Invited Forum: Providing Evidence Against the Presence of a (Meaningful) Effect
*Tutorial*

Advances in Methods and
Practices in Psychological Science
2018, Vol. 1(2) 270–280

DOI: 10.1177/2515245918771304
www.psychologicalscience.org/AMPPS

# Rejecting or Accepting Parameter Values in Bayesian Estimation

**John K. Kruschke**
Department of Psychological and Brain Sciences, Indiana University

**Abstract**
This article explains a decision rule that uses Bayesian posterior distributions as the basis for accepting or rejecting null values of parameters. This decision rule focuses on the range of plausible values indicated by the highest density interval of the posterior distribution and the relation between this range and a region of practical equivalence (ROPE) around the null value. The article also discusses considerations for setting the limits of a ROPE and emphasizes that analogous considerations apply to setting the decision thresholds for *p* values and Bayes factors.

**Keywords**
Bayesian, credible interval, Bayes factor, equivalence testing, hypothesis testing, meta-analysis, open materials

Received 12/1/17; Revision accepted 3/16/18

In everyday life and in science, people often gather data to estimate a value precisely enough to take action. We use sensory data to decide that a fruit is ripe enough to be tasty but not overripe—that the ripeness is "just right" (e.g., Kappel, Fisher-Fleming, & Hogue, 1995, 1996). Scientists measured the position of the planet Mercury (among other things) until the estimate of the parameter $\gamma$ in competing theories of gravity was sufficiently close to 1.0 to accept general relativity for applied purposes (e.g., Will, 2014).

These examples illustrate a method for decision making that I formalize in this article. This method, which is based on Bayesian estimation of parameters, uses two key ingredients. The first ingredient is a summary of certainty about the measurement. Because data are noisy, a larger set of data provides greater certainty about the estimated value of measurement. Certainty is expressed by a confidence interval in frequentist statistics and by a *highest density interval* (HDI) in Bayesian statistics. The HDI summarizes the range of most credible values of a measurement. The second key ingredient in the decision method is a range of parameter values that is good enough for practical purposes. This range is called the *region of practical equivalence* (ROPE). The decision rule, which I refer to as the HDI+ROPE decision rule, is intuitively straightforward: If the entire HDI—that is, all the most credible values—falls within the ROPE, then accept the target value for practical purposes. If the entire HDI falls outside the ROPE, then reject the target value. Otherwise, withhold a decision.

In this article, I explain the HDI+ROPE decision rule and provide examples. I then discuss considerations for setting the limits of a ROPE and explain that similar considerations apply to setting the decision thresholds for *p* values and Bayes factors.

## Disclosures

Files available at the Open Science Framework (OSF; https://osf.io/jwd3t/) provide complete R code for the two-group example in Figure 2. This code can be trivially modified for other sets of two-group data. The Supplement file available at the same URL discusses the following topics:

- ROPE limits for regression coefficients in logistic regression
- Highest-density intervals versus equal-tailed intervals

**Corresponding Author:**
John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington, IN 47405-7007
E-mail: johnkruschke@gmail.com

- Decision-theoretic properties of the HDI+ROPE decision rule, including its asymptotic consistency and a loss function for which the decision procedure may be a Bayes rule
- A decision rule based on the ROPE without the HDI
- Comparison of the HDI+ROPE decision rule with frequentist equivalence testing, null-hypothesis significance testing (NHST), and Bayes factors
- Application of the HDI+ROPE decision rule to meta-analysis and comparison with meta-analysis using Bayes factors

## Bayesian Parameter Estimation

Bayesian inference is merely reallocation of credibility across possibilities, according to the mathematics of conditional probability. In formal data analysis, the possibilities are parameter values in a model of the data. For example, suppose we are measuring the systolic blood pressure (in units of millimeters mercury) of a group of people who have been exposed to a stressor. We may choose to describe the set of blood pressures with a normal distribution, which has two parameters: the location parameter, $\mu$, which characterizes the central tendency, and the scale parameter, $\sigma$, which characterizes the variability across people. We start with a prior distribution, a reasonable probability distribution over possible values of the parameters. Note that the prior distribution is a joint distribution over the space of $(\mu,\sigma)$ parameter value combinations. (The prior distribution is not a distribution over data, nor is the prior distribution a sampling distribution of test statistics.) After measuring the people's blood pressures, we reallocate probability to values of $\mu$ and $\sigma$ that are consistent with the observed measurements. The result is a posterior probability distribution over the joint space of all possible combinations of the parameter values, $(\mu,\sigma)$. *Bayesian inference* computes the reallocation using a simple formula called Bayes rule, named after Thomas Bayes (Bayes & Price, 1763). (For nontechnical introductions to Bayesian data analysis, see Kruschke & Liddell, 2018a, 2018b; for an accessible book-length tutorial, see Kruschke 2015).

### *Probability distribution over parameter values*

There is uncertainty about the parameter values because many parameter values are reasonably consistent with whatever data we may have. In a Bayesian framework, the uncertainty in parameter values is represented as a probability distribution over the space of parameter values. Parameter values that are more consistent with the data have higher probability than parameter values that are less consistent with the data. If we are uncertain about the parameter values, perhaps because we have very few data, then the probability distribution over the parameter space is spread out. With more data, the distribution becomes more peaked over a narrower range of values, reflecting our increased certainty in the estimate.

### *The HDI*

In the case of continuous parameters, the height of the distribution at a given value is called the *probability density* for that value (for discrete-valued parameters, the term *probability mass* is used). The width of the parameter distribution indicates our uncertainty in the parameter value. A useful summary of the width is the *95% HDI*. Any parameter value inside the HDI has higher probability density than any value outside the HDI, and the total probability of values in the 95% HDI is 95%. Parameter values with higher density are interpreted as more credible than parameter values with lower density. Therefore, we can describe values inside the 95% HDI as "the 95% most credible values of the parameter." (For further discussion, see the section titled Equal-Tailed Intervals Vs. Highest-Density Intervals in the Supplement file at the OSF, https://osf.io/jwd3t/).

## Making a Decision Based on the Relation Between the HDI and ROPE

Discrete decisions should be avoided if possible, because such decisions encourage people to ignore the magnitude of the parameter value and its uncertainty (e.g., Cumming, 2014; Kruschke & Liddell, 2018b; Wasserstein & Lazar, 2016, and many references cited therein). Such black-and-white thinking leads to misinterpretation and confusion. Despite this admonition against black-and-white thinking, there may be some situations in which an analyst needs to make a discrete decision about a parameter value such as a null value. In medical applications, for example, decisions to recommend a treatment or not must be made.

### *The ROPE*

There are many possible decision rules, but here I focus on one that requires the analyst to consider whether all the most credible parameter values are sufficiently far away from the null value that the null value can be rejected, or whether all the most credible parameter values are sufficiently close to the null value that the null value can be accepted. This

decision rule is made concrete by defining proximity to the null value using the ROPE, which specifies the range of parameter values that are equivalent to the null value for practical purposes. The notion of the ROPE appears in the literature under many different names, such as indifference zone, range of equivalence, equivalence margin, margin of noninferiority, smallest effect size of interest, and good-enough belt (e.g., Carlin & Louis, 2009; Freedman, Lowe, & Macaskill, 1984; Hobbs & Carlin, 2008; Lakens, 2014, 2017; Serlin & Lapsley, 1985, 1993; Spiegelhalter, Freedman, & Parmar, 1994).

### *The HDI+ROPE decision rule*

Consider a ROPE around a null value of a parameter. If the 95% HDI of the parameter distribution falls completely outside the ROPE, then one should reject the null value, because the 95% most credible values of the parameter are all not practically equivalent to the null value. If the 95% HDI of the parameter distribution falls completely inside the ROPE, then one should accept the null value for practical purposes, because the 95% most credible values of the parameter are all practically equivalent to the null value. If the 95% HDI is neither completely outside nor completely inside the ROPE, then one should remain undecided, because some of the most credible values are practically equivalent to the null but others are not. This HDI+ROPE decision rule has been described in several previous publications (Kruschke, 2010, 2011a, 2011b, 2013, 2015; Kruschke, Aguinis, & Joo, 2012; Kruschke & Liddell, 2018a, 2018b; Kruschke & Vanpaemel, 2015).

Figure 1 illustrates different relationships between an HDI and ROPE, and the decisions to which they lead. Figure 1a shows a case in which the HDI falls completely outside the ROPE, and therefore the null value is rejected because all the most credible values

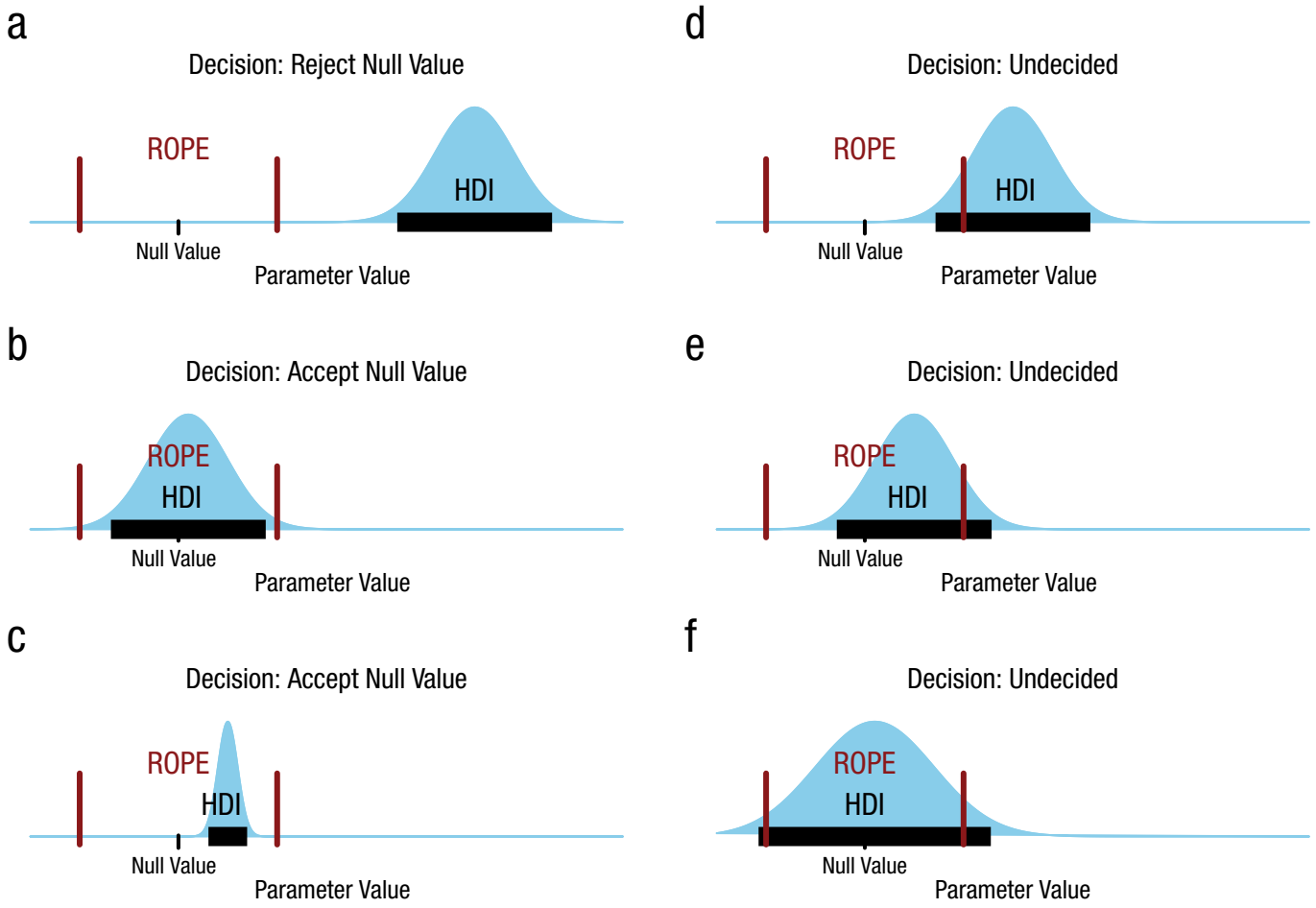


**Fig. 1.** Examples of different relationships between a highest density interval (HDI) and the region of practical equivalence (ROPE), and the decisions to which they lead. In each panel, the unmarked vertical axis is probability density, the HDI is marked by the horizontal bar, and the ROPE limits are marked by the two vertical bars.

are not practically equivalent to the null. Figure 1b shows a case in which the HDI falls completely inside the ROPE, and therefore the null value is accepted for practical purposes because all the most credible values are practically equivalent to the null.

Figure 1c also shows a case in which the null value is accepted for practical purposes—here, despite the fact that the null value is not itself within the HDI. This case is important because it contrasts the meaning of the HDI (from Bayesian inference) with the meaning of the ROPE (from decision making). Accepting the null value for practical purposes does *not* mean that the null value is among the most credible values of the parameter distribution. Accepting the null value for practical purposes means merely that all the most credible values are practically equivalent to the null value.

The remaining panels in Figure 1 show cases in which we should remain undecided. In all three panels, some of the HDI falls outside the ROPE and some of the HDI falls inside the ROPE. Notice that we do not reject the null value in the situation depicted in Figure 1d despite the fact that the null value falls outside the HDI, because some of the HDI is practically equivalent to the null. This decision contrasts with the analogous situation in NHST: A null value is rejected if it falls outside the 95% confidence interval. We do not accept the null value in the case of Figure 1e despite the fact that the null value falls within the HDI, because some of the HDI is not practically equivalent to the null. We do not accept the null value in the case of Figure 1f despite the fact that the HDI spans the ROPE, because some of the most credible values are not equivalent to the null value.

Notice that accepting a landmark parameter value, as in the situations illustrated in Figures 1b and 1c, is not the same thing as treating the accepted value as the best estimate of the parameter value. On the contrary, the Bayesian posterior distribution indicates the estimate of the parameter value, and typically the most probable (modal) parameter value is treated as the best estimate of the parameter value. When we accept a landmark parameter value, we are merely saying that the estimate of the parameter value is close enough to the landmark value, with high enough precision, that we can treat the landmark value as good enough for practical purposes. In other words, accepting a landmark parameter value means that the best estimate of the parameter value is *practically equivalent to* the landmark value, not that the best estimate of the parameter value *is* the landmark value.

The Supplement file at the OSF (https://osf.io/jwd3t/) describes some decision-theoretic properties of the HDI+ROPE decision rule.

## *Numerical example*

I illustrate this decision rule by applying it to a comparison of two groups for whom we have metric data (as opposed to ordinal or categorical data). Suppose the data are IQ scores from participants who have been given a placebo and participants who have been given a drug intended to make them smarter. The data within each group might have outliers, so we describe the groups with distributions that have optionally heavy tails (namely, mathematical *t* distributions). The model therefore has central-tendency parameters for the two groups, denoted $\mu_1$ and $\mu_2$; scale parameters for the two groups, denoted $\sigma_1$ and $\sigma_2$; and a normality parameter, denoted $\nu$, that has large values for nearly normal distributions and small values for heavy-tailed distributions. The analysis begins with a broad prior distribution on the joint space of these five parameters. The broad prior is designed to have minimal influence on the form of the posterior distribution (see Kruschke, 2013, for complete details).

The data for this example were created as random numbers from normal distributions, and the sample sizes were arbitrary. The data are represented by the histograms in the upper right panels of Figure 2. The other panels of Figure 2 show aspects of the five-dimensional posterior distribution; that is, they show different perspectives of the single joint distribution. The parameter distributions were derived with Markov chain Monte Carlo methods (MCMC; see chap. 7 of Kruschke, 2015) and computed using the JAGS software (Plummer, 2003, 2017) with the runjags package in R (Denwood, 2016). HDI limits were computed from the MCMC chain using the method explained in Section 25.2.3 of Kruschke (2015) and with an effective sample size that exceeded 10,000, as recommended in Section 7.5.2 of Kruschke (2015). Complete computer code for this example is available at the OSF (https://osf.io/jwd3t/).

