ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 2, *pp.* 449–469

Discussion on the meeting on 'Signs and sizes: understanding and replicating statistical findings'

Jane L. Hutton (*University of Warwick, Coventry*)

I thank the authors for their interesting approaches to p-values.

Rice, Bonnett and Krakauer remind us to 'specify which exact question is being addressed' and then to focus on 'testing a hypothesis about the sign of a real univariate parameter θ' rather than on estimation. It is useful to think through the costs of decisions, as they do for the Bonferroni approach to multiple testing, in which non-decisions are cheap, active decisions are expensive and the tests are conservative.

Held proposes sceptical prior distributions to assess whether a finding is credible, through downweighting the p-value of the replication study. He puts analysis before design in the title. Perhaps this explains the statement that there is 'no established standard for the statistical evaluation of replication', and his focus on how 'claims of new discoveries can be confirmed'. As Professor Held knows, drug and medical device regulations have established standards by which claims of efficacy or equivalence are assessed.

Assessing whether 'the absolute value of the effect estimate' overlaps some target fails to address the problem of an original effect estimate which was derived by excluding 30% of participants because they did not conform to the researchers' assumptions (Hardisty and Weber, 2009; Lim, 2019). In considering the intrinsic credibility of unexpected findings, I recommend considering selection biases as well as publication biases. If 'p-hacking' is suspected, bias due to within-study selection of the most significant' results can be estimated (Hutton and Williamson, 2000; Hahn *et al.*, 2000).

An obscure old article proposed a Bayesian approach to sample size estimation that is relevant in circumstances in which a parameter has been precisely estimated by one scientist, but is not believed by a wider audience, and a further study is planned (Hutton and Owens, 1993). The method considers how much information is required when there are conflicting prior distributions, allowing for uncertainty about the 'true' parameter. The power calculations can aim at convergence or overlap of posterior credible intervals for rational scientists with different priors (Hutton and Owens, 1993). The sceptical priors that are proposed by Held are similar in spirit, but the replication study estimate is required to conflict with the *post hoc* prior distribution. This is implicitly a one-sided approach, though more stringent than testing the sign of the parameter (Rice and his colleagues). The old and the new Bayesian approaches both inflate the sample size that is required for a confirmation study (Hutton and Owens (1993) and Held).

However, the essential question is *not* 'To p or not to p?' (Leek and Peng, 2015).

A convenient timeline of the replication crisis in social science is given by Andrew Gelman in his 2016 blog 'What has happened down here is the winds have changed' (`https://statmodeling.stat.columbia.edu/2016/09/21what-has-happened-down-here-is-the-winds-have-changed/`). Null hypothesis significance testing, rather than good design, dominated decisions on publication. Eventually, concern about 'false positive psychology' and lack of replication resulted in the Open Science Collaboration which tried to replicate 100 studies published in 2008 in three high ranking psychology journals (Open Science Collaboration, 2015). This collaboration still focused on p-values.

How can a study be replicated if the published article does not give the aims, study population, design, summary statistics or estimated effects? A checklist developed from the guidelines in Wilkinson (1999) was used to evaluate a random sample of 30 of the 100 Open Science Collaboration (2015) studies. The maximum score was 38% (Lim, 2019). Estimating how many of the remaining 70 studies would score more than 40% is left as an exercise for the reader.

Were all 100 Open Science Collaboration (2015) studies worth replicating? Academics are under pressure to publish and might find little time to ask 'why am I doing this?'. One replicated study was a series of experiments on remembering sounds, with only three or four subjects, two of whom were the authors (Demany *et al.*, 2008). The authors asserted that the subjects were 'four audiometrically normal listeners'. Normal?: surely variations in hearing, for example with age, are important. Without a specified target population, the relevance of results is unclear. The conclusion was that 'our data suggest that change detection is based on different mechanisms in audition and vision'. That a 'high ranking psychology journal' publishes a conclusion about vision based only on listening is worrying. Ethical professional actions are dependent on the quality of the underlying knowledge (Hutton, 1995).

0964–1998/20/183449

Good research requires a clear aim, appropriate design, thorough data management, exploratory data analysis, model fitting and checking. Only then can summaries of effect sizes be provided, and the implications for future studies considered.

Social science researchers would benefit from considering the principles behind medicines regulation, and the EQUATOR Network guidelines for reporting of medical research (`www.equator-network.org`). Guidelines for assessing association and causation in epidemiology would also be relevant in psychology (Hill, 1965).

The vote of thanks was passed by acclamation.

Peter J. Diggle (*Lancaster University*)

One property of a sceptical p-value is that it can be no smaller than its corresponding standard p-value. What would be the relationship between the two if the original and replication studies were identical, both in design and in the results that they produce? If a reproducibility study confirms the original study's significant findings, should not that reinforce, rather than dilute, the original findings?

Most statisticians would agree that estimation is often more interesting than testing. What would be a suitable definition of a sceptical confidence interval, and how would the duality between standard confidence intervals and p-values translate to their sceptical counterparts?

