The HDI+ROPE decision rule is logically incoherent but we can fix it

Alexander Etz[a,b,c], **Adriana F. Chávez De la Peña**[b,c], **Luis Baroja**[b,c], **Kathleen Medriano**[b,c], **and Joachim Vandekerckhove**[b,c,d,*]

Draft of February 20, 2024.

The Bayesian HDI+ROPE decision rule is an increasingly common approach to testing null parameter values. The decision procedure involves a comparison between a posterior highest density interval (HDI) and a pre-specified region of practical equivalence (ROPE). One then accepts or rejects the null parameter value depending on the overlap (or lack thereof) between these intervals. Here we demonstrate, both theoretically and through examples, that this procedure is logically incoherent. Because the HDI is not transformation invariant, the ultimate inferential decision depends on statistically arbitrary and scientifically irrelevant properties of the statistical model. The incoherence arises from a common confusion between probability density and probability proper. The HDI+ROPE procedure relies on characterizing posterior densities as opposed to being based directly on probability. We conclude with recommendations for alternative Bayesian testing procedures that do not exhibit this pathology and provide a "quick fix" in the form of quantile intervals.

Highest-density interval | ROPE | Incoherence | Bayesian inference | Bayes factor

The crisis of confidence in psychological science has reinflamed historical controversies surrounding the enterprise of statistical hypothesis testing. Classical null hypothesis significance testing (NHST) has been the target of the majority of these criticisms, including that it is overly dichotomous (Gibson, 2021), easily "hackable" (Simmons, Nelson, & Simonsohn, 2011), it can only reject and not accept the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009), that null hypotheses are false *a priori* (Cohen, 1994; McShane, Gal, Gelman, Robert, & Tackett, 2019), and that it answers the wrong question because estimation is more useful (Cumming, 2014).

Alternative methods have been proposed. One prominent example is the so-called HDI+ROPE decision rule (henceforth, HRDR) introduced by Kruschke (2011, 2013) as a superseding alternative to classical NHST.

While we are strong proponents of the Bayesian statistical paradigm (Etz & Vandekerckhove, 2018; Vandekerckhove, Rouder, & Kruschke, 2018) in which HRDR is based, we will argue here that HRDR is flawed and should be avoided, on grounds similar to the above objections to NHST. Specifically, HRDR can lead to inconsistent inferences that depend critically on highly arbitrary choices that must be made by the analyst about how to represent the model.

In what follows we will first informally describe the HRDR with a fictional scenario that highlights its problematic nature. The example will demonstrate that multiple researchers using the HRDR can come to different conclusions despite employing mathematically equivalent models, priors, and data.

To give insight into why inconsistent inferences using the HRDR occur, we will discuss the formal decision-theoretical properties of the method. The crucial flaw we highlight is that *mathematically equivalent representations of a hypothesis test do not necessarily lead to equivalent inferences*. We show that this pathology is due to the HRDR's reliance on probability density to determine which parameter values are "most plausible." However, unlike probability *proper*—which requires that equivalent sets must have equal probabilities—probability *density* values depend on how we label the parameters of a model. Differently put, *sets of parameter values with high density in one representation of the model may have low density in another*. As a result, HRDR can simultaneously conclude that the null hypothesis *is* and *is not* to be rejected, which is logically incoherent.

We provide multiple examples of statistical models with arbitrary parameterizations that can lead to incoherence if the HRDR is applied. Finally, we propose an easy-to-implement modification of the HRDR that resolves the current pathology and achieves coherence. The solution is to use a test that is based on probability rather than probability density.

Introduction to the HRDR

The HRDR is similar in procedure to the broader category of equivalence tests (Berger & Hsu, 1996; Lindley, 1998; Rogers, Howard, & Vessey, 1993), which can be characterized as extensions of point-null tests into tests of regions of practical equivalence (ROPEs). However, the HRDR only uses the ROPE as a means to test a point null and does not entail acceptance or rejection of the entire region. The HRDR is therefore more similar in logic and application to other point null tests.

The HRDR is conducted as follows (Kruschke, 2011). First, determine a parameterized model for the data and specify prior distributions for all parameters. Then, specify the null hypothesis of interest as one particular parameter value: some specific value of the parameter that is considered especially important, such as a correlation being zero. Then, specify a "region of practical equivalence" (ROPE) around the null value, containing those parameter values that are considered only negligibly different from it

[a]Department of Psychology, University of Texas, Austin; [b]Department of Cognitive Sciences, University of California, Irvine; [c]Department of Statistics, University of California, Irvine; [d]Department of Logic and Philosophy of Science, University of California, Irvine
All authors contributed to the final draft.
[*]Correspondence concerning this article should be addressed to Joachim Vandekerckhove (joachim@uci.edu).
This work was supported by National Science Foundation grants #1658303, #1850849, and #2051186. The authors declare no conflicts of interest. This work is based on a part of AE's doctoral dissertation at UC Irvine. A version of it was previously published on PsyArXiv Preprints with DOI 10.31234/osf.io/5p2qt.

for practical purposes. Then, collect data and obtain the posterior distribution of the parameter of interest. For this step, it is necessary to choose a parameterization of the model to be used for the test. From this posterior, then construct a $(100\alpha)\%$ highest-density interval (HDI), which is an interval in which every value has higher posterior density than any outside the interval and which contains $(100\alpha)\%$ of the posterior mass.

Then, finally, if the HDI and ROPE do not overlap, reject the null hypothesis for practical purposes. If the ROPE encompasses the entire HDI, accept the null for practical purposes. If the ROPE and HDI partially overlap, reserve judgment about the status of the null hypothesis.

A fictional example

Avery, Blair, and Cassidy are triplets who are training their pet hamster to detect by smell whether a piece of cheese is safe to eat or not. After months of training they have decided to run a rigorous, blinded experiment, in which they will present their hamster with cheese and record how many it identifies correctly.

To determine if their hamster has been successfully trained to detect the safety of cheese, the triplets decide to implement an HRDR as outlined above for their analysis. They specify that the success or failure of a given cheese identification is modeled as an independent Bernoulli trial with probability of success θ. The three decide to to use a uniform distribution from 0 to 1 as their prior distribution for θ, a default specification suggested by Jeffreys (1961).

Next, the triplets need to determine their null hypothesis and ROPE. A natural null hypothesis in this case is that the hamster is responding at chance level: $\theta = .50$. After deliberation, the triplets agree that their hamster would be considered "practically guessing" if its success rate is within 3% of chance. Thus, they specify a ROPE that spans from $\theta = .47$ to $\theta = .53$.

During the experiment, the hamster correctly determines the safety of a piece of cheese $z = 32$ times out of $N = 47$. When the triplets are ready to present their results, however, they disagree about how that should be done...

The psychologist. It turns out that Avery is a psychology researcher, and feels that the most intuitive scale for the results is the probability scale θ. Avery produces the plot in Figure 1(a), showing that the HDI for the posterior of θ spans $.542$ to $.800$ and does not overlap with the ROPE. Thus, argues the psychologist, the null hypothesis can be rejected with room to spare. Their hamster really can tell when cheese is safe to eat!