In this application to two groups, it is natural to want to know the typical IQ score in each group (i.e., the magnitudes of $\mu_1$ and $\mu_2$), the spread of scores in each group (i.e., the magnitudes of $\sigma_1$ and $\sigma_2$), the difference in magnitude and spread between the two groups, and the uncertainty of all those estimates. We are interested in the magnitude of the difference between the means because that indicates how much IQ scores have been shifted by the smart drug, on average. We are interested in the magnitude of the difference between the spreads because that indicates how much the consistency of the scores has been affected by the smart drug. It is known, for example, that stressors can increase variability across people, as some people improve in response to a stressor whereas others decline (e.g.,

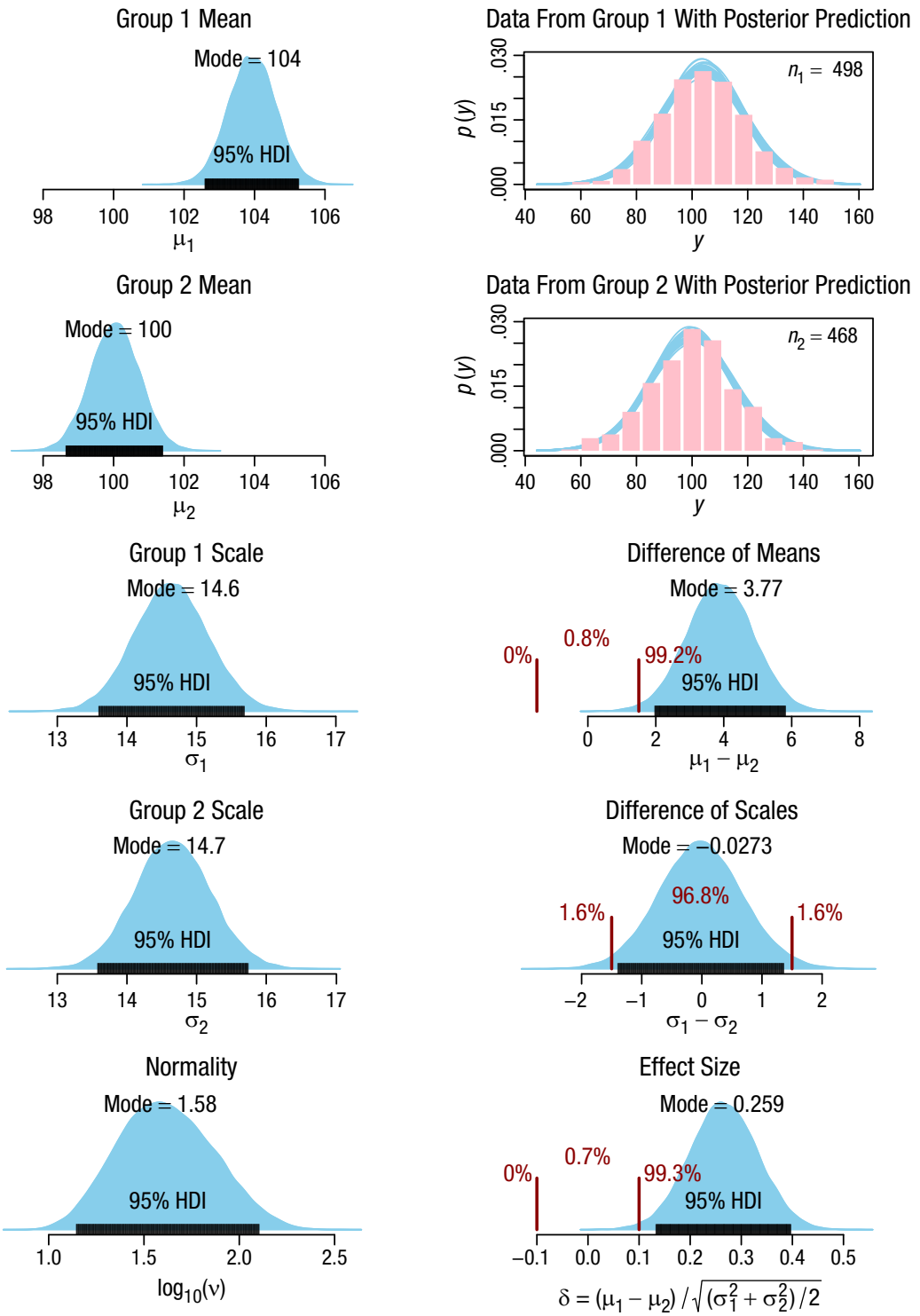


**Fig. 2.** Applying the decision rule to compare two groups. The data from the groups (i.e., the IQ scores, denoted here by the generic label *y*) are shown as histograms in the two upper panels of the right column. Superimposed on the data histograms are *t* distributions predicted from the posterior distribution. The left column shows the (marginal) posterior distributions of the individual parameters in the model. The lower three panels of the right column show aspects of the posterior distribution with regions of practical equivalence (ROPEs), delimited by the vertical bars (see the main text for how the ROPE limits were selected). These panels show the percentages of the posterior distributions below the low limit of the ROPE, within the ROPE, and above the high limit of the ROPE. The 95% highest density intervals (HDIs) are indicated by the black horizontal bars. In this example, the ROPE+HDI decision rule rejects a mean difference of zero because the 95% HDI falls completely outside the ROPE, accepts a scale difference of zero because the 95% HDI falls completely inside the ROPE, and rejects an effect size of zero because the 95% HDI falls completely outside the ROPE.

Lazarus & Eriksen, 1952). And of course, we are interested in the uncertainty of those estimates, so that we know how much confidence to place on their values.

The lower three panels of the right column in Figure 2 show, respectively, the posterior distribution of the difference between the means (i.e., $\mu_1 - \mu_2$), the posterior distribution of the difference between the scales (i.e., $\sigma_1 - \sigma_2$), and the posterior distribution of the effect size (i.e., the standardized difference between the means, calculated as $\delta = (\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$; Cohen, 1988).

We can establish ROPEs on the parameters (or combinations of parameters) to make decisions. I discuss methods for setting ROPE limits later in this article, but here, for purposes of illustration, I set a ROPE on the effect size at half of Cohen's conventional definition of a small effect, that is, at $\delta = \pm 0.1$. To establish a ROPE on the difference between the means, one option is to start with the same convention as for $\delta$ and translate it to an analogous value on $\mu_1 - \mu_2$. Thus, if we assume that the standardized population value of $\sigma$ is 15.0, then we can calculate the corresponding ROPE as follows: $\mu_1 - \mu_2 = \pm 0.1 \times 15 = \pm 1.5$. A different option is to derive the ROPE from a real-world consideration, such as a change in mean IQ that would imply a negligible change in gross domestic product (GDP) per capita. Rindermann and Thompson (2011, p. 761) reported that a change of 1 IQ point in the population mean predicts a change of $229 in GDP per capita. If we suppose that a $687 change is practically equivalent to zero, then the ROPE would again have a width of 3 IQ points (i.e., ±1.5). Finally, for the ROPE on the difference between scales, I again used a half-width of 1.5, because $\sigma$ is on the same scale as $\mu$. This setting is merely a fallback position in the absence of specific knowledge about the utility of changes in variability, as distinct from changes in central tendency. These ROPEs are indicated in Figure 2. We decide to reject a zero difference between the means because the 95% most credible values are outside the ROPE. We also decide to reject a zero effect size. But we decide to accept a zero difference between the scales ($\sigma$s) because the 95% most credible values of the difference are all practically equivalent to zero.

In summary, the posterior distribution is a multidimensional distribution on the joint parameter space, and various parameters (and combinations of parameters) can be compared simultaneously with relevant ROPEs. It is important to keep in mind that the full posterior distribution is the information delivered by Bayesian analysis, as summarized by the mode and 95% HDI of the distribution. The discrete decisions using ROPEs are secondary conclusions. Notice that to make these decisions using the HDI+ROPE rule, we must explicitly consider the magnitudes and uncertainties of the parameters; in contrast, *p* values and Bayes factors do not indicate the magnitudes and uncertainties of the parameters.

## More About the ROPE

### *The ROPE in theory testing*

The concept of the ROPE is useful for implementing a solution to a paradox from Meehl (1967, 1997). Theories pursued by NHST posit merely any nonnull effect and are therefore confirmed merely by rejecting the null value of the parameter, regardless of the actual magnitude of the parameter. Assuming that most variables of interest have some small but nonzero correlation with any other variable of interest, the correlation will be detected if the data set is large enough, and then the anything-but-null theory will be confirmed. Thus, anything-but-null theories incur a methodological paradox: Such theories become easier to confirm with larger sample sizes, rather than easier to disconfirm, and this is not the way scientific theories are supposed to work. By contrast, quantitatively predictive theories become easier to disconfirm with larger sample sizes because reality will almost always be somewhat discrepant from any quantitatively specific prediction. For example, the specific quantitative predictions of the Newtonian theory of gravity were disconfirmed by precise measurements of the orbit of the planet Mercury (e.g., Schiff, 1960; Will, 2014).

But how can quantitatively predictive hypotheses be confirmed? Serlin and Lapsley (1985, 1993) explained that a decision to confirm a quantitative prediction requires a ROPE (what they called a "good-enough belt") around the predicted value. If the observed value is within the ROPE, the hypothesis is confirmed for the current practical purposes. The ROPE is a decision boundary that reflects the precision needed to distinguish current theories. If two theories make very similar predictions, then a narrow ROPE is needed to distinguish them. If two theories make rather different predictions, then a wider ROPE can be used. The ROPE also should take into account the practical meaning of the magnitude of discrepancy. In this way, when an observed value of a parameter falls within the ROPE of the predicted value, the prediction is said to be confirmed for current practical purposes.

### *The ROPE in equivalence testing and noninferiority testing*

The concept of the ROPE is essential to frequentist *equivalence testing* (e.g., Lakens, 2017). In equivalence

testing, the analyst specifies a ROPE around the null value and decides that the estimated parameter is statistically equivalent to the null value if the confidence interval falls entirely within the ROPE (e.g., Westlake, 1976, 1981). This decision rule follows naturally from the meanings of the confidence interval and ROPE: The confidence interval is the range of parameter values that are not rejected (e.g., Cox, 2006), and if all the unrejected values fall within the ROPE, then they are all practically equivalent to the null value.[1]

The notion of the ROPE is also central to *noninferiority testing* (e.g., Lesaffre, 2008; Wiens, 2002), although only the low end of the ROPE is emphasized. In noninferiority testing, the analyst specifies a value below the null value that represents the largest decrease from the null value that is, nevertheless, negligible for practical purposes. The estimated value of the parameter is declared to be noninferior to the null value if that estimated value is significantly above the low end of the ROPE.

## *Specifying ROPE limits*

How does one specify the limits of a ROPE? Because the ROPE is a decision threshold that captures practical equivalence, its limits are influenced by practical considerations, which might change through time as risks are reassessed and as theories are refined. Any decision rule must be calibrated to be useful to the audience of the analysis and to the people who are affected by the decision, and this is also true of decision rules based on $p$ values and Bayes factors.

Equivalence testing has been used extensively in medical research, and the U.S. Food and Drug Administration (FDA) has set guidelines for the decision boundaries in equivalence testing (e.g., U.S. FDA, Center for Drug Evaluation and Research, 2001; U.S. FDA, Center for Veterinary Medicine, 2016). Recent FDA guidance for bioequivalence studies recommends ROPE limits of 0.8 and 1.25 for the ratio of means in the two groups (U.S. FDA, Center for Veterinary Medicine, 2016, p. 16). Contemporary industry standards use ROPE limits around ±20% for applications with moderate risk, but the ROPE may be narrower (i.e., ±5% to ±10%) when the risks are high, or the ROPE may be wider (i.e., ±26% to ±50%) when the risks are low (Little, 2015, Table 1).

Standards for the decision boundary of noninferiority testing have also been established by the FDA, and their recent guidance emphasizes that great care must be taken to establish the noninferiority limit because of the tremendous real-world costs and benefits of drugs and therapies (U.S. FDA, Center for Drug Evaluation and Center for Biologics Evaluation and Research, 2016). Walker and Nowacki (2011) explained that one conventional setting of the noninferiority limit is at half of "the lower limit of a confidence interval of the difference between the current therapy and the placebo obtained from a metaanalysis" (p. 194).

In many fields of science, competing theories make detailed quantitative predictions. For example, a parameter called $\gamma$ should be exactly 1.0 in the theory of general relativity, but 0 in Newtonian gravity and other values near 1 in other theories (see Will, 2014, Fig. 5, p. 43, for a summary of the progression of 90 years of experiments measuring $\gamma$). A recent experiment established a value of $1 \pm 0.00001$ (Bertotti, Iess, & Tortora, 2003). This experiment does not merely reject Newtonian gravity ($\gamma = 0$), but confirms general relativity ($\gamma = 1.0$) even if one is using very narrow ROPEs.

In the social sciences, Cohen (1988) defined measures of effect size for different sorts of parameters and proposed conventional values for small, medium, and large effects typically observed in social-science research. In the case of the effect size of a mean, defined as $\delta = (\mu - \mu_0)/\sigma$, Cohen suggested that 0.2 is a "small" effect, and therefore we might say that an effect is practically equivalent to zero if it is less than, say, half the size of a small effect and falls within a ROPE of ±0.1. This conventional limit was used for Figure 2.

It must be emphasized that "half the size of a small effect" is merely a fallback convention when there is no way to calibrate effects by their real-world consequences. In the case of IQ points, for instance, there might be applications for which a 0.1 effect implies nonnegligible practical consequences. A study of the GDP of 90 nations as a function of IQ and other variables found that "an increase of 1 IQ point in the intellectual class [the IQ at the 95th percentile] raises the average GDP [per capita] by $468 U.S." (Rindermann & Thompson, 2011, p. 761). (The influence of IQ is weaker at the mean than at the 95th percentile, as mentioned earlier in the context of Fig. 2.) Thus, an increase of average IQ of the intellectual class from 130 to 131, for example, might have important consequences for GDP because that increase is multiplied across millions of people, even though an increase of 1 IQ point in any one person may be negligible for that person.

A different approach to setting the limits of a ROPE was described by Lakens (2017, p. 359), who pointed out that the maximum sample size a researcher is willing to collect data from implies, for any specific desired power, the minimal effect size that can be reliably detected. Implicitly, the sample size indicates the minimal effect size that the researcher is willing to treat as not practically equivalent to zero. This minimal effect

size, in turn, implies corresponding ROPE limits for an equivalence test. I think, however, that this approach will yield ROPEs that are too wide when sample sizes are small (e.g., when research is underpowered; Maxwell, 2004) and will yield ROPEs that are too narrow when sample sizes are large (e.g., with "big data"; Adjerid & Kelley, 2018). ROPEs should be set according to the demands of competing theories and the practical implications of decisions, not by the measurement precision implied by sample size. Falling objects do not hit the ground more softly if they are measured with less precise instruments. A new drug is not more equivalent to an existing drug if it is tested for equivalence using a smaller sample size. Moreover, there are often moderate-*N* studies (which individually yield only moderate-precision estimates) that are worth doing even when the ROPE is relatively narrow, because future meta-analyses of multiple moderate-*N* studies may find a narrow meta-analytic HDI. Indeed, for random-effects models in meta-analyses, usually greater precision can be achieved by many moderate-*N* studies than by a few large-*N* studies, because hierarchical shrinkage of estimated parameter values operates more effectively (e.g., Kruschke & Liddell, 2018b; Kruschke & Vanpaemel, 2015). In meta-analyses, there is no foreknowledge of which studies will be uncovered for inclusion (from database searches of published studies and social-network searches of unpublished studies), so an analyst cannot anticipate the samples sizes or the number of studies. The ROPE must be defined from other considerations.

For parameters that have the same scale as the data, it is relatively straightforward to think about a ROPE. For example, in the case of IQ scores with a normal distribution, the mean, μ, is on the IQ scale, and its ROPE limits are in IQ points. Other models may have parameters that are less directly related to the scales of the data, and therefore ROPE limits may need to be derived more indirectly. Consider linear regression. We might want to say that a regression coefficient, $\beta_x$, is practically equivalent to zero if a change across the "main range of $x$" produces only a negligible change in the predicted value, $\hat{y}$. Suppose we specify a negligible change in $\hat{y}$ as $\pm 0.1 S_y$, where $S_y$ is the standard deviation of $y$ (a range that may be motivated by the convention that $0.1S$ is half of a "small" effect), and we specify the "main range of $x$" as $M_x \pm 2S_x$ (because if $x$ were normally distributed, this range would cover just over 95% of the distribution). Given these specifications, a regression coefficient is practically equivalent to zero when a change of $x$ from $M_x - 2S_x$ to $M_x + 2S_x$ yields a change of $\hat{y}$ only from $M_y - 0.1S_y$ to $M_y + 0.1S_y$, which implies ROPE limits of $\beta_x = \pm 0.05$ for standardized variables. Similar considerations apply to logistic regression, as explained in the Supplement file at the OSF (https://osf.io/jwd3t/).

## *ROPE limits are like decision thresholds for* p *values and Bayes factors*

In general, ROPE limits are defined by considering what counts as practically equivalent to the null value, by quantifying acceptable uncertainty as constrained by competing theories or real-world utilities. It can be challenging to specify a definitive ROPE, but one should not delude oneself into thinking that it is any more straightforward to specify a definitive decision threshold for a $p$ value. Some people have grown comfortable with .05 as the decision threshold for a $p$ value because it is a conventional value that statistical rituals are designed to comply with. But the convention hides the fact that there is vigorous debate about an appropriate decision threshold for $p$. In a recent article, Benjamin et al. (2018) argued that the threshold $p$ value for the social sciences should be changed to .005. In physics, the contemporary conventional threshold $p$ value corresponds to $5\sigma$, which requires $p < .00000029$ for significance. Decision thresholds for $p$ values are on no firmer ground than ROPE limits.

Bayesian null-hypothesis testing involves a decision statistic called the Bayes factor (BF). The specification of decision thresholds for BFs is as fraught as the specification of ROPEs and decision thresholds for $p$ values. Jeffreys (1961) attached decision-strength labels to ranges of BFs as follows: 3.16 through 10.0 is "substantial," greater than 10.0 through 31.6 is "strong," greater than 31.6 through 100.0 is "very strong," and greater than 100.0 is "decisive." A subsequent influential article by Kass and Raftery (1995) suggested that BFs of 3.0 through 20.0 are "positive" evidence, BFs greater than 20.0 through 150.0 are "strong" evidence, and BFs greater than 150.0 are "very strong" evidence. In the psychological sciences, many proponents of BFs have routinely used 3 as the decision threshold (e.g., Dienes, 2016). On the other hand, Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017) recommended a BF of 10 for mature confirmatory research but other limits for nascent research, and those authors also pointed out that different BF thresholds may apply to different types of hypothesis tests. Rouder, Morey, and Province (2013) emphasized that an extremely large BF is needed to reject null hypotheses that have a large prior probability, such as the null hypothesis that people cannot foretell the future through temporally reversed causality. Again, do not be lulled into thinking that establishing a decision threshold for Bayes factors is any easier than establishing ROPEs for HDIs.

Regardless of the decision statistic being used ($p$ value, BF, or HDI), decision thresholds should ultimately take into account the utilities (i.e., costs and benefits) of the decisions. Unfortunately, the utilities are often unavailable. Regardless of the availability of utilities, the decision criteria should be established before the data are observed, to prevent biased decisions (e.g., Lakens et al., 2017).

## Conclusion

Deciding to accept or reject a null value is dangerous, as it engenders fallacious black-and-white thinking. But when it is necessary to make such a decision, the fallacy might be fended off by focusing on explicit estimates of parameter magnitude and uncertainty. The HDI+ROPE decision method does exactly that: The analyst explicitly examines the probability distribution over parameter values and considers the relationship between the most credible parameter values and a region of practical equivalence to the null value. On the other hand, $p$ values and BFs hide the parameter's magnitude and uncertainty, which makes it easier to slip into specious black-and-white thinking. Setting the limits of a ROPE is no more difficult in principle than setting the decision threshold for a $p$ value or for a BF, so researchers should be no more uncomfortable setting a ROPE than setting these other decision thresholds.

### Action Editor

Daniel J. Simons served as action editor for this article.

### Author Contributions

J. K. Kruschke is the sole author of this article and is responsible for its content.

### Acknowledgments

For comments on an early version of this article, the author gratefully acknowledges Brad Celestin and Torrin Liddell. In review and production, constructive comments were provided by Rogier Kievit, two anonymous reviewers, and Michele Nathan.

### Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

### Open Practices

The R code for the example in Figure 2 has been made publicly available via the Open Science Framework and can be accessed at https://osf.io/jwd3t/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245918771304. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

### Note

1. The equivalence-testing procedure, involving a confidence interval and ROPE, is mathematically equivalent to the method of *two one-sided tests* (TOST; Schuirmann, 1987). With TOST, the analyst checks whether the estimated parameter is significantly below the high end of the ROPE and significantly above the low end of the ROPE. If both directional tests are passed, the analyst concludes that the parameter is statistically equivalent to the null value. Because these tests are one sided, using $1 - \alpha$ tests will achieve the same Type I error rate as using a $1 - 2\alpha$ confidence interval in equivalence tests.

### References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*. Advance online publication. doi:10.1037/amp0000190

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, *53*, 370–418. doi:10.1098/rstl.1763.0053

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E (2018). Redefine statistical significance. *Nature Human Behavior*, *2*, 6–10. doi:10.1038/s41562-017-0189-z

Bertotti, B., Iess, L., & Tortora, P. (2003). A test of general relativity using radio links with the Cassini spacecraft. *Nature*, *425*, 374–376. doi:10.1038/nature01997

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge, England: Cambridge University Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25. doi:10.18637/jss.v071.i09

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. doi:10.1016/j.jmp.2015.10.003

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

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

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Kappel, F., Fisher-Fleming, R., & Hogue, E. J. (1995). Ideal pear sensory attributes and fruit characteristics. *HortScience*, *30*, 988–993.

Kappel, F., Fisher-Fleming, R., & Hogue, E. J. (1996). Fruit characteristics and sensory attributes of an ideal sweet cherry. *HortScience*, *31*, 443–446.

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

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 658–676. doi:10.1002/wcs.72

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

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS* (1st ed.). Burlington, MA: Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*, 573–603. doi:10.1037/a0029146

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Burlington, MA: Academic Press.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*, 722–752. doi:10.1177/1094428112457829

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*, 155–177. doi:10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178–206. doi:10.3758/s13423-016-1221-4

Kruschke, J. K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In J. R. Busemeyer, J. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 279–299). Oxford, England: Oxford University Press.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701–710.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological & Personality Science*, *8*, 355–362.

Lakens, D., Adolfi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., . . . Zwaan, R. (2017). *Justify your alpha*. Retrieved from https://psyarxiv.com/9s3y6

Lazarus, R. S., & Eriksen, C. W. (1952). Effects of failure stress upon skilled performance. *Journal of Experimental Psychology*, *43*, 100–105. doi:10.1037/h0056614

Lesaffre, E. (2008). Superiority, equivalence, and noninferiority trials. *Bulletin of the NYU Hospital for Joint Diseases*, *66*, 150–154.

Little, T. A. (2015). Equivalence testing for comparability. *BioPharm International*, *28*(2), 45–48.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 395–425). Mahwah, NJ: Erlbaum.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Retrieved from https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Plummer, M. (2017). *JAGS Version 4.3.0 user manual*. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/jags user manual.pdf/download

Rindermann, H., & Thompson, J. (2011). Cognitive capitalism: The effect of cognitive ability on wealth, as mediated through scientific achievement and economic freedom. *Psychological Science*, *22*, 754–763.

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

Schiff, L. I. (1960). On experimental tests of the general theory of relativity. *American Journal of Physics*, *28*, 340–343. doi:10.1119/1.1935800

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657–680.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). *Rational appraisal of psychological research and the good-enough principle*. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Mahwah, NJ: Erlbaum.

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.

U.S. Food and Drug Administration, Center for Drug Evaluation and Research. (2001). *Guidance for industry: Statistical approaches to establishing bioequivalence*. Retrieved from https://www.fda.gov/downloads/drugs/guidances/ucm070244.pdf

U.S. Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation

and Research. (2016). *Non-inferiority clinical trials to establish effectiveness: Guidance for industry*. Retrieved from https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm202140.pdf

U.S. Food and Drug Administration, Center for Veterinary Medicine. (2016). *Guidance for industry: Bioequivalence: Blood level bioequivalence study VICH GL52*. Retrieved from https://www.fda.gov/downloads/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/UCM415697.pdf

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*, 192–196.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133. doi:10.1080/00031305.2016.1154108

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*, 741–744.

Westlake, W. J. (1981). Response to bioequivalence testing—a need to rethink. *Biometrics*, *37*, 591–593.

Wiens, B. L. (2002). Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled Clinical Trials*, *23*, 2–14.

Will, C. M. (2014). The confrontation between general relativity and experiment. *Living Reviews in Relativity*, *17*, Article 4. doi:10.12942/lrr-2014-4

Version of February 25, 2018.

# *Supplement to* Rejecting or accepting parameter values in Bayesian estimation

John K. Kruschke
Indiana University, Bloomington, USA

Contents

## Specifying ROPE limits for logistic regression

In the section titled "Specifying ROPE limits" in the main article, I described a method for defining conventional ROPE limits for linear-regression coefficients. Here the method is applied to logistic regression. The predicted variable $y$ is dichotomous (e.g., success/failure), and a regression coefficient refers to a change in the log-odds of the predicted outcome probabilities (e.g., Kruschke, 2015, Ch. 21). Formally, denote the predicted probability of "success" as $\hat{\pi}$ and the log-odds as $\text{logit}(\hat{\pi}) \equiv \log(\hat{\pi}/(1-\hat{\pi}))$. We start by defining a negligible change in $\hat{\pi}$. Toward that end, suppose we are conducting a political poll in a population with preferences for candidates A and B split approximately 50/50, and we are predicting preference as a function of income. Suppose we specify that a negligible change in $\hat{\pi}$ over the main range of $x$ is $\hat{\pi} = 0.50 \pm 0.03$. If the main range of $x$ is $M_x \pm 2S_x$ (as in the example of linear regression in the main article), then a negligible standardized regression coefficient has a ROPE of $|\beta_x| < (\text{logit}(0.53) - \text{logit}(0.47))/4 \approx 0.06$. Suppose instead we are studying occurrence of heart attack during a five-year window as predicted by the patient's average systolic blood pressure, and we specify that a negligible change in $\hat{\pi}$ over the main range of $x$ $\hat{\pi} = 0.020 \pm 0.002$. Then a negligible standardized regression coefficient has a ROPE of $|\beta_x| < (\text{logit}(0.022) - \text{logit}(0.018))/4 \approx 0.05$.

## Equal-tailed intervals vs highest-density intervals

As was explained in the main text, the highest-density interval treats the scale as a meaningful reference. If, on the other hand, the scale of a parameter can be arbitrarily non-linearly transformed (while preserving order) then density is not very meaningful because the relative densities change across non-linear transformations. In such a Dali-esque world, practitioners may use the 95% *equal-tailed interval* (ETI), which extends from the 2.5 percentile of the distribution to the 97.5 percentile. Differences between an HDI and an ETI are evident in skewed distributions. Examples with skewed distributions are shown in Figure 12.2, p. 342, of Kruschke (2015). In this article I have assumed that the parameter scale is a meaningful reference, and therefore used the 95% HDI as a summary of the 95% most credible parameter values. The only exception was the normality parameter in the lower-left panel of Figure 2, which used a logarithmic transformation merely to make the display more readable.

## Decision-theoretic properties of the HDI+ROPE procedure

A full development of any proposed decision rule would couch the rule in the formal framework of decision theory (e.g., Berger, 1985; Robert, 2007). Such a framing is beyond the intended scope of this article. Nevertheless, this section will provide a couple of suggestions regarding some decision-theoretic aspects of the HDI+ROPE rule.

*Consistency: HDI+ROPE decision is correct as sample size approaches infinity*

One desirable property of a decision rule is that it makes the correct decision as the sample size of the data approaches infinity. This property is called "consistency." The HDI+ROPE rule is consistent. To support this claim, recall that "accepting" a landmark value means that the parameter value is *practically equivalent to* the landmark value, not that the parameter value *is* the landmark value. Thus, if the true parameter value is anywhere inside the ROPE, then a correct decision is to "accept" the landmark value. The HDI+ROPE rule will converge to this correct decision as the sample size goes to infinity because the 95% HDI converges to the true value of the parameter. Moreover, according to the HDI+ROPE rule, "rejecting" a landmark value means that the parameter value is outside the ROPE. If the true parameter value is outside the ROPE, then the HDI+ROPE rule will converge to the decision of rejecting the landmark values because the 95% HDI converges to the true value of the parameter.

*Is there a loss function for which the Bayes rule is the HDI+ROPE procedure?*

The HDI+ROPE decision rule is motivated directly from the meanings of the intervals: Reject the null when all the most credible values are *not* practically equivalent to the null, and accept the null when all the most credible values *are* practically equivalent to the null. Here we seek a formal expression of that intuition. Essentially, we seek to define a *loss function* that captures what costs are avoided when making decisions that way. Technically, a *Bayes rule* is a decision rule that minimizes the loss when integrated (averaged) over the posterior distribution. In this section I will offer one suggestion for a loss function for which the Bayes rule may be the HDI+ROPE decision rule. My main goal is to show that the intuitive rule may have a formal expression, which some readers may find less unpalatable than a mere heuristic. A secondary goal is to elicit future development of related formalisms.

The loss will be a joint function of the decision made and the choice of how to summarize the parameter distribution. Define discrete variables for the decisions, $d_R, d_A \in \{0, 1\}$. The decision variables have values such that $d_R = 1$ indicates the null is rejected and $d_R = 0$ otherwise, and $d_A = 1$ indicates the null is accepted and $d_A = 0$ otherwise. Notice these two decision values are independent until the loss function implies that they are mutually exclusive. Also define continuous variables for the HDI limits on the parameter, $a, b \in \boldsymbol{\Theta}$. The continuous variables $a$ and $b$, with $a < b$, are meant to be the limits of the interval that summarizes the parameter distribution, which for an appropriate loss function

will be an HDI. A candidate loss function is

$$\begin{aligned} L(a,b,d_A,d_R) = & \underbrace{(b-a)}_{\substack{\text{HDI}\\ \text{width}}} + \underbrace{c_P \cdot \mathbf{1}(\theta \notin [a,b])}_{\substack{\text{cost of param. not}\\ \text{being in HDI}}} \\ & + \underbrace{c_R d_R \cdot \mathbf{1}([a,b] \cap [\theta_0 - r, \theta_0 + r])}_{\substack{\text{cost of reject if HDI}\\ \text{overlaps ROPE}}} + \underbrace{c_A d_A \cdot \mathbf{1}([a,b] \setminus [\theta_0 - r, \theta_0 + r])}_{\substack{\text{cost of accept if HDI}\\ \text{is not in ROPE}}} \\ & + \underbrace{c_N (1-d_R)(1-d_A)}_{\substack{\text{cost of}\\ \text{no decis.}}} \end{aligned} \tag{1}$$

where $c_P$, $c_R$, $c_A$, and $c_N$ are positive cost coefficients with $0 < c_N < c_R, c_A$, where $r$ is the radius (half-width) of the ROPE, and where the function $\mathbf{1}(s) = 1$ if $s$ is true, and $\mathbf{1}(s) = 0$ if $s$ is false. The first two terms of Equation 1 come from the loss function for the HDI limits as derived by Schervish (1995, pp. 328–329). The first two terms by themselves imply a Bayes rule for which $[a, b]$ is a HDI. The second two terms of Equation 1 capture costs of decisions, inspired by Rice, Lumley, and Szpiro (2008, Section 4.1), and the final term applies a cost for no decision (without which setting $d_R = 0$ and $d_A = 0$ would trivially minimize loss).

To illustrate the operation of the loss function, let us first presume that $[a, b]$ is the HDI. Suppose now that the HDI is completely inside the ROPE. Then $\mathbf{1}([a,b] \cap [\theta_0 - r, \theta_0 + r]) = 1$ and $\mathbf{1}([a,b] \setminus [\theta_0 - r, \theta_0 + r]) = 0$, and the last three terms of Equation 1 become $c_R d_R + c_N(1-d_R)(1-d_A)$, which is minimized (with loss zero) by setting $d_R = 0$ and $d_A = 1$, that is, deciding to accept. Suppose instead that the HDI is completely outside the ROPE. Then $\mathbf{1}([a,b] \cap [\theta_0 - r, \theta_0 + r]) = 0$ and $\mathbf{1}([a,b] \setminus [\theta_0 - r, \theta_0 + r]) = 1$, and the last three terms of Equation 1 become $c_A d_A + c_N(1-d_R)(1-d_A)$, which is minimized (with loss zero) by setting $d_R = 1$ and $d_A = 0$, that is, deciding to reject. Finally, suppose that the HDI overlaps the ROPE, with some of HDI inside of the ROPE and some outside. Then $\mathbf{1}([a,b] \cap [\theta_0 - r, \theta_0 + r]) = 1$ and $\mathbf{1}([a,b] \setminus [\theta_0 - r, \theta_0 + r]) = 1$, and the last three terms of Equation 1 become $c_R d_R + c_A d_A + c_N(1-d_R)(1-d_A)$, which is minimized (with loss $c_N$) by setting $d_R = 0$ and $d_A = 0$, that is, deciding to neither reject nor accept.

The loss function in Equation 1 formally expresses the intuitive costs for making decisions that are not consistent with the relation of HDI to ROPE. The specific form in Equation 1 is merely suggestive and there may be many other possible formal expressions of the intuition. The loss function of Equation 1 mixes units of interval width with units of (weighted) discrete decisions and consequently might lead to "paradoxical behavior" (Casella, Hwang, & Robert, 1993). It might be that the loss function could be re-expressed so that interval overlaps are measured in the same units as interval width (cf. Rice, 2011).

It is not my goal here to prove that the Bayes rule for the loss in Equation 1 produces

the intuitive HDI+ROPE decision rule. Indeed it might not be the Bayes rule if the loss-minimizing values of $a$ and $b$ are influenced by the decisions $d_R$ and $d_A$. My goal is to point out that formal expressions are possible for the loss implicit in the intuitive HDI+ROPE rule, because formally marking the costs can be valuable even if subsequent formalizations differ in the exact quantification of those costs.

## Decision rule based on ROPE alone

Some authors (e.g., Wellek, 2010) prefer to consider the proportion of the posterior distribution that falls within the ROPE as the statistic for decision making. For example, we might reject the null value if less than 5% (say) of the distribution falls within the ROPE, and we might accept the null value if more than 95% of the distribution falls within the ROPE. Notice that this rule ignores the probability density of parameter values inside or outside the ROPE.

One appeal of the ROPE-only rule is its invariance under arbitrary monotonic transformations of the parameter. For example, if we take the logarithm of $\mu$, then the distribution on $\log(\mu)$ has the same percentage within the logarithm-ROPE limits. But because the ROPE-only rule ignores probability density, it can lead to some counter-intuitive conclusions in some cases. For example, consider a broad prior distribution that has only a small percentage of the distribution inside the ROPE. According to the ROPE-only rule, the null value would be rejected already. But the HDI+ROPE rule would not reject the null value in the prior because the HDI overlaps the ROPE. There are other distributions that could arise that are strongly skewed. In such cases of skewed distributions it is possible for the ROPE to contain 95% of the distribution but for the HDI to extend beyond the ROPE. In these cases perhaps we should remain undecided because some relatively high-density parameter values remain outside the ROPE.

## Comparison with frequentist equivalence testing and NHST

In frequentist equivalence testing, a parameter value is deemed to be statistically equivalent to the null value if the parameter's 90% confidence interval (CI) falls entirely within the ROPE, and the null value is rejected when the 95% CI excludes the null value (i.e., standard NHST). As mentioned in a footnote in the main text, equivalence testing uses the $(1-2\alpha)$ CI because it amounts to two mutually exclusive one-sided tests (TOST), each of which has $\alpha$ in its tail. I will refer to this decision procedure as TOST+NHST.

TOST+NHST is analogous to HDI+ROPE in many respects. In particular, they both make decisions by considering a ROPE and its relation to an interval estimate of the parameter. In many applications their decisions will be the same. But because of their different motivations and meanings, there are cases in which their decisions will differ,

sometimes quite substantially. In this section I will point out those differences, but the reader should keep in mind that often the two techniques will make the same decisions.

The first difference between TOST+NHST and HDI+ROPE is the meanings of the intervals that define the parameter estimate. The frequentist 95% CI is the range of parameter values that would not be rejected by a two-tailed test with $p < .05$ (Cox, 2006). There is no probability distribution over parameter values, rather, a parameter value either is in the CI or is not. On the other hand, the Bayesian 95% HDI is the range of most credible parameter values that have 95% probability. There is inherently a probability distribution over parameter values because that's what defines the HDI. Thus, the meaning of a CI falling within the ROPE is different than the meaning of the HDI falling within the ROPE. The TOST indicates that all values outside the ROPE can be rejected, while the HDI+ROPE procedure indicates that the most credible parameter values fall inside the ROPE.

Another computational difference between a CI and an HDI is that HDI's are straight forward to compute from MCMC for all standard models, while CIs can be challenging to determine without specialized software. For example, in the simple situation of estimating the probability of "success" in a dichotomous outcome, computing a CI can be complicated (Dunnett & Gent, 1977). For more elaborate models (e.g., hierarchical logistic regression) CI's are usually only roughly approximated, typically too narrowly, in most software.

There can also be conflicting decisions within TOST+NHST, such that TOST declares that the parameter is "statistically equivalent" to the null while NHST simultaneously declares that the parameter is "significantly different" from the null. This happens when the CI is so narrow that it falls between the null value and a ROPE limit, as shown in Panel C of Figure 2 of the main article. This semantic conflict of being simultaneously equivalent and different occurs because TOST uses the ROPE limits to make a decision while NHST does not. It would make more sense, I think, if statistical difference from the null were declared only if the CI fell completely outside the ROPE, but that is not the conventional frequentist procedure. On the other hand, the HDI+ROPE procedure was specifically designed so that accepting the null and rejecting the null would never conflict, because both decisions use the same ROPE limits. For more details, see `http://doingbayesiandataanalysis.blogspot.com/2017/02/equivalence-testing-two-one-sided-test.html` or `https://osf.io/q686c/`. A closely related difference between NHST and HDI+ROPE is seen when the HDI or CI excludes the null value but still overlaps the ROPE, as in Panel D of Figure 2 of the main article. In this case, NHST would declare that the null value is rejected, but HDI+ROPE would remain undecided because some of the values inside the HDI are practically equivalent to the null value.

Importantly, both TOST and NHST can be greatly changed when there are multiple

tests (Rogers, Howard, & Vessey, 1993, p. 562). Because $p$ values increase as more tests are considered, the CI limits expand as the number of intended tests increases (e.g., Maxwell & Delaney, 2004). Consider, for example, a factorial analysis of variance (ANOVA) in which several groups are compared against each other (pairwise, or in complex comparisons). If we consider a single test of two groups, the CI on the difference of means might be narrow enough to fall inside a ROPE and to decide that the groups are statistically equivalent. But if the comparison of the two groups is part of a larger set of multiple tests, then the CI expands and the two groups might *not* be statistically equivalent. By contrast, HDIs are unchanged by different sets of intended tests because HDIs are not based on sampling distributions of hypothetical data but are instead based on only the actually observed data.

The key goal that TOST+NHST specifically attempts to achieve is control of error rates, which HDI+ROPE can not do directly. This difference between the procedures is the core difference of frequentist vs Bayesian approaches. Error rates can only be defined in terms of imaginary data that might have been generated by some hypothetical state, whereby sampling distributions are constructed. Bayesian inference does not consider error rates, but instead derives a probability distribution over parameter space conditional on the actually observed data. Error rates of Bayesian decision rules can be considered (e.g., Kruschke, 2015, Ch. 13) but the error rates are not the usual basis for Bayesian decision making.

## Comparison with Bayes factors

A different Bayesian approach to assessing null values is a special case of Bayesian model comparison. In general Bayesian model comparison, two or more models, with their own parameters and prior distributions, are set under a model-index parameter. The model-index parameter has value "1" to indicate that model 1 is consistent with the data, and value "2" to indicate that model 2 is consistent with the data, and so on for other models. This model-index parameter is analogous to any other parameter, such as the parameter $\mu$ in a normal model. The model-index parameter exists in a joint parameter space with all the other parameters in the individual models. According to Bayesian inference, credibility is re-allocated across all the parameters simultaneously. In particular, credibility is re-allocated across the model-index parameter values, such that the posterior probabilities of values 1, 2, etc., indicate the posterior probabilities of the models.

In Bayesian null hypothesis testing, the null hypothesis is expressed as a restricted model relative to the full model under consideration. For example, consider a situation where the full model is a normal distribution that has parameters $\mu$ and $\sigma$ with a broad prior distribution. A null hypothesis, that parameter $\mu$ has value 100.0, could be formalized by making the prior distribution on $\mu$ be a narrow "spike" around $\mu = 100.0$, while the

prior on $\sigma$ remains relatively broad. Then the null model and full model are placed under a model-index parameter and the entire hierarchy of parameters undergoes Bayesian inference. More extensive explanations of this framework are provided by Kruschke and Liddell (2017a, 2017b) and by Kruschke (2015).

A key statistic in Bayesian hypothesis testing is the *Bayes factor* (BF), which indicates how much the odds of the models have shifted from prior to posterior. Formally, if we rename the full model as the alternative hypothesis, "alt," then the BF for the null hypothesis is

$$\mathrm{BF}_{\mathrm{null}} = \frac{p(\mathrm{null}|D)}{p(\mathrm{alt}|D)} \Big/ \frac{p(\mathrm{null})}{p(\mathrm{alt})} \tag{2}$$

where $D$ is the observed data and $p(\mathrm{null}|D)$ indicates the posterior probability of the null hypothesis. We could instead define the BF with respect to the alternative hypothesis, which is simply the reciprocal: $\mathrm{BF}_{\mathrm{alt}} = 1/\mathrm{BF}_{\mathrm{null}}$. Notice that when credibility is re-allocated away from the alternative model toward the null model, then $\mathrm{BF}_{\mathrm{null}} > 1$ and $\mathrm{BF}_{\mathrm{alt}} < 1$. When credibility is re-allocated away from the null model, toward the alternative model, then $\mathrm{BF}_{\mathrm{null}} < 1$ and $\mathrm{BF}_{\mathrm{alt}} > 1$.

Importantly, notice that the BF does *not* indicate the posterior probabilities of the models. Instead, the BF indicates the *shift* in the model odds. If the prior probability of the null hypothesis is only $p(\mathrm{null}) = 0.01$, then $\mathrm{BF}_{\mathrm{null}} = 10$ indicates that the posterior probability of the null is only $p(\mathrm{null}|D) \approx 0.11$. Despite the fact that the BF does not indicate the posterior probabilities of the models, the usual decision rule for Bayesian null-hypothesis testing considers the magnitude of the BF relative to a decision threshold $C$. If $\mathrm{BF}_{\mathrm{null}} > C$ then the null hypothesis is accepted and the alternative is rejected, but if $\mathrm{BF}_{\mathrm{null}} < 1/C$ (i.e., $\mathrm{BF}_{\mathrm{alt}} > C$) then the alternative hypothesis is accepted and the null is rejected, otherwise we remain undecided. The value of the decision threshold, $C$, is set by practical considerations, just like the decision threshold for $p$ or the limits of a ROPE. Various recommendations for the setting of $C$ were discussed in the main article, at the end of the section on Specifying ROPE Limits. Typical values for the decision threshold $C$ are 3 or 10.

Figure S.1 shows an example of Bayesian hypothesis testing. The upper panel of nine plots shows the prior distribution. Within the upper panel of nine plots, the top row shows the null-hypothesis prior. Notice the "spike" on $\mu$, which in this case is actually a narrow distribution bounded by the ROPE limits. See Morey and Rouder (2011) for a related treatment of interval null hypotheses. The spike on $\mu$ induces a corresponding spike for the prior distribution of the effect size, $\delta$. The middle row of the upper panel shows the alternative-hypothesis prior, which has a broad distribution over $\mu$ instead of a spike. This broad prior indicates that many values of $\mu$ are possible, unlike the null hypothesis prior.

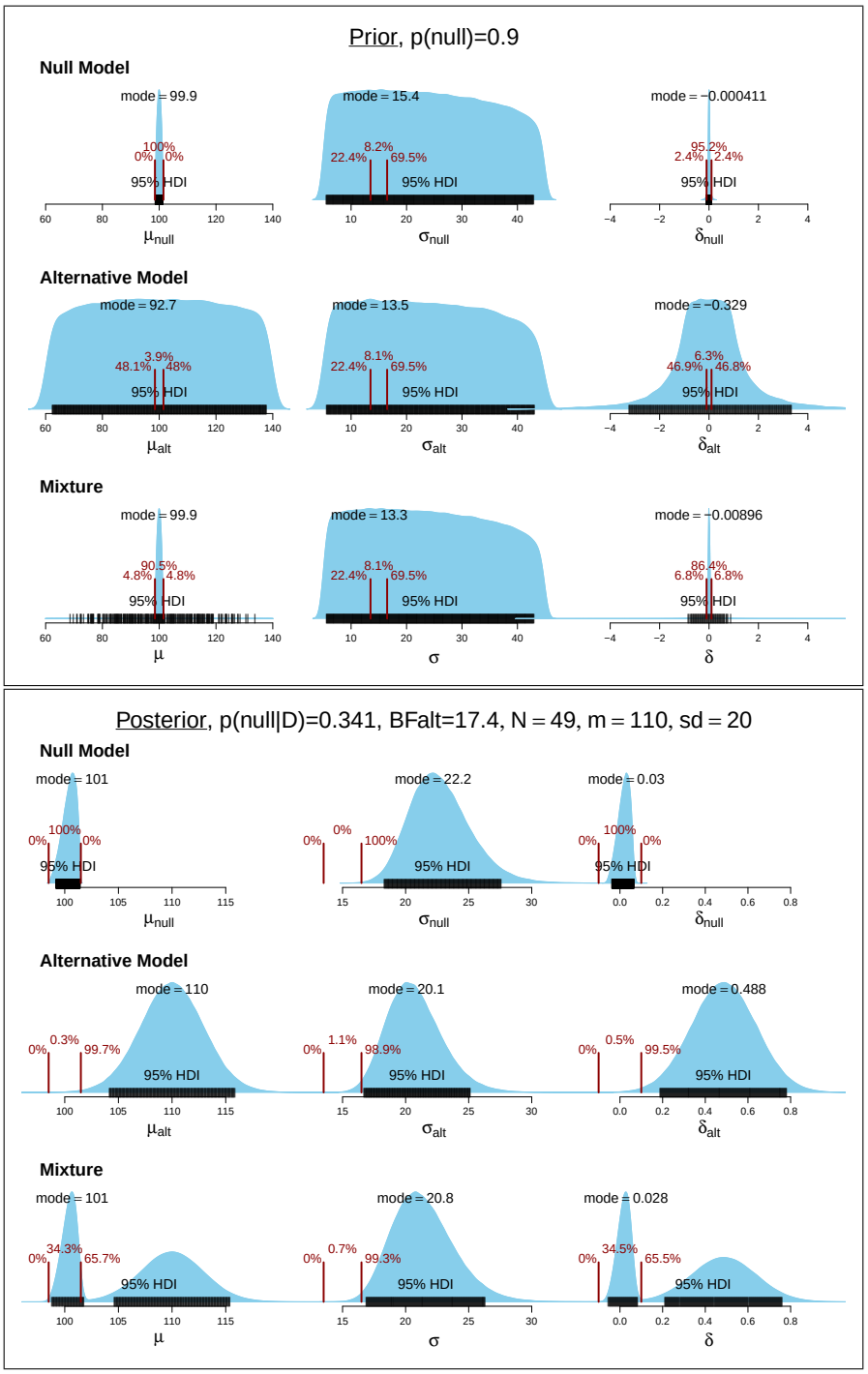


*Figure S.1.* An example of Bayesian hypothesis testing. Upper panel of nine plots shows the prior distribution. Lower panel of nine plots shows the posterior distribution. Here the BF *rejects* the null hypothesis on $\mu$, but the mixture estimate of $\mu$ has substantial mass and density *inside* the ROPE (even if the prior probability of the null is set at 0.50, not shown).

As shown in the title of the upper panel, for this example the prior probabilities on the models are set to $p(\text{null}) = 0.9$ and $p(\text{alt}) = 0.1$.

The lower row (of the upper panel Figure S.1) shows the mixture of the two models, weighted by their probabilities. The weighted mixture creates a "spike and slab" prior on $\mu$ (and on $\delta$) which expresses the prior distribution implied by the overall model structure (cf. Rouder, Haaf, & Vandekerckhove, 2018). In the general model-comparison framework, the different submodels can have different parameters and consequently it would not make sense to collapse across the models (in terms of the parameter space, but it would be perfectly reasonable to integrate across predictions of the models as in Bayesian model averaging [BMA]). In the case of null-hypothesis testing, however, the two submodels have the same parameters and therefore the model index can be collapsed (i.e., integrated over). The key point for our purposes is that Bayesian inference on the mixture prior yields the same result as Bayesian inference on a hierarchical prior that has an explicit model index. The mixture distribution is revealing because it shows explicitly the distribution on the parameters when both submodels are taken seriously as contending simultaneously to describe the data, instead of considering only one submodel at a time. The mixture distribution is useful for considering HDI and ROPE relationships when the null and alternative models are treated as simultaneously viable models, contributing according to their abilities, instead of as mutually exclusive competitors.

On the other hand, and importantly, collapsing across the model index loses the ability to compute a Bayes factor because the model distinction is obliterated. The mixture distribution does not uniquely indicate which submodels may have created it. The submodels may have been a spike and a slab as shown in Figure S.1. Or, the submodels that created the mixture may have been interested in testing non-inferiority when the null value is plausible, such that one hypothesis' prior distribution is values of $\mu$ below the low limit of the ROPE and the other hypothesis' prior distribution is values of $\mu$ above the low limit of the ROPE including a spike at the null value (i.e., a left-side prior distribution on $\mu$ shaped like "_" below the ROPE, and a right-side prior distribution on $\mu$ shaped like "L"). The BF for those submodels will be quite different than the BF for spike-null vs slab-alternative, yet both pairs of models have the same mixture distribution. To reiterate, while the mixture distribution shown in Figure S.1 is useful for considering intervals when the two models are treated as simultaneous contributors instead of as mutually exclusive competitors, the mixture distribution also renders the BF irrelevant, or at least not unique.

The lower panel of nine plots in Figure S.1 shows the posterior distribution after considering a set of data with mean and standard deviation as indicated in the title of the panel. The data mean is quite a bit bigger than the null-hypothesis value of $\mu$, and therefore credibility is shifted away from the null hypothesis toward the alternative hypothesis, that

is, $BF_{alt} >> 1$, as indicated in the title of the panel. According to the usual decision rule, we would therefore reject the null hypothesis. Notice, however, that in this case the posterior probability of the null hypothesis is still fairly large, which suggests we should *not* reject the null hypothesis. Moreover, the mixture distribution has substantial probability mass inside the ROPE. The mixture distribution also shows that the HDI is split into two subintervals, one of which falls within the ROPE.

It may be difficult to consider the posterior HDI from a spike-and-slab prior because the density of the spike can be arbitrarily large, in both the prior and posterior, depending on the narrowness of the spike. For very narrow spikes, the posterior will virtually always have some narrow subinterval of the HDI within the ROPE. If the spike is *arbitrarily* extremely dense, it is not meaningful to consider its density in the posterior because it remains arbitrary. In this case, it may only be meaningful to consider the probability mass within the ROPE. The moral is that the spike null, in all its detail, must express a meaningful model and not merely a convenient default. In general, model comparison is only as meaningful as the models being compared.

When the null hypothesis and alternative hypothesis use identical prior distributions except for the spike on one parameter, then the Bayes factor can be computed using the Savage-Dickey density ratio (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). This ratio can be approximated by considering the mass of the target parameter's distribution inside the ROPE of the alternative hypothesis: The prior mass inside the ROPE divided by the posterior mass inside the ROPE is very nearly the BF for the alternative. For example in the alternative model of Figure S.1, the prior mass inside the ROPE of $\mu$ is displayed as 3.9% and the posterior mass is 0.3% (both rounded to only one decimal place), and their ratio is just under 17 when more decimal places are considered. This conceptualization is useful for understanding that the BF in null hypothesis testing focuses on how much the probability near the null value shifts from prior to posterior. The ratio fo ROPE masses does not indicate the posterior mass, and the ratio of ROPE masses does not indicate relation of HDI to ROPE.

Figure S.2 shows another example of Bayesian hypothesis testing, but in this case on parameter $\sigma$ instead of $\mu$. The point is to demonstrate that every hypothesis test is a distinct model comparison. Notice in the upper row of Figure S.2 that the null model has a spike prior on $\sigma$ but a broad prior on $\mu$. The middle row (of the upper panel) shows that the alternative-model prior is the same as in Figure S.1. Using the same data as in Figure S.1, the lower panel of Figure S.2 shows that posterior distribution within the alternative model is the same as in Figure S.1 (except for random variation in the MCMC chain). The BF favors the alternative model over the null model, which would lead to a decision to *reject* the null hypothesis, even though the posterior probability of the alternative model is only

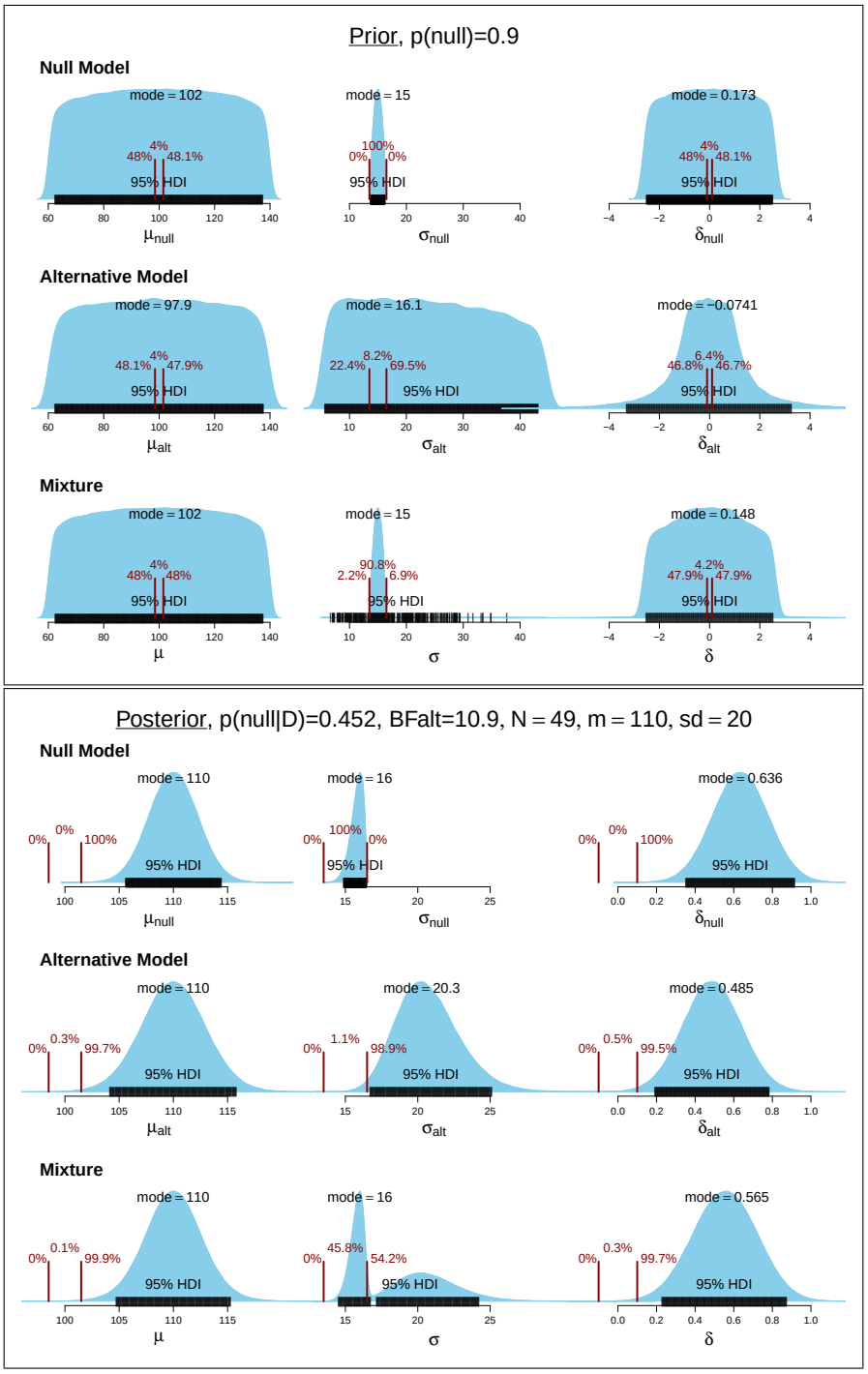


*Figure S.2.* An example of Bayesian hypothesis testing on $\sigma$. Here the BF rejects the null model of $\sigma$, but the mixture estimate of $\sigma$ has values *inside* the ROPE (even if the prior probability of the null is set at 0.50, not shown). The null model prior is the same but with a narrow "spike" prior within the ROPE of $\sigma$.

a little greater than 50%, and the mixture distribution shows substantial probability mass and density within the ROPE, which would suggest we should remain undecided.

Figure S.3 shows an example in which the BF accepts the null hypothesis. The prior distribution is the same as in Figure S.1, except here the prior probability of the null hypothesis is set to $p(\text{null}) = 0.1$. Notice this results in a spike-and-slab mixture prior on $\mu$ (and $\delta$). The lower panel of Figure S.3 shows the posterior distribution from data that are consistent with the null hypothesis. The BF favors the null hypothesis, such that the usual decision rule would accept the null hypothesis and reject the alternative hypothesis. Notice, however, that in this case the posterior probability of the null hypothesis is barely any bigger than the posterior probability of the alternative hypothesis, which suggests that we should *not* accept the null hypothesis. Moreover, the mixture distribution has substantial probability mass outside the ROPE. The mixture distribution also shows that the HDI extends outside the ROPE.

Bayesian hypothesis testing has some appealing features, especially relative to frequentist NHST. In particular, Bayesian hypothesis testing explicitly specifies competing hypotheses and computes posterior probabilities of those hypotheses. Consequently, Bayesian hypothesis testing can show support in favor of the null hypothesis or against it. But, as noted previously, Bayesian hypothesis testing is only as meaningful as the prior distributions in the models, and the prior probabilities of the models. If the model priors are merely caricatures of imaginary possibilities, then the BF is only telling you the shift in credibilities from dragons to unicorns. In particular, a spike-and-slab prior, as exemplified in Figures S.1, S.3, and S.2, does not solve Meehl's paradox (which was described in the section titled, The ROPE in theory testing). Meehl's paradox arises when the alternative model expresses "anything but null," that is, when the alternative is a slab across a broad range of parameter values. Bayesian model comparison could solve Meehl's paradox by instead using model priors that express competing specific predictions with only modest uncertainty. Unfortunately this is rarely done.

Another potential advantage of Bayesian null hypothesis testing is that it gives special weight to the null value. This can be seen graphically by comparing the mixture prior distribution with the alternative-model prior distribution in Figure S.3. The alternative-model prior distribution is smooth and relatively flat over the null value, whereas the mixture prior has a narrow and tall bump at the null value. The mixture prior expresses the idea that there is extra prior probability at the null value, making the null value uniquely probable relative to other parameter values. Ultimately, the choice of a smooth prior distribution or a bump-at-null prior distribution depends on theories being tested and the previous data that can inform the prior. Technically, however, the posterior distribution from a broad smooth prior tends to be very stable against changes in the breadth of the prior, while the posterior

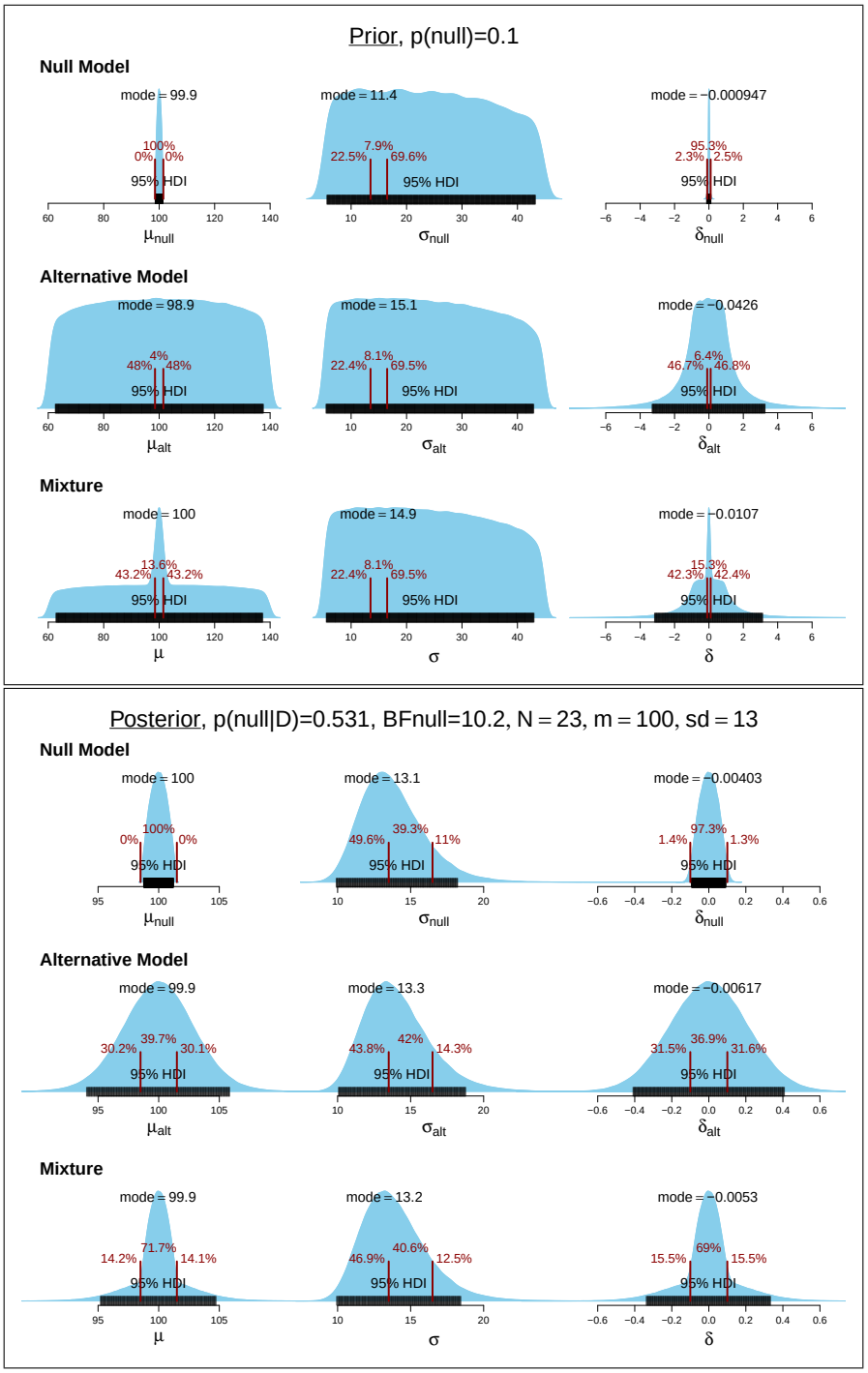


*Figure S.3.* In this example of Bayesian hypothesis testing, the BF *accepts* the null model of $\mu$ when the data have small $N$, but the mixture estimate of $\mu$ has high-density values and substantial mass *outside* the ROPE (even if the prior probability of the null is set at 0.50, not shown).

mass near a bump depends strongly on the breadth of the surrounding prior. Therefore if a Bayesian null hypothesis test is conducted, the prior distributions in both models must be carefully set so they are meaningful, and the prior probabilities of the models must be meaningfully set, and the results must be checked for robustness against small changes in the priors.