Even in the context of randomized trials where significance testing plays a central role for regulatory approval purposes, I would much prefer sample size calculations to be couched in terms of precision of estimation rather than power of testing, for two reasons. Firstly, this avoids the need to specify a seemingly arbitrary value for the required power at the clinically significant difference (CSD); where did the 80% convention originate? Secondly, I have never understood why a significant result in a conventionally powered trial can be regarded as a positive result. If the CSD represents an effect size that is sufficiently large to be of material benefit, should a positive result not be one that conveys some degree of confidence that the actual effect size is at least equal to the CSD?

Sheila M. Bird (*Medical Research Council Biostatistics Unit, Cambridge*)

I congratulate both speakers on their excellent presentations: especially helpful as I had not pre-read their papers.

I offer brief remarks about rules of thumb. First, I generally have a relatively strong prior expectation that any replication effect size will be lower—especially if the discovery study was itself a first. My personal rule of thumb is that effect size may halve so that four times as many subjects may be needed in the replication (Bird and Hutchinson, 2003) as in the discovery study (Seaman *et al.*, 1998). This being so, if the replication effect size is larger than the discovery effect size, should I be surprised ... unless I can 'excuse' the finding as due to better adjustment for covariates (Pierce *et al.*, 2018) or some other rationalization?

Professor Diggle asked about the origin of preference for 80% power. My reasons are ethical and pragmatic. Typically, the achieved 95% confidence interval is informative when the trial design has delivered 80% power in respect of a plausible prior effect size; and two such trials deliver over 90% power even for testing at the 1% level of significance. Better still, for 80% power and 5% significance level, with 50:50 randomization and a measured outcome, the required number of patients per randomized group is easily derived as

$$(0.84 + 1.96)^2 \text{ or } 8 \text{ times} \{2 \times \text{common variance}/(\text{target difference in means})^2\}.$$

80% power with a multiplier of 8 is a useful back-of-envelope mnemonic, especially as a multiplier of 4 *or* $(0 + 1.96)^2$ gives the number to be randomized for 50% power with 5% level of significance. And no-one should embark on a randomized controlled trial that does not have at least 50% power to give the right answer (Turner *et al.*, 2013) ....

Christian Hennig (*University of Bologna*)

I enjoyed reading the paper by Rice, Bonnett and Krakauer, and I agree that it gives a new view on two-sided tests that can contribute to their better understanding. What the paper develops, however, is its own method, closely related to tests and p-values, but not the same. Therefore I think that it is wrong to claim that is gives a new 'motivation' for two-sided tests and p-values.

I disagree with the (not very elaborate) criticism that 'point null hypotheses are usually scientifically implausible and hence only a straw man' that motivates the claim that two-sided tests lack a proper motivation. Starting from the idea that 'all models are wrong', it should be clear that, when testing a point

null, we cannot attempt to confirm its truth (I therefore object to the term 'accept' for interpreting non-rejection). Rather the question is whether the observed data are compatible with data that are generated from the null hypothesis (in the sense defined by the test statistic), in which case they cannot be used to claim evidence for anything else. This is what the two-sided test addresses, and I do not have issues with it that would require Rice and his colleagues' 'motivation' to sort them out. I do not see why considering a point null hypothesis is worse than relying on a parametric model in the first place; if this is required to hold literally, it is also clearly unrealistic.

I think that a major issue with statistical tests is that they are usually interpreted in terms of deciding what the 'true model' is (either the null or the alternative), rather than taking explicitly into account that models are helpful tools for thinking rather than 'true' reflections of reality. The authors' Bayesian approach does not help with that, because it is based on a prior and posterior, often interpreted misleadingly as encoding probabilities that certain models are true.

The method presented by Rice and his colleagues has its own merits and interest; however, in my view two-sided tests are just fine without it, if interpreted carefully.

Nick Longford (*Imperial College London*)

I wholeheartedly welcome the proposal of Rice, Bonnett and Krakauer to engage decision theory in problems that are at present addressed by hypothesis testing. The end product of a hypothesis test of any consequence is the choice of a course of action: either that associated with the hypothesis, or that with the alternative. That is a decision, and we all agree that hypothesis testing is neither suitable nor intended for making decisions. In variance with Rice and his colleagues I endorse the statement of Cox (1982) as factually correct, but I regard it as grounds for dismissing hypothesis testing from statistical practice altogether.

Unfortunately, Rice and his colleagues do not involve in their discourse the client for whom the analysis is conducted, and who wants to inform their agenda by statistical analysis. An analysis is distinctly second rate if it ignores the client's perspective, value judgements, priorities or remit. Outside the confines of statistics, we make decisions in the presence of uncertainty by weighing the (hypothetical) probabilities of inappropriate choice, as done in hypothesis testing, and integrating them with the consequences (ramifications) of the (two) kinds of error that may be committed.

