this is not a problem. When priors are only weakly informed, then relatively small amounts of data will overwhelm the prior, and the specific form of the prior has little influence on the posterior (for typical moderate-scale models). Examples in subsequent sections will illustrate this fact. Lee and Wagenmakers (2005) provide a summary of mathematical formalizations of vague priors.

Priors for scientific data analysis are not idiosyncratically capricious and covert. Priors for an analysis are overtly specified and deemed reasonable by the scientific community, and in particular by skeptical reviewers of articles in peer-reviewed journals. Moreover, because priors are explicitly specified, it is straightforward to do Bayesian analysis with different plausible priors and report the invariance of the resulting posteriors.

Most importantly, it is crucial to recognize that it is *un*reasonable *not* to consider prior beliefs. For example, suppose we are trying to determine the underlying probability that a particular coin comes up heads. We know it is a regular coin manufactured at an official government mint. We flip it 1000 times and it comes up heads 535 times. While this result deviates somewhat from the 500 heads we would expect from a fair coin, we would be reluctant to declare that the coin is biased, because of the strong prior belief that the coin is fair. On the other hand, suppose that we are parachuted into a random foreign city where two candidates are running for mayor. We have no familiarity with this culture, so we have only the most vague prior belief regarding what the underlying preferences in the population might be. Suppose we poll 1000 people at random and find that 535 prefer candidate A. From this sample preference, which shows a seven percentage point advantage for candidate A, we would be fairly confident in asserting that there is indeed a real preference in the population, because our prior beliefs were weak.

These two situations, i.e., the strong-prior coin toss and the weak-prior election poll, are handled identically by NHST. If the intention was to flip the coin or conduct the poll until $N = 1000$, then $p = 0.015$ and we reject the null; i.e., the coin is biased and the population has a preference. If the intention was to flip the coin or conduct the poll until the end of the day (and there just happened to be 1000 outcomes in the sample) then p is different, but still identical for the two situations. NHST treats the coin and the poll identically, ignoring prior knowledge, but irrationally relying on the intended stopping rule for data collection. Bayesian inference, on the other hand, does not use the data collector's arbitrary and unknowable intentions, but does rationally incorporate the scientific community's prior knowledge.

An example of Bayesian data analysis

This section uses an extended example to illustrate some of the results and benefits of Bayesian data analysis. First, the example illustrates how parameters are estimated and how the posterior distribution on the parameters explicitly indicates which values are credible. In particular, "null" values may or may not be among the credible values. (A subsequent section describes a different Bayesian approach to testing a null value.) Next, a Bayesian ANOVA is conducted to show how multiple comparisons can be made without fear of inflated false alarm rates. Each comparison is merely a different perspective on the multi-dimensional posterior distribution. This section also describes how the posterior distribution inherently reveals correlations among credible param-

eter values, and how the posterior distribution can be used by other researchers to encourage cumulative growth of knowledge.

For purposes of illustration, consider data from a simple learning experiment. In this experiment, a person sees common words such as "radio" and "ocean" on a computer screen, and the person must learn which words indicate which keys to press in response. A trial consists of the words appearing, the learner pressing a key, the computer displaying the correct response, and the learner studying the feedback and then pressing the space bar for the next trial. In a *highlighting* design, the learner initially is trained on cases of two words, denoted PE and I, indicating outcome E. Later in training, two words, PL and I, indicate a different outcome L. The word labels are mnemonic: E is the early-learned outcome, PE is the perfect predictor of E, L is the later-learned outcome, PL is the perfect predictor of L, and I is an imperfect predictor. Notice that the early-trained $PE.I \rightarrow E$ and late-trained $PL.I \rightarrow L$ cases are symmetric: Each has a perfectly predictive cue word, and the cases share the imperfect predictor. If people learn the simple symmetry, then I should be equally strongly associated with both outcomes, and PE should be as strongly associated with E as PL is associated with L. These predictions are assayed by subsequent testing of novel cue combinations. It turns out that when people are tested with cue I by itself, there is a strong tendency to select response E. This is not merely a generic primacy bias, however, because when tested with cue combination $PE.PL$, there is a strong tendency to select response L. Readers interested in the cognitive theory underlying this perplexing phenomenon are referred to the overview provided by Kruschke (2010). The emphasis here is on methods for analyzing the data.

For present purposes, the focus is on the test trials after learning. There were several different cue combinations presented at test, such as I by itself, the combination $PE.PL$, PE and PL by themselves, and various others. For each test item, the learner's choice and response time were measured. The design can therefore be described as a repeated measures, within-subject design: Every subject provided data for every test item, repeated several times. Because repetitions of the same item were randomly interspersed among many other test items, we will assume that responses on each trial were independent of other trials. This is merely a simplifying assumption made for the analysis; it is typical of both NHST and Bayesian analyses. In principle, Bayesian models can incorporate trial-to-trial dependencies, but this is not undertaken here.

The emphasis of highlighting experiments has been on choice preferences, but Lamberts and Kent (2007) manipulated response time (RT) deadlines to explore whether (presumably slow) rule-application processes might underlie the preferences. Lamberts and Kent (2007) also examined RTs for unspeeded preferences, and modeled those RTs with the ADIT model (Kruschke, 1996). Because RTs for the various probes have implications for process models, it can be useful to compare their RTs. Because there are many different probes to compare, this immediately raises the issue of multiple comparisons, which will be addressed presently from a Bayesian perspective.

The analysis of RT used here is an "off the shelf" hierarchical Bayesian ANOVA model (Gelman, 2005; Gelman, Hill, & Yajima, 2009; Qian & Shen, 2007). For our illustrative purposes, we will model $\log_{10}(RT)$ by a normal distribution. (A more accurate model of RT distributions is provided by Rouder, Lu, Speckman, Sun, & Jiang, 2005) In the hierarchical Bayesian ANOVA, the goal is to estimate (a) the overall baseline RT, (b) the deflection away from that baseline due to each test item, and (c) the deflection away from that baseline due to each subject. The deflections are constrained to sum to zero. For example, suppose that the baseline RT is 1.5 sec. The test item PE is a single perfect predictor, and therefore is responded to quickly, which means it has a negative deflection relative to the baseline RT. The test combination $PE.PL$, however, consists of two conflicting cues, and therefore is responded to slowly, which means that it has a positive deflection relative to the baseline RT. Different individual subjects are faster or slower than average. The deflection from baseline for fast subjects is negative, and the deflection for slow subjects is positive. The test-item and subject deflections are modeled as coming from higher level distributions centered at zero. The variances of the higher-level distributions are estimated from the data.

The prior beliefs regarding values for the baseline, treatment, and subject effects are shown in Figure 3. The priors are only vaguely informed by the general background knowledge that response times are in range of single seconds. Thus, the prior belief for the baseline, denoted β_0 on a $\log_{10}(RT)$ scale, is a normal distribution centered at zero (i.e., at $0 = \log_{10}(1 \text{ sec})$), with a standard deviation of about 2. This prior only very weakly emphasizes RTs near 1 sec., expressing (perhaps unrealistically) huge uncertainty in the baseline RT. The purpose of using such a vague prior is to be as uncontroversial as possible. The prior distribution on the deflection due to the test item, a.k.a. the treatment effect and denoted β_j, is centered at zero but given a broad range so as not to presume a null effect. Similarly, the prior distribution on the deflection due to subjects, denoted β_i, is centered at zero but very broad. These rather non-committal priors have very little influence on the posterior, as will be seen.

The posterior for the baseline and various test-item deflections is shown in Figure 4. Like the prior, these posterior distributions are represented by a large random sample from the continuous underlying posterior.[2] Consider the top-left distribution in Figure 4, which shows the believable values for the baseline. The mean of the believable baseline values is $\log_{10}(RT) = 0.175$, which corresponds to a baseline response time of 1.50 sec. Notice that the breadths of the posterior distributions are tiny compared to the diffuse breadth of the prior, and that the variation from one test effect to another is

[2] The program for generating the sample was written in the R language (Ihaka & Gentleman, 1996), using the BRugs interface (Thomas, 2004) to the OpenBUGS version (Thomas, O'Hara, Ligges, & Sturtz, 2006) of BUGS (Gilks, Thomas, & Spiegelhalter, 1994).

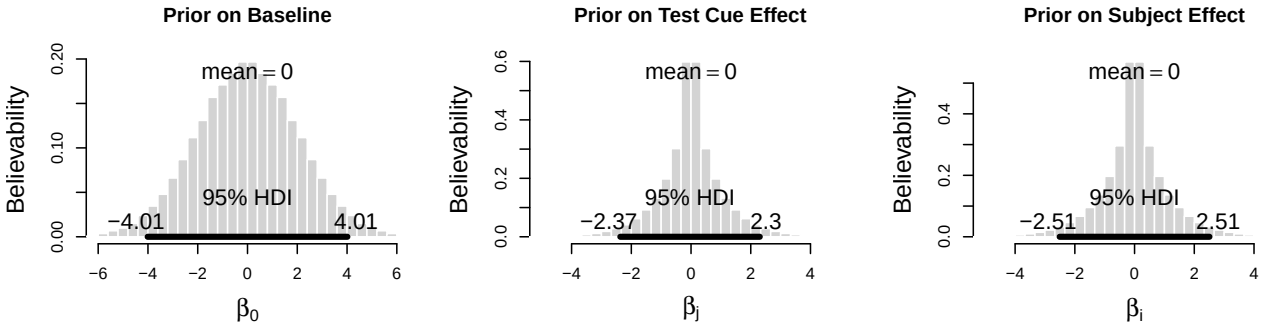

Figure 3. Prior probability distributions for parameters in Bayesian ANOVA. The left distribution is the prior for the baseline RT, denoted β_0. The β_0 values are in units of $\log_{10}$(sec.). The mean of the distribution is at $\log_{10}$(0sec.) = 1sec., indicating a prior belief that RTs are about 1 sec. in duration. Notice that the scale is broad, which indicates that the prior is only mildly informed by the knowledge that human response times are not on the order of nanoseconds or millennia. The middle distribution is the prior for all seven test cue effects, denoted β_j. The right distribution is the prior for all 64 subject effects, denoted β_i. Both the middle and right distributions indicate *deviations* from baseline, hence a prior mean of 0 indicates a mild preference for null effects. The dark bar labeled "95% HDI" indicates the highest density interval, i.e., the interval that contains 95% of the distribution such that parameter values outside the HDI have less believability than parameter values inside the HDI. These histograms were generated by a large random sample from the continuous and symmetric underlying distribution.

tiny compared to the breadth of the prior. This is one indication that the specific form of the prior had little constraint on the posterior. Indeed, if the priors are made more diffuse, the posterior hardly changes at all.

The posterior distribution explicitly reveals the extent to which various baseline RTs and cue effects are credible, given the data. Unlike NHST, we do not have merely a dichotomous decision to reject or fail-to-reject the null value, nor do we have a p value that mutates according to the intentions of the data collector. The posterior shows a continuous distribution of believability across the spectrum of parameter values. It can be summarized by values such as its mean and 95% highest density interval (HDI), which is the interval that contains 95% of the distribution such that all points inside the interval have higher believability than points outside the interval.[3]

Multiple comparisons. It is important to understand that the posterior is a *joint* probability distribution in a high-dimensional parameter space, and that what is shown in Figure 4 is only the *marginal* distribution of individual parameters, like pressing a flower between the pages of a heavy book. In other words, the posterior specifies the credible *combinations* of all the parameter values. Because believable parameter values might be correlated, the marginals of different parameters should not be directly compared to assess differences between parameters. Instead, the differences between parameter values are explicitly computed at only the believable combinations of parameter values. Figure 5 shows examples of these differences. For example, the top row's middle distribution shows the difference between the *PL* and *PE* deflections for all the believable combinations of β_{PL} and β_{PE}. The distribution reveals that most of the believable differences are greater than zero, and the 95% HDI does not span zero. Therefore it is quite credible that response times for *PL* are faster than response times for *PE*. The distribution of credible differences takes into account the fact that the estimated effects of *PE* and *PL* are negatively correlated, with $r = -0.187$. In other words, the uncertainty of the difference, shown in Figure 5, is appropriately larger than would be suggested by the individual distributions of *PE* and *PL* in Figure 4.

The joint posterior distribution on test-cue deflections can be summarized in many informative ways. For example, we might be interested in knowing whether ambiguous tests, including *PE.PL*, *I*, and *PE.PLo*, take longer on average than unambiguous tests, including *PE*, *PL*, *PE.I*, and *PL.I*. This comparison is easy to conduct: For every believable combination of the test deflections, we simply compute the average of the ambiguous-test deflections and subtract the average of the unambiguous-test deflections. Then we inspect the distribution of the differences, which is shown in the lower-right panel of Figure 5. The posterior distribution of differences is far above zero, and therefore it is highly credible that ambiguous tests take longer than unambiguous tests.

All the comparisons displayed in Figure 5 are merely different perspectives on the same joint posterior distribution. We can make these comparisons, and any others we care to think of, without worrying about inflated false alarm rates, because the posterior distribution does *not* change when we consider other comparisons. (Recall that in NHST the sample space *does* change when we consider other comparisons.) The joint posterior distribution specifies the believable parameter combinations, given the data. Any comparison we make is simply looking at that high-dimensional posterior distribution from a different perspective.

[3] What I call the highest density interval (HDI) is often referred to as the highest posterior density (HPD) interval. I use the more general term, HDI, so that it can be applied either to the prior distribution or to the posterior distribution.

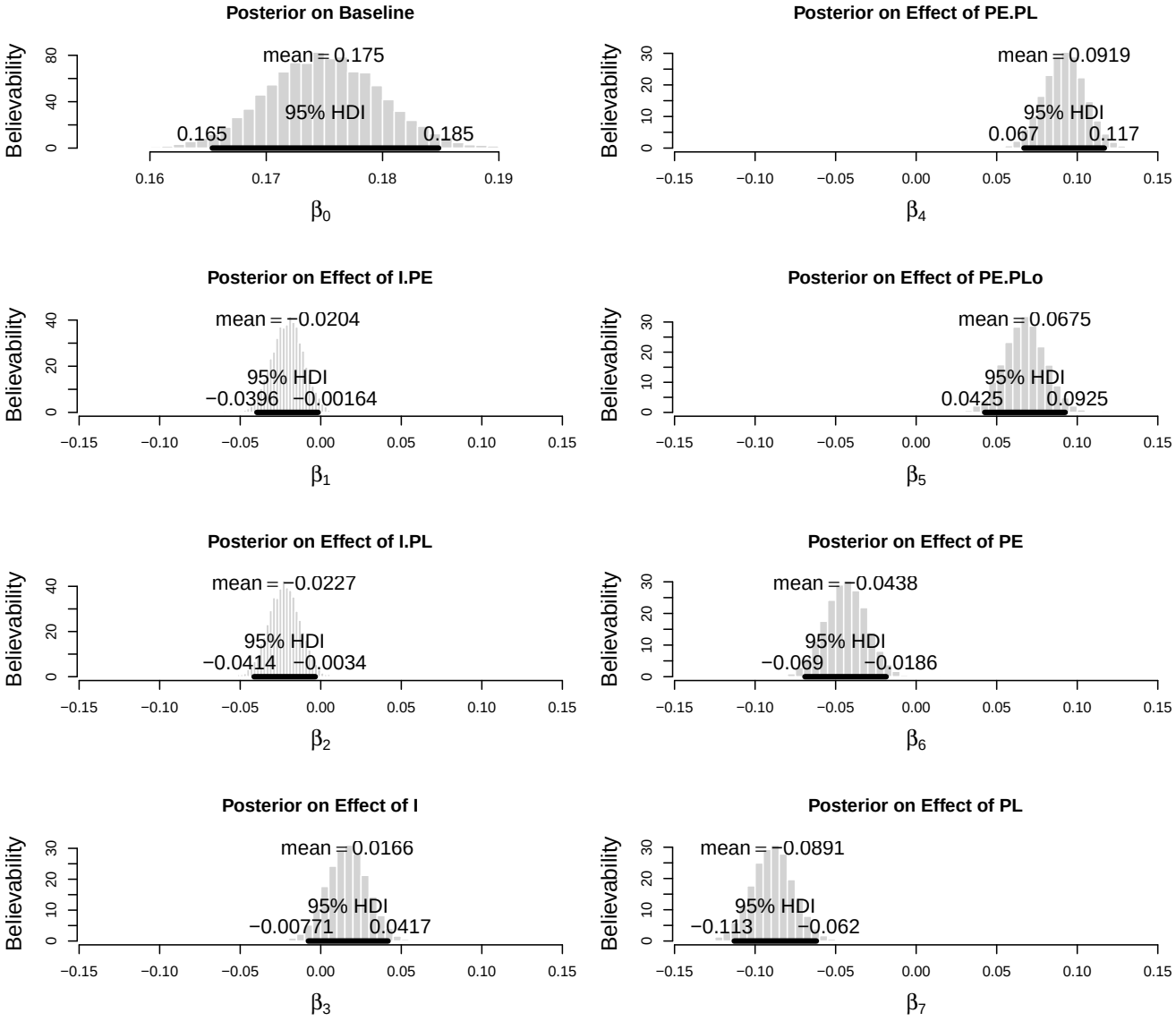

Figure 4. Posterior distributions for the test cue effects (compare with priors in Figure 3). The β values are in units of $\log_{10}(sec.)$; for example, the mean baseline $\beta_0 = 0.175$ corresponds to $10^{0.175} = 1.50$ *sec*.

A Bayesian analysis is not immune to false alarms. False alarms are unavoidable by *any* analysis because data are a random sample, and some random samples will accidentally comprise a coincidence of outliers. But Bayesian analysis says it is irrational to make decisions based on the probability of false alarms (i.e., the *p* value), because that probability is ill-defined: It depends on experimenter's intentions. Bayesian analysis instead says that the posterior distribution is the best we can do, given the data we have.

How, then, does a Bayesian analysis mitigate false alarms? Answer: By incorporating knowledge into the prior distribution. In the case of multiple groups, the prior can include structural knowledge whereby data from different groups mutually inform and constrain each other's estimates. For example, in the Bayesian ANOVA model, the group deflections come from an overarching distribution. The variance of that overarching distribution is estimated from the data. If the data from most groups are near each other, then the estimate of the variance of the overarching distribution is small, which in turn acts to attenuate the estimated deflections of outlying groups. Thus, estimates of outlying conditions are shrunk toward the grand mean, and this shrinkage attenuates false alarms (e.g., Berry & Hochberg, 1999; Gelman, 2005; Gelman et al., 2009; Lindquist & Gelman, 2009; Meng & Dempster, 1987). A clear example of shrinkage is provided later in the article, accompanying Figure 6. Shrinkage is a benefit of Bayesian ANOVA that is not in NHST ANOVA. Notice that shrinkage is not a mere kludge

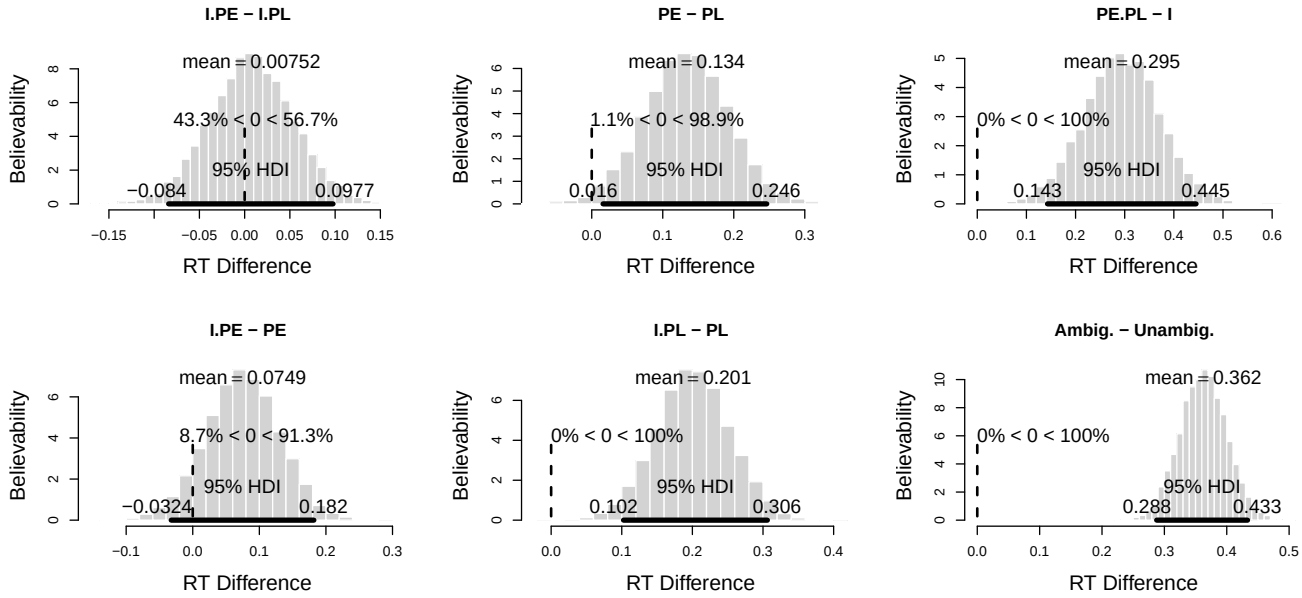

Figure 5. A variety of comparisons of RTs for various test cues. These are the differences of the posterior estimates of parameter values shown in Figure 4, transformed back to the original RT scale of *sec*. instead of $\log_{10}(sec)$. Importantly, these differences take into account any correlations in believable parameter values.

appended onto the analysis like an NHST post-hoc correction. Rather, the higher-level structure in the prior is an explicit formalization of the prior knowledge that the different treatments are mutually informative.

Correlated parameters inherently revealed. The posterior distribution reveals *combinations* of believable parameter values in the high-dimensional joint parameter space. In other words, the posterior distribution inherently reveals correlations among believable parameter values. Correlated parameters are frequently found in analyses such as multiple linear regression. The intercept and slope parameters are correlated for data not centered at zero, and slopes on different regressors can be correlated especially when the values for the regressors are correlated in the data. These situations pose no special problem for Bayesian analysis, because the consequences for parameter estimation are explicit in the conjoint posterior distribution. Of course, it is up to the analyst to carefully examine the joint posterior distribution.

Posterior distributions can be used by subsequent researchers. Posterior distributions of high-dimensional models can be examined in myriad ways, not all of which can be mentioned in any single research report. The posterior distribution of an analysis can be posted online, whereby subsequent researchers can examine the posterior any way they wish. Another benefit of Bayesian analysis is that if future researchers conduct similar experiments and analyze their data with similar models, then the posterior of one experiment can inform the prior of the next. In other words, if a researcher were to replicate an experiment, then the posterior of the first analysis could be used as the prior for the subsequent analysis, or at least as a strong informant for the prior of the subsequent analysis.

Is there an effect? Two Bayesian approaches

The examples in the previous section showed the natural application of parameter estimation to judging whether a null value is among the credible values. This is a straightforward approach to assessing the credibility of null values, and it is presented in many textbooks (e.g., Berry, 1996; Bolstad, 2007; Carlin & Louis, 2009; Gelman, Carlin, Stern, & Rubin, 2004; Gelman & Hill, 2007; Lynch, 2007). The approach begins with a question of estimation, such as, "What is the magnitude of the effect?", or, "What is the magnitude of the difference of effects?" To answer the question, we establish a prior on the effects, that may be informed by the situation but must be agreeable to a skeptical scientific audience. An example of such a prior was shown in Figure 3. Bayesian inference yields a posterior distribution that indicates the believabilities of different magnitudes of effect, for example as shown in Figure 5. It is then natural to summarize the form of the posterior relative to any landmark value(s) of interest, such as a null value. For example, the lower right panel of Figure 5 shows that the 95% HDI falls far from zero, a situation that we interpret as indicating that a difference of zero is not credible.

An infelicity of deciding to reject the null value if it is excluded by an HDI is that the procedure can lead to false alarms when the null is true, even for large sample sizes. It turns out that, if the null is true, then the 95% HDI will exclude the null value 5% of the time, regardless of the amount of data. Despite this behavior, (sometimes referred to as "in-

consistency"), it is also true that the HDI gets narrower and gets closer to the true value as the amount of data increases. Therefore, the large-sample false-alarm rate can be reduced to zero by a simple fix: We reject the null value only if the HDI falls outside a *range of practical equivalence* (ROPE) around the null value. The actual size of the ROPE is determined by situation-specific practical considerations. The ROPE can be arbitrarily small, in principle, to solve the technical false-alarm problem, but in practice real-world costs and benefits can be taken into account (for examples see Carlin & Louis, 2009, where the ROPE is called an "indifference zone"). In the case of response times for perceptual categorization tasks, the ROPE might be reasonably set at $[-0.005, +0.005]$ sec. Our decision rule is, therefore, if any part of the 95% HDI falls within 5 msec. of the null value, we do not decisively reject the null. But mere overlap of the HDI with the ROPE does not mean that we accept the null; acceptance is suggested when the HDI falls entirely within the ROPE. A notion similar to ROPE was introduced by Freedman, Lowe, and Macaskill (1984, who called the ROPE the "region of equivalence") and developed by Spiegelhalter, Freedman, and Parmar (1994) and Hobbs and Carlin (2008). The approach is used subsequently in this article in the section regarding power and replication probability.

It should be kept in mind that any discrete decision regarding a null value is an incomplete summary of the actual posterior distribution. It is always the actual posterior distribution that specifies what we ought to believe. The use of the HDI with the ROPE is a convenient way to summarize whether the null value is among the credible values.

A second approach to judging the veracity of a null value begins with a different question. Instead of asking to estimate an effect size, it asks to adjudicate between two distinct prior distributions. One prior distribution expresses the hypothesis that the null value is true. For this prior, all belief is loaded onto a sharp spike over the null value (typically zero). A second prior distribution expresses the alternative hypothesis that any value is possible. In this alternative prior, beliefs are diffused over a broad range (for discussions see, e.g., Edwards, Lindman, & Savage, 1963; Wagenmakers, 2007). Typically in this approach, the alternative prior is chosen to represent a minimally informed "automatic" prior that does not require consideration of prior knowledge, and that respects technical mathematical considerations such as consistency, which means that as the sample size approaches infinity, the null hypothesis is favored if it is true (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009). In another approach, a spectrum of limited-range vague priors is considered (Gallistel, 2009). In any case, once an automatic alternative prior is established, then Bayesian model comparison can be used to derive the posterior relative believabilities of the two priors.

As an example, consider data reported by Solari, Liseo, and Sun (2008, Table 3) regarding an experiment that measured ascorbic acid content of tomatoes, in nine groups that were given different types of manuring. The third group appeared post-hoc to differ from the other eight. To decide whether the apparent difference was believable, the analysts conducted a Bayesian model comparison for null and alternative hypotheses. The null model used the same central-tendency parameter for all nine groups. The alternative model had a distinct central-tendency parameter for the third group but a shared central-tendency parameter for the other eight groups, and a diffuse prior distribution on the difference between the third group and the other eight. The Bayesian model comparison showed that the alternative model was relatively much more probable than the null model.

The null-versus-alternative comparison is difficult to interpret, however, because either way the result points, we will not necessarily want to actually believe the favored model. For the tomato-manuring data, the favored model is the alternative, but the alternative model asserts that eight of the groups are identical to each other. While this might accidentally happen to be true for these particular data, it is more likely to be the case that the other eight groups are different from each other, at least somewhat, because they had eight different treatments. To better illustrate this point, recall the comparison of Ambiguous versus Unambiguous probe items, shown in the lower-right panel of Figure 5. This comparison of Ambiguous versus Unambiguous cases could instead have been posed as a comparison of two models: The null model would put all probes equal to each other. The alternative model would put all the ambiguous probes equal to each other, and put all the unambiguous probes equal to each other, and set a diffuse prior distribution on the difference between those two central tendencies. The Bayesian model comparison would, no doubt, favor the alternative model, which states that ambiguous probes are not equal to unambiguous probes, but all ambiguous probes *are* equal to each other and all unambiguous probes *are* equal to each other. We would not want to believe the preferred model, because clearly the unambiguous probes are *not* all equal (e.g., *PE* versus *PL* in the upper middle panel of Figure 5) and the ambiguous probes are *not* all equal (e.g., *PE*.*PL* versus *I* in the upper right panel of Figure 5).

To address this problem of which combinations of group means are equal to each other, the model-comparison approach may be applied to a combinatorial hierarchy of all possible treatment groupings, to tease apart which combinations of groups are credibly different (e.g., Berry & Hochberg, 1999; Mueller, Parmigiani, & Rice, 2007). Scott and Berger (2006) presented a model that simultaneously estimates the deflections of each group, and the probability that the deflection is non-zero. The method naturally incorporates Bayesian shrinkage on the estimates of non-zero probabilities, but the authors note that the estimates are very sensitive to prior assumptions, and therefore informed priors should be used. Gopalan and Berry (1998) considered a hierarchy of all possible combinations of group-mean equalities, using a Dirichlet process prior. It should be kept in mind, however, that this latter approach entertains hypotheses that different groups might or might not derive from identical underlying means, and this sort of hypothesis structure is not universally applicable; it should only be tested when it is genuinely meaningful to hypothesize that different groups actually have identical means.

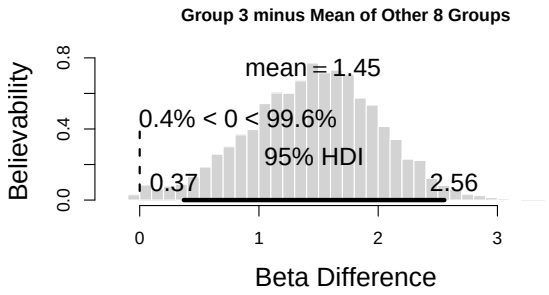

Figure 6. Posterior estimate for difference of groups in Solari et al. (2008). The third group is credibly different from the mean of the other eight groups.

The effect-estimation approach, on the other hand, provides a complete multi-dimensional posterior estimate of all the parameters simultaneously, which can be meaningfully examined for any comparison of interest, as was exemplified earlier in Figure 5. Consider again the tomato acidity data from Solari et al. (2008). We can conduct a Bayesian ANOVA, and then compute a contrast in which the third group is compared with the mean of the other eight. Figure 6 shows the resulting posterior distribution of the contrast, where it can be seen that a difference of zero has little credibility. The conclusion from the contrast agrees with the model comparison that group 3 is different from the others, but the contrast estimate in Figure 6 is made while the other eight groups have their own (correlated) individual mean estimates, unlike the model comparison which assumes equal means in the other groups.

The contrast estimate in Figure 6 also illustrates *shrinkage* of the group estimates. Because eight of the nine groups happen to have little between-group variation, the estimate of the variance between the group means is small, causing the estimate of the mean of the third group to shrink toward the overall baseline, away from that group's sample mean. Specifically, the mean of the posterior estimate for group 3 is 6.17, which is noticeably less than group 3's actual sample mean of 6.77. Shrinkage is a natural consequence of priors expressed by model structure: The model assumes that groups come from an overarching distribution, therefore data from one group inform the estimate of the other group. As described earlier, shrinkage is a rational way to mitigate false alarms.

The Bayesian model-comparison approach is sometimes touted as a way to garner evidence in favor of the null hypothesis (e.g., Gallistel, 2009; Rouder et al., 2009), unlike traditional NHST which can only reject the null hypothesis and never accept it. Nevertheless, when a Bayesian model comparison concludes by preferring the null hypothesis, the analyst should be careful when interpreting the result. Consider again the tomato acidity experiment. Suppose that a model comparison showed that two types of manuring yielded results so similar that the null model is preferred. Should we believe the null hypothesis that there is no difference between the two types of manuring? Probably not; after all, the two types of manuring are, by definition, different. What we really want is an estimate of the magnitude of the difference. In this situation, when we compute the posterior estimate, it will probably span zero, from which we conclude that zero is among the credible differences, but we will also have a posterior distribution that reveals our uncertainty in that estimate, including other small but credible differences.

There are even situations in which a null-versus-alternative model comparison will *favor* the null, but the posterior estimate of the effect will have an HDI that *excludes* the null value (e.g., Stone, 1997). This situation can arise when there is a confluence of two conditions. First, the data indicate that the true effect is *near* the null value, but the sample size is so large that there is only a narrow range of *non*-null effect magnitudes that accommodate the data. Second, the alternative prior is so diffuse that its density is very slight at the effect magnitudes indicated by the data. The consequence of these conditions is that model comparison prefers the null hypothesis because the alternative prior is so diluted, but effect estimation excludes the null value because the sample size is so large. This uncomfortable situation highlights a general point, that the outcome of a Bayesian model comparison can be very sensitive to the choice of priors in the two models (e.g., Liu & Aitkin, 2008). One way to try to address the sensitivity is by considering a spectrum of alternative priors, and deciding that the null hypothesis should be preferred if the model comparison favors the null across the entire range of alternative priors (Gallistel, 2009). Unfortunately, the spectrum of alternative priors that is usually entertained in these approaches does not include the most plausible alternative hypothesis, namely one that is informed by previous research. When the alternative prior is actually informed by previous data and by tenable theory based on those data, rather than being an untenable caricature of "objectivity", then the alternative prior will probably place fairly high believability on values that are consistent with new data. In this situation, null-versus-alternative comparison will have a much more difficult time favoring the null hypothesis.

In general, Bayesian model comparison can be an excellent procedure for judging the relative veracity of genuinely viable models, as long as the priors in each model are carefully justified and the conclusion is examined for sensitivity to the priors (Jefferys & Berger, 1992; Kass & Raftery, 1995; Liu & Aitkin, 2008). This general assessment applies to the special case of comparing null and alternative hypotheses. In other words, the Bayesian model-comparison approach to null hypothesis testing should be used only when the null hypothesis is a genuinely tenable belief in the context of the actual situation (Rouder et al., 2009, p. 235), and when the alternative hypothesis is also formulated as a genuinely tenable belief, preferably as informed by previous research. Otherwise the model comparison reveals merely which unbelievable prior is more unbelievable. If a researcher is specifically interested in deciding between the null and alternative models because those models have genuine theoretical meaningfulness, then the Bayesian model comparison procedure is quite appropriate. In some clinical settings where

Analysis Type	Data Generator	Data	Prior for Bayesian Analysis	Posterior
Actual	Real World	Observed Once	Skeptical Audience	Actual Posterior
1. Prospective Power	Hypothesis	Simulated Repeatedly	Skeptical Audience	Anticipated
2. Retrospective Power	Actual Posterior	Simulated Repeatedly	Skeptical Audience	Anticipated
3. Cumulative Replication Probability	Actual Posterior	Simulated Repeatedly	Actual Posterior	Anticipated

Figure 7. Three types of Bayesian replication probability.

a yes/no decision is demanded, rather than a parameter estimate, the model comparison procedure may have better operating characteristics (e.g. Johnson & Cook, 2008). It is up to the researcher to determine whether s/he seeks a posterior estimate of effect magnitudes, or a judgment about two vying hypotheses.

Can the effect be replicated? A natural Bayesian answer

The previous sections have emphasized that the result of Bayesian analysis is a posterior distribution that reveals a joint estimate of all the parameter values, given the data. The posterior estimate helps the researcher achieve a goal of the research. Sometimes the goal is to demonstrate that a null value is incredible. Other times, such as in a political poll, the goal is simply to get an estimate with a minimal degree of precision. Greater precision can always be obtained (on average) by collecting more data. With the research goal specified, and an intended data-collection procedure proposed, we can ask how likely it is to achieve the goal.

The Bayesian framework provides a natural way to address the question (unlike NHST, which suffers serious problems; see Miller, 2009). The posterior distribution on the parameters indicates the most credible values, given the data. These parameter values shape the model that we use to describe the data. The model, with its credible parameter values, can also be construed as a machine for generating random data that should be distributed like the actual data. We use the generative model as a mechanism for anticipating results from repeating the experiment.

Consider, for example, the posterior that was shown in Figure 4. To generate data from the posterior, we start by randomly selecting a believable baseline value (from the upper left distribution in Figure 4), then randomly selecting a believable treatment effect (from the pertinent treatment distribution in Figure 4), then randomly selecting a subject effect (not shown in Figure 4), and finally randomly selecting RTs based on those believable parameter values.[4] The posterior thereby implies what believable data should look like, incorporating our uncertainty regarding the underlying parameter values.

We use the posterior distribution to simulate data from many replications of an intended experiment. From the simulated replications, we tally how often the research goal is achieved. For example, in the simulated replications, we can tally how often the HDI excluded the ROPE, or we can tally how often a desired precision was achieved.

Figure 7 lists three types of replication probability in a Bayesian framework. The first row of the table expresses an actual Bayesian analysis for a set of data observed from the real world. The prior for the analysis is whatever the analyst and the skeptical scientific audience can agree upon. The result of the analysis is the *actual posterior* distribution.

The remaining rows of Figure 7 describe ways to define the probability that an intended data-collection procedure will accomplish a desired research goal. In all cases, data are repeatedly simulated from an assumed generative distribution that models the world, and the simulated data are examined by a Bayesian analysis. What differs among the three ways is the assumptions regarding the generative distribution and the prior used for the Bayesian analysis of the simulated data.

A *prospective power* analysis uses a researcher's hypothetical belief distribution over model parameters to generate simulated data. Each set of simulated data is examined with a Bayesian analysis that uses a skeptical-audience prior. The resulting posterior either does or does not accomplish the research goal, and a tally is kept across the many simulated data sets, as an estimate of the probability that the research goal will be accomplished.

A *retrospective power* analysis proceeds the same way as a prospective power analysis except that the data generator is the actual posterior distribution from a previously conducted experiment. The idea is that the researcher's best estimate of the world is the posterior from the previous experiment, and therefore that posterior is used to generate simulated data. The simulations are tallied for whether or not the research goal is achieved.

Finally, a *cumulative replication probability* is computed by using the actual posterior as the data generator, and the actual posterior as the prior for the Bayesian analysis. The idea here is that the simulated data are treated as cumulative along with the actual experiment's data, so that the simulated data represent a simulated cumulative replication of the original experiment.

To clarify these concepts, what follows is a specific example of a retrospective power analysis. Consider another learning experiment, this one involving learning which keys to press in response to simple geometric figures. The figures were rectangles that had different heights, containing a vertical interior segment that could have different lateral positions. Different groups of learners had to learn different correspondences of rectangles to key responses. The two *filtration* correspondences allowed perfect classification by attending to height alone or to segment position alone; the other dimension could be filtered out. The two *condensation* correspondences required attention to both stimulus dimensions to achieve perfect accuracy. The primary goal of the

[4] Terminology: The distribution of data generated by the posterior distribution of parameters is sometimes called the "posterior predictive distribution".

experiment was to assess whether learning accuracy was better in the filtration conditions than in the condensation conditions, as predicted by theories that posit learned attention to dimensions (see Kruschke, 1993, for more details).

The results can be examined with a hierarchical Bayesian data analysis. Each individual provides a number correct out of the total number of training trials. This number correct was, for purposes of illustration, simplistically modeled as a random draw from a binomial distribution. (In other words, changes of accuracy across trials were not modeled; merely the overall accuracy was modeled.) The binomial distribution has a parameter that describes the probability of being correct on a trial; every individual was allowed to have a different value for this parameter. The individual parameters were modeled as coming from an overarching distribution for the group. (The model of individual differences used here is a simplistically unimodal beta distribution; see Lee and Webb (2005) for more details regarding individual differences within these groups, and see Navarro, Griffiths, Steyvers, and Lee (2006) for a Bayesian approach to finding clusters of individuals, analogous to the method of Gopalan and Berry (1998) for finding clusters of groups in ANOVA.) The overarching distribution has two parameters that describe its mean and narrowness. The mean parameter for each group is the primary descriptor of interest.

Figure 8 shows results of the Bayesian analysis. The top row shows the prior distributions of the group accuracy parameters for the four groups. The mildly informed prior gave greater believability to accuracies greater than chance, because these learning tasks were known to be fairly easy. The second row shows the prior on selected contrasts of the group mean parameters, as implied by simply taking differences of the prior mean parameters. The third row shows the posterior on the group mean parameters. Notice that their range is much smaller than the range of priors, which suggests that the mildly informed prior had very little influence on the posterior. (Repeating the analysis with other priors confirms that the posterior is negligibly affected by this prior.) The bottom row shows various comparisons of the group parameters. In particular, the lower-right distribution shows that the average of the filtration groups is much higher than the average of the condensation groups; 100% of the posterior falls well above a difference of zero. On the other hand, the lower-left distribution shows that two filtration groups are only marginally different from each other, depending on how wide a ROPE is defined. The lower-middle distribution shows that the most credible difference between the two condensation groups is essentially zero, and the distribution also shows the range of uncertainty around the estimate.

Now that the actual posterior distribution has been established, we can ask about replication probability. In particular, I will illustrate a retrospective power analysis. The posterior shown in Figure 8 constitutes our best estimate of the parameters that describe people's accuracy in these groups. We can therefore use these estimates to predict what would happen if we had run the experiment using different sample sizes. To do this, we generate simulated data for a subject as follows: First, randomly sample representative parameter values from the posterior to simulate the subject, then use that subject's parameter values to randomly generate simulated data. Repeat this process for every simulated subject. Then, conduct a Bayesian analysis on the simulated data and assess whether the goal was achieved.

In the present application, there are at least three different goals we may want to achieve. First and foremost, we may want to show that the mean accuracy of the filtration groups is credibly higher than the mean accuracy of the condensation groups. We will declare this goal to be achieved if the 95% HDI of the μ differences excludes a ROPE extending from -0.01 to $+0.01$. Second, we may want to show the more subtle difference between the two filtration groups is credible. We will declare this goal to be achieved by the same criterion as for the first goal. Notice that this criterion is barely exceeded in the lower-left panel of Figure 8. Third, because the two condensation were not expected to differ much, we may want to achieve a specific *minimal precision* on the estimate of the difference between the two condensation groups. We will declare this goal to be achieved if the width of the 95% HDI is less than 0.15, as it is, for example, in the lower-middle panel of Figure 8. The goal of minimal precision is more robust than the goal of excluding a ROPE from an HDI, because minimal precision can always be achieved (in principle) with a larger sample size, but excluding a ROPE from an HDI can only be achieved with high probability if the data-generating distribution actually has a fairly certain non-null effect. [For other goals involving the HDI, see Adcock (1997), De Santis (2004), De Santis (2007), Joseph, Wolfson, and du Berger (1995a), Joseph, Wolfson, and du Berger (1995b), and Wang and Gelfand (2002). For goals involving model comparison, see Weiss (1997) and De Santis (2004).]

We also must decide what prior to use in the Bayesian analysis of the simulated data. For the present demonstration, we will suppose that the analysis of the simulated data is to be presented to an audience that knows only of the new data, and therefore the Bayesian analysis will use the same prior that we used for the initial analysis. This is a case of *retrospective power analysis* as listed in Figure 7.

There were 500 simulated replications run. When the sample size was $N = 40$ per group, as in the original experiment, there was 100% success in showing that filtration is more accurate than condensation. But there was only 23% success in showing that the two filtration groups differed from each other, i.e., that the 95% HDI of the difference excluded the ROPE of $[-.01, +.01]$. In other words, the marginal difference observed in the actual data (lower-left panel of Figure 8) had a relatively small probability of exceeding the ROPE. Finally, the probability is 60% that the width of the 95% HDI of condensation differences achieves 0.15 or less. If either of the latter two goals was of primary concern, we would want to increase the sample size to achieve a higher probability a achieving the goal. If, on the other hand, the primary goal is merely showing that filtration is more accurate than condensation, we can get away with using fewer subjects. It turns out that when the sample size is only $N = 8$ per group, the probability that the 95% HDI of filtration minus condensation exceeds the ROPE is still 89%.

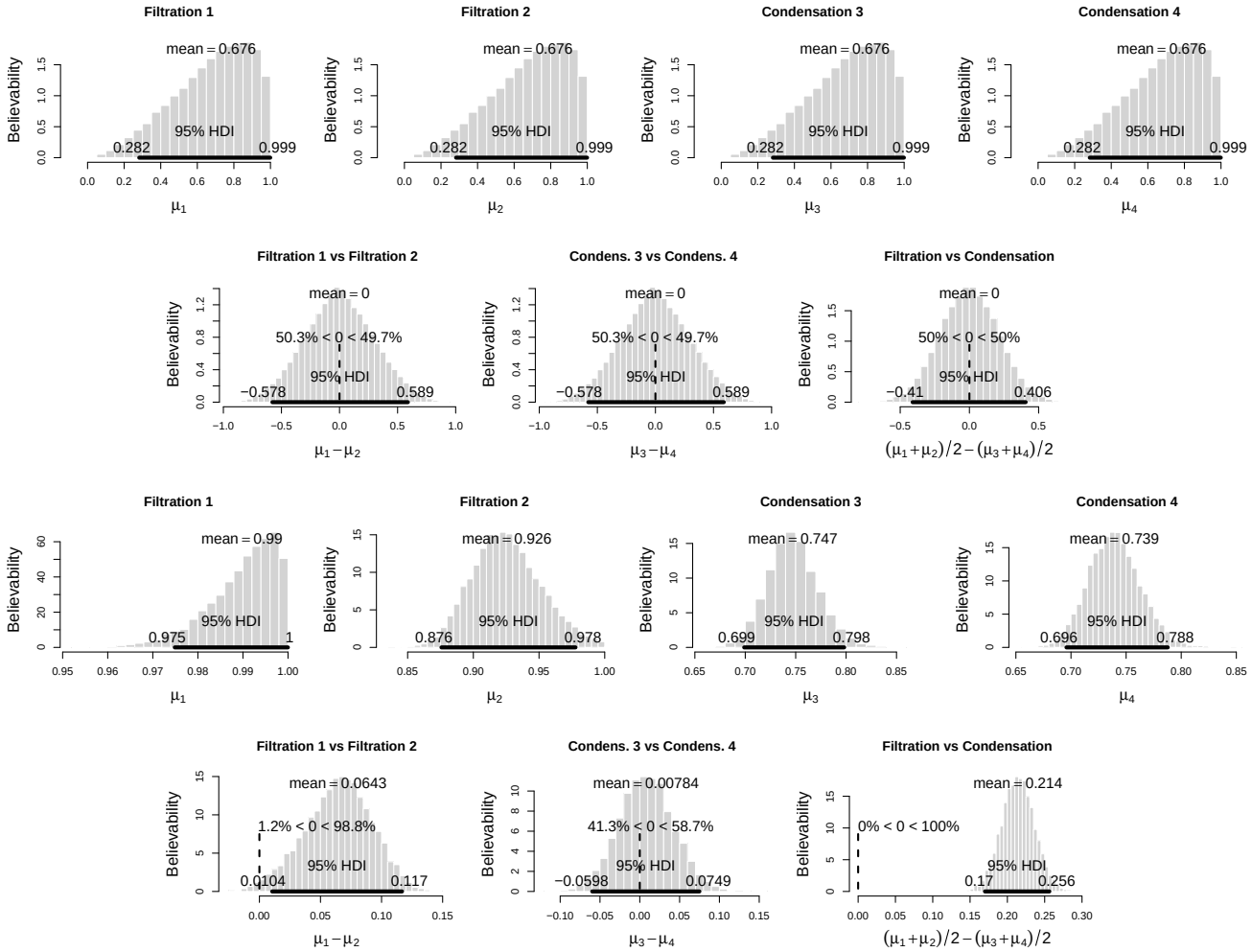

Figure 8. Group accuracies, parameterized as μ_1 through μ_4, for the filtration/condensation experiment. First row: Prior distributions on the four group accuracies. The prior is mildly informed by the knowledge that accuracies will tend to be better than chance. Second row: The priors on various differences of group accuracies, implied by the priors in the first row. Third and fourth rows: Posterior distributions, corresponding to the first two rows.

Of course, the probability of achieving the other two goals drops dramatically, to only 5% and 0% respectively.

A researcher might instead be interested in computing the probability of achieving the goal in a follow-up or continuation of the already obtained data. In this case, the posterior of the original data analysis is the prior for the simulated replication, to determine what is anticipated for accumulated data rather than de novo data. This is called the *cumulative replication probability* and is listed in the bottom row of Figure 7.

The specific types of goal-achievement probabilities listed in Figure 7 use novel nomenclature, but the simulation approach described here was also presented by Wang and Gelfand (2002), who, like their predecessors, distinguished between data-generating distributions and analysis priors. The approach described here has assumed a data-generating distribution based on a posterior from a single experiment, but De Santis (2007) described how to mix posteriors from several previous experiments to create a data-generating distribution.

It is important to recognize that the use of simulated data for computing replication probabilities does *not* fall victim to the criticisms of NHST presented earlier in this article. NHST uses intention-based simulated data to interpret the significance of actual data via a p value, but the computation of replication probability does not. Replication probability uses intention-based simulations exactly as appropriate, viz, to anticipate probable results if the intended experiment were conducted. But all the simulated data, and the actual data, are analyzed in a Bayesian fashion.

Conclusion

This article began by pointing out that NHST is based on the intentions of the researcher and analyst: The p value depends entirely on the assumed intention. If data are collected

for a certain duration instead of for a certain sample size, the p value changes. If some data are lost by accident or attrition or declaration of outliers, the p value changes. If the analyst wants to be thorough and investigate multiple comparisons, the p value changes. If the researcher might possibly collect more data in the future that could be compared with the present data, then the p value of the present data changes.

Bayesian data analysis does not suffer those dependencies on the researcher's intentions. Bayesian data analysis computes what we actually want to know: The believability of candidate values given the data we have actually observed. Bayesian data analysis yields natural ways to assess the credibility of null values, and the probability of achieving a research goal.

Bayesian data analysis can and should endure regardless of whether Bayesian models of cognition endure. As modelers of mind, we do not know what models and parameters the mind might be using, and so it is an open question as to whether we will be able to create models that accurately mimic human cognition (Jacobs & Kruschke, 2010). But Bayesian data analysis is based on generic descriptive models (e.g., linear regression, ANOVA, etc.) that are useful for summarizing trends in data regardless of the underlying natural processes that generated those data. As data analysts, we get to define the models and parameters of interest to us, and, having done so, then the rational way to allocate beliefs among those models and parameters is via Bayesian analysis.

The advantages of Bayesian analysis have been recognized by many scientific disciplines, including archeology (Litton & Buck, 1995), astrophysics (Loredo, 1992), conservation biology (Wade, 2000), ecology (Ellison, 2004), epidemiology (Dunson, 2001), evolutionary biology (Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001), meteorology (Berliner, Royle, Wikle, & Milliff, 1999), political science (Jackman, 2004), etc. It is time that cognitive science does too.

References

Adcock, C. J. (1997). Sample size determination: a review. *The Statistician*, *46*, 261–283.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Berliner, L. M., Royle, J. A., Wikle, C. K., & Milliff, R. F. (1999). Bayesian methods in the atmospheric sciences. In J. M. Bernardo, J. O. BErger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6: Proceedings of the sixth Valencia international meeting, June 6-10, 1998* (pp. 83–100). Oxford, UK: Oxford University Press.

Berry, D. A. (1996). *Statistics: A Bayesian perspective*. Belmont, CA: Duxbury Press / Wadsworth.

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1-2), 215–227.

Bolstad, W. M. (2007). *Introduction to Bayesian statistics* (2nd ed.). Hoboken, NJ: Wiley.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind*. Oxford, U.K.: Oxford University Press.

Chater, N., Tenenbaum, J. B., & Yuille, A. (Eds.). (2006, July). Special issue: Probabilistic models of cognition. *Trends in Cognitive Sciences*, *10*(7), 287–344.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 530–572.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, *124*, 121–144.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A*, *170*, 95–113.

Doyle, A. C. (1890). *The sign of four*. London: Spencer Blackett.

Dunson, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, *153*(12), 1222–1226.

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

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, *7*(6), 509–520.

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, *40*, 575–586.

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

Gelman, A. (2005). Analysis of variance — why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, Florida: CRC Press.

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

Gelman, A., Hill, J., & Yajima, M. (2009). *Why we (usually) don't have to worry about multiple comparisons.* (March, 2009: http://www.stat.columbia.edu/~gelman/ research/ unpublished/ multiple2.pdf)

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*(1), 169–177.

Gopalan, R., & Berry, D. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association*, 1130–1139.

Hobbs, B. P., & Carlin, B. P. (2008, January). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(1), 54–80.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, *294*(5550), 2310–2314.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 299–314. (With other contributors listed at http://www.r-project.org/)

Jackman, S. (2004). Bayesian analysis for political research. *Annual Review of Political Science*, *7*, 483–505.

Jacobs, R. A., & Kruschke, J. K. (2010). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, **(**), **–**.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72.

Johnson, V. E., & Cook, J. D. (2008, July). *Bayesian design of*

single-arm phase II clinical trials with continuous monitoring. University of Texas, MD Anderson Cancer Center Department of Biostatistics Working Paper Series. The Berkeley Electronic Press. (Working Paper 47, http://www.bepress.com/ mdandersonbiostat/ paper47)

Joseph, L., Wolfson, D. B., & du Berger, R. (1995a). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician*, *44*, 143–154.

Joseph, L., Wolfson, D. B., & du Berger, R. (1995b). Some comments on Bayesian sample size determination. *The Statistician*, *44*, 167–171.

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

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 3–26.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210-226.

Kruschke, J. K. (2010). Attentional highlighting in learning: A canonical experiment. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. **, pp. **–**). **: Academic Press. (Pre-print available at author's website, http://www.indiana.edu/~kruschke)

Lamberts, K., & Kent, C. (2007). No evidence for rule-based processing in the inverse base-rate effect. *Memory & Cognition*, *35*(8), 2097–2105.

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

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*(3), 662–668.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

Lindquist, M. A., & Gelman, A. (2009). Correlations and multiple comparisons in functional imaging – a statistical perspective. *Perspectives in Psychological Science*, *4*(3), 310–313.

Litton, C. D., & Buck, C. E. (1995). The Bayesian approach to the interpretation of archaeological data. *Archaeometry*, *37*(1), 1–24.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Loredo, T. J. (1992). The promise of Bayesian inference for astrophysics. In E. D. Feigelson & G. J. Babu (Eds.), *Statistical challenges in modern astronomy* (pp. 275–297). New York: Springer-Verlag.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

Meng, C. Y. K., & Dempster, A. P. (1987). A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics*, *43*(2), 301–311.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*(4), 617–640.

Mueller, P., Parmigiani, G., & Rice, K. (2007). FDR and Bayesian multiple comparisons rules. In J. M. Bernardo et al. (Eds.), *Bayesian statistics 8.* Oxford, UK: Oxford University Press.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.

Qian, S. S., & Shen, Z. (2007). Ecological applications of multilevel analysis of variance. *Ecology*, *88*(10), 2489–2495.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Sadiku, M. N. O., & Tofighi, M. R. (1999). A tutorial on simulation of queueing models. *International Journal of Electrical Engineering Education*, *36*, 102–120.

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of statistical planning and inference*, *136*(7), 2144–2162.

Solari, F., Liseo, B., & Sun, D. (2008). Some remarks on Bayesian inference for one-way ANOVA models. *Annals of the Institute of Statistical Mathematics*, *60*, 483–498.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A*, *157*, 357–416.

Stone, M. (1997). Discussion of papers by Dempster and Aitken. *Statistics and Computing*, *7*, 263–264.

Thomas, A. (2004). *BRugs user manual (the R interface to BUGS).* http://mathstat.helsinki.fi/openbugs/ data/ Docu/ BRugs%20Manual.html.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006, March). Making BUGS open. *R News*, *6*(1), 12–17.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32.

Wade, P. R. (2000). Bayesian methods in conservation biology. *Conservation Biology*, 1308–1316.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wang, F., & Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, *17*, 193–208.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician*, *46*, 185–191.

Young, K. D., & Lewis, R. J. (1997). What is confidence? Part 2: Detailed definition and determination of confidence intervals. *Annals of emergency medicine*, *30*(3), 311–318.