The biostatistician. Blair and Cassidy take issue with Avery's presentation of the results. Blair is a biostatistician with extensive experience interpreting log-odds in the context of clinical trials. Thus, to Blair, it seems obvious that the results should be presented on the scale of $\mathrm{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, shown in Figure 1(b). In this parameterization, the test value would be 0, with a ROPE ranging from $\mathrm{logit}(.47) = \log\left(\frac{.47}{1-.47}\right) \approx -.12$ to $\mathrm{logit}(.53) \approx .12$. The lower bound of the HDI for the posterior of $\mathrm{logit}(\theta)$ is just outside the ROPE: This version of the test allows one to reject the null hypothesis, if only just barely. Maybe the evidence is not so strong after all, concludes the biostatistician.

The physician. Cassidy is a practicing physician, and is used to presenting the uncertainty of diagnoses to patients using odds of occurrence. Thus, to Cassidy, it seems only natural to present the results on the scale of $\mathrm{odds}(\theta) = \frac{\theta}{1-\theta}$, shown in Figure 1(c). In this parameterization, the test value would be 1, with a ROPE ranging from 0.887 to 1.128. The HDI for $\mathrm{odds}(\theta)$ spans 1.02 to 3.61, intersecting with the ROPE. It seems clear to Cassidy that more data is still needed – and so the physician concludes that judgment should be withheld about their hamster's abilities.

This example serves to highlight the critical weakness of the HRDR: whether the null hypothesis is rejected depends on what is essentially an historical and cultural accident. That is, if the individuals who collected the data had come from a different tradition then the same data would have led them to different conclusions. Avery and Blair are able to reject the null hypothesis due to their statistically arbitrary preference for a parameterization in terms of probability or log-odds of success, respectively, but Cassidy must withhold judgment owing to their preference for framing results in terms of odds of success. With the same model, the same prior information, and the same data, the triplets come to different conclusions. In other words, their conclusions do not *cohere*.

It is important to emphasize here that incoherence is a constant property of the HRDR. Even in cases where one chooses and stays with a single parameterization of the model—that is, one never transforms parameters from one space to another—the issue persists: *Any* choice of parameter space is statistically arbitrary and conclusions from the HRDR depend critically on that choice.

While the hamster scenario is of course a fiction, this is not an unusual or selectively presented pattern of data – the numerical values in this section were taken from Figure 1 in the paper that first described the HRDR procedure (Kruschke, 2011).

Formal description of the HRDR

The HRDR is informally described as an assessment of how two intervals overlap. In order to describe the HRDR's problem of incoherence more precisely, we will use the language of statistical decision theory. We will begin this section with a brief introduction to decision-theoretic ideas and then discuss how these ideas apply to the HRDR.

Using the language of statistical decision theory, a hypothesis testing procedure is a type of decision rule in which the decision is a choice between one of the candidate hypotheses. If we use δ_T to represent the *decision* made for test T, then the value of δ_T corresponds to the hypothesis chosen. For example, in the classic Neyman-Pearson testing framework one sets up a test to choose between the null hypothesis H_0 and the alternative hypothesis H_1. A test statistic (X), acceptance region (R_0) and rejection region (R_1) are defined. If the test statistic falls in the predefined rejection region, then $\delta_{NP} = 1$ and we choose H_1; if the test statistic falls in the acceptance region, then $\delta_{NP} = 0$ and correspondingly we choose H_0.

The HRDR is fundamentally a hypothesis testing decision procedure and thus it can also be represented as a decision process. With the HRDR we have the same hypotheses as usual: H_0 states that the parameter is (practically) equal to a postulated null value θ_0; H_1 states that the parameter takes some value other than θ_0. To test these hypotheses using HRDR we define the ROPE, then carry out estimation of the parameter and examine if its posterior density function shows considerable overlap with the ROPE. If the areas of high density mainly lie within the ROPE, H_0 is accepted. If the areas of high density are largely outside the ROPE,

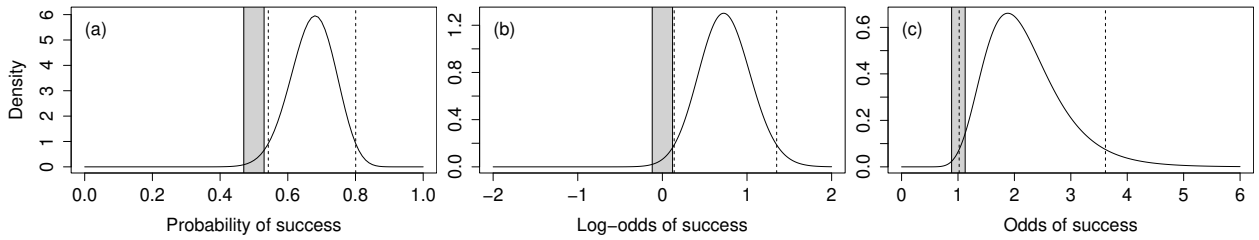

Fig. 1. An illustration of the HRDR. The shaded region indicates the ROPE, and the dashed line indicates the 95% HDI of the respective posterior distributions. **(a)** θ parameterization. The test value is $.50$, with a ROPE from $.47$ to $.53$. The HDI does not intersect the ROPE, leading to rejection of the null hypothesis. **(b)** $\text{logit}(\theta)$ parameterization. Test value is $\text{logit}(.50) = 0$, with a ROPE from $\text{logit}(.47) \approx -.12$ to $\text{logit}(.53) \approx .12$. The HDI does not intersect the ROPE, leading to rejection of the null hypothesis. **(c)** $\text{odds}(\theta)$ parameterization. Test value is $\text{odds}(.50) = 1$, with ROPE from $\text{odds}(.47) = .887$ to $\text{odds}(.53) = 1.128$. The HDI intersects the ROPE, leading to withheld judgment.

H_1 is accepted. The intuition behind this procedure is seemingly straightforward: If the most plausible parameter values are practically equivalent to the null value, then it makes sense to accept it for practical purposes. Likewise, if the most plausible parameter values are not practically equivalent to the null, then reject it.

The overlap of the ROPE and the posterior distribution is formally determined by constructing a 95% highest density interval (HDI) for the test parameter. An HDI is a set consisting of 95% of the posterior mass, with the specific property that every parameter value in the interval has higher posterior density than any value outside the set. Formally, an HDI consisting of $100\alpha\%$ of the posterior mass is defined as the set

$$\text{HDI}_\alpha = \{\theta : p(\theta|D) > k_\alpha\} \quad [1]$$

with k_α chosen such that $P(\text{HDI}_\alpha) = \alpha$ (Druilhet & Marin, 2007).

Formally, we define the decision rule associated with the HRDR as follows:

$$\delta = \begin{cases} 1 & \text{if HDI} \cap \text{ROPE} = \emptyset \\ -1 & \text{if HDI} \subset \text{ROPE} \\ 0 & \text{otherwise.} \end{cases}$$

In the HRDR decision rule, $\delta = 1$ corresponds to rejection of $H_0 : \theta = \theta_0$ and $\delta = -1$ refers to its acceptance. $\delta = 0$ refers to the case there is partial overlap of the sets and one must withhold judgment about the status of the null hypothesis and (if possible) collect more data until one of the other conditions is met.