Hypothesis tests and any other method that is oblivious to these consequences is inconsequential, and that disqualifies them from modern statistical practice. Arguably the use of non-informative priors is another example of such a disqualification—is there really no prior information? The rationale for being Bayesian is to use such information, but then we declare such information to be vacuous. The eye of the analysis is firmly off the inferential ball.

Treating all biomedical applications with one α and those in physics with another (Section 1.1) is too crude and simplistic (Longford, 2013, 2016). If adopted, the proposal of Rice and his colleagues will ossify into a new computational ritual unless we engage the client in the formulation of the loss function. This is not a technical or methodological problem, but a problem of the prevailing scientific culture in which we abdicate the responsibility to tailor the details of the analysis to the client's perspective and fail to conclude the analysis by a proposal for what to do next—to choose one of the contemplated courses of action. Hypothesis testing is of immense historical importance in statistics, like the steam engine is for locomotion. That is a good reason for finding for it a prominent place—in a museum.

Maya B. Mathur and Tyler J. VanderWeele (*Harvard University, Cambridge*)

To assess replication success, Held quantifies the extent to which a Bayesian prior on the original study's effect would need to be 'sceptical' to shift the credible interval to include the null, and then he considers evidence from the replication against this prior. His methods sensibly represent asymmetrical information between the original study and replication, and they account for the magnitude of the replication's point estimate, not just its 'statistical significance'. We consider these methods best suited to replications of inherently implausible original findings; otherwise they may be conservative. A prior mean of 0 is suitable when, *a priori*, the most likely effect size is 0, and effects in either direction are equally plausible. In these cases, the replication may indeed provide the only source of evidence against the sufficiently sceptical prior. However, in many replication contexts, the original studies are likely to have true effect sizes larger than 0; Open Science Collaboration's (2015) replications suggested $r = 0.197$ on average. Increasing the prior mean accordingly would increase apparent replication success, so we suggest using several choices, ideally informed by existing replication efforts in similar domains.

Held's power analyses aid principled determination of which original studies are sufficiently informative on their own to be good candidates for replication. Interestingly, his analyses suggest that, in the context of

Table 1. Proportion of successful replication based on different definitions in the Open Science Collaboration (2015) study

Statistically significant ($p < 0.05$) result in the same direction in the replication study	36%
Statistically significant ($p < 0.05$) result in meta-analysis of original and replication study	68%
Original effect sizes are in the 95% confidence interval of the replication study effect	47%
Subjective assessment that the replication study replicated the original result	39%
Effect size in replication study at least as large as (or larger than) in the original	15%
Replication success (*per* Held)	15%†

†Based on analysis of 73 of the 100 topics by Held by using the meta-analytic subset of Johnson *et al.* (2016). Only six of the 11 successful replications have an effect size in the replication study that is larger than the effect size in the original study. Therefore, even though the replication success criterion and the criterion of asking for as large or larger effect size appear equally demanding, they select for very different studies.

inherently implausible original findings, those with p-values above 0.01 are poor candidates for replication; near that point, the sample size required for a well-powered replication study asymptotes to ∞.

Other considerations are also relevant to replication success. One can assess the extent to which the original study is statistically consistent with the replication study, in that its data are compatible with the same underlying population parameter that is unbiasedly estimated in the replications (Mathur and VanderWeele, 2019). This method assesses whether the replication and original studies' point estimates are more distant than expected by chance, accounting for statistical uncertainty in both, and hence are conceptually different from Held's reverse Bayesian sensitivity analysis on the original study. For example, the consistency metric accommodates 'non-significant' original studies and those with *a priori* plausible or implausible findings; it also applies readily to multisite replication studies, accounting for effect heterogeneity (Mathur and VanderWeele, 2019). With multisite replications, one can also reassess the strength of evidence for the effect under investigation by meta-analytic methods to estimate the proportion of true effect sizes in the replications that are stronger than a chosen threshold of scientific importance (Mathur and VanderWeele, 2019).

John P. A. Ioannidis (*Stanford University*)
Held proposes an ingenious method to assess the replication success for a replication study. The method builds nicely on analysis of credibility and more specifically on the concept of the scepticism limit (Matthews, 2018). The method proposed adds a sceptical member in an already long list of commonly applied criteria in judging whether replication has been achieved. Table 1 shows how different the conclusions can be depending on how one judges replication in the Open Science Collaboration (2015) project. As shown, the replication success criterion of Held is eventually as demanding as asking the replication study to have at least as large or larger effect size estimate compared with the original study—even though statistically significant original discoveries already tend to have inflated results (winner's curse).

There is large variability of opinion about the extent and causes (if present) of the 'reproducibility crisis'. Investigators whose findings are challenged as non-reproducible and those who have advocated for methods whose performance is severely challenged in the current reproducibility checks' environment might espouse mostly the lower non-replication rate estimates. Many other investigators see much larger problems of non-reproducibility reflected in the accumulated evidence. Held's method may be more popular mostly with the latter group.

The following contributions were received in writing after the meeting.

Christine P. Chai (*Microsoft Corporation, Redmond*)
In Held's paper about replication studies, my initial reaction is 'What is the unique contribution?' because of the title words 'a new standard'. Most new standards provide limited benefits because they simply add to the competing pool (Randall Munroe, `https://xkcd.com/927`). But, after reading the full paper, I found useful perspectives from the sceptical p-value.

One concept I like is the intrinsic credibility. Before evaluating the replication study results, we should check the original study's credibility to ensure possible reproducibility. Ioannidis (2005) expressed concerns that many published research findings are false, so we should be cautious about the existing literature's correctness. Another aspect I like is the potential extension to multiple replication studies using a sequential

approach: compare the original study with the first replication, and combine them into the *new original effect estimate* to evaluate the second replication. Hence, if the chain of replication studies fails, we could quickly identify the pipeline leak.

A future extension can have many original studies and one replication study from the same experiment or observation. A meta-analysis of the original studies is needed to create the benchmark. For example, consider the relevant studies of 'Drug A is more effective than drug B'. Study 1 shows that drug A is 50% more effective, but study 2 shows only 30%. Our replication study shows 60%, so we need to know which result(s) to believe. Since Professor Held has established how to evaluate replication studies, so the framework would benefit the science development. It is important to have statistical approaches that are better aligned with scientific practice (Goodman, 2016).

The paper of Rice, Bonnett and Krakauer is a theoretical paper on two-sided tests. I agree with the controversy of *p*-values (Wasserstein and Lazar, 2016), but I am looking for applications of the decision-making framework. Moreover, I do not understand the scenario where we ignore the extremely weak data and report no decision. 'No decision is still a decision' (Farrell, 2008; Hubbard, 2013) If an enterprise does not make a decision, they would continue the same production process. The same conclusion is made when they fail to reject the null in hypothesis testing. I would also like to see the details about assigning loss values, since not everything has a monetary measure. The example of physics's 5σ-threshold for discovery (Lamb, 2012) is far from the day-to-day work of statisticians. An A–B testing example in digital marketing would be more appropriate to demonstrate the loss function.

(The opinions and views expressed here are those of the author and do not necessarily state or reflect those of Microsoft.)

David L. Dowe (*Monash University, Clayton*)

An early criticism of *p*-values is due to C. S. Wallace (later the originator of the information theoretic Bayesian minimum message length principle (Wallace and Boulton, 1968; Wallace and Freeman, 1987; Wallace and Dowe, 1999; Wallace, 2005)) in 1954 (as relayed by him in 2003) (Dowe (2008), section 1, pages 549–550), where we hear of a *p*-value of the order of 1% on a small data set but which became insignificant as more data were collected. Wallace was later recorded commenting words to the effect that '... publication outlets which will not publish negative results (except in response to an earlier claim of a positive result) are introducing a bias' (Dowe (2008), section 0.2.5, pages 538–539).

Given that seemingly significant *p*-values often drop on collection of further data—and possibly also because of publication biases alluded to above—for Held's proposed new standard, 'the sceptical *p*-value p_S increases with decreasing absolute replication effect estimate relative to the original effect estimate' (Section 3.1).

Regarding Held's discussion (Section 6) of original, replication and extending to several replication studies, I would initially next like to see Section 3.2, equations (12) and (13) revisited (as a starting point) with original and two replications. As a further exercise, if there are more replications, then can anything further be said about how many studies will have to be significant as a necessary requirement for replication success—and will (and should) this depend on the order of replications?

Rice, Bonnett and Krakauer seem to simplify the hypothesis testing problem somewhat by reducing it to determining the *sign* of a (one-dimensional) parameter. In the cases where they have 'no decision' ($d = N$), I would be interested in whether they might sometimes be able and willing to say something about a confidence or a probability—and, more specifically, what might be able to be said about how such a confidence or probability might change (or why it should not change) over the range for which $d = N$. In case there is freedom in choosing loss functions, some arguments advocating the logarithm of probability as a loss function are given in Dowe (2008), footnotes 175 and 176, Dowe (2011), section 3, and Dowe (2013), section 4.1.

For those interested in the minimum message length approach to hypothesis testing—which is Bayesian—see, for example, Dowe (2008), section 0.2.5, page 539, and section 0.2.2, page 528, and Dowe (2011), pages 919 and 964.

John Ferguson and Nicola Fitz-Simon (*National University of Ireland Galway*)

We thank Held, and Rice, Bonnett and Krakauer for interesting and stimulating papers. At a general level, we suggest that ingenious attempts to redefine *p*-values and to justify thresholds do not address the real problem: that *p*-values are misused in practice. Instead of being viewed as one of many useful summaries of the evidence provided by a study, they are made less informative by thresholding, and then treated as the primary result.

More specifically, Professor Held discusses desirable properties of the proposed sceptical p-value; we find that is also has some strange properties. In particular, we find the property that $p_S \geqslant \max\{p_o, p_r\} \geqslant p_o$ quite counterintuitive. Effectively, this says that, no matter how strong the evidence of an effect in the replication experiment, the hypothesis of no treatment effect cannot be rejected at a level α if $p_o \geqslant \alpha$. For instance, assuming equal variances of the effect estimates in the two studies, if $p_o = 0.049$, p_r needs to be smaller than 10^{-97} to attain 'sceptical significance' ($p_S < 0.05$). Concluding no evidence of a treatment effect in spite of definitive evidence in a well-designed replication study is a peculiar property. The pair $\{p_o = 0.005, p_r = 0.005\}$ gives a smaller sceptical p-value; that this provides stronger evidence of a treatment effect raises doubts about the utility of p_S as a measure of evidence. We feel that it is more natural to place higher emphasis on the p-value from the replication test, which is by assumption a test of a preselected hypothesis.

We view Professor Rice and colleagues' paper as an interesting theoretical contribution, but we question the helpfulness of a decision theory paradigm to practitioners. In practice, quantifying the relative losses for different decisions may be time consuming and difficult. It is easy to specify that incorrectly concluding $\theta_1 > \theta_2$ is 20 times worse than making no decision but, unless one can clearly dictate the consequences of each decision, the choice of 20 seems arbitrary. Furthermore, the connection between p-values and posterior probabilities is asymptotic; and linking the Bayesian and frequentist paradigms in typical small studies is problematic. Given such a study, where a sceptical prior is indicated, a small p-value might be incorrectly interpreted as strong evidence of an effect. In practice, researchers wishing to apply these methods should work from a fully Bayesian paradigm with carefully elicited priors; using these decision theoretic thresholds in frequentist tests might lead to even more false positive results.

Tim Friede and Christian Röver (*University Medical Center Göttingen*)

Congratulations go to Professor Held on a very timely and principled contribution to the debate of reproducibility (Held, 2019a). Here, we highlight an issue that we consider relevant, but which has not had the attention that it deserves in this context.

In practical applications, researchers commonly encounter a certain amount of inconsistency (beyond measurement error) in experiments that are actually aiming to estimate the same parameter. When such data are combined in a meta-analysis, this is typically dealt with by introducing an additional variance component, the *between-study heterogeneity* (e.g. Hedges and Olkin (1985), Kontopantelis *et al.* (2013) and Turner *et al.* (2015)).

In Held's paper the model underlying the sceptical p-value may be written as

$$\hat{\theta}_i = \theta + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma_i^2) \tag{1}$$

with two independent error terms ϵ_i with known variances σ_i^2, where $i \in \{o, r\}$. Introducing a random effect h_i for the heterogeneity, we recover a special case (for $k = 2$ estimates) of the normal–normal hierarchical model given by

$$\hat{\theta}_i = \theta + h_i + \epsilon_i \qquad h_i \sim N(0, \varsigma^2) \tag{2}$$

(Hedges and Olkin, 1985). The conditional variance here is $\mathrm{var}(\hat{\theta}_i|\varsigma) = \sigma_i^2 + \varsigma^2$. If ς is unknown with some prior distribution, then the marginal variance is given by $\mathrm{var}(\hat{\theta}_i) = \sigma_i^2 + E[\varsigma^2]$.

With this generalization comes the notion of 'study-specific effects' $\theta_i = \theta + h_i$ that may in general differ from the 'overall effect' θ for each individual study i. Assuming the more general random-effects model, it may now not be sufficient for a study to show a significant *study-specific effect* alone ($\theta_i > 0$); rather, one may want to demonstrate more generally a convincing *overall effect* ($\theta > 0$), or even a condition of the type $\theta > z\varsigma$ (for some $z > 0$) may be desired.

Adding the heterogeneity variance component raises the bar for the evidence required for 'replication' substantially. What makes things worse is that the actual amount of heterogeneity ς is usually highly uncertain, and the two estimates $\hat{\theta}_i$ provide only very little evidence (Friede *et al.*, 2017). The issue, however, is not unique to the sceptical p-value approach proposed by Professor Held but arises more generally in problems of reproducibility.

A. P. Grieve (*UCB Pharma' Slough*)

I enjoyed the interesting paper by Professor Held and have a story to share.

In Grieve (1992) I described applications of Bayesian ideas in the context of determining the LD50 to classify the toxicity of agrochemical products. One application concerned a claim by a national authority that a new formulation of the pesticide Basudin had an LD50 of the order of 200 mg kg^{-1} or less. Historically, three studies in rats had established the LD50 in the range 780–1015 mg kg^{-1} with fiducial limits

Table 2. Predictive distribution of r deaths out of 10 animals receiving 200 mg kg^{-1} of Basudin

r deaths out of 10 animals	*Predictive probability*
0	0.930
1	0.062
2	0.007
3	0.001
>3	0.000

which excluded 200 mg kg^{-1} so we judged it highly unlikely that the LD50 could be as low as 200 mg kg^{-1}. This was confirmed by Bayesian analyses using uninformative priors giving posterior probabilities of less than 0.0001 that the LD50 was less than 200 mg kg^{-1} in each study.

Prima facie this was a clear case of non-replication given that there was no biological rationale for a small change in formulation having such a huge effect. Let me describe what we did.

Given pressures to reduce the numbers of animals and under the strong conviction that no substantial change in the toxicity had taken place, it was deemed inappropriate to perform a full LD50-study. We chose a predictive approach in which a dose of 200 mg kg^{-1} taken from each of 10 batches was given to 10 rats. Using the posterior distributions of the three studies we predicted the number of deaths from the 10 tested rats per batch. One of the three predictive distributions is shown in Table 2.

The results from the 10 batches conformed to expectation: no deaths in each cohort of animals. We concluded that the new formulation did not have an increased toxicity. This analysis rested on the assumption that all experiments—those original, those on which the claims for increased toxicity were based and the new experiments—were conducted in a similar fashion and under similar conditions. The importance of this assumption was underlined when it was subsequently revealed that the regulatory authority that had claimed that the LD50 of Basudin had changed used mice instead of rats.

My question to Professor Held is how could the approach proposed in the paper be modified to make it more applicable to the kind of context I have just described?

Kuldeep Kumar (*Bond University, Gold Coast*)

Two-sided tests are often used by 'experts' as well as 'non-experts' to verify a certain hypothesis. I congratulate Rice, Bonnett and Krakauer for rejuvenating this topic by giving an alternative decision theoretic approach by viewing the signs of an underlying parameter and making a decision based on whether the parameter is positive or negative. The whole procedure looks very complicated and it seems that plenty of subjectivity is involved, which is a little difficult to interpret. For example, how do we set an optimal value of a which minimizes the risk in testing the significance? Killeen (2006) proposed a decision-theory-based approach for hypothesis testing which calculates the expected utility of an effect on the basis of the probability of replicating it. However, it seems like it was only a theoretical contribution.

Rice and his colleagues have not given any examples to show how the results may be different from conventional statistical tests or a p-value approach. Conventional statistical tests look more transparent and can lead to further analysis like confidence intervals, which are lacking in this approach.

Alexander Ly (*University of Amsterdam and Centrum Wiskunde und Informatica, Amsterdam*)

The paper by Rice, Bonnett and Krakauer provides an alternative motivation for the p-value. I am honoured to be given the opportunity to praise the work and, as is customary, to raise some concerns.

The paper identified the decision theoretical framework in which the decision $\min\{\mathbf{P}(\theta < \theta_0), \mathbf{P}(\theta > \theta_0)\} < \alpha/2$ is Bayes. Unfortunately, by excluding the null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ this framework does not allow for the statistical confirmation of invariances including the speed of light and the cosmological constant. More mundane examples include the absence of mediation effects, and the notion of independence in Gaussian graphical models. Point nulls are essential to address the universal scientific question whether or not variation can be considered random. In general, point nulls provide much-needed protection against overfitting.

To take the null serious requires it to be assigned positive prior probability (Wrinch and Jeffreys, 1919, 1921, 1923): a crucial insight that resulted in the development of Bayes factors (e.g. Ly *et al.* (2019)). These Bayes factors quantify the evidence provided by the observed data, which can be specified independently of utilities concerning effort or money. I submit that a researcher's primary concern is evidence; combined with prior beliefs and utilities, this may ultimately result in a decision. But to take decisions without first quantifying the evidence is a precarious undertaking.

Furthermore, I am curious about the authors' practical recommendations. Do they recommend using $\mu \propto 1$ for the z-test? This probability matching prior might provide further frequentist justification for their procedure. By doing so, however, the classical z-test is retrieved, which I believe the authors wanted to improve on. Similarly, the authors show that an (inverse (?)) gamma distribution on σ^2 affects the risk of the t-test, and I am curious which prior they recommend. Could the right-Haar prior $\sigma \propto \sigma^{-1}$, which is necessary for t-tests to be safe under optional stopping and continuation (Grünwald *et al.*, 2019; Hendriksen *et al.*, 2018) be recommended?

Ulrich Mansmann (*Ludwig Maximilian's University, Munich*)

I congratulate Professor Held for his elegant proposal to design and analyse replication studies. He establishes a concept for the development of a successful 'replication study' that does not need additional information based on a significant 'discovery study'. Professor Held introduces the test statistics t_{Box} that integrates the information from the discovery study. Replication success on the level α is equivalent to $t^2_{\text{Box}} > z^2_{\alpha/2}$. The concept proposed integrates Bayesian arguments with frequentist components. Professor Held presents his results in the form of a classical frequentist point of view: p-value; confidence interval; sample size planning.

The important finding that his paper gives me as a frequentist is not to compare the effect estimate of a replication study with a classical null hypothesis. A non-classical null hypothesis challenges the replication study and formulates the asymmetric relationship between the discovery and its replication. It is intuitive that the non-classical test depends on the sample size of the discovery study and the size of the effect found in it.

Using the expression S in formula (1), $S_{z_{1-\alpha/2}}$ may define the upper bound of the non-classical null hypothesis. Performing the related frequentist test, a replication success at the level α occurs if an inequality analogous to term (5) holds. Term (5) must be modified by a multiplication on the right-hand side with the expression $1 + 2[\{(t_0/Z_{\alpha/2})^2 - 1\}/c]^{0.5}$.

More intuitively, one could use $s_0 z_{1-\alpha/2}$ as the upper bound of the null hypothesis. The standard error s_0 of the test statistic of the discovery study reflects a possible challenge to the replication study. The value $s_0 z_{1-\alpha/2}$ may alternatively represent the upper bound of noise around zero that would not allow discovering a signal in the discovery study. The size of the effect estimate in the discovery study and its 95% confidence interval allow determining the effect size that is needed to plan the replication study. This way, the frequentist adopts a kind of prior belief on the effect strength provided by the discovery study.

This is a first idea to produce a frequentist analogue to the strategy proposed by Professor Held. It needs refinement and assessment to prove usefulness for the design and analysis of replication studies.

On the one hand, the search for good replication concepts is omnipresent. The corresponding literature and commentaries, on the other hand, testify to a certain helplessness (Dirnagl, 2019; Piper *et al.*, 2019). This mirrors the achievement of Professor Held's work.

Jorge Mateu (*University Jaume I, Castellón*)

It is my pleasure to congratulate Professor Held on this interesting, timely and certainly attractive paper on the assessment of replication studies. I say timely because, now that we are heading and living in the big field of data science and data analytics, the concept, evaluation and treatment of replications is essential. As stated in the paper, replicability of research findings is crucial to the credibility of all empirical domains of science. With the current availability (and the easy access to produce great amounts) of data, we need to be able to assess to what extent new discoveries can be confirmed in independent replication studies that match the original study as much as possible. The paper focuses on proposing a new standard for the evidential assessment of replication studies in a, say, classical inferential problem. I would like to reinforce the necessity of such an approach when the problem comes into the area of spatial statistics.

In particular I draw the attention of Professor Held to this challenge in the context of mineral engineering. Consider the analysis of the spatial distribution of bubbles under three specific frother concentrations and three levels of volumetric air flow, in a flotation experiment with the aim of analysing whether there are significant main and interaction effects of these two factors when explaining the spatial patterns of bubbles.

The images of bubbles can be regarded as marked spatial point patterns. In each combination of frother concentration levels and volumetric air flow rate levels we have a number of replications. We have thus a classical factorial design experiment but where the observations are spatial point patterns. In this context recently proposed permutation tests (see Gonzalez *et al.* (2019)) require exchangeable units (functions), and this assumptions is difficult to guarantee although needed to have real replications that reproduce the overall experiment. The tests involved are a sort of functional descriptor of the spatial structure that produces functional means, and where partial and grand means together with variances are used as a basis of the testing procedure. There is currently no clear way to check for accurate replications and for the replication success. I wonder whether an adapted concept of a sceptical p-value (reducing the risk of small 'false positive' findings) as an indirect measure of the degree of replication success is applicable here. Many industrial problems, and a number of experiments that deal with spatial events, would benefit from this.

Robert A. J. Matthews (*Aston University, Birmingham*)

The inversion of Bayes's theorem as a means of resolving issues concerning priors was first proposed as a 'very useful technique' about 70 years ago (Good (1950), page 35). Only relatively recently, however, has it been recognized as a means of addressing contentious issues concerning, *inter alia*, clinical trials (Matthews, 2001; Spiegelhalter *et al.*, 2004), epidemiology (Greenland, 2006) and statistical education (Matthews, 2019).

Professor Held has recently made the inversion technique the basis of a new approach for bringing novel findings lacking prior support within the scope of Bayesian analysis (Held, 2019b). He has now extended this to address issues raised by the so-called replication crisis. One outcome is the proposal of a new metric for replication success: the *sceptical p-value* p_S. This can be calculated directly from standard data summaries and leads to a quantitative measure of replication success at a prespecified level α. As with conventional p-values, the precise interpretation of this metric is not intuitive, however. With p-values now seen as a major cause of misinterpretation of study findings (see for example Wasserstein and Lazar (2016)), the introduction of a similar concept to assess replications appears somewhat problematic.

However, p_S leads to a potential reframing of p-values as a source of insight into the replication of unprecedented findings. For such findings, p_S acquires a clear and practical interpretation via $p_{rep} = 1 - p_S/2$ where p_{rep} is the probability of an identical independent replication giving the same direction of effect as the original study (see Held (2019b), section 4). Professor Held also provides evidence that p_S is more reliable than conventional p-values in gauging whether a spectacular one-off finding is a freak result, or a 'null' finding might be worth another shot.

Most striking, however, are Professor Held's findings concerning the power of attempted replications. Simply put, unless a novel finding achieves a conventional p-value of less than 0.0056, an identically designed replication is no more reliable than a coin toss. As such, Professor Held provides a principled argument for why so many replications of novel findings fail, and for introducing a threshold for such claims of $p \leqslant 0.005$ (e.g. Benjamin *et al.* (2018)).

The consequent need for larger studies performed to higher inferential standards also adds weight to Altman's celebrated view that 'We need less research, better research, and research done for the right reasons' (Altman, 1994). As such, I very much welcome this paper.

Beat Neuenschwander (*Novartis Pharma, Basel*) **and Marcel Zwahlen** (*University of Bern*)

Professor Held's paper introduces the sceptical p-value as a new metric for replication studies. The approach is inspiring but conceptually and practically challenging.

The concept builds on a sceptical prior such that the combined evidence of the prior and the original data (D_1) is 'just significant' (e.g. at the one-sided 2.5% level); we use only one-sided probabilities here. Then, for replication success, a sufficiently large conflict between this prior and the replication data (D_2) is required. This may entail situations that are difficult to interpret.

The first example in Table 3 assumes an original trial of size $n_1 = 100$ and a highly significant treatment benefit ($p_1 = 0.001$). The sceptical prior is very informative and in clear conflict with the data ($p_{Box} = 0.008$). The replication study is of the same size and shows a less pronounced but still significant benefit ($p_2 = 0.02$). The conflict between prior and replication data is not sufficient to declare replication success at level 2.5% ($p_{Box} = 0.056$; sceptical p-value 0.044). The conclusion is surprising since both studies are significant. Similarly, failure to replicate occurs in the other examples also. Example 2, with reversed p-values ($p_1 = 0.02$; $p_2 = 0.001$), shows that a modestly significant original trial is difficult to replicate; even a p-value of 0.001 is insufficient. In example 3, the p-value constellation ($p_1 = p_2 = 0.01$) for a small original

Table 3 Four examples failing to declare replication success at the one-sided 2.5% level, showing one-sided tail probabilities for comparisons between H_0 (the null hypothesis), D_1 (the original data), D_2 (the replication data) and the sceptical prior SP for an original and replication trial of size n_1 and n_2 respectively†

Comparison	*Results for example 1, $n_1 = 100$, $n_2 = 100$*	*Results for example 2, $n_1 = 100$, $n_2 = 100$*	*Results for example 3, $n_1 = 50$, $n_2 = 250$*
H_0 versus D_1	*0.001*	*0.020*	*0.010*
H_0 versus SP $+ D_1$	*0.025*	*0.025*	*0.025*
SP *versus* D_1	0.008	0.270	0.105
H_0 versus D_2	*0.020*	*0.001*	*0.010*
H_0 versus SP $+ D_2$	*0.096*	*0.002*	*0.013*
SP *versus* D_2	0.056	0.178	0.261
Sceptical p-value	0.044	0.044	0.098

†Comparisons with the null hypothesis are shown in italics.

and a large replication study is less extreme. Again, replication fails at level 2.5%, although both studies are significant and even the combined evidence of prior and replication data meets the 2.5% threshold ($p = 0.013$).

We think that the choice of the sceptical prior explains the puzzling conclusions. In example 1, the sceptical prior is in clear conflict with the original data. So why would such a prior be relevant for the replication study? In example 2, the sceptical prior is rather vague, and, therefore, a conflict with the replication data is difficult to achieve.

The examples show that the approach proposed (based on a metric for prior–data conflict) makes it very difficult to achieve replication at the standard 2.5% level even if both studies are significant at this level (with or without the addition of the sceptical prior). To avoid the risk of a loss in translation (which we may have here), a metric based on the posterior probability for treatment benefit (combining a more sensible sceptical prior with the data) seems more principled to assess replication success.

Luis Pericchi (*University of Puerto Rico, Rio Piedras*)

The interesting procedures of Matthews and of Held are based on intervals and not on the probability of hypotheses. If it is desired to stick with intervals to achieve consistency it is necessary to make thresholds decrease with the sample size, as in Pérez and Pericchi (2014). I propose a general alternative and based on probabilities of hypothesis and the corresponding Bayes factors.

Our starting point is the intrinsic prior, defined as (Berger and Pericchi (2004), equation (6))

$$\pi^{\mathrm{I}}(\theta|\psi, m) = \int \pi\{\theta|X^*(m), \psi\} f_0\{X^*(m)|\theta_0, \psi\} \mathrm{d}X^*(m), \tag{3}$$

where f_0 is the null likelihood, ψ are potential nuisance parameters and X^* are theoretical training samples of size m. This is general and solves the weakest link of Bayes factors: their necessity for a specific (conditionally) proper prior. The methods of the present paper seem to rely on a particular conjugate prior, but why conjugate and how can it be generalized? Still, in equation (3) the training sample size m must be assessed, and it is there where the reverse Bayes approach appears naturally in this context.

Consider the first example of Professor Held's paper, for which equation (3) yields

$$\pi^{\mathrm{I}}(\mu|m) = N(\mu|0, 2\sigma_0^2/m), \tag{4}$$

where σ_0^2 is the (known) variance. (Here the intrinsic method yