Running head: NO FREE LUNCH

The $p < .05$ Rule and the Hidden Costs of the Free Lunch in Inference

Jeffrey N. Rouder

University of Missouri

Richard D. Morey

University of Groningen

Josine Verhagen

University of Amsterdam

Jordan M. Province

University of Missouri

Eric-Jan Wagenmakers

University of Amsterdam

Jeff Rouder

rouderj@missouri.edu

Abstract

The field of psychology, including cognitive science, is vexed in a crisis of confidence. Although the causes and solutions are assuredly varied, we focus here on a common logical problem in inference. The default mode of inference is significance testing, and significance testing and inference by confidence intervals have a *free lunch property* where researchers need not make detailed assumptions about the alternative to test the null hypothesis. We present the argument that there is no free lunch, that is, valid testing requires that researchers test the null against a well-specified alternative. We show how this requirement follows from the basic tenets of conventional and Bayesian probability. Moreover, we show in both the conventional and Bayesian framework that not specifying the alternative leads to rejections of the null hypothesis with scant evidence. We review both frequentist and Bayesian approaches to specifying alternatives, and show how such specifications improves inference. The field of cognitive science will benefit—consideration of reasonable alternatives will undoubtedly sharpen the intellectual underpinnings of research.

The $p < .05$ Rule and the Hidden Costs of the Free Lunch in Inference

Prelude: The Infamous Case of Sally Clark

In fields such as medicine and law, statistical inference is often a matter of life or death. When the stakes are this high, it is crucial to recognize and avoid elementary errors of statistical reasoning. Consider, for instance, the British court case of Sally Clark (Dawid, 2005; Hill, 2005; Nobles & Schiff, 2005). Clark had experienced a double tragedy: her two babies had both died, presumably from cot death or sudden infant death syndrome (SIDS). If the deaths are independent, and the probability of any one child dying from SIDS is roughly 1/8543, the probability for such a double tragedy to occur is as low as $1/8543 \times 1/8543 \approx 1$ in 73 million. Clark was accused of killing her two children, and the prosecution provided following statistical argument as evidence: Because the probability of two babies dying from SIDS is as low as 1 in 73 million, we should entertain the alternative that the deaths at hand were due not to natural causes but rather to murder. And indeed, in November 1999, a jury convicted Clark of murdering both babies, and she was sentenced to prison.

Let us reflect on what happened. Forget the fact that the deaths may not be independent (due to common environmental or genetic factors), suppress any worries about how the rate of 1/8543 was obtained, and focus solely on the inferential logic used in the case. Assuming that the probability of two babies dying from SIDS is indeed as low as 1 in 73 million, to what extent does this incriminate Clark? The prosecution followed a Popperian line of statistical reasoning similar to that used throughout the empirical sciences: one postulates a single hypothesis (i.e., the null hypothesis: "Sally Clark is innocent") and then assesses the unlikeliness of the data under that hypothesis. In this particular case, the prosecution felt that the probability of two babies dying from SIDS

was sufficiently small to reject the null hypothesis of innocence, and thereby accept the hypothesis that Sally Clark is guilty. Case closed.

The flaw in the reasoning above is the consideration of only one hypothesis: the null hypothesis that Sally Clark is innocent and her babies died from SIDS. Hence, the only datum deemed relevant is the low background probability of two babies dying from SIDS. But what of the background probability of two babies dying from murder? In 2002, President of the Royal Statistical Society Peter Green wrote an open letter to the Lord Chancellor in which he explained that "The jury needs to weigh up two competing explanations for the babies' deaths: SIDS or murder. The fact that two deaths by SIDS is quite unlikely is, taken alone, of little value. Two deaths by murder may well be even more unlikely. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation." (Nobles & Schiff, 2005, p. 19). Statistician Phil Dawid (2005, p. 8) agreed: "(...) if background evidence of double-death-rates due to SIDS (or other natural causes) is relevant, then surely so too should be background evidence of double-death-rates due to murder. To present either of these figures without some assessment of the other would be both irrational and prejudicial." Ray Hill (2005, p. 15) showed how different the statistical conclusions are when one takes into account both sides of the coin: "Obtaining reliable estimates based on limited data is fraught with difficulty, but my calculations gave the following rough estimates. Single cot deaths outnumber single murders by about 17 to 1, double cot deaths outnumber double murders by about 9 to 1 and triple cot deaths outnumber triple murders by about 2 to 1. (...) when multiple sudden infant deaths have occurred in a family, there is no initial reason to suppose that they are more likely to be homicide than natural."

It is perhaps ironic that Sally Clark's case was tainted by such a fundamental error of statistical logic; after all, the need to compare both hypothesis under consideration (i.e., innocence vs. guilt) is aptly symbolized by the two scales of Lady Justice. Knowing the

weight on only one of the scales is "of little value", as the relevant information is provided by the balance between the scales. The main lesson from the Sally Clark case, however, is not that lawyers sometimes make mistakes when they have to reason with numbers. Instead, the main lesson is that the very statistical error that plagued the Sally Clark case also besets statistical inference in the empirical sciences. To see why, consider a revision of the letter by Peter Green in which we have replaced the terms that are specific to the Sally Clark case with more general statistical concepts: "The researcher needs to weigh up two competing explanations for the data: The null hypothesis or the alternative hypothesis. The fact that the observed data are quite unlikely under the null hypothesis is, taken alone, of little value. The observed data may well be even more unlikely under the alternative hypothesis. What matters is the relative likelihood of the data under each hypothesis, not just how unlikely they are under one hypothesis."

Statistical Inference in Psychology and the Desire for a Free Lunch

In psychology, statistical inference is generally not a matter of life or death. Nonetheless, large-scale adoption of methods that focus on the null alone without recourse to well-specified alternatives does have deleterious long term consequences that are becoming ever more apparent. Inferential logic that is ill-suited for determining the truth of Sally Clark's guilt will be equally ill-suited to finding the answer to any other question.

Recent work suggests that psychological science is facing a "crisis of confidence" (Yong, 2012; Pashler & Wagenmakers, 2012) fueled in part by the suspicion that many empirical phenomena may not replicate robustly (e.g., Carpenter, 2012; Roediger, 2012; Yong, 2012; for examples see Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013; Huizenga, Wetzels, van Ravenzwaaij, & Wagenmakers, 2012; LeBel & Campbell, in press; Shanks et al., 2013). Note that psychology shares this crisis of confidence with other fields; for instance, pharmaceutical companies recently

complained that they often fail to replicate preclinical findings from published academic studies (Begley & Ellis, 2012; Osherovich, 2011; Prinz, Schlange, & Asadullah, 2011), with some replication rates as low as 11%. Another concern is that reputable journals have recently published findings that are highly implausible, the most prominent example featuring a series of studies on extra-sensory perception (Bem, 2011; but see Francis, 2012; Galak, LeBoeuf, Nelson, & Simmons, 2012; Judd, Westfall, & Kenny, 2012; LeBel & Peters, 2011; Ritchie, Wiseman, & French, 2012; Rouder & Morey, 2011; Schimmack, 2012; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; Wagenmakers, Krypotos, Criss, & Iverson, 2012).

The current crisis of confidence assuredly has many causes (Nosek & Bar–Anan, 2012; Nosek, Spies, & Motyl, 2012), but one contributing cause is that the field is relying on flawed and unprincipled methods of stating evidence, and our everyday, bread-and-butter practices need revision. Just as in medicine and law, one of the main goals of psychological scientists is to assess and communicate the evidence from data for competing positions. For example, in the simplest hypothesis-testing framework, the analyst is trying to assess the evidence for the null and alternative hypotheses. The prevailing approach used in the field is what we term the "$p < .05$ rule", that is, effects are considered to be demonstrated if the associated p values are less than .05. Most researchers hold complicated views about the wisdom of this rule. On one hand, most of us realize that data analysis is a bit more thoughtful and more organic than blind adherence to the $p < .05$. On the other, many would argue that the $p < .05$ rule has served the community well as a rough-and-ready safeguard from spurious effects (Abelson, 1997). Researchers by and large seem to believe that although the $p < .05$ rule is not perfect in all applications, it is useful as a conventional guide for experimenters, reviewers, editors, and readers.

We argue here is that the $p < .05$ rule does not serve us well at all, and in fact, its

use contributes to our methodological turmoil. At its core, the $p < .05$ rule assumes what we term here the *free lunch of inference*. The free lunch refers to the ability to assess evidence against the null hypothesis without consideration of a well-specified alternative. In significance testing, for instance, the analyst does not need to commit to any particular alternative to assess the evidence against the null. A contrasting approach is inference with power analyses. With a power analysis, a researcher can choose an appropriate balance between Type I and Type II errors. But to do so, the alternative must be specified in detail: the effect size under the alternative is set to a specified value.

Most of us would rather have a free lunch. We may have the intuition that our conclusions are more valid, more objective, and more persuasive if it is not tied to a particular alternative. This intuition, however, is dangerously wrong. We argue here that statistical evidence may be properly understood and quantified only with reference to detailed alternatives. Moreover, and more importantly, inference with and without an alternative yield markedly different results. Inference without judicious consideration of alternatives is capriciously tilted toward concluding that there are effects in data, and those who adopt it require too low a standard of evidence to make bona-fide claims. Specifying alternatives certainly requires more thought and effort, but it will lead to more principled and more useful evaluations of results. As the saying goes, "there ain't no such thing as a free lunch".[1]

The message that there is no free lunch in testing originates from the 1963 *Psychological Review* article by Edwards, Lindman, and Savage. This 50-year old paper correctly identified the methodological problems we are dealing with today, and suggested the appropriate remedy: namely, that lunch must be paid for. The article by Edwards et al. was largely ignored by psychologists, but has been influential in statistics through the work of Berger (e.g., Berger & Delampady, 1987; Sellke, Bayarri, & Berger, 2001), Raftery (e.g., Raftery, 1995, 1999) and others. The arguments from Edwards et al. have recently

been reimported back to psychology by Dienes (2008), Gallistel (2009), Rouder, Speckman, Sun, Morey, and Iverson (2009), and Wagenmakers (2007), among others. There is sage wisdom in Edwards et al. that remains relevant, important, and under-appreciated; the reader will benefit from a careful rereading of this classic. One might consider our message to be a half-century commemoration of Edwards et al.'s ideas and impacts.

Our argument that there is no free lunch in testing is motivated by logic and probability. We start with the strong logic of falsification, in which alternatives are not needed. We note that extending strong falsification to the probabilistic case fails, and then show how one must "pay for lunch" under both frequentist and Bayesian probability frameworks. To make the case using frequentist probability, we rely on the fact that sure knowledge is defined in the large-data limit; for example, in as data collection continues, sample means converge to true means. Frequentist testing should be perfect in the large-sample limit; that is, with an infinite amount of data, testing should provide *always* for the correct decision. This perfection in the large-sample limit is called *consistency* and, unfortunately, significance testing is not consistent. Consistent frequentist testing is possible, but it requires that analysts specify an alternative. Bayesian probability implies a different constraint on inference: prior beliefs must by updated rationally in light of data. This constraint leads immediately to inference by Bayes factor, and the Bayes factor requires priors. It is through these priors that the analyst specifies alternatives. The message is that regardless of whether one is a frequentist or Bayesian, in principle one must specify alternatives.

The critique that researchers must specify alternatives is most easily seen for hypothesis testing. Although hypothesis testing is nearly ubiquitous, there have been repeated calls for replacing hypothesis testing with estimation. The proponents of estimation suggest that researchers report point estimates of parameters of interest (e.g., effect size) and confidence intervals (CIs) (Cumming, 2014; Grant, 1962; Loftus, 1996). In

our view, there are two different uses of CIs in practice: one in which they are used to make hypothesis testing decisions by noting whether they cover zero, and a second one in which they are used descriptively without making decisions about the presence of effects. The first, testing by CI is not only recommended (Tryon, 2001), but is widely believed to be an appropriate use of confidence intervals (Hoekstra, Morey, Rouder, & Wagenmakers, in press). We include a critique of testing by CIs because it is testing without specification of alternatives. We address the more descriptive uses of estimation and its relationship to testing in the conclusion.

The Need for Alternatives: Starting from Logic

Logical inference provides the foundation for scientific inference. Propositional logic offers a means for rejecting strong theories through *modus tollens*. For example, if a theory predicts a certain experimental result could not occur, and the result does occur, then the theory may be rejected; i.e.,

(Premise) If Hypothesis H is true, then event X will not occur.

(Premise) Event X occurred.

(Conclusion) Hypothesis H is not true.

Here, note that we do not need to posit an alternative to Hypothesis H. The above argument is valid, and the conclusion is justified if the premises are certain. Yet, in almost all cases of empirical inquiry, we are not certain in our premises, and when there is noise or error contraindicative events may occur even when Hypothesis H holds. When premises involve uncertainty, we may invoke probability theory. The modified version of the argument above adapted for significance testing is

(Premise) If Hypothesis H is true, then Event X is unlikely.

(Premise) An Event in set X has occurred.

(Conclusion) Hypothesis H is probably not true.

In this argument, the impossibility of Event X in the first premise has been replaced by rarity: that is, the event will *usually* not occur, assuming the Hypothesis $H$is true. In the case of NHST, the rare event is observing that the p value is lower than a certain level, usually .05. This argument, however, is NOT valid. The conclusions do not follow, and belief in the conclusion is not justified by belief in the premises. Pollard and Richardson (1987) demonstrate the invalidity of the argument by the following example:

(Premise) If Jane is an American, then it will be unlikely that she is a U. S. Congressperson.

(Premise) Jane is a U. S. Congressperson.

(Conclusion) Jane is probably not an American.

The argument is obviously invalid, and yet it is exactly the same form as the NHST argument used through the sciences (Cohen, 1994; Pollard & Richardson, 1987). The reason why it is invalid is that it fails to consider the alternative to the hypothesis "Jane is an American." If Jane were not an American, the observed event – that she is a Congressperson – would be impossible. Far from calling into doubt the hypothesis that Jane is an American, the fact that she is a Congressperson makes it certain that she is an American. Because the form of the above argument is invalid, the argument underlying NHST is also invalid, and for exactly the same reason. Alternatives matter for valid inference; there is no free lunch.

In the following sections, we further develop the argument that there is no free lunch from the basic definitions of probability. This development shows not only the need for alternatives but how alternatives may be specified and integrated into inference. When alternatives are considered, inference changes appreciably—often more evidence is needed to claim that there is an effect. The proper interpretation of this fact is that inference with a free lunch overstates effects, and, by extension, that cognitive scientists – and all other users of NHST – have sometimes been claiming effects with scant evidence.

No Free Lunch: The Frequentist Argument

Following the lead of Fisher (1925, 1955), significance tests are most often used without consideration of alternatives. The logic of significance testing described in the previous section is often called "Fisher's disjunction": either the null hypothesis is false, or a rare event has occurred. Since rare events typically do not occur, one can supposedly infer that the null hypothesis is probably false. Fisher's philosophical rival, Jerzy Neyman, opposed the testing of null hypotheses without reference to an alternative. Neyman (1956) goes so far as to state with respect to frequentist hypothesis tests that "the main point of the modern theory of testing hypotheses is that, for a problem to make sense, its datum must include not only a hypothesis to be tested, but in addition, the specification of a set of alternative hypotheses..." Neyman noted that to call an observation "rare" it must be in some region where observations do not tend to occur. But if the region is small enough, then any region could do, even the middle region. For instance, t-value between -.06 and .06 are quite rare and occur less than 5% of the time, yet these are not used to reject the null. Neyman argued that to know which region may serve as rejection rejection requires an implicit specification of an alternative.

Although we appreciate Neyman's argument, we think there is a much stronger case to be made from consideration of the foundations of probability theory. In the frequentist paradigm, probabilities are defined as long-run proportions. The probability of an event – say, that a coin lands heads up – is the proportion of times a coin lands heads up in the limit of infinitely many flips. This so-called large sample limit can be used to assess the adequacy of frequentist methods. In small data sets, analysis will be error prone due to random variation. But as sample size increases, good frequentist methods become more accurate, and in the large-sample limit, they reach the correct answer to arbitrary precision. This property is called *consistency*.

Because the central tenet of frequentist probability is convergence in the large

sample limit, it is important that hypothesis testing and model comparison be consistent, that is, they reach the correct answer always in the large-sample limit. From a frequentist point-of-view, consistency is a minimal property for good inference. Is significance testing consistent? If it is we expect to make the correct decision in the large-sample limit regardless of whether the null is true or false. First consider the case that the null hypothesis is false. In this case and in the large-sample limit, the null hypothesis will always be correctly rejected, which is consistent behavior. The difficulty arises when the null is true. In this case, by construction, test statistics will lead to a rejection with a set Type I error rate, usually 5%. The problem is that this probability does not diminish with sample size; regardless of how large one's sample is, the analyst is condemned to make Type I errors at the preset level. These errors, however, violate consistency. Therefore, significance testing violates the core principle of frequentist probability that knowledge is certain in the large-sample limit.

Confidence intervals are advocated as an alternative to significance testing (Cumming, 2008; Grant, 1962; Loftus, 1996), and researchers are prone to use these as a test of effects (Hoekstra et al., in press). Consider the behavior of a researcher who computes the sample mean and draws the associated 95% confidence interval, in order to determine what values are plausible. Whatever the true value is, and however large the sample size, the confidence interval will have the same 5% probability of excluding the true value, even in the large sample limit (Neyman, 1937). Hence, when estimation is to state that a particular value is implausible, it behaves like significance testing. Using estimation in this way is inconsistent and thus should be avoided.

It may seem that these violations are small and inconsequential—after all, what are the consequences of a 5% error rate? To appreciate some of these consequences, consider the study by Galak et al. (2012); in seven experiments, the authors tried to replicate the ESP results reported in Bem (2011). The seventh and last experiment featured $2,469$

participants, and the result yielded $t(2468) = -0.23$, $p = .59$. These results are clearly not significant, and it may be tempting to conclude that the statistical results from such an extremely large sample must be highly dependable. But this is a mistake – given that the null hypothesis is true and ESP does not exist, there was a 5% chance of obtaining a significant result. In other words, regardless of how many participants one tests, whether ten or ten million, with a fixed α level of .05 there is a 5% chance of falsely rejecting a true null hypothesis. So collecting more data does not increase one's chances of drawing the correct conclusion when the null hypothesis is true. In fact, using significance testing for high-powered replication experiments where the null hypothesis is true resembles a game of Russian roulette. In the case of Experiment 7 from Galak et al. (2012), the field of psychology dodged a bullet that had a 5% chance of hitting home ("Psychologists prove existence of ESP: results significant with 2,469 participants!"). If the researchers had used confidence intervals, it would have been no better ("Psychologists prove existence of ESP: narrow confidence interval excludes zero!")

To ameliorate the problem, we could make a seemingly slight change in significance testing to assure consistency. Instead of letting α be fixed at .05, we let it vary with sample size such that, in the limit, $\alpha \to 0$ as $N \to \infty$. We expand the notation, and let α_N denote the Type I error rate at a particular sample size. In significance testing, bounds are set such that $\alpha_N = .05$ for all N. Consider just a modest variant of the standard approach where we set the critical bound such $\alpha_N = \min(.05, \beta_N)$, where β_N is the Type II error rate at sample size N (the Type II error rate is the probability of failing to detect the alternative when it is true, and is the complement of power). Here we never let the Type I error exceed .05, just as in the standard approach, but we also decrease it as the sample size increases so that the Type I error rate never exceeds the Type II error rate. With this schedule, both α_N and β_N necessarily decrease to zero in the large sample limit, and consequently inference is consistent. This new rule meets two requirements: (1)

It is clearly consistent, as both Type I and Type II error rates decrease to zero in the large sample limit; and (2) Type I errors should not be any more plentiful than 1-in-20 if the null is indeed true.

This variant approach differs from $p < .05$ significance testing in a few crucial aspects. One is that it requires the analyst to pay for lunch. The computation of β_N and α_N requires the specification an effect size under the alternative: that is, a fixed value for the effect size that is compared to the null hypothesis. For the purpose of demonstration, we set ours to .4 in value. We understand that some researchers may not like making this commitment to an alternative, but researchers who insist on using significance testing violate the minimum frequentist obligation to be consistent in inference.

Although this variant rule is similar to the original $p < .05$ rule, the resulting conclusions may differ markedly. Displayed in Figure 1 are the critical effect-size values needed to reject the null, and these are plotted as a function of sample size. The wide, light-colored line is for standard significance testing where $\alpha_N = .05$. As can be seen, the needed critical effect size falls with sample size, and will fall to zero in the large sample limit. The dashed line shows the critical effect size for $\alpha_N = \min(.05, \beta_N)$ rule for an alternative effect size of .4 in value. Critical effect sizes reach a bound of .2 and decrease no further. With this small variation, we have eliminated researchers' ability to reject the null with small effects; indeed effects need to be larger than .2 to reject the null. The value of .2 is no coincidence, and it is easy to show that this value is half the value of the specified alternative (Faul, Erdfelder, Lang, & Buchner, 2007). We also computed critical effects sizes for the rule that Type II errors are 5 times as large as Type I errors ($\alpha_N = \beta_N/5$; see the thin solid line). This rule leads to consistent inference as both Type I and Type II errors decrease to zero in the large sample limit. It too requires that the value of effect size be set under the alternative. Adopting this rule leads to the same results that small effects cannot be used as evidence to reject the null no matter the

sample size. The idea of balancing Type I and Type II errors is not new; Neyman and Pearson (1933) suggested that researchers strike some sort of balance between the two sorts of errors, based on the research context. This advise, however, has been ignored.

Significance testing as is currently practiced is inconsistent. Because consistency is a minimal requirement for frequentist inference, we recommend that the field adopt a consistency criterion in everyday practice. Consistent inference holds only when the analyst specifies an alternative, and inference with an alternative not only leads to different conclusions than without, but raises the bar for declaring effects. This difference is most pronounced for for larger sample sizes where small observed effects are no longer sufficient to reject the null.

No Free Lunch: The Bayesian Argument

In the Bayesian framework, probability describes a degree of belief. It is a subjective concept from first principles, and different analysts are expected to arrive at different probabilities for events depending on their background knowledge. The key principle in Bayesian statistics is that beliefs, represented as probability, should be revised optimally in light of data. The steps to optimal revision are well-known—the analyst simply follows Bayes' rule. Perhaps the primacy of Bayes' rule in Bayesian analysis is stated most directly by Bradley Efron in his presidential address to the American Statistical Association (Efron, 2005). Efron writes, "Now Bayes rule is a very attractive way of reasoning, and fun to use, but using Bayes' rule doesn't make one a Bayesian. *Always* using Bayes' rule does, and that's where the practical difficulties begin." (p. 2, italics in original).

Figure 2A shows an example of updating by Bayes' rule for the case where data are assumed to be normally distributed governed by some true effect size. The broad dashed line denotes *prior beliefs* about the true effect size, and these are the beliefs the analyst may have before seeing the data. As can be seen, these are broadly distributed reflecting

that a wide range of values of the effect size are plausible before seeing the data. With this prior, analysts may make statements such as, "My belief that the effect size will be between -1 and 1 is .84," which is true for the shown prior because indeed 84% of the area is between -1 and 1. After the data are observed, these beliefs are updated, and the solid line shows beliefs, called the *posterior beliefs*, after observing 40 observations with a sample effect-size of .4. As can be seen, the data have caused a shift in beliefs where larger values of effect size are now more plausible than they were with prior beliefs. One way of characterizing these beliefs is to compute a central interval that contains a set area of the posterior distribution. The dark solid interval marker with a star is the so-called 95% credible interval for the shown posterior.

The main question at hand is whether there is an effect in a set of data. One seemingly reasonable approach is to see whether the 95% credible interval contains zero. If it does not, then a decision to reject the null may be made, and indeed, there is evidence at some level, perhaps .05, to support an effect. Figure 2B shows a number of credible intervals. Credible Interval A excludes zero, so, accordingly, there is evidence for an effect. Credible Interval B includes zero, so the evidence is not as strong. This approach is analogous to inference by confidence intervals, with the main difference being that the confidence intervals are replaced by Bayesian credible intervals. The intuitive appeal of this position is quite strong, and we reluctantly admit that we have used it on occasion to establish effects (Rouder et al., 2008). One of the earliest advocates of this position was Lindley (1965). A real-world example in cognition is provided by Storm, Tressoldi, and Utts (2013) who note that their credible interval on effect-size excludes chance levels as evidence of telepathy. A second use of credible intervals comes from Kruschke (2012), who asks researchers to consider a prespecified small region, often around zero, called the *region of posterior equivalence* or ROPE. Such a ROPE is also shown in Figure 2B as a shaded region. Credible interval iv is entirely contained in the ROPE, and consequently,

for this credible interval Kruschke would claim evidence for a null effect.

Kruschke takes a nuanced view of the use of credible intervals. He advocates that researchers graph and inspect their credible intervals, the zero point, and ROPE, and use such a graph to report estimates of effect size and associated precision. The zero-point and the ROPE serve as guides. In this regard, Kruschke's approach is similar to Gelman's (2005), and both ask researchers to emphasize visual inspection over hard-and-fast decision rules. Given the unfortunately long history of using credible intervals to make inferential decisions, we suspect that this nuanced view will be lost on most researchers and they will use credible intervals for testing hypotheses. Indeed, Storm et al. (2012) cite Kruschke for their support of telepathy from a credible intervals.

Inference by credible intervals seems to have a nearly free lunch property. While it is true that Bayesians specify priors, the posteriors are relatively stable even when the priors vary considerably so long as there is a reasonable amount of data. Figure 2C shows the case when the prior has a standard deviation of 10 rather than 1. This prior is so broad that most of the mass is off the graph. The data is the same as that for Figure 2A, that is, the observed effect size is .4 across 40 observations. Even though the priors differ markedly, the resulting posterior densities are nearly identical. Most critically, the resulting credible intervals are about the same, and more generally, inference based on whether they contain zero or fall into a ROPE is hardly affected by the choice of prior so long as the sample size is not too small and the prior is sufficiently broad. It is in this sense that inference by credible interval may be made without much thought given to the specification of the alternative. Given that the posterior is largely invariant to the choice of prior, many researchers might seemingly prefer very broad priors, even those that are flat everywhere, because they seem to imply a position of *a priori* ignorance.

The critical question for a Bayesian is not whether inference by credible interval is robust to prior specification but whether it follows from Bayes' rule. On one hand, it

might be argued that because the posteriors and credible intervals in Figure 2A-C come about from Bayes' rule, all inference about the presence or absence of an effect automatically follows Bayes' rule. This argument, however, is wrong. Bayes' rule in this context describes how to update the plausibility of each value of effect size. Take for example, the value of $\delta = .5$ in Figure 2A, for which there are open points on the curves. Here the posterior density is 5.5 times higher than the prior density indicating that the data have caused a increase in plausibility by a factor of 5.5. Another example is provided for $\delta = 0$, which is indicated by the filled points. Here, the data has led to a decrease in plausibility by a factor of 3.5. For Figure 2A the credible intervals roughly tell us which areas have increased posterior plausibility relative to prior plausibility. Unfortunately, this concordance is fortuitous and does not hold generally. In Figure 2C, the credible interval does not include zero, and one might be tempted to think that the data render this value less plausible. Yet, this statement is false; in fact, the posterior density at zero is higher than the prior density is! The data make the value of zero *more* plausible afterwards than before. Using the credible intervals for inference distorts the picture because the interval makes it seem as though zero is not plausible when in fact the data lead to an increase in plausibility.

The discordance between Bayes' rule and inference by credible intervals may be shown with the ROPE as well. Figure 2D shows a prior and a posterior that is centered around a null value but extends well past a ROPE. According to credible-interval logic, the experiment has failed to achieve the requisite precision to show support for the null. The situation is different when viewed from Bayes' rule. Before observing the data, the prior probability that δ is in the ROPE is about .08, which is not large. After observing the data, the probability has increased to about .48, and the ratio of increase is 6-to-1. Hence, the data has forced us to revise our beliefs upward by a factor of 6 even though the credible intervals are not well contained in the ROPE.

One of the key points in the above analysis is that with Bayes' rule the updating factor is a function of the prior. Take the above case in Figure 2D where the plausibility of a true effect being in the rope was increased by 6. In this case, the prior was spread but not excessively so. Suppose the prior were very broad having a variance of one-million. In this case, the prior area within the ROPE is tiny, .00016 but the posterior area in the rope is still sizable, .047. The updating factor is now quite large, almost 3000-to-1 in favor of the null. This factor holds even though the credible intervals are much broader than the ROPE. As can be seen, the prior does matter, and flat priors with arbitrarily large variances lead to arbitrarily large support for any reasonable value, including the null. The problem is not with Bayes rule; rather, it is with the broad priors. As pointed out by De Groot (1982), the flat prior does not mean a commitment to no particular alternative; the flat prior means a commitment to arbitrarily large alternatives, and thus it is not surprising that the null hypothesis looks very good in comparison. The upshot is that updating by Bayes' rule requires reasonable and judiciously-chosen priors, and it is here that we see that the specification of reasonable alternatives to the null hypothesis matters.

The differences between credible intervals and Bayes' rule were perhaps shown most starkly by Lindley (1957) and Bartlett (1957). These statisticians noted that one could find data such that 95% of the posterior mass might lie to one side of the 0, favoring the rejection of the null, yet the probability of the null relative to a reasonable alternative is high, say 20-to-1, favoring the rejection of the reasonable alternative in favor of the null. The discordance is known as Lindley's paradox, yet it is not problematic Checking whether a credible interval covers a value or is in a range does not follow from Bayes' rule; it is rather a heuristic approach borrowed from frequentist inference by confidence intervals. The only reason that it appears to be a paradox is a disagreement between intuition and what can plainly be shown using Bayes' rule. There is no reason, however, why intuitions built on frequentist logic should be consistent with Bayes' rule.

In the previous section, we proposed a variant of significance testing to gain consistency. In this section, we discuss how analysts can meet Bayes' rule in inference. The answer is straightforward—use Bayes' rule directly. To do so, the analyst must place an *a priori* belief on whether a particular model holds, and then update this belief in light of data using Bayes' rule. Laplace first expressed the view that probabilities may be placed on models to answer questions of inference (Laplace, 1829), and Jeffreys (1961) provided the first systematic treatment. We demonstrate this using two models, a null model denoted $\mathcal{M}_0$ and an alternative denoted $\mathcal{M}_1$. The null model specifies that the true effect size is zero; the alternative specifies that the true effect size is nonzero, and our *a priori* uncertainty about its value is described by a standard normal as in Figure 2A. Our datum is the observed effect size, denoted by d.

We begin by noting that one can describe the *a priori* relative plausibility of Model $calM_1$ to Model $\mathcal{M}_0$, by the ratio $Pr(\mathcal{M}_1)/Pr(\mathcal{M}_0)$, called *the prior odds*. Direct application of Bayes' rule yields the posterior odds

$$\frac{Pr(\mathcal{M}_1|d)}{Pr(\mathcal{M}_0|d)} = \frac{Pr(d \mid \mathcal{M}_1)}{Pr(d \mid \mathcal{M}_0)} \times \frac{Pr(\mathcal{M}_1)}{Pr(\mathcal{M}_0)}. \tag{1}$$

The term $Pr(d \mid \mathcal{M}_1)/Pr(d \mid \mathcal{M}_0)$ is the *Bayes factor*, and it describes the extent to which the data have updated the prior odds (Kass & Raftery, 1995). We denote the Bayes factor by B_{10}, where the subscripts indicate which two models are being compared. A Bayes factor of $B_{10} = 5$ means that prior odds should be updated by a factor of 5 in favor of model $\mathcal{M}_1$; likewise, a Bayes factor of $B_{10} = 1/5$ means that prior odds should be updated by a factor of 5 in favor of model $\mathcal{M}_0$. Bayes factors of $B_{10} = \infty$ and $B_{10} = 0$ correspond to infinite support of one model over the other with the former indicating infinite support for model $\mathcal{M}_1$ and the later indicating infinite support for model $\mathcal{M}_0$.

The specification of the alternative enters when computing the probability of the data under the models. The probability of the data under the null, $Pr(d \mid \mathcal{M}_0)$, is

straightforward; it is the probability of observing a given sample effect size when the true effect size is zero which follows a scaled t-distribution. The probability of the data under the alternative may be expressed as:

$$Pr(d \mid \mathcal{M}_1) = \int_{\delta} Pr(d \mid \delta)\pi(\delta)d\delta. \tag{2}$$

This conditional probability of data is also called *the marginal likelihood* for $\mathcal{M}_1$. It is the weighted average of the likelihood over all possible parameter values, with the weights determined by the prior $\pi(\delta)$. If the prior is broad, then $\pi(\delta)$ is very small and this weighted average is low. If the prior is more realistic, and the data are where there is some prior mass, then the weighted average is increased. If we let m_1 denote this weighted average, and m_0 denote the probability of the data under the null hypothesis, then the Bayes factor is simply the ratio m_1/m_0. It is in this manner that the Bayes factor reflects specification of the alternative. Moreover, Bayes factors penalizes flexible alternatives: alternatives with broad priors that account for a wide range of data will yield low average probability for any observed data. This penalty flows naturally from Bayes' theorem without counting of parameters or asymptotic arguments (Jefferys & Berger, 1991).

For $\mathcal{M}_0$ and $\mathcal{M}_1$, it may be shown that the Bayes factor B_{10} is ratio of densities depicted by the points in Figure 2A and 2C (Dickey, 1971; Morey, Rouder, Pratte, & Speckman, 2011; Verdinelli & Wasserman, 1995; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). The null is evidenced if the plausibility of $\delta = 0$ increases; the alternative is evidenced if the plausibility decreases.

The Bayes factor has many advantages: it allows researchers to state evidence for or against models, even nulls, as dictated by the data, it provides a report of evidence without recourse to reject or fail-to-reject, avoiding the dichotomous thinking inherent in significance testing (Cumming, 2014), and provides a natural penalty for model complexity. Nonetheless, we think the most important advantage is that it is

principled—it is the only approach that meets the Bayesian obligation to update rationally. In this case, beliefs about models themselves are updated rationally through Bayes' rule. Other methods of stating evidence for or against positions necessarily violate Bayes' rule and are not ideal. As shown in Figure 2, these other methods can deviate substantially from Bayes' rule.

Edwards et al. (1963) were perhaps the first to correctly identify these issues. The force of their argument stems from the elevation of Bayes' rule to a foundational principle in inference. The Bayes factor is the logical consequence of Bayes' rule. Hence, if the Bayes factor requires a commitment to a well-specified alternative, then this is the requirement of being principled and coherent. In Edwards et al.'s hands, Lindley's paradox showed the necessity of specifying alternatives, and any method that allowed for inference without prior sensitivity was unprincipled because it implied that posterior beliefs about the models were not being updated rationally. In effect, the Bayes factor is not "too" sensitive to the prior; instead, other methods that do not require a specified alternative are not sensitive enough.

We find it helpful to trace the thoughts of Lindley himself on these issues. After stating the paradox in 1957, Lindley at first recommended the credible interval approach (Lindley, 1965). But by the 1980s, Lindley reverses course completely and advocates Bayes factor as the principled Bayesian approach to inference. Lindley has repeated this position several times (Lindley's 1982 comment on Shafer, 1982; Lindley, 1990; Lindley, 2000). Lindley (2000), in particular, is an easily readable introduction to his later views on statistical philosophy.

As with frequentist methods, one can ask if it matters appreciably if one uses inference by credible interval or inference by Bayes factors. The wide, lightly-colored line in Figure 3 is the critical effect size needed from credible intervals to exclude zero. Also shown are the effect-size values needed for Bayes factors evidence in favor of the

alternative of 1-to-1, 3-to-1, and 10-to-1 (solid, dashed, and dotted thin lines, respectively). These Bayes factors were evaluated for the alternative $\mu \sim \text{Normal}(0, 1)$, that is, the true effect size is from a standard normal of mean zero and variance one. This prior, though informative, is quite reasonable for experimental psychology. It is also called the *unit-information prior*, and underlies the popular BIC model-selection statistic (Raftery, 1995). The credible interval method requires little evidence to deem the null implausible, and for no sample size is the evidence more than 3-to-1 in favor of an alternative. In fact, for sample sizes as small as 50, the credible-interval rejection of the null corresponds to a Bayes factor of equivocal evidence, and for larger sample sizes, the rejection is made for data that actually favor the null more than the alternative. When alternatives are judiciously specified, larger observed effect sizes are needed to state evidence for an effect.

How To Specify The Alternative, A Real-World Example

We have argued that researchers must commit to specific alternatives to perform principled inference. Many researchers, however, may feel unprepared to make this commitment. We understand the concern. The good news is that making judicious commitments is not as difficult as it might seem, because researchers have more information than they may realize (Armstrong & Dienes, 2013; Dienes, 2011, Guo, Li, Yang, & Dienes, 2013). The argument goes as follows:

The key to specification of the alternative is consideration of effect-size measures, which are widely understood and can be used in almost all cases. Importantly, effect sizes have a natural scale or calibration. For example, an effect size of 10.0 is very large, and effects of this size would be obvious with just a handful of observations. Likewise, effect sizes of .01 are too small to be discriminated in most experiments, in fact, trying to do so would exhaust our subject pools. The four panels on the left side of Figure 4 captures these constraints. In all four panels, the priors have common properties: 1. the prior is

symmetric about zero reflecting ignorance about the direction of the effect; 2. there is decreasing plausibility with increasing effect size. The difference between them is in the scale or range of effect sizes, and these may be seen in the x-axis values. The prior labeled "Too Wide," shows a specification of the alternative that makes a commitment to unrealistically large values of effect size. Evidence for the null would be overstated in this case because the alternative is too broad. There are three additional panels, labeled"Wide", "Narrow," and "Too Narrow," that show other specifications. The specification "Too Narrow" shows commitments to very small effect sizes, and would not be useful in most applications in experimental psychology. The priors "Wide" and "Narrow" define the outer points of reasonable specification—priors more dispersed than "Wide" seem too dispersed, and priors more concentrated than "Narrow" seem too concentrated. These specifications are referenced by a single quantity, the scale of effect size, denoted σ_δ, which ranges in these specifications from 10, for "Too Wide," to .02, for "Too Narrow." A reasonable range for r is from .2, the narrow prior, to 1, the wide prior. In practice, we often use $\sigma_\delta = .5$ and $\sigma_\delta = \sqrt{2}/2$, depending on our prior expectations.

These different specifications will lead to different Bayes factors, but the degree of variation is far less than one might expect. The right side of Figure 4 shows the effect of the prior scale of the alternative, σ_δ, on Bayes factor for three values of the t-statistic as a sample size $N = 50$. The highlighted points in green show values for $\sigma_\delta = .2$ and $\sigma_\delta = 1$, which define a reasonable range. The red points are from scales considered too extreme. Prior scale does matter, and may change the Bayes factor by a factor of 2 or so, but it does not change the order of magnitude. The priors used on the left side of Figure 4 with specified scale on effect size serves as an example of a *default prior*, a prior that may be used broadly, perhaps with some tuning across different contexts. We have recommended default priors in our advocacy of Bayes factors (Morey & Rouder, 2011; Morey et al., 2013; Rouder et al., 2009; Rouder et al., 2012; Rouder et al., 2013; Rouder & Morey 2012;

Wetzels, Grassman, & Wagenmakers, 2012; Wetzels & Wagenmakers, 2012), and we implement them for *t*-tests, regression, and ANOVA, as well in our web applets (`pcl.missouri.edu/bayesfactor`) and BayesFactor package for `R` (Morey & Rouder, 2014). Dienes (2011) and Gallistel (2009) recommend alternative subjective priors that capture more contextual information, and these constitute judicious approaches for researchers who wish to make alternative choices in analysis.

For concreteness, we illustrate the Bayes factor approach to paying for lunch with an example from the moral reasoning literature. One of the current themes in understanding how people make moral judgements is that there are two systems, one emotional and another utilitarian (Greene & Haidt, 2002). Accordingly, moral judgements may vary even within a person depending on the degree to which each system is engaged. Rai and Holyoak (2010) for example attempted to manipulate judgements in the following moral dilemma: A runaway trolley car with brakes that have failed is headed towards five workers on the train track. Participants are put into the shoes of a bystander with the option to pull a lever which redirects the train onto a side track. Unfortunately, there is a worker on this side track as well. The bystander has two choices: either do nothing and let five workers be killed or pull the lever and actively kill one worker. Rai and Holyoak (2010) reasoned that if the emotional system was engaged, the participant might not actively kill one worker, but if the utilitarian system was engaged, the participant might actively kill the one worker to save the other five. To manipulate the engagement of these systems, Rai and Holyoak (2010) asked participants to declare reasons for pulling the lever and killing the one and saving the five. Listing reasons should put more emphasis on the utilitarian system, and therefore prime the more utilitarian judgment of pulling the lever. In one condition, participants were asked to list only two reasons, in the other condition they were asked to list as many reasons as they could, up to seven reasons. The prediction is that those in the seven-reason condition would be more utilitarian in their moral

judgments than those in the two-reason condition.

Rai and Holyoak (2010) considered a second, alternative theory from consumer psychology. In consumer psychology, giving more reasons for a choice often decreases rather than increases the probability of making that choice (Schwarz, 1998). The reason for this apparent paradox is straightforward—as people attempt to list more reasons, they either fail to do so or list reasons of lower quality than their primary reasons. A focus on either this failure or on these reasons of lower quality results in a less favorable view of the choice. This explanation, called here the *framing alternative*, predicts that asking participants for many reasons for pulling the lever results in a lower probability of a moral judgment to do so.

We consider the results of Rai and Holyoak (2010), Experiment 1, where 124 participants first provided reasons for pulling the lever, and then indicated on a 4-point scale whether they agreed with the judgment of pulling the lever. Half of the participants were given the two-reason instructions, the remaining half were given the seven-reason instructions. The critical contrast is whether participants agreed with pulling the lever more in the two-reason than seven-reason condition. Before assessing this critical contrast, Rai and Holyoak first assessed whether the instruction manipulation was effective in generating more reasons. Participants provided an average of 3.2 and 1.7 reasons, respectively, in the seven- and two-reason conditions, respectively. The corresponding t-value, 5.46, is large and corresponds to a small p-value well under .01. To assess the evidence in a principled fashion, we compared the null to a wide default alternative with $\sigma_\delta = 1$ (see Figure 4), and the resulting Bayes factor is about 50,000-to-1 in favor of the alternative. Therefore, there is compelling evidence that the instruction manipulation did indeed affect the numbers of reasons provided.

The critical contrast is the difference in moral judgements in the seven-reason condition than in the two-reason conditions. Rai and Holyoak (2010) report more

agreement with pulling the lever in the two-choice condition than in the seven-choice condition (2.17 vs. 1.83 on the 4-point agreement scale). The t value for the contrast is $t(122) = 2.11$, which is just significant at the .05 level. Based on this significance, the authors conclude, "We found that people are paradoxically less likely to choose an action that sacrifices one life to save others if they are asked to provide more reasons for doing so." We are less convinced.

Given the researchers' intent to test competing theories, we chose the following three models: The first is the null—there is no effect of instructions on the moral judgements—which is certainly plausible. We chose two different alternatives: the one-sided alternative from the dual-process theory that the utilitarian judgement is greater in the seven-choice condition, and the other one-sided alternative from the framing alternative that the utilitarian judgment is less in the seven-choice condition. The prior scale on effect size in these alternatives is $\sigma_\delta = .7$, an in-between value reflecting that effect sizes in framing manipulations and moral-reasoning dilemmas tend to vary. Relative to the null, the Bayes factors are 15.1-to-1 against the dual-process hypothesis and 2.77-to-1 in favor of the framing alternative. The evidence from the data stand in opposition to the dual-process predictions, but they do not provide much support for the framing alternative. We think the evidence in this regard is marginal at best, and would be far more persuasive if the evidence for the framing alternative was larger relative to the null hypothesis.

Conclusion

In this paper we have argued that if one is to test hypotheses, or more generally compare models, then as a matter of principle one must pay the price of specifying a reasonable alternative. Moreover, we argue that paying the price for principled inference may have material effects on conclusions. When alternatives are reasonable, small

observed effects in large samples do not readily serve as evidence for true effects. In this section, we comment on why specifying alternatives, that is paying for lunch, would lead to a better psychological science.

Is Specifying Alternatives Too Subjective?

The $p < .05$ rule purportedly provides field-wide protection: supposedly, when we, the scientific community, insist on $p < .05$ rule, we implement a safeguard against which individual researchers cannot exaggerate their claims. Our view is that the field will give up no safeguards whatsoever by adopting a specification viewpoint. In fact and to the contrary, having researchers state *a priori* their expectations of effect sizes under the alternative will vastly improve the transparency and integrity of analysis. Analyses will be more transparent because researchers will need to state and defend their choices of the alternative, and researchers and readers can easily evaluate these defenses much as they evaluate all other (subjective) aspects of research including the operationalization of variables and the link between experimental results and theoretical objectives. In this regard, analysis is put on an equal footing with all other aspects of research. Analyses will have more integrity because researchers can no longer hide behind a manufactured rule which stands only on convention without any intellectual basis (Gigerenzer, 1998). Instead, analysts must invest some thought into possible alternatives, and defend their analysis decisions. We believe that psychological scientists are up to the task, and when they do so, the benefits will be broad and obvious.

Should We Test Hypotheses?

Some critics have objected to hypothesis testing on the grounds that the null model is never true (Cohen, 1994; Meehl, 1978). Those who advocate this position recommend that researchers replace estimation of effect sizes, and, presumably, the attentive reader might wonder if estimating effect sizes avoids the need for specifying alternatives. There

are three basic issues: that the null is never true, that estimation should replace testing, and that estimation requires no commitment to nulls or alternatives. We discuss them in turn.

We find the position that "the point null is never true" difficult on several grounds (Iverson, Wagenmakers, & Lee, 2010; Morey & Rouder, 2011; Rouder & Morey, 2012). However, it is not necessary to consider point nulls to show the necessity of specifying alternatives. Each of our examples may be restated using interval null hypotheses, such as $-.1 < \mu < .1$, without any loss. The same inferential logic that applies to point nulls also applies to bounded intervals. The "null is never true" argument does not obviate the need to specify alterantives.

A second question is whether estimation should replace testing as is advocated by Cumming (2014). We have no issue with estimation as part of the scientific process. However, there are questions that estimation cannot answer; for instance, "Does the Higgs Boson exist?" "Do all humans come from a single individual?" Or, pedestrianly, "How concordant is a certain set of data with a certain theory?" These questions are about evidence for hypotheses (Morey, Rouder, Verhagen, & Wagenmakers, in press). Although estimation has been suggested as a replacement for testing for several decades now, testing remains nearly ubiquitous. Perhaps testing retains its popularity because it answers questions that researchers ask. Estimation fails because researchers do not wish to simply catalog effect sizes across tasks and manipulations. They want to understand them in theoretical terms.

Finally, there is the question whether estimation entails specification of alternative models. It certainly appears that the estimate of effect size is model free. To see if this is the case, consider the following hypothetical where a wealthy donor who decides to test the skills of a data analyst. The donor tells the analyst that she will donate \$10,000 to the analyst's favorite charity if the analyst can predict exactly the number of heads in 1000

flips of a coin. If the analyst cannot predict the exact number, the donor will deduct an amount from the total according to a schedule: she will deduct $1 from the total if the analyst is off by a single count; $4 if the analyst is off by two counts, $9 if the analyst is off by three counts and so on. The donor warns the analyst that the coin may or may not be fair. To compensate for this complication, the donor allows the analyst to observe 1000 flips to form an opinion, and then has to predict the outcome of the next 1000 flips. We consider three different possible outcomes of the first 1000 flips. **a.** Suppose there are 750 heads in the first 1000. We should be certain that the coin is not fair and the best prediction is that another 750 heads will occur in the next 1000 flips. Here it seems as if specification of models is unnecessary. **b.** Suppose instead that there were 510 heads in the first 1000 flips. It seems likely that this value is far more concordant with a fair coin than with an unfair coin, and if this is the case, the better prediction might be 500. The 10-head deviation from 500 in the first set is likely noise and if this is the case, then predictions based on it will be more error prone than those at the 500 value. Indeed, if the unfair-coin-alternative is the hypothesis that all probabilities are equally likely, the outcome of 510 heads supports the fair coin null over this alternative by a factor of 20-to-1. **c.** Perhaps the most interesting case is if there are 540 heads in the first 1000. This value is intermediate, and in fact, is equally probable under the null as under the above alternative. A judicious analyst may wish to take an average of 500 and 540, the estimates under each equally-plausible model. In fact, the best prediction in this case is 520, but the analyst had to specify the alternative to obtain it.

The above example with the donor and coin is not far fetched. Most important theoretical questions are about regularity, lawfulness, and invariances, which correspond to null hypotheses like the fair coin. To obtain an estimate, analysts must be willing to commit to an alternative and weight the estimates from the null and alternative models, just as the analyst did in the above example. Currently, most researchers do not take

advantage of model averaging, and, as a consequence, we suspect most are overstating the effect sizes in their studies. The critical point is that estimation is not model-free, and researchers who make commitments are acting with more principle than those who do not. Researchers may choose to use summary statistics like sample means. They should note, however, that this choice excludes rather than includes lawful invariances, just as the analyst who predicts 510 heads after observing 510 heads has excluded the possibility of a fair coin. Researchers who make such a choice should be prepared to defend it, though, in truth, it seems rarely defensible in experimental situations.

Bayesian Probability Works Well In Scientific Contexts

Bayesian probability has many theoretical and practical advantages. In the Bayesian view, probabilities are the expression of belief rather than long-run proportions. Probabilities are held by observers rather than a physical properties of the system, and can be changed or updated rationally in light of new information. The notion that probabilities are mutable beliefs corresponds well with scientific reasoning where opinions are revised through experiments (Rozenboom, 1960). Because Bayesian analysis has the built-in concept of prior belief, it is natural and transparent to specify alternatives, and Bayes' rule provides a mechanisms to understand both why paying for lunch is necessary, and how the particular prior choices affect inference.

The Adverse Effects of Free-Lunch Myth

In the beginning of this paper, we argued that the free-lunch myth—the belief that we may learn about the null without specifying alternatives—has led to a mind-set and culture that is academically counterproductive. From a pragmatic perspective, free-lunch inference overstates the evidence against the null and leads to rejections with too low a threshold. This low threshold fools researchers into thinking their results are more reliable than they truly are, much as the jury rejected the null hypothesis about Sally Clark's

innocence. Yet, we are equally concerned with a broad perspective. Significance testing has reduced data analysis to a series of prescribed procedures and fixed decision rules. Inference has become an intellectual dead zone that may done by algorithms alone (Gigerenzer, 1998). We need only provide *pro forma* evaluation of the outputs. And in doing this *pro forma* evaluation, we abandon our responsibility to query the data. We use significance testing to reify results without appreciation of subtlety or qualification, and to manufacture a consensus even when this consensus is unwarranted or meaningless. We live down to the $p < .05$ rule, and, in this process divorce what we may know about our data from what we can tell others about them. It is in this space, where hypothesis testing is simultaneously a bureaucratic barrier and the sanctioned ritual of truth, that we take our short cuts, be they minor infelicities at the margins or major failures of outright fraud. And it is here where we nāively convince ourselves and the community at large of the veracity of claims without anything close to principled evidence.

There are a number of current proposals on how to remedy our statistical practices (e.g., Simmons, Nelson, Simonsohn, 2011). Although many of these proposals make sense, most miss the critical point that alternatives must be specified. This is a shame. If the field is ready to do the hard intellectual work of specifying alternatives, then assuredly better science will result.

References

Abelson, R. P. (1997). On the suprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.

Armstrong, A. M., & Dienes, Z. (2013). Subliminal understanding of negation: Unconscious control by subliminal processing of word pairs. *Consciousness & Cognition*, *22*, 1022–1040.

Bartlett, M. S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, *44*, 533-534.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531-533.

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

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

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558-1561.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286-300.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*.

Dawid, A. P. (2005). Statistics on trial. *Significance*, *2*, 6–8.

DeGroot, M. H. (1982). Lindley's paradox: Comment. *Journal of the American Statistical Association*, *77*(378), 336-339.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204-223.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave MacMillan.

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

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

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

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, *100*(469), 1–5.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, *39*, 175-191.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 69-78.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151-156.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate Psi. *Journal of Personality and Social Psychology*, *103*, 933–948.

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

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

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54–61.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*, 517–523.

Guo, X., Li, F., Yang, Z., & Dienes, Z. (2013). Bidirectional transfer between metaphorical related domains in implicit learning of form-meaning connections. *PLoS ONE*, *8*, e68100.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE*, *8*, e72467.

Hill, R. (2005). Reflections on the cot death cases. *Significance*, *2*, 13–15.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (in press). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review*.

Huizenga, H., Wetzels, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2012). Four empirical tests of unconscious thought theory. *Organizational Behavior and Human Decision Processes*, *117*, 332–340.

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, *15*, 172–181.

Jefferys, W. H., & Berger, J. O. (1991). *Sharpening Ockham's razor on a Bayesian strop.* Technical Report.

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

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54–69.

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

Kruschke, J. K. (2012). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General.*

Laplace, P.-S. (1829). *Essai philosophique sur les probabilités.* Brussels: H. Remy.

LeBel, E. P., & Campbell, L. (in press). Heightened sensitivity to temperature cues in highly anxiously attached individuals: Real or elusive phenomenon? *Psychological Science.*

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379.

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

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian point of view, part 2: Inference.* Cambridge, England: Cambridge University Press.

Lindley, D. V. (1990). Good's work in probability, statistics and the philosophy of science. *Journal of Statistical Planning and Inference*, *25*, 211–223.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *49*, 293-337.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current directions in psychological science*, *5*, 161-171.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology,*, *46*, 806–834.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*, 68-75.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406-419.

Morey, R. D., & Rouder, J. N. (2014). *BayesFactor 0.9.6.* Comprehensive R Archive Network.

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368-378.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (in press). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*, 333-380.

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *Journal of the Royal Statistical Society. Series B (Methodological)*, *18*(2), 288-294.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289-337.

Nobles, R., & Schiff, D. (2005). Misleading statistics within criminal trials: The Sally Clark case. *Significance*, *2*, 17–19.

Nosek, B. A., & Bar–Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Osherovich, L. (2011). Hedging against academic risk. *Science–Business eXchange*, *4*.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Pollard, P., & Richardson, J. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*, 159-163.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, *27*, 411–427.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science*, *34*, 311–321.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE*, *7*, e33423.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682-689.

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877-903.

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes-factor meta-analysis of recent ESP experiments: A rejoinder to Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*, *139*, 241-247.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374.

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 and Review*, *16*, 225-237.

Rozenboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple–study articles. *Psychological Methods*, *17*, 551–566.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of

declarative and experiental information in judgment. *Personality and Social Psychology Review*, *2*, 87–99.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, *55*, 62-71.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, *77*(378), 325-334.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., et al. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Storm, L., Tressoldi, P. E., & Utts, J. (2013). Testing the Storm et al. (2010) meta-analysis using Bayesian and frequentist approaches: Reply to Rouder et al. (2013). *Psychological Bulletin*, *139*(1), 248-254.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis significance tests. *Psychological Methods*, *6*, 371-386.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, *90*(430), 614–618.

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

Wagenmakers, E.-J., Krypotos, A., Criss, A., & Iverson, G. (2012). On the interpretation

of removable interactions: a survey of the field 33 years after Loftus. *Mem Cognit*, *40*(2), 145-60.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158–189.

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

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for ANOVA designs. *American Statistician*, *66*, 104–111.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064.

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*, 298-300.

Author Note

Jeff Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211, rouderj@missouri.edu. This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

Footnotes

[1]According to Wikipedia, the earliest known written usage of this phrase is from a 1938 *El Paso Herald-Post* newspaper article entitled "Economics in Eight Words."

Figure Captions

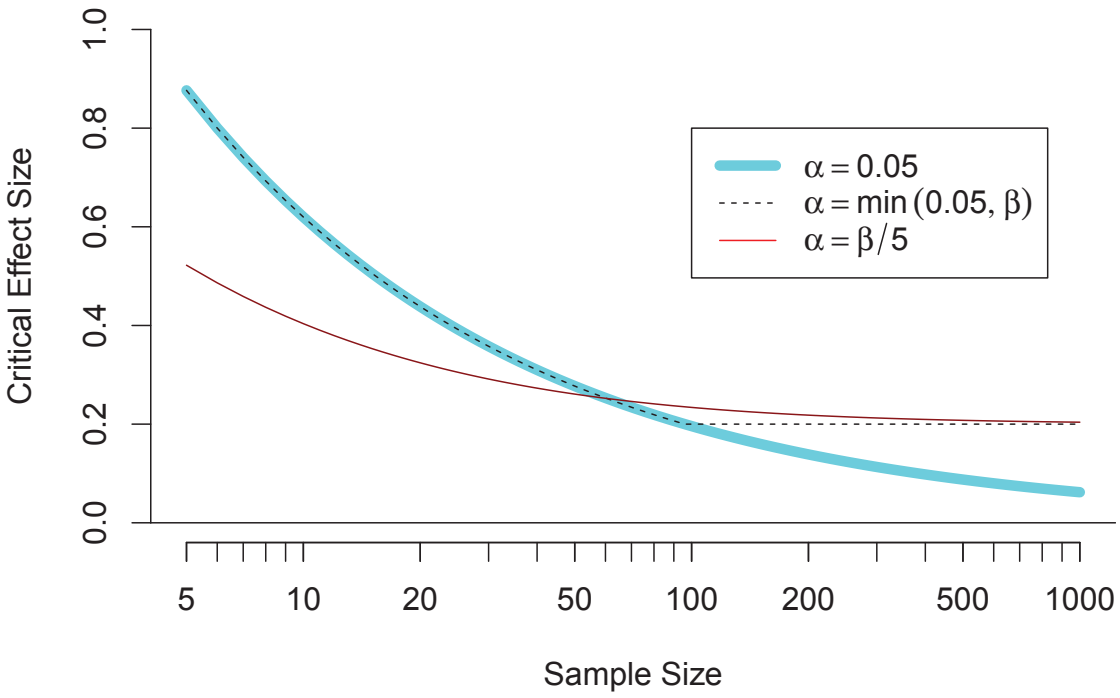
Figure 1. Critical effect sizes needed to reject the null as a function of sample size for a few rules. The wide, light-color line shows the case for the $p < .05$ rule; the dashed line shows the case for the $\alpha_N = \min(.05, \beta_N)$ rule; the thin solid line shows the case for the $\alpha_N = \beta_N/5$ rule. For the two later rules, an alternative must be assumed, and effect size was set to .4.

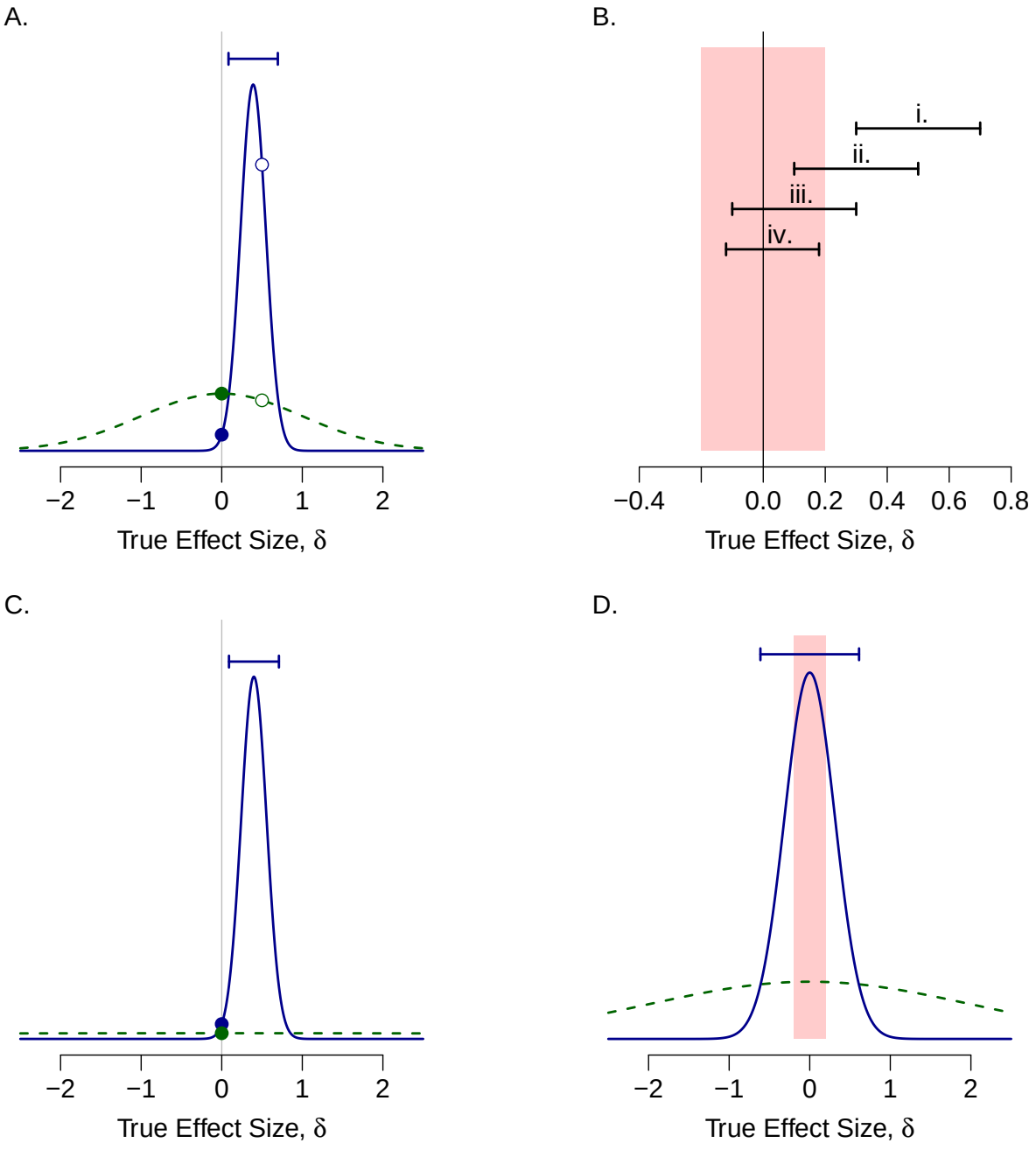
Figure 2. Bayesian inference by credible intervals and Bayes rule may not agree. **A.** Prior and posterior distributions on true effect size. The posterior reflects updating when the observed effect size is .4 with 40 observations. The credible interval excludes zero, and indeed, Bayes rule reveals that the plausibility that $\delta = 0$ is decreased (filled points) and the plausibility of $\delta = .5$ is increased (open points). **B.** Inference by credible intervals. Intervals i. and ii. show support for an effect with interval i. showing more support than interval ii. Intervals iii and iv do not provide support for an effect, and interval iv. is fully contained in the *region of posterior equivalence* (ROPE) indicating support for the null. **C.** With broad priors, inference by credible intervals exclude zero but Bayes rule shows increased plausibility from the data for this null value. **D.**

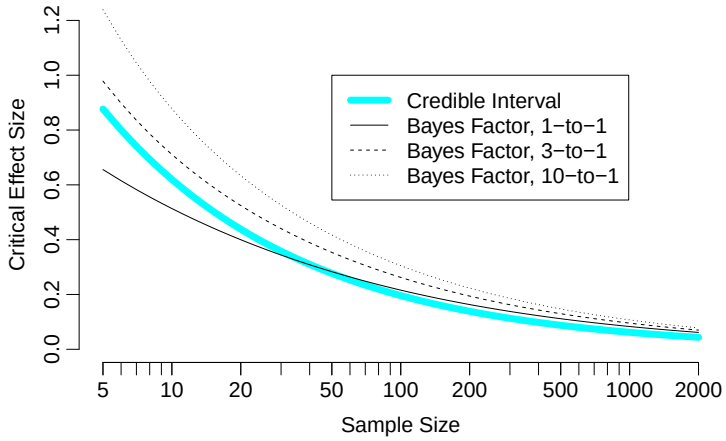
Figure 3. Critical effect sizes needed to state evidence for an effect. The wide, light-color line shows the effect size such that the 95% credible interval excludes zero. The solid, dashed and dotted lines show critical effect sizes needed to maintain a 1:1, 3:1 and 10:1 Bayes factors in favor of the alternative, respectively. For moderate sized samples, inference by credible leads to the statement of effects at even when the evidence favors the null. The priors used in computing the Bayes factors are that the effect size is from a standard normal, which is also called the unit-information prior and underlies BIC.

Figure 4. Alternatives may be specified on effect-size measures. **Left.** Four possible

specifications. These specifications share the common property of being symmetric around zero and specifying that smaller effect sizes are more common than larger ones. The difference between them is the scale, denoted σ_δ. Scales of 10 and .02, top left and bottom right panels, respectively, are too extreme to be useful. Scales of 1 and .2, top right and bottom left panels respectively, reflect limits of reasonable specifications in most contexts. **Right.** Bayes factor as a function of prior specification (r) shows that while the specification matters as it must, the changes across reasonable specifications are not great, especially when compared to the sensitivity of the Bayes factor on the data (t-values).


Critical Effect Size
0.0
0.2
0.4
0.6
0.8
1.0
5
10
20
50
100
200
500
1000
Sample Size
α = 0.05
α = min (0.05, β)
α = β/5


A.
−2
−1
0
1
2
True Effect Size, δ
B.
i.
ii.
iii.
iv.
−0.4
0.0
0.2
0.4
0.6
0.8
True Effect Size, δ
C.
−2
−1
0
1
2
True Effect Size, δ
D.
−2
−1
0
1
2
True Effect Size, δ


Critical Effect Size
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Credible Interval
Bayes Factor, 1−to−1
Bayes Factor, 3−to−1
Bayes Factor, 10−to−1
5
10
20
50
100
200
500
1000
2000
Sample Size

No Free Lunch, Figure 4