Let us now revisit the hamster example using this new language. In presenting their results, Avery showed that the HDI and ROPE did not intersect for testing $\theta = .50$, and thus made the decision $\delta = 1$ and rejected the null hypothesis. Blair came to a similar conclusion for testing $\text{logit}(\theta) = 0$, but the evidence did not appear as conclusive. Cassidy concluded that the HDI and ROPE partially overlapped for testing $\text{odds}(\theta) = 1$ and thus made the decision $\delta = 0$ and withheld judgment. Thus, despite having the same information, the triplets come to different conclusions about logically equivalent hypotheses. In the next section, we will provide some critical background on probability theory and use it to explain why the HRDR leads to such instances of incoherence.

Why is the HRDR incoherent?

We will now explain the cause of the incoherent behavior exhibited by the HRDR. A central idea is that the procedure requires the user to find a set of parameters with "high" posterior density, but that there is no unique set of parameter values with the highest density. Density is a property that is determined by the parameterization of the model, which is an explicit choice made by the user of the test. Indeed, as Bernardo (2005) points out, "*any* feature of the posterior which is not invariant under reparametrization is completely illusory, since the parametrization is arbitrary" (p. 374, emphasis original). As we have seen in our examples above, different choices of parameterization lead to different regions of the model with high density, which in turn leads to different conclusions from the procedure.

To understand why this happens, we will first review the fundamental relationship between probability and probability density. Then, we will show the difference in the way probability and probability density behave when converting from one parameterization to another. Finally, we will return to the HRDR and show how it is affected by these considerations.

Density and probability. As an illustrative example, imagine that we are testing whether the mean of a normal distribution is equal to zero or not. The posterior distribution we obtain for θ from our analysis happens to be the standard normal distribution. The cumulative distribution function (CDF) for a standard normal is shown in Figure 2(a). This function tells us for any given candidate θ, the posterior probability that the "true" θ^* is less than θ. For example, for $\theta = 0$, we have $P(\theta^* < \theta|D) = .50$. This is seen in the graph by the height of the CDF curve for $\theta = 0$ being at .50. We will use $F(\theta|D)$ as notation for the CDF.

Recall that probability theory is fundamentally a collection of rules for assigning numbers between 0 and 1 to various sets. The CDF tells us how much probability is associated with the set of values that fall below a candidate θ. Using this function, we can find the probability of other sets; the probability that θ is between any two limits θ_1 and θ_2 ($\theta_1 < \theta_2$) can be obtained by finding how much the CDF increases as we move from θ_1 to θ_2; we subtract $F(\theta_1|D)$ from $F(\theta_2|D)$ to get $P(\theta_1 < \theta^* < \theta_2|D)$. For instance, we take $\theta_2 = .6$ and $\theta_1 = .4$, then $F(.6|D) = .73$ and $F(.4|D) = .66$, giving $P(.4 < \theta^* < .6|D) = .07$.

When the parameter space is continuous, the probability of any individual point is zero. We can then look at the probability density, which tells us how much probability is concentrated "near" a given parameter value. The idea of probability density is directly analogous to physical concepts of density, in that it tells us how much (probability) mass exists in a given region of some space. Consider our example with the standard normal posterior for θ. We could take a small window $\Delta\theta$ around each θ that spans .1 below and .1 above, and find the probability that θ^* is between those limits. Some examples are drawn on Figure 2(a). Regions that have more probability around them have greater density.

"Near" $\theta = -1$, there is 5% of our total probability. "Near" $\theta = 0$

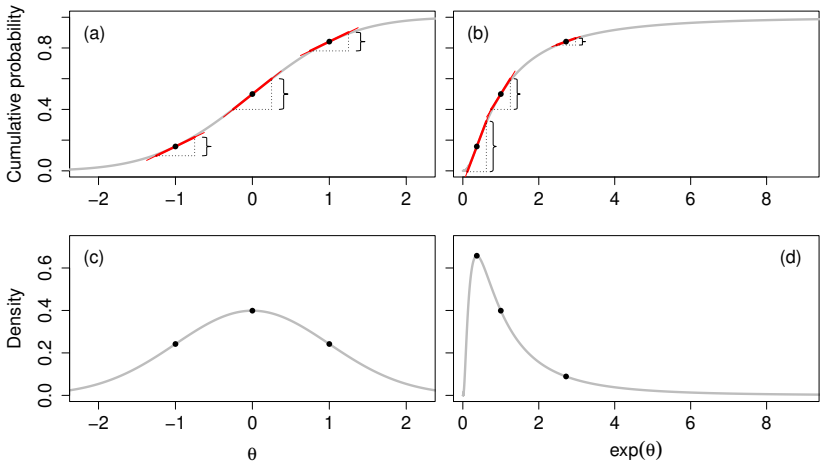

Fig. 2. An illustration of how transformation affects the CDF and PDF. Thin red lines show tangent vectors at the given points. Brackets indicate the change in the CDF within the neighboorhood of the highlighted parameter values. **(a)** The CDF of a normally distributed θ. **(b)** The CDF resulting from applying the transformation $\exp(\theta)$. **(c)** The PDF of a normally distributed θ. **(d)** The PDF resulting from applying the transformation $\exp(\theta)$.

there is 8% of our total probability. This distribution is symmetric around 0, so "near" $\theta = 1$, there is again 5% of our total probability. If we take these amounts and divide them by the length of the window $\Delta\theta$, we have an idea of how densely the probability is packed around each of these values. We can compute the density of these windows around -1, 0, and 1 to be .25, .4, and .25, respectively.

Thinking in terms of the CDF, how dense a region is around a parameter corresponds to how steeply the function rises near that point. The idea of probability density follows this line of thinking into the limit of smaller and smaller windows around θ values; we look at how much the CDF's value is changing with an infinitesimally small change around the parameter value. In other words, we can look at the slope of the CDF at a given parameter value to find its density. Thus, the posterior density function is the derivative (slope, or rate of change) of the cumulative distribution function. In this way, an individual parameter value can have a non-zero density, which critically distinguishes probability density from probability.

To summarize, the density of candidate parameter values is determined by first defining a window length to be applied to the regions of θ (i.e., $\Delta\theta$), finding the probability mass within that window, and then passing on to a limit. This results in taking the derivative of the CDF, telling us how much probability exists in a very small window "near" each candidate θ value. As we can see for our standard normal example, the highest-density point (i.e., the mode) is at $\theta = 0$ and the density gradually and symmetrically decreases in either direction.

The difference in the way transformations act on probabilities and probability densities is central to our argument. Equivalent sets of parameters must naturally have equal probabilities, but their densities can be quite different.

Reparameterization and transformation of variables. Probability density quantifies how much probability mass is "near" a point in the support of a probability distribution. When the probability distribution is a prior or posterior distribution of a parameter, it is tempting to think of the density as a measure of how "plausible" each parameter value is. However, this line of thinking fails when we realize that a parameterization of a model is merely an indexing system of a family of distributions that could have generated the data (Bernardo, 2005). Usually when we talk about inference, we make reference to parameter values such as θ, but what is being tested is the data generating process for the data, $f(x|\theta)$. For example, in a t-test we may make an inference on whether the difference between group means μ_1 and μ_2 equals zero, but this is merely an expedient way of determining whether $f(x|\mu_1)$ is the same function as $f(x|\mu_2)$. Parameters do not exist outside of a model for some data generating process.

When we think of parameterizations as indices of data generating processes, it becomes clear that we have to make a choice of which index to use in any given case. It may be that we choose based on convention or ease of interpretability, but we must make a choice; there is no God-given parameterization. And in making our choice, we must recognize that other analysts may make different choices. That is, we may prefer to think of the model in terms of a θ, but someone else may prefer to think in terms of ψ that is some function of θ. For example, in the case of generalized linear models we estimate parameters on one scale and interpret them on another. In the case that this function is bijective (both one-to-one and onto), we can say that ψ is a *reparameterization* of the model based on θ.

Let us now consider what happens to our standard normal posterior distribution when we reparameterize the model in terms of $\exp(\theta)$. This transformation is commonly applied to the output of generalized linear models to ease interpretation of the parameters. The top right panel shows the CDF of the posterior distribution for this new parameterization, known as the log-normal distribution. Note that the only change is to scrunch and stretch the axis in different regions. The height of the curve has no need of adjustment. Naturally, the probability that $\theta < 0$ is just the probability that $\exp(\theta) < 1$. This makes sense, as all we have done is assigned a new labeling scheme for the model, going from $f(x|\theta)$ to $f(x|\psi)$. For example, the data generating processes that have a θ smaller than 0 are precisely those that have a ψ less than 1.

In contrast to the simple way probability is maintained through transformation, the new density is not as straightforward. Recall that density is telling us how much probability is "near" different parameter candidates, and what is considered "near" a parameter value is defined as a window of width Δ. The critical problem for thinking about plausibility in terms of density is this: *data generating*

processes that are "near" one another in terms of θ may not be nearby in terms of ψ.

Consider the three θ values of -1, 0, and 1. In this space (represented by the horizontal axes in Fig. 2(a) and 2(c)), the data generating process corresponding to $\theta = -1$ and $\theta = 1$ are equally far from the one corresponding to $\theta = 0$. But the data generating process corresponding to $\psi = \exp(0) = 1$ is now much closer to the one corresponding to $\psi = \exp(-1) = .37$ than it is to $\psi = \exp(1) = 2.7$. However, critically, the probability mass of this interval must stay the same, as equivalent sets must have equal probabilities. Thus, the probability mass in the smaller region of ψ between $\exp(-1)$ and $\exp(0)$ must be more dense than the equally probable but more spread out region between $\exp(0)$ and $\exp(1)$. This is the central idea of density: the same amount of mass in a smaller region is more dense.

The story is the same as we shrink the window $\Delta\psi$ in the limit to find the density function. Because the probability mass has been stretched and scrunched differently in different regions of parameter space, some data generating processes that had high (low) density when expressed in terms of θ may have low (high) density when expressed in terms of ψ. From the ψ perspective, the densest region of parameter space is at $\exp(-1)$. The data generating process corresponding to $\psi = \exp(-1)$ has the most probability "around" it.

So far we have given an intuition of the idea of how posterior density changes when we transform our parameterization or representation of the model. Essentially, we have to ensure that we account for the differential stretching being done to the parameter space; regions that stretch out must become less dense, and regions scrunching up must become more dense. In the limit, as we look at smaller neighborhoods around the parameter values, we have to make finer grained adjustments to the density. The limiting adjustment is a factor called the *Jacobian*, and it takes a central role in our argument here. The Jacobian corrects the density up or down to the extent that the neighborhood around the given parameter is shrinking or expanding.

The Jacobian (described in more detail in the subsection "The Jacobian and coherence") is generally going to be a function of the new parameter, and can have a drastically different effect on different regions of the new parameter space depending on the transformation. It is perhaps instructive to see how the Jacobian works for a few points in the transformation of the parameter spaces for Avery and Cassidy from our hamster example earlier. In transforming from θ to $\mathrm{odds}(\theta)$, calling $\gamma = \theta(1-\theta)^{-1}$, the Jacobian is $(1+\gamma)^{-2}$. The point $\theta = .50$ for Avery goes to the new point $\gamma = 1$ for Cassidy, and in doing so it undergoes a density adjustment factor of $(1+1)^{-2} = 1/4$. The point $\theta = .70$ for Avery goes to the point $\gamma = 7/3$ for Cassidy, with a Jacobian adjustment of $(1+7/3)^{-2} = .09$. Finally, the point $\theta = .90$ goes to the new point $\gamma = 9$ with a Jacobian adjustment of $1/100$. The Jacobian is acting more strongly on the larger values of the parameter because the transformation of the space becomes more exaggerated as we approach the boundary of θ. Consider that $\theta = .9$ goes to $\gamma = 9$, $\theta = .95$ goes to $\gamma = 19$, and $\theta = .99$ goes to $\gamma = 99$. The probability around these points is being stretched across larger and larger regions of odds, so the density must be adjusted downwards accordingly. This differential adjustment due to the Jacobian will generally lead to a different set of parameter values having the highest density.

Critically, if we were to make the mistake of thinking of plausibility in terms of probability *density* (rather than probability proper), we would have to interpret the transformation from one parameterization to another as providing some new information – it would be as if the Jacobian were another source of data. But of course the adjustment due to the Jacobian does not introduce any new information; its sole purpose is to ensure that the probability density function integrates to 1.

The Jacobian and coherence. We now give a formal definition of the Jacobian and demonstrate it in the context of the log-normal example above. If a univariate random variable X has probability density function $f_X(x)$, then the density function of the random variable $Y = g(X)$, for a smooth function g with inverse $X = h(Y)$, is

$$f_Y(y) = f_X(h(y))\,|J(y)|\,. \qquad [2]$$

The factor $J(y) = dh(y)/dy$ is the Jacobian of the transformation, and rescales the density function to ensure that the probability density function integrates to 1. The Jacobian factor is a function of the new parameter, meaning that its value depends on where in the parameter space we are; neighborhoods that are shrinking will have a Jacobian greater than 1, and neighborhoods that are expanding will have a Jacobian less than 1.

Let us continue with our example of transforming from a normal random variable to a log-normal. Let X follow a normal distribution with mean μ and variance σ^2, that is, $X \sim N(\mu, \sigma^2)$. The density function of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

for $-\infty < x < \infty$. The mode of this distribution is that value of x that maximizes $\exp\{-(x-\mu)^2\}$, which is $x = \mu$.

Table 1 shows the Jacobian for many common transformations. The change of variable $Y = \exp(X)$ corresponds to the fourth row of the table with $a = 1$. Thus, we have inverse $X = \log(Y)$ and a Jacobian equal to $1/Y$. Equation 2 tells us that the density function of Y is given by

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\Big((\log(y)) - \mu\Big)^2\right\} \cdot \left|\frac{1}{y}\right| \\ &= \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\Big(\log(y) - \mu\Big)^2\right\}, \end{aligned}$$

for $0 < y < \infty$. It can be shown that the mode of this new density is at $y = \exp(\mu - \sigma^2)$.

If we were to incorrectly interpret density as an indication of "plausibility," then the point with highest density should be considered the "most plausible" value of a random variable (i.e., the mode). We would then have to simultaneously believe that the most plausible value of X is μ, but that the most plausible value of $Y = \exp(X)$ is not $\exp(\mu)$, but the potentially very different $\exp(\mu - \sigma^2)$. These two beliefs are contradictory, so treating density as a measure of plausibility leads to logical absurdity and is not a coherent system of reasoning.

Issues with the density-based HRDR. As shown in our discussion of density and transformation, regions of parameter space having "high" or "low" density depends on how we choose to parameterize the data generating process. The relative ordering of density can change drastically across reparameterizations of the model, meaning any inferences based directly on density are essentially artifacts of a statistically arbitrary parameterization choice.

The HRDR relies on the determination of overlap between a high-density region and an "equivalency" region of the parameter

Table 1. Some common transformations and their corresponding Jacobian factors. The first column lists various transformations in the form $Y = g(X)$. The second column lists the inverse functions, $X = h(Y)$. The third column lists the Jacobian factors corresponding to each transformation, obtained by taking the derivative of the corresponding inverse function (see the explanation in the main text). Note that Φ is the normal CDF and ϕ is the normal PDF. Finally, the rightmost column indicates whether the transformation is linear or nonlinear – the scale dependence of the HRDR occurs only with nonlinear transformations.

Transformation	Inverse function	Jacobian	Linear
$Y = g(X)$	$X = h(Y)$	$J(Y) = \frac{dh(Y)}{dY}$	
$Y = aX + b$	$X = \frac{Y-b}{a}$	$J(Y) = \frac{1}{a}$	Yes
$Y = X^a$	$X = Y^{1/a}$	$J(Y) = \frac{1}{a} Y^{\frac{1-a}{a}}$	No
$Y = X^{1/a}$	$X = Y^a$	$J(Y) = aY^{a-1}$	No
$Y = e^{aX}$	$X = \frac{1}{a}\log(Y)$	$J(Y) = \frac{1}{aY}$	No
$Y = a\log(X)$	$X = e^{Y/a}$	$J(Y) = \frac{1}{a} e^{Y/a}$	No
$Y = \frac{X}{1-X}$	$X = \frac{Y}{1+Y}$	$J(Y) = \frac{1}{(1+Y)^2}$	No
$Y = \frac{X}{1+X}$	$X = \frac{Y}{1-Y}$	$J(Y) = \frac{1}{(1-Y)^2}$	No
$Y = \log\left(\frac{X}{1-X}\right)$	$X = \frac{1}{1+e^{-Y}}$	$J(Y) = \frac{e^{-Y}}{(1+e^{-Y})^2}$	No
$Y = \frac{1}{1+e^{-X}}$	$X = \log\left(\frac{Y}{1-Y}\right)$	$J(Y) = \frac{1}{Y(1-Y)}$	No
$Y = \Phi^{-1}(aX + b)$	$X = \frac{\Phi(Y)-b}{a}$	$J(Y) = \frac{1}{a}\phi(Y)$	No

space. The issue with the procedure is now apparent: regions of parameter space can have high density in one parameterization and low density in another, meaning the location of the HDI relative to the ROPE is dependent on an arbitrary parameterization choice. Thus, the inference one draws from the HRDR is not invariant to reparameterization of the model.

In fact, the non-invariance of the HRDR to reparameterization suggests that the way we have written the decision rule associated with the procedure in the earlier section is incomplete; the entire decision process should be indexed by the parameterization choice made. Thus, the HDI should be explicitly HDI_θ, the ROPE should explicitly be ROPE_θ, and the decision rule should be explicitly δ_θ. Thus, in terms of decision theory, the problem with the HRDR is that with the same data and prior, it is not the case that the decision δ_θ will necessarily be equal to the decision δ_ψ associated with a reparameterization of the model. Different analysts with the same information can come to different conclusions based solely on their choice of parameterization. Note that the incoherence problem does not involve moving from one parameterization to another – the changing choice of parameters simply serves to illustrate the fact that the results of the HRDR depend on this statistically arbitrary choice.

Computational demonstration of transformation incoherence. To help develop a deeper understanding of these technical points, we will now demonstrate how the transformation incoherence of the HRDR can be seen in practice using modern computational tools. Modern Bayesian analysis is done using computational tools such as MCMC (van Ravenzwaaij, Cassey, & Brown, 2018). These computational methods are able to generate samples from a posterior distribution, which can then be used to draw inferences from the data. We can obtain the posterior samples from a reparameterized version of a model by applying the transformation directly to the original posterior samples (treating them as derived parameters). For example, if we can obtain samples from the posterior distribution of a binomial rate θ, and we wish to draw inferences about some transformation $g(\theta)$, we can simply apply the transformation to each sample from the original posterior to obtain the posterior samples for the derived $g(\theta)$.

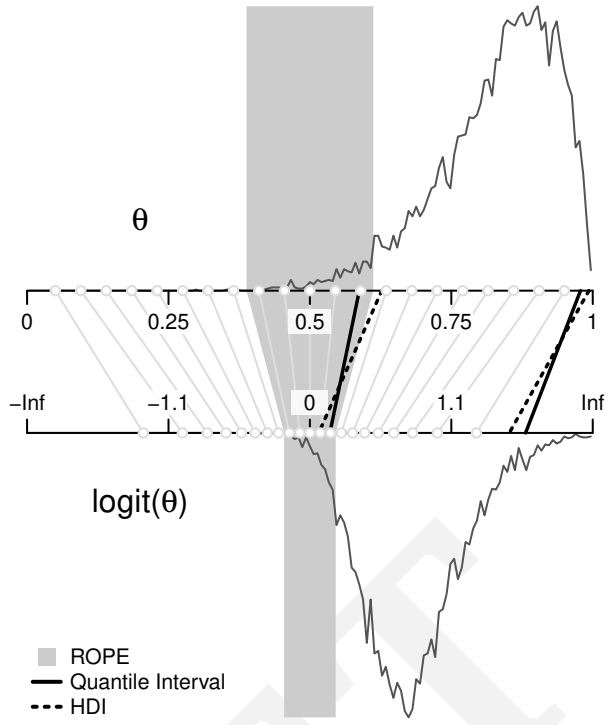

Fig. 3. A computational illustration of transformation incoherence. The distribution at the top contains samples of the posterior distribution over the rate of success θ, while the inverted distribution at the bottom contains samples of the posterior distribution over the transformation $\text{logit}(\theta)$. The connected points (open circles) highlight the uneven shrinking of the logit transformation, as does the gray polygon marking the ROPE limits in both scales. The 95% HDI, indicated with dashed black lines, switches from non-overlapping in the original θ scale to overlapping after the logit transformation. By contrast, the 95% quantile interval, shown with solid black lines, must retain its original overlap with the ROPE since the relative order of the posterior samples is preserved due to the monotonicity of the logit transformation.

Suppose we have 1000 posterior samples for θ, denoted θ_k for $k = 1, 2, \ldots, 1000$, ordered from smallest to largest. The 95% highest-density interval for θ can be found in two steps. Start by taking the first sample and the 950th sample and check the length of the interval they form, $\theta_{950} - \theta_1$. Then, check $\theta_{951} - \theta_2$, $\theta_{952} - \theta_3$, and so on in sequence until the length of all such possible intervals are computed. The shortest such interval will be the highest density interval.

Consider the reparameterization $\text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$. To obtain posterior samples from $\text{logit}(\theta)$, we simply apply the transformation to each θ_k. The reparameterization will not change the rank order of the samples, it will simply put them into a new scale. The effect of this transformation is shown in Figure 3. Here we are implementing the HRDR to test $\theta = .5$ with a shaded ROPE of $(.40, .60)$. The top panel shows the posterior distribution of 1000 samples taken from a $\text{beta}(10, 2)$ distribution. The bottom panel shows the distribution of the posterior samples after applying the logit transformation. The lines connecting the two panels show how values of θ are mapped into the space of $\text{logit}(\theta)$. Notice that equally spaced values of θ do not remain equally spaced after the reparameterization: values near $\text{logit}(\theta) = 0$ are closer to each other than points at the extremes. The effect of the reparameterization on the ROPE is shown by the shaded region connecting the panels.

When we obtain the HDI for the reparameterization, the procedure operates the same way on the transformed samples. But, critically, the relative distances between the transformed samples will not be preserved due to the fact that samples near the middle of the scale are now closer to each other than samples near the extremes. This means that some intervals that were relatively longer in the original parameterization may be shorter in the new one, and vice versa. The effect of this transformation is shown by the dashed lines connecting the two panels of Figure 3. Notice the shearing effect when compared to the rest of the traced lines. This shearing force is the Jacobian in action.

Because the HDIs must account for the density adjustment due to the Jacobian, they transform at a different angle. The transformation incoherence of the HRDR is demonstrated where the lower bound of the HDI crosses into the ROPE area.

The thick solid lines connecting the two panels show the 2.5% and 97.5% quantiles of the posterior distributions, which together form a central 95% quantile interval. As we explain below, quantiles are based on probabilities and therefore inherit coherence. The transformation coherence of quantile intervals can be seen by them operating at the same angle of the rest of the transformation. The quantile interval lines cross with the sheared HDI lines because the quantile intervals are unaffected by the Jacobian factor.

Kruschke (2010) provides code to produce HDIs from MCMC output and produce figures summarizing the results. In Figure 4 we provide code that can be used to replace the HDIs with quantile-based intervals. This simple fix brings the procedure back to a state of coherence while allowing the user to continue using the computational framework provided by Kruschke (2010) with minimal interruption.

```
QIofMCMC <- function(sampleVec,credMass=0.95){
  # Computes the quantile interval from
  #   (1-credMass)/2 to 1-(1-credMass)/2, where
  #   credMass is the target percentage of
  #   posterior samples to be included in the
  #   interval.
  # The typical value credMass=0.95 results in
  #   an interval from the 2.5% to the 97.5%
  #   quantile.
  alp <- (1-credMass)/2
  lim <- quantile(sampleVec,probs=c(alp,1-alp))
  return(lim)
}
```

Fig. 4. R code to compute quantile intervals instead of HDIs. This function has the same interface and can be used in the same way as the `HDIofMCMC` function provided by Kruschke (2010, p. 628-629).

Additional examples

Recognizing that it is possible to expose the scale sensitivity of the HDI by deriving the Jacobian of the transformation, we can now easily show other examples of models used in psychological science that are susceptible to incoherence if HRDR is used. In this section, we provide additional examples of statistical models whose varying parameterizations lead to incoherence.

Three parameterizations of the Rasch model. The Rasch model (Rasch, 1960) is one of the most common models in item response theory, with many useful applications and theoretical results well known to psychometricians. Its most typical parameterization is one where the probability of person p answering correctly on item i is a logistic function of the difference between p's ability θ_p and i's difficulty β_i:

$$P(X=1) = \left[1+e^{-(\theta_p-\beta_i)}\right]^{-1}.$$

However, at least two other parameterizations exist and have their own specific use cases (see, e.g., Batchelder, 1998; Crowther, Batchelder, & Hu, 1995):

$$P(X=1) = \frac{\alpha_p}{\alpha_p+\delta_i}$$

and

$$P(X=1) = \frac{a_p b_i}{a_p b_i + (1-a_p)(1-b_i)}.$$

Here, the parameters map to one another in the following ways: $\delta_i = e^{\beta_p}$, $\alpha_p = e^{\theta_p}$, $a_p = \left[1+e^{-\theta_p}\right]^{-1}$, and $b_i = \left[1+e^{-\beta_i}\right]^{-1}$. The Jacobians associated with these transformations are shown in Table 1. Since the associated Jacobians are a function of the parameters themselves (i.e., the stretching of the scale is not constant everywhere but depends on the value of the parameter itself), these transformations distort the scale in a way that causes incoherence when using the HRDR. Hence, making the innocuous switch from one of these parameterizations to another, keeping everything else the same, can cause a switch in the statistical decision that results.

Kimura phylogenetic model. Stepping outside of the field of psychology, a second example concerns DNA evolution. Models of DNA evolution can be thought of as descriptions of the process of nucleotide substitution (Zwickl & Holder, 2004). One family of DNA evolution model known as the "Kimura family model" assumes that all four nucleotides will be equally common, that the eight types of transversions will occur at one rate, and that the four types of transitions will occur at a second rate. Within the confines of these assumptions, the Kimura model family can generate predictions ranging from all substitutions being transversions to all being transitions. A key parameter in this model family is the transition/transversion rate ratio κ, which is used to describe this range of possibilities. The two ends of the spectrum of predicted substitution patterns are specified with $\kappa = 0$ (all transversions) and $\kappa = +\infty$ (all transitions). An alternative parameterization of this model family instead characterizes the proportion ϕ of substitutions that are transitions. This parameter can be written as a function of κ: $\phi = \kappa(2+\kappa)^{-1}$. The Jacobian corresponding to this reparameterization is $J(\phi) = 2(1-\phi)^{-2}$, which is again not a constant but a function of the parameter ϕ.

The circular drift-diffusion model. The circular drift-diffusion model (CDDM) is a process model used to describe response and response time data collected in tasks where the decision space is a circle (Smith, 2016). The model belongs to the broad category of sequential sampling models that assume that individuals accumulate information about the stimuli presented from the moment the trial starts until they are ready to input a response.

There are two common parameterizations of the model: one that describes the drift vector in terms of drift angle θ and drift length δ, and the other describing it in terms of horizontal drift μ_x and vertical drift μ_y. Moving between these two parameterizations involves transforming between Cartesian and polar coordinates:

$$\begin{aligned}\mu_x &= \delta\cos(\theta)\\ \mu_y &= \delta\sin(\theta).\end{aligned}$$

The case of reparameterization with multiple interdependent parameters is slightly more complicated than our previous examples. We do not go into further detail here, but the scaling factor that we need is the absolute value of the determinant of the transformation, which is:

$$\begin{aligned} ||J|| &= \left|\left| \begin{matrix} \frac{\partial \mu_x}{\partial \delta} & \frac{\partial \mu_x}{\partial \theta} \\ \frac{\partial \mu_y}{\partial \delta} & \frac{\partial \mu_y}{\partial \theta} \end{matrix} \right|\right| \\ &= \left|\left| \begin{matrix} \frac{\partial \delta \cos\theta}{\partial \delta} & \frac{\partial \delta \cos\theta}{\partial \theta} \\ \frac{\partial \delta \sin\theta}{\partial \delta} & \frac{\partial \delta \sin\theta}{\partial \theta} \end{matrix} \right|\right| \\ &= |[\cos(\theta) \cdot \delta \cos(\theta)] - [-\delta \sin(\theta) \cdot \sin(\theta)]| = \delta. \end{aligned}$$

Again the Jacobian is a function of a parameter.[1]

The three examples above serve to illustrate the ubiquity of nonlinear reparameterizations possible for scientific models across psychology and related fields. Despite a popular preference to represent models using certain "canonical" parameterizations (e.g., Kruschke, 2018), ultimately *the epistemic content of a model is not in its parameters but in the distributions it generates over data* (Villarreal, Etz, & Lee, 2023). There is always an alternative parameterization that could be meaningfully understood; good methods should give us the same conclusions across them all.

Broadening our scope from these example models, we conclude by pointing out that the ubiquitous framework of generalized linear modeling relies on the frequent application of "link functions" that map parameters between scales – often from one scale that is natural for the parameter (e.g., a probability that lives between 0 and 1) to one that is convenient for regression (e.g., its logit transform that lives on the full real line). All nontrivial link functions are nonlinear, and hence the HRDR will be incoherent in those common cases.

Coherent alternatives to the HRDR

The incoherence of the HRDR lies entirely in the choice of the HDI as a reference to compare with the ROPE. Because the HDI is defined with regard to the specific parameterization of the posterior distribution, and because density of any given data generating process can be high or low depending on parameterization, the set that we end up comparing to the ROPE depends on a statistically arbitrary choice. This means that any conclusions we draw from our hypothesis testing procedure depends critically on a choice of how the model is described.

The transformation incoherence we have highlighted here can only occur when sufficient posterior mass lies inside or outside of the ROPE. By definition, a 95% HDI is a set of parameter values with probability .95. This probability will remain inside/outside the ROPE no matter how we choose to reparameterize. Thus, we cannot have a case where the conclusion from an HRDR flips from rejection to acceptance – the maximum extent of the incoherence is to change a rejection or acceptance into an inconclusive result (or vice versa). This is still not very comforting, because any opportunity for inferences to change arbitrarily based on statistically irrelevant choices indicates a fundamental weakness of a procedure. In the remainder of this paper we will suggest some changes that could be made to the HRDR test to achieve coherence.

[1] The multidimensionality of the CDDM leads to additional complications. We could continue and derive the multidimensional posterior distribution corresponding to the new parameterization, but how to conduct the HRDR at that point is less immediate. We suspect most researchers would prefer to perform the HRDR using the marginal posterior of the parameter of interest. This involves integrating out any other non-focal parameters, such as δ. This adds more layers of complexity to the transformations involved, and we do not discuss them further here. Suffice it to say that the multiparameter HRDR is fraught with nonlinear relationships.

The HRDR suffers from transformation incoherence precisely because it uses density-based interval estimates. As we have demonstrated, the regions of parameter space with high density in one parameterization can have low density in another. This weakness can be overcome with a simple modification of the procedure: transitioning to probability-based decision rules. Because probability theory itself is coherent, methods derived directly from it will inherit that coherence.

Quantile intervals. The simplest transition to probability-based inference is to use quantile intervals, which already happen to be a very common output in Bayesian software. These intervals are constructed by taking the inner $100\alpha\%$ of the probability mass of the posterior distribution, leaving $\left(\frac{1-\alpha}{2}\right)\%$ of the mass outside the interval on either end.

Quantile-based intervals will naturally be transformationally coherent, because they are based directly on probability. Using the CDF, a quantile interval can be constructed by finding the parameter values with heights equal to .025 and .975. As we have seen, the only change that occurs in a CDF during a transformation is to stretch and scrunch the points along the x-asis – the function values at those points remain unchanged. Thus, tests based on quantile-interval endpoints will be transformationally coherent.

Transitioning to probability-based intervals is an easy fix to the problem of transformation incoherence because they are already in broad use. Most software that performs Bayesian inference, such as JAGS (Plummer, 2003) and Stan (Carpenter et al., 2017), will by default produce these intervals in a model summary. These are computationally easy to produce because all one needs to do is note which posterior samples correspond to the appropriate sample quantiles. Almost no software is producing density-based intervals by default, probably because it involves additional computational steps to find the shortest such interval.

Posterior mass in the ROPE. Once we move to a quantile-based interval for use in the HRDR test, an even easier alternative solution presents itself. Why not simply compute the posterior probability that the parameter is in the ROPE? The interval comparison required by the ROPE test is superfluous at this point. Recall that a 95% quantile interval is constructed by taking 2.5% and 97.5% quantiles of the CDF. But most posterior distributions are unimodal, meaning this interval is generally going to be a contiguous set of parameter values. Thus, for the interval to not intersect the ROPE, the probability the parameter is in the ROPE must be at most 2.5%. Thus, we could simply use a decision rule that rejects the null hypothesis when the probability it is in the ROPE is sufficiently low, and accept it when it is sufficiently high.

Indeed, directly computing a probability is the natural Bayesian approach to such a problem (Wellek, 2002). If we look back at our example with the triplets and their hamster, the set of parameters in the HDIs changed across the three parameterizations. However, the probability of the ROPE is approximately .02, sufficiently small to cast doubt on the null hypothesis regardless of one's choice of parameterization. We will next present an example of how this approach would impact the conclusions of a recently published result in the next section.

Alternatively to estimating the size of the parameter of interest, one may still be interested in evaluating the statistical evidence for or against the nullity of the parameter. To this end, one can compute the Bayes factor associated with the hypothesis that the parameter is inside versus outside the ROPE. Such a Bayes factor would compare the posterior odds that the parameter is inside the

ROPE to the corresponding prior odds, which is computationally convenient if the posterior is obtained via sampling.[2] This Bayes factor would tell us the extent to which the data are making us more or less confident that the parameter is inside the ROPE. Critically, this Bayes factor would be based on probabilities and thus would trivially maintain coherence.

Solution example. Newton et al. (2018) present a cluster randomized trial of the efficacy of an intervention program designed to reduce cannabis usage among Australian high school students. A total of 2190 students were randomly assigned to one of three possible intervention conditions or a control condition, with measures related to cannabis usage, related harms, and knowledge taken at baseline, and 6-9 months, 12, 24, and 36 months after the baseline (i.e., post-intervention measures).

The data analysis presented by the authors focused mainly on the differences between the control group and all intervention conditions. The primary statistical analysis was conducted using multilevel mixed-effects linear models that incorporated individual and school-level random effects. Using a classical hypothesis test, Newton et al. found no statistical differences in cannabis usage and related harms between the control and intervention groups. The authors then implemented an HRDR to quantify how much evidence the data provided in favor or against the null hypothesis. They defined a test value for the odds ratio $\theta = 1$ and a ROPE of $[0.9, 1.1]$. The authors note that the HDI for the odds ratio overlapped partially with the ROPE and concluded that their initial non-significant results were ambiguous and not indicative of a lack of true effect.

The data in this real-world example allow us to contrast the HRDR results with those achievable through using a direct probability approach. Table 2 presents results reported by Newton et al., their Table A2, using a collection of alternative probability-based measures corresponding to the data collected at each moment in time. In Table 2, the second column presents the median odds ratio and the range of the 95%HDI as reported by the authors; the third column lists the probability that the odds ratio value is contained in the ROPE $P(\text{ROPE})$; the fourth column shows the probability that the odds ratio is smaller than 0.9 $P(\theta < 0.9)$; and the rightmost column shows the probability that the odds ratio is larger than 1.1 $P(\theta > 1.1)$. Summing the fourth and the fifth columns we obtain the probability that the odds ratio falls outside of the ROPE (i.e., being either smaller than 0.9 or larger than 1.1), which by the complement rule of probability theory is the same as $1 - P(\text{ROPE})$. Probability-derived measures are coherent by definition.

As we can see, there is a decline over time in the probability that the odds ratio coincides with the ROPE. More specifically, by 24 months, the probability that the odds ratio is practically equivalent to 1 is only .061, and at 36 months, the probability is only .04. Critically, we note that despite the wide HDIs found at 24 and 36 months, we are relatively confident that the odds ratios at these times are outside the ROPEs.

This real-world example helps to illustrate the advantages of working with direct probability measures, which preserve the intuitive reasoning of the HRDR while also guaranteeing logical coherence.

[2] Note that the odds in question are equal to $P(\text{ROPE})/P(\neg\text{ROPE})$. If one obtains the posterior by simulation, this would correspond to the proportion of posterior samples inside versus outside the ROPE.

Table 2. A selection of results of the logistic regression analysis from Newton et al. (2018). The key statistic is the odds ratio θ.

Time	Median θ (95% HDI)	P(ROPE)	$P(\theta<0.9)$	$P(\theta>1.1)$
6m	0.87 (0.74 to 1.02)	.351	.646	.003
12m	0.76 (0.54 to 1.02)	.138	.851	.011
24m	0.58 (0.27 to 1.01)	.061	.913	.026
36m	0.45 (0.12 to 0.99)	.040	.929	.031

Conclusion

We have highlighted a critical flaw with the HDI+ROPE or HRDR method: transformation incoherence. This flaw can lead different researchers with the same priors and data to draw different conclusions because they have arbitrarily chosen different parameterizations of the problem. We have shown that the root cause of this incoherence lies with the choice of using highest density intervals in the test. Because these intervals are constructed with reference to a specific density function from one specific parameterization, the change from one chosen parameterization to another causes the set of highest density points to change; where the original set may have excluded the ROPE, the new set may intersect with the ROPE.

We are not the first to point out the problems with thinking of density as a measure of "plausibility." However, most discussion tends to focus on the consequences this thinking has on coherent specification of priors. Perhaps the most famous objection to this line of thinking was made by Fisher (Lehmann, 2011), who argued against the use of uniform priors to represent ignorance because they will only be uniform in one parameterization of a model (for a thorough demonstration see Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017). Zwickl and Holder (2004) provide a clear illustration of the problem with taking uniform priors to represent ignorance in the context of the General Time-Reversible Model, in which it common to reparameterize in terms of either κ or $\phi = \kappa/(2+\kappa)$. Uniform priors in either κ or ϕ lead to highly informative priors on the other scale.

Similar considerations led Jeffreys (1946) to develop the now famous invariant Jeffreys prior rule. Druilhet and Marin (2007) propose a class of prior densities that lead to invariant highest density sets, avoiding transformation incoherence of the HDI. However, their solution imposes specific choices of Jeffreys-type priors that may not be desirable or may not even exist in practice for models with sufficient complexity. This led Druilhet and Pommeret (2012) to develop new invariant conjugate families of priors that can be applied when the model is a member of the exponential family.

In addition to issues in prior specification, issues with using density-based intervals in estimation and testing have been pointed out by others (Bernardo, 2005; García & Oller, 2006; Robert, 1996; Shalloway, 2014). For instance, Bernardo argues against using density-based intervals in favor of a parameterization invariant "intrinsic" intervals based on the structure of the likelihood function. Our argument may be interpreted as an extension of Bernardo's to the specific case of the HRDR test. However, we prefer probability-based intervals over "intrinsic" intervals, if for no other reason than that they are computationally and conceptually simpler and thus easier to adopt in practice. More broadly, for the reasons we discuss in this paper and those presented by Bernardo and others, we do not believe HDIs are useful tools for Bayesians.

We have focused primarily on the simplest case where this problem arises, namely, models with a single parameter. The potential for incoherent inferences only increases when considering

models with simultaneous hypothesis tests of multiple parameters, as the multivariate Jacobian factor must account for stretching and scrunching across all dimensions of the parameter space.

Critically, all parameterizations of a model are equally valid—there is no "one true parameterization" for any model—so the inference from the HRDR test critically hinges on what may be considered an arbitrary choice made by the researcher. We suggest a change to the HRDR test to remedy this incoherence: use probabilities instead of densities. A direct probability approach is both conceptually and computationally simpler than one based on densities. Probabilities, by their very construction, must lead to coherent inferences.

References

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344.

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 283–302.

Bernardo, J. M. (2005). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test*, *14*(2), 317–384.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*.

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

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement–theoretic analysis of the fuzzy logical model of perception. *Psychological Review*, *102*, 396–408.

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

Druilhet, P., & Marin, J.-M. (2007). Invariant HPD credible sets and MAP estimators. *Bayesian Analysis*, *2*(4), 681–691.

Druilhet, P., & Pommeret, D. (2012). Invariant conjugate analysis for exponential families. *Bayesian Analysis*, *7*(4), 903–916.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34.

García, G., & Oller, J. M. (2006). What does intrinsic mean in statistical estimation?(invited article with discussion: Jacob burbea, joan del castillo, wilfrid s. kendall and steven thomas smith). *SORT-Statistics and Operations Research Transactions*, *30*(2), 125–170.

Gibson, E. W. (2021). The role of p-values in judging the strength of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research*, *13*(1), 6–18.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461.

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

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial introduction with R and BUGS*. Burlington, MA: Academic Press.

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

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. Springer Science & Business Media.

Lindley, D. V. (1998). Decision analysis and bioequivalence trials. *Statistical Science*, *13*(2), 136–141.

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, *80*, 40–55.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245.

Newton, N. C., Teesson, M., Mather, M., Champion, K. E., Barrett, E. L., Stapinski, L., … Slade, T. (2018). Universal cannabis outcomes from the climate and preventure (CAP) study: a cluster randomised controlled trial. *Substance Abuse Treatment, Prevention, and Policy*, *13*(1), 1–13.

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*. Vienna, Austria.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedogogiske Institut.

Robert, C. P. (1996). Intrinsic losses. *Theory and decision*, *40*, 191–214.

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

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

Shalloway, D. (2014). The evidentiary credible region. *Bayesian Analysis*, *9*, 909-922.

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

Smith, P. L. (2016). Diffusion theory of decision making in continuous report. *Psychological Review*, *123*(4), 425–451. doi:

van Ravenzwaaij, D., Cassey, P., & Brown, S. (2018). A simple introduction to Markov chain Monte-Carlo sampling. *Psychonomic Bulletin and Review*, *25*, 143–154.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

Villarreal, M., Etz, A., & Lee, M. D. (2023). Evaluating the complexity and falsifiability of psychological models. *Psychological Review*.

Wellek, S. (2002). *Testing statistical hypotheses of equivalence*. Chapman and Hall/CRC.

Zwickl, D. J., & Holder, M. T. (2004, 12). Model Parameterization, Prior Distributions, and the General Time-Reversible Model in Bayesian Phylogenetics. *Systematic Biology*, *53*(6), 877-888.