There are various reasons that one might prefer decisions using HDI+ROPE from a smooth prior distribution instead of using Bayes factors. One reason is that estimation from a smooth prior yields the answer to all tests simultaneously. By contrast, the BF approach requires a distinct prior distribution for every null hypothesis tested. Much like in frequentist hypothesis testing, there is a hierarchy of nested (null) models, each of which requires a distinct BF prior and computation. For example, the tests of $\mu_0 = 100.0$ and $\sigma_0 = 15.0$ require two distinct "spike" null models, as was illustrated in Figures S.1 and S.2. Notice in Figures S.1 and S.2 that the posterior distribution in the alternative model is identical (except for MCMC noise).

Various others cautions regarding Bayesian hypothesis testing were outlined by Kruschke and Liddell (2017a). First, Bayes factors can change dramatically with seemingly innocuous changes in the prior distributions of the models, and default priors are often meaningless. Therefore it is important to use genuinely meaningful prior distributions and model probabilities.

The Bayes factor ignores the prior probabilities of the models, and a more sensible decision rule would be based on the posterior probabilities of the models, that is, the BF multiplied by the prior odds. The examples in Figures S.1, S.2, and S.3 used prior probabilities on the models that are not 50/50 specifically to demonstrate the difference between the BF and the posterior probabilities of the model. The main creator of the BayesFactor package (Morey & Rouder, 2015) has said, "It should also make clear that the Bayes factor is not really the useful decision statistic; rather, the posterior odds are." (`http://bayesfactor.blogspot.com/2015/01/on-verbal-categories-for-interpretation.html`) A prime illustration of what a blunder it is to ignore prior probabilities comes from disease diagnosis. For a diagnostic test with a positive outcome, the BF is the ratio of the correct-detection rate to the false-alarm rate (i.e., the ratio of the likelihoods under the two models). But the posterior odds are usually much smaller than the BF because the prior probability of having the disease is usually small. Details can be found at `http://doingbayesiandataanalysis.blogspot.com/2015/12/lessons-from-bayesian-disease-diagnosis_27.html` or `https://osf.io/r9zfy/`.

The BF can accept the null value even when the estimate of the parameter value has poor precision. Figure S.3 showed an example. This can happen when the alternative hypothesis is vague and can therefore be rejected by a relatively small set of data near the null. If the alternative hypothesis is more narrowly specified and is near the null value, then

it takes a large data set to distinguish the models.

Finally, using the BF encourages black-and-white thinking more than using the HDI+ROPE because the information provided by the BF completely ignores the parameter estimation. The decision by BF alone never looks at the parameter estimates. While the parameter distributions in Figures S.1, S.2, and S.3 are explanatory, they are superfluous for computing the BF and for making a decision based on the BF. Thus, decision making by the BF alone (especially using default priors) can easily devolve into Gigerenzer's "mindless ritual" of hypothesis testing (Gigerenzer, 2004; Gigerenzer & Marewski, 2015). On the other hand, decisions based on HDI+ROPE inherently forefront the posterior estimates of the parameters and their uncertainty.

## Application to meta-analysis

This section demonstrates a realistic application to meta-analysis. Meta-analysis is crucially important for establishing replicability of findings in both basic and applied research. Meta-analysis is aimed not merely at establishing the presence or absence of an effect across studies, but is concerned with establishing the typical magnitude of the effect across studies, the variability of the magnitude across studies, and our uncertainty in those estimates. Meta-analysis is a central emphasis of the "New Statistics" endorsed by the journal *Psychological Science* (Cumming, 2014; Eich, 2014; Lindsay, 2015; Kruschke & Liddell, 2017b).

The examples of meta-analysis in this section illustrate specification of ROPE limits and the importance of prior probabilities of the models in Bayesian hypothesis testing. The examples also illustrate concrete applications of null-value assessment for an effect parameter and for a scale parameter (for random-effects vs fixed-effect models), analogous to $\mu$ and $\sigma$ in the abstract examples of the previous section. Finally, these examples demonstrate the importance of taking into account the prior probabilities of hypotheses in Bayesian hypothesis testing.

We will consider two application domains. In one application domain, patients who experienced heart attacks were randomly assigned to a control group or a group who received a heart-muscle relaxant called a beta-blocker. The dependent variable was death or survival (during some limited duration of treatment). There were several studies at different hospital sites. The number of patients in the control group at site $s$ is denoted $n_{C[s]}$ and the number of deaths is denoted $z_{C[s]}$. Typically $z_{C[s]}/n_{C[s]}$ is about 9%. The number of patients and deaths in the treatment group are denoted $n_{T[s]}$ and $z_{T[s]}$, respectively. If the treatment is successful, it should be the case on average that $z_{T[s]}/n_{T[s]} < z_{C[s]}/n_{C[s]}$. We will use data from Yusuf, Peto, Lewis, Collins, and Sleight (1985) as reported in Gelman et al. (2013, Sec. 5.6).

In a second application domain, patrons of hotels who stayed more than one night were randomly assigned to a control group or a group who received a notice in the hotel room that most patrons reused their towels (to conserve resources wasted by needless laundering). The number of people who reuse their towels is $z$, and if the treatment is successful then on average it should occur that $z_{T[s]}/n_{T[s]} > z_{C[s]}/n_{C[s]}$. We will use data from Scheibehenne, Jamil, and Wagenmakers (2017). (For critical commentary and reply, see Carlsson, Schimmack, Williams, and Bürkner (2017) and Scheibehenne, Gronau, Jamil, and Wagenmakers (2017).)

*The meta-analytic model*

In a meta-analysis, the data from experiments at different sites are analyzed together to extract the typical effect across sites. There are two main questions: First, how big is the average effect across sites and is it practically different from zero? Second, how different are the effects at different sites and are they similar enough in magnitude that we could decide they are practically equivalent? The second question can be re-phrased as asking if we should describe the variation across sites as different, with a "random-effects" model, or should we describe the variation across sites as negligible and instead use a "fixed-effect" model? This latter question is usually answered with advice always to use the random-effects model (e.g., Field, 2003; Hunter & Schmidt, 2000), but here we will consider a Bayesian hypothesis test because that approach was advocated by Scheibehenne, Gronau, et al. (2017).

To address the questions, we must formalize the model. The model that will be reported here is extended from the model used by Kruschke and Liddell (2017b) for the heart-attack data, and which was also applied to towel re-use at `http://doingbayesiandataanalysis.blogspot.com/2016/11/bayesian-meta-analysis-of-two.html` or `https://osf.io/eps5f/`. Here I will describe a subset of the parameters that are most relevant to the discussion, and the full details are in an appended subsection. Let $\theta_{C[s]}$ represent the underlying rate of occurrence in the control condition at site $s$; that is, $z_{C[s]}/n_{C[s]}$ is a random sample from rate $\theta_{C[s]}$. Analogously, $\theta_{T[s]}$ is the rate of occurrence in the treatment group at site $s$. The effect of treatment at site $s$ is denoted $\rho_s$, and is on the scale of log-odds: $\rho_{[s]} = \text{logit}(\theta_{T[s]}) - \text{logit}(\theta_{C[s]})$ where $\text{logit}(\theta) = \log(\theta/(1-\theta))$. Re-arranging, we have $\theta_{T[s]} = \text{logistic}(\rho_{[s]} + \text{logit}(\theta_{C[s]}))$ where logistic is inverse logit, which means that at each site, the rate of occurrence in the treatment condition depends on that site's rate of occurrence in the control condition and that site's effect.

The meta-analytic model describes the distribution of effects $\rho_{[s]}$ across sites as a normal distribution with mean $\mu_\rho$ and standard deviation $\sigma_\rho$. The parameter $\mu_\rho$ indicates the typical magnitude of the treatment effect across sites, and the HDI of the posterior distribution of $\mu_\rho$ is the uncertainty in that typical magnitude of treatment. Usually we

are most interested in $\mu_\rho$. If $\mu_\rho$ is zero, then the treatment has zero effect on average across sites, although some sites may have positive treatment effect while other sites have negative treatment effect. Setting $\mu_\rho \equiv 0$ is the null-effect model, but remember this allows individual sites to have non-zero effects: $\rho_{[s]} \neq 0$.

The parameter $\sigma_\rho$ is the standard deviation of the treatment effects across sites. If $\sigma_\rho$ is zero, the model assumes that the treatment effect is the same at every site, $\rho_{[s]} \equiv \mu_\rho$, which is the fixed-effect model. The fixed-effect model allows there to be a non-null treatment effect, but insists those effects are identical at every site. Moreover, the fixed-effect model allows the rates of occurrence to vary across sites, because the distribution of $\theta_{C[s]}$ across sites is modeled as a beta distribution with mode $\omega_{\theta C}$ and concentration $\kappa_{\theta C}$. The parameter $\omega_{\theta C}$ is the typical rate of occurrence in the control condition across sites. The rates of occurrence in the control and treatment conditions can vary across sites even if the effect of treatment is identical at all sites. As mentioned above, full details of the model are provided in the final subsection.

In the analyses presented below, I will first report results from estimation with moderately broad priors on the parameters. Then I will report results from a Bayesian hypothesis test for the null-effect model, involving a spike prior on $\mu_\rho$. And I will report results from a Bayesian hypothesis test for the fixed-effect model, involving a spike prior on $\sigma_\rho$.

*Setting ROPEs*

To set the ROPEs, consider first the application to deaths after heart attack. Brophy, Joseph, and Rouleau (2001) pointed out that treatment effects should be calibrated by their clinical implications in terms of lives saved per 100 deaths (for a posterior distribution on lives saved, see Figure 9 of Kruschke & Liddell, 2017b). Let us suppose that one life saved per hundred patients is the minimum clinically important effect (relative to costs and deleterious side effects). For a control-condition death rate of 10%, this minimum effect translates into $\rho = \log(.10/(1-.10)) - \log(.09/(1-.09)) \approx 0.1$. Therefore we will set the ROPE limits on $\mu_\rho$ at $0 \pm 0.1$. To keep the illustration simple, I will use the same ROPE for the towel-reuse data. A typical rate of towel reuse in the control condition is about 60%, and $\mu_\rho = 0.1$ implies an increase to approximately 63%. This constitutes a fairly small increase to qualify as not practically equivalent to zero, and a demandingly narrow criterion for equivalence. For the ROPE on $\sigma_\rho$, notice that its scale is the same as the scale for $\mu_\rho$. Therefore, in lieu of specific knowledge of costs associated with *variability* of the treatment effect, I will tentatively use the same ROPE limit on $\sigma_\rho$.

*Estimation of parameters*

Figure S.4 shows the result of estimation of the parameters from moderately broad priors (on a single model without any Bayesian hypothesis testing). The first two panels of

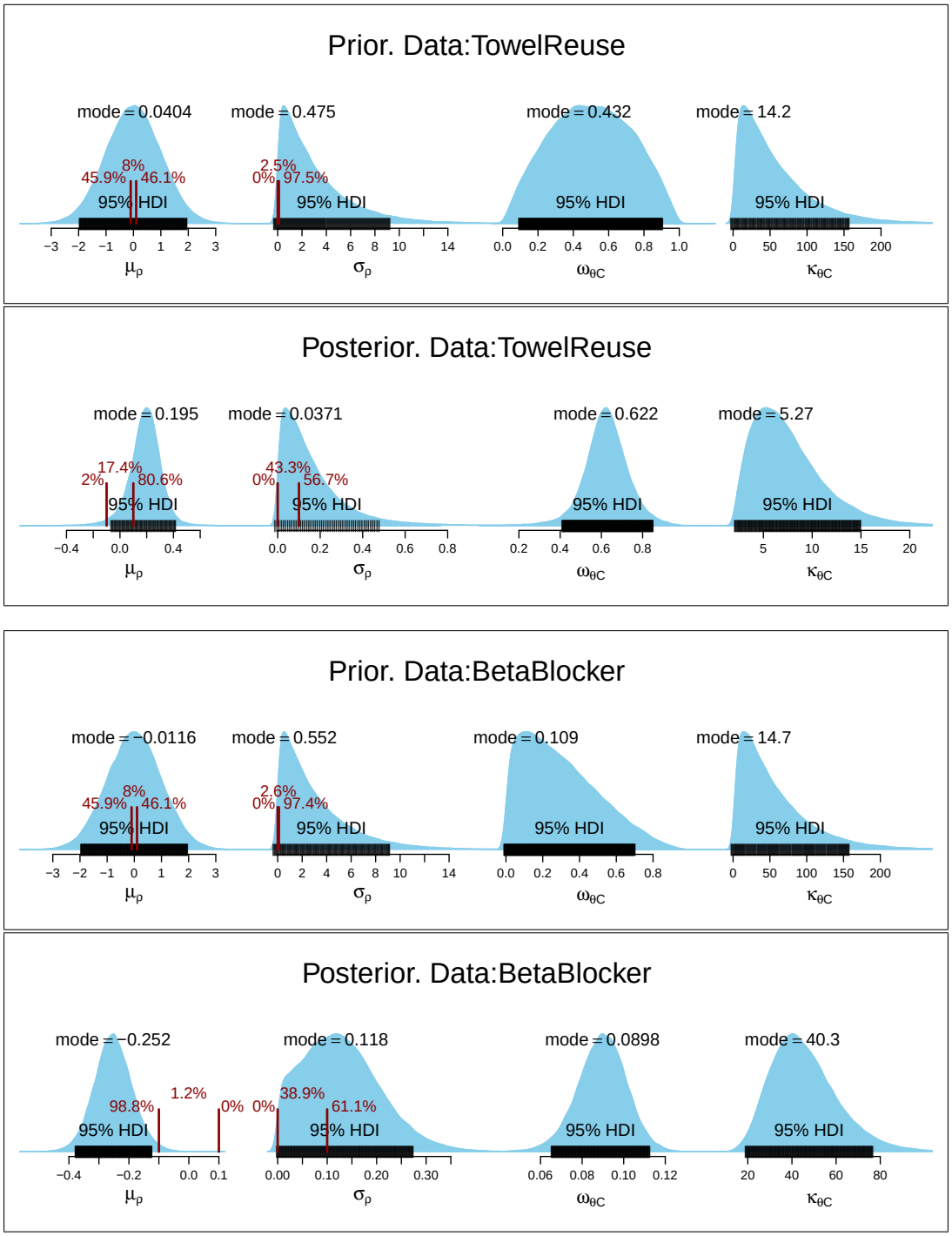


*Figure S.4.* Meta-analysis of towel reuse data and beta-blocker data. Emphasis is on the parameter $\mu_\rho$ in the left column, which is the mean effect across sites.

Figure S.4 show the prior and posterior distributions for the towel-reuse data. Each panel shows the marginal distributions of the four key parameters that describe the tendency across sites. In particular, the posterior distribution of $\mu_\rho$ has a positive mode, suggesting some increase in towel reuse due to the treatment, but its 95% HDI is wide and overlaps the ROPE and indeed overlaps an effect of zero. (This posterior distribution matches the results previously reported at `http://doingbayesiandataanalysis.blogspot.com/2016/11/bayesian-meta-analysis-of-two.html` or `https://osf.io/eps5f/`.) Therefore we would neither reject the null value nor accept it, and we remain undecided about the treatment effect. Figure S.4 also shows that the 95% HDI of the posterior distribution on $\sigma_\rho$ includes values inside and outside the ROPE, indicating that we would neither reject nor accept the value zero for $\sigma_\rho$.

The second two panels of Figure S.4 show the prior and posterior distributions for the beta-blocker (heart attack) data. The posterior distribution of $\mu_\rho$ has a negative mode, indicating a decrease in heart attacks due to the treatment, and its 95% HDI excludes the ROPE. (This posterior distribution matches the results previously reported by Kruschke and Liddell (2017b, Fig. 8).) Therefore we decide to reject the null value. Figure S.4 also shows that the 95% HDI of the posterior distribution on $\sigma_\rho$ includes values inside and outside the ROPE, indicating that we would neither reject nor accept the value zero for $\sigma_\rho$.

*Bayesian hypothesis tests*

Figure S.5 shows the prior and posterior distributions for a Bayesian hypothesis test of the null effect across sites, for the towel-reuse data. Notice the spike prior on $\mu_\rho$ in the null model. In this implementation, the spike is truly infinitesimally narrow, unlike the examples of Figures S.1 and S.3. The alternative-model prior is the same as was used for the previous meta-analysis in Figure S.4 (except for MCMC noise in the graphs). As a default, the prior probability of the null model was arbitrarily set at 0.5. Setting the prior probability this way is lazy and dangerous; a serious researcher of this topic would have knowledge of publicly accessible previous research to inform the prior probability. For example, there might be research in other applications of norm interventions to suggest that there is probably some effect of the norm intervention, however small but non-zero (otherwise the experiments would not have been conducted). Indeed, in a strict logical sense it is virtually impossible for the norm intervention to have exactly zero effect. Thus, the default setting of $p(\text{null}) = 0.5$ is probably too large. The lower panel of Figure S.5 shows the posterior distribution. Notice that the posterior distribution within the alternative model is the same as in Figure S.4 (except for MCMC noise). The title of the panel annotates the BF which slightly favors the null hypothesis, though not decisively. In the mixture distribution, there is substantial probability mass both inside and outside the ROPE.

Figure S.6 shows the prior and posterior distributions for a Bayesian hypothesis test of a fixed effect across sites, for the towel-reuse data. Notice the spike prior on $\sigma_\rho$ in

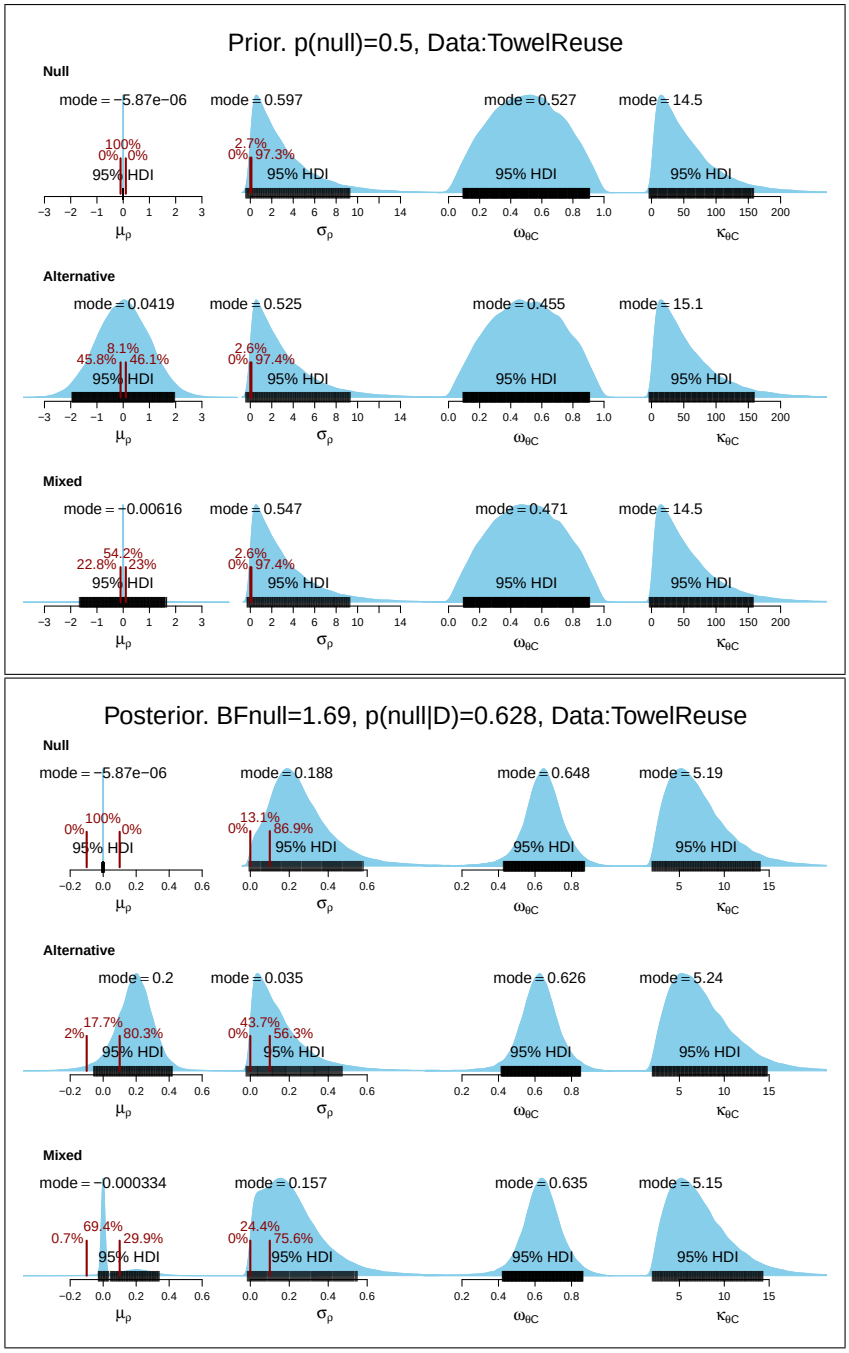


*Figure S.5.* Meta-analysis of towel reuse data, with Bayesian hypothesis test of null effect. Emphasis is on the parameter $\mu_\rho$ in the left column, which has a spike prior in the null model.

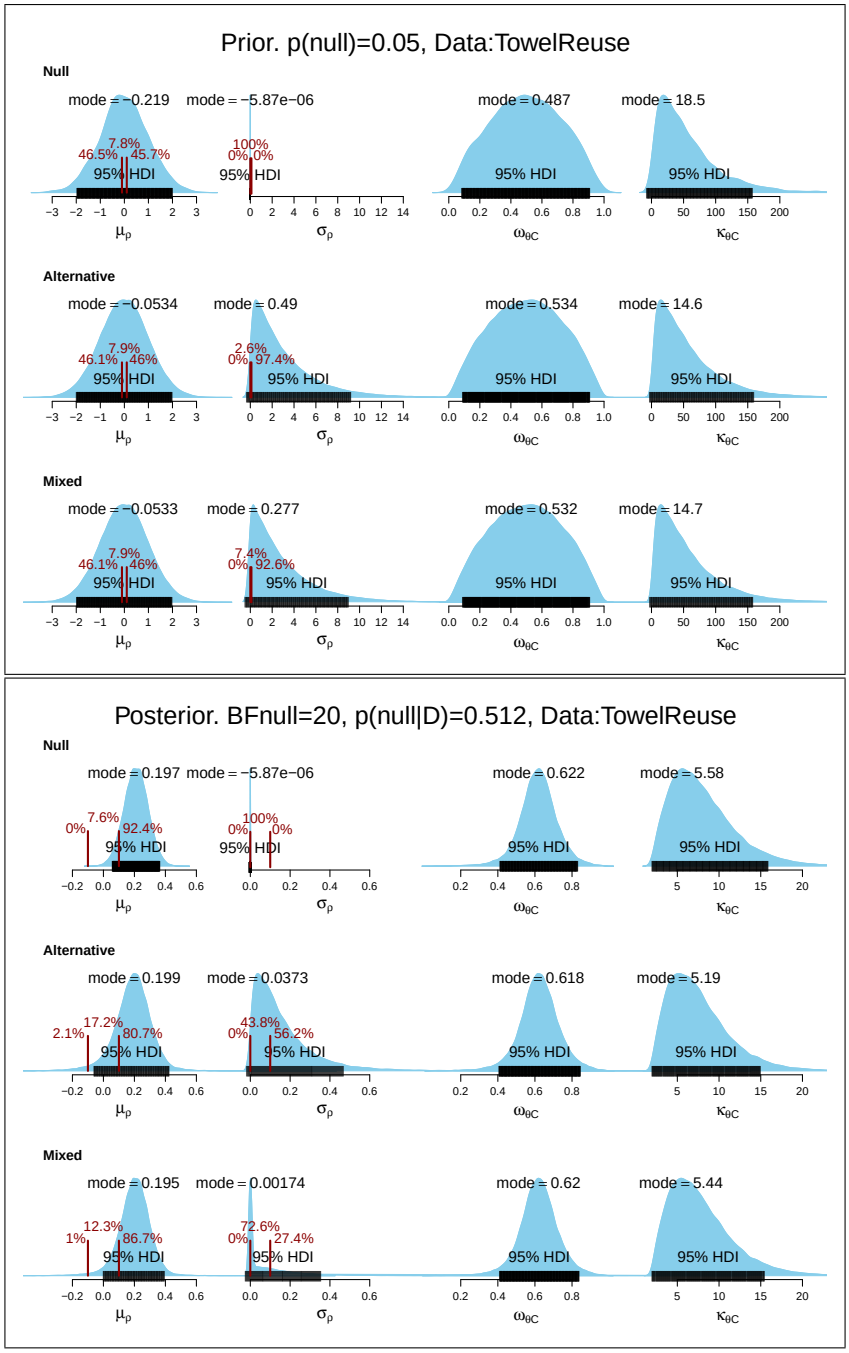


*Figure S.6.* Meta-analysis of towel reuse data, with Bayesian hypothesis test of fixed effect. Emphasis is on the parameter $\sigma_\rho$ in the second column, which has a spike prior in the null model.

the null model. In this implementation, the spike is truly infinitesimally narrow, unlike the example of Figure S.2. The alternative-model prior is the same as was used for the previous meta-analysis in Figure S.4 (except for MCMC noise in the graphs). The prior probability of the fixed-effect model was set at 0.05 because it is highly implausible that the effect of the norm intervention should be the same for the clienteles of different hotels (for instance, it is implausible that the intervention would have the same effect on patrons of a 2-star campground motel as patrons of a 5-star cosmopolitan hotel). The lower panel of Figure S.6 shows the posterior distribution. Notice that the posterior distribution within the alternative model is the same as in Figures S.4 and S.5 (except for MCMC noise). The title of the panel annotates the BF which favors the fixed-effect model. But the posterior probability of the fixed-effect model is only middling (because its prior probability was small). In the mixture distribution, there is substantial probability mass both inside and outside the ROPE.

Importantly, Figure S.6 also reveals an influence of the fixed-effect assumption on the estimate of treatment effect, $\mu_\rho$. Examine the posterior distributions of $\mu_\rho$ within the Null (i.e., fixed-effect) model and within the Alternative (i.e., random-effects) model. Notice that the distribution on $\mu_\rho$ is narrower for the fixed-effect model than for the random-effects model. In particular, the 95% HDI on $\mu_\rho$ in the fixed-effect (Null) model does not include zero, but the 95% HDI on $\mu_\rho$ in the random-effects (Alternative) model does include zero. Intuitively, the distribution on $\mu_\rho$ is wider for the random-effects model because of greater uncertainty produced by estimating distinct effects at all the different sites. The narrower distribution for the fixed-effects model can be intuitively thought of as artificially certain from the implausible assumption of no variation in treatments across sites (cf. Field, 2003; Hunter & Schmidt, 2000).

Figure S.7 shows the prior and posterior distributions for a Bayesian hypothesis test of the null effect across sites, for the beta-blocker (heart attack) data. Notice the spike prior on $\mu_\rho$ in the null model. In this implementation, the spike is truly infinitesimally narrow, unlike the examples of Figures S.1 and S.3. The alternative-model prior is the same as was used for the previous meta-analysis in Figure S.4 (except for MCMC noise in the graphs). As a default, the prior probability of the null model was arbitrarily set at 0.5. As mentioned above, setting the prior probability this way is lazy and dangerous; a serious researcher of this topic would have knowledge of publicly accessible previous research to inform the prior probability. For example, there might be research in other applications of beta-blockers to suggest that there is probably some effect on heart attack, however small but non-zero (otherwise the experiments would not have been conducted). Indeed, in a strict logical sense it is virtually impossible for the beta blocker to have exactly zero effect. Thus, the default setting of $p(\text{null}) = 0.5$ is probably too large. The lower panel of Figure S.7 shows the

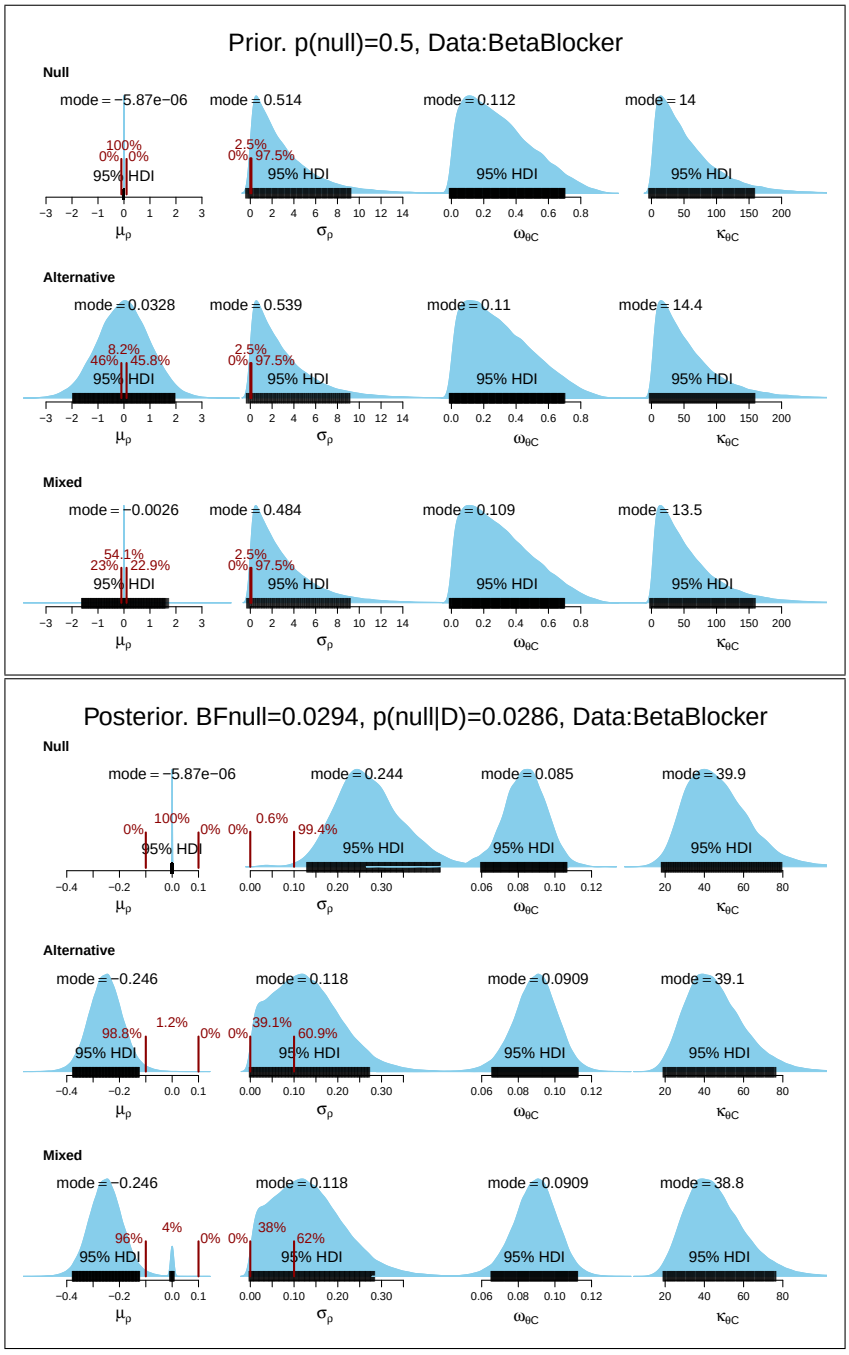


*Figure S.7.* Meta-analysis of the beta-blocker (heart attack) data, with Bayesian hypothesis test of null effect. Emphasis is on the parameter $\mu_\rho$ in the left column, which has a spike prior in the null model.

posterior distribution. Notice that the posterior distribution within the alternative model is the same as in Figure S.4 (except for MCMC noise). The title of the panel annotates the BF which strongly favors the alternative hypothesis (i.e., $\mathrm{BF}_{null}$ is very small). In the mixture distribution, there is relatively little probability mass inside the ROPE, and there would be even less mass inside the ROPE if the prior probability of the null-effect model were less.

Figure S.8 shows the prior and posterior distributions for a Bayesian hypothesis test of a fixed effect across sites, for the beta-blocker (heart attack) data. Notice the spike prior on $\sigma_\rho$ in the null model. In this implementation, the spike is truly infinitesimally narrow, unlike the example of Figure S.2. The alternative-model prior is the same as was used for the previous meta-analysis in Figure S.4 (except for MCMC noise in the graphs). The prior probability of the fixed-effect model was set at 0.05 because it is highly implausible that the effect of the beta-blocker should be the same for the patient intake of different hospitals (for instance, it is implausible that the treatment would have the same effect on patients from a generally unhealthy region as patients from a healthy community). The lower panel of Figure S.8 shows the posterior distribution. Notice that the posterior distribution within the alternative model is the same as in Figures S.4 and S.7 (except for MCMC noise). The title of the panel annotates the BF which favors the fixed-effect model. But the posterior probability of the fixed-effect model is only middling (because its prior probability was small). In the mixture distribution, there is substantial probability mass both inside and outside the ROPE.

*Discussion of Bayesian hypothesis tests.* Despite the detailed considerations regarding the priors in these Bayesian hypothesis tests, the models being compared were, in some respects, as meaningless as dragons and unicorns. In particular, the alternative hypotheses had prior distributions on $\mu_\rho$ and $\sigma_\rho$ that were merely defaults that expressed neutral uncertainty. Realistically, the priors should have been informed by application-specific prior knowledge. Thus, the prior on $\mu_\rho$ for towel reuse should perhaps have had a mean around 0.2 (not 0.0) and a standard deviation perhaps around 0.3 (not 1.0), while the prior on $\mu_\rho$ for the beta-blockers should have had a mean around $-0.2$. These different settings of the priors would produce very different BF's, but not very different posterior distributions within the alternative model.

The examples of meta-analysis have re-emphasized various points from the previous abstract examples. The meta-analyses also showed how, when considering tests of the fixed-effect model, the BF alone is a misleading decision statistic. Instead, the posterior probabilities of the models should be examined. The posterior probabilities of the models rely on the choice of prior probabilities for the models. Moreover, the posterior probabilities of the models also depend strongly on the variance of the prior distribution in the alternative

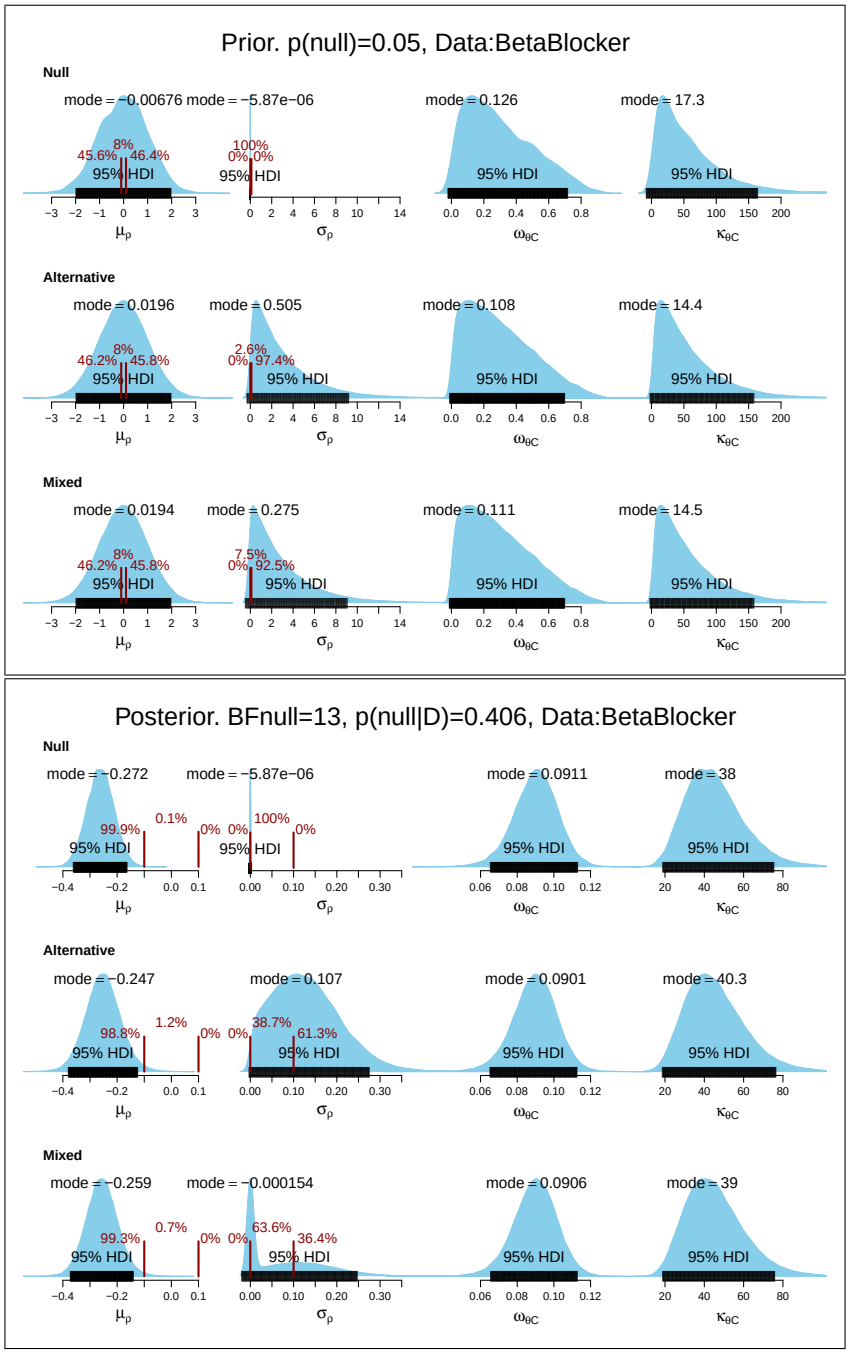


*Figure S.8.* Meta-analysis of beta-blocker (heart attack) data, with Bayesian hypothesis test of fixed effect. Emphasis is on the parameter $\sigma_\rho$ in the second column, which has a spike prior in the null model.

model, although this was not demonstrated here. The results indicate that fixed-effect models are implausible, contrary to arguments by Scheibehenne, Jamil, and Wagenmakers (2017) and Scheibehenne, Gronau, et al. (2017), and consistent with arguments of Carlsson et al. (2017).

*Formal details of the meta-analysis model*

The previous section defined some of the key parameters of the model used for meta-analysis, and this subsection assumes the reader has already scanned those details. The formal specifications below use standard notation for expressing Bayesian models; in particular, the notation $x \sim \text{distrib}(\epsilon)$ means that variable $x$ is randomly distributed according to probability distribution "distrib" which has parameter $\epsilon$. The complete model is expressed by the following relations:

$$
\begin{aligned}
z_{C[s]} &\sim \text{binomial}(\theta_{C[s]}, n_{C[s]}) && \text{\# data likelihood} \\
z_{T[s]} &\sim \text{binomial}(\theta_{T[s]}, n_{T[s]}) && \text{\# data likelihood} \\
\theta_{T[s]} &= \text{logistic}(\rho_{[s]} + \text{logit}(\theta_{C[s]})) && \text{\# } \rho_{[s]} \text{ is effect at site } s \\
\theta_{C[s]} &\sim \text{beta}(\omega_{\theta C}, \kappa_{\theta C}) && \text{\# } \omega\text{=mode, } \kappa\text{=concentration} \qquad (3) \\
\rho_{[s]} &= \phi \cdot \mu_\rho + (1-\phi) \cdot \rho^*_{[s]} && \text{\# } \phi = 1 \text{ for fixed effect} \qquad (4) \\
\phi &\sim \text{bernoulli}(P_\phi) && \text{\# prior prob of fixed effect} \qquad (5) \\
\rho^*_{[s]} &\sim \text{normal}(\mu_\rho, \sigma_\rho) \\
\mu_\rho &= \nu \cdot 0 + (1-\nu) \cdot \mu^*_\rho && \text{\# } \nu = 1 \text{ for null effect} \qquad (6) \\
\nu &\sim \text{bernoulli}(P_\nu) && \text{\# prior prob of null effect} \qquad (7) \\
\mu^*_\rho &\sim \text{normal}(M_{\mu_\rho}, S_{\mu_\rho}) \qquad (8) \\
\sigma_\rho &\sim \text{gamma}(S_\sigma, R_\sigma) \qquad (9) \\
\kappa_{\theta C} - 2 &\sim \text{gamma}(S_\kappa, R_\kappa) \\
\omega_{\theta C} &\sim \text{beta}(M_\omega, K_\omega)
\end{aligned}
$$

The effect at the meta level is expressed by $\mu_\rho$. When the non-null effect model is being used, $\nu = 0$ in Equation 6, and $\mu_\rho$ becomes $\mu^*_\rho$ which has prior expressed in Equation 8. The constants, $M_{\mu_\rho}$ and $S_{\mu_\rho}$, are set to express prior knowledge about the typical effect. The prior probability of the null effect is $P_\nu$ in Equation 7.

The variability across sites is expressed by the standard deviation of the effects, $\sigma_\rho$, which has a prior described as a gamma distribution in Equation 9 when the random-effects model is used. The broad prior is used when the model index $\phi = 0$ in Equation 4, otherwise when $\phi = 1$ all the $\rho_{[s]}$ values are set to the same value, namely $\mu_\rho$. The prior probability of the fixed effect is $P_\phi$ in Equation 5.

For the prior distribution of $\theta_{C[s]}$ in Equation 3, $M_\omega$ is mildly informed by the data as a proxy for asking the user what a typical occurrence rate is: $M_\omega$ is set at $\sum_s z_{C[s]} / \sum_s n_{C[s]}$. $K_\omega$ is arbitrarily set at 4. The shape and rate constants of the gamma distributions are set at reasonably broad and noncommittal values for the scale of the data.

## References


Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis, 2nd edition*. Springer.

Brophy, J. M., Joseph, L., & Rouleau, J. L. (2001). $\beta$-blockers in congestive heart failure: A Bayesian meta-analysis. *Annals of Internal Medicine*, *134*, 550–560.

Carlsson, R., Schimmack, U., Williams, D. R., & Bürkner, P.-C. (2017). Bayes factors from pooled data are no substitute for Bayesian meta-analysis: Commentary on Scheibehenne, Jamil, and Wagenmakers (2016). *Psychological Science*, *28*(11), 1694–1697. doi: 10.1177/0956797616684682

Casella, G., Hwang, J. G., & Robert, C. (1993). A paradox in decision-theoretic interval estimation. *Statistica Sinica*, *3*, 141–155.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge, UK: Cambridge University Press.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*(1), 7–29.

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 x 2 tables. *Biometrics*, 593–602.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*(1), 3–6.

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, *2*(2), 105–124.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, third edition* (3rd ed.). Boca Raton, Florida: CRC Press.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*(4), 275–292.

Kruschke, J. K. (2015). *Doing Bayesian data analysis, Second Edition: A tutorial with R, JAGS, and Stan*. Burlington, MA: Academic Press / Elsevier.

Kruschke, J. K., & Liddell, T. M. (2017a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, **(**), **–**. Retrieved from `https://link.springer.com/article/10.3758/s13423-016-1221-4` doi: 10.3758/s13423-016-1221-4

Kruschke, J. K., & Liddell, T. M. (2017b). The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, **(**), **–**. Retrieved from `https://link.springer.com/article/10.3758/s13423-016-1221-4` doi: 10.3758/s13423-016-1221-4

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/`

`package=BayesFactor` (R package version 0.9.12-2)

Rice, K. M. (2011). *Making peace with p's: Bayesian tests with straightforward frequentist properties.* Retrieved from `http://faculty.washington.edu/kenrice/riceihme.pdf` (Talk presented April 6, 2011)

Rice, K. M., Lumley, T., & Szpiro, A. A. (2008). *Trading bias for precision: decision theory for intervals and sets* (Tech. Rep.). Retrieved from `http://biostats.bepress.com/uwbiostat/paper336/`

Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation, 2nd edition.* Springer.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, **(**), **–**. (in press)

Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? a resolution through model averaging: Reply to Carlsson, Schimmack, Williams, and Bürkner (2017). *Psychological Science*, *28*(11), 1698–1701. doi: 10.1177/0956797617724426

Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (2017). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*, *27*(7), 1043–1046. doi: 10.1177/0956797616644081

Schervish, M. J. (1995). *Theory of statistics.* Springer.

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.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority, second edition.* Boca Raton: Chapman and Hall/CRC Press.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*, *27*(5), 335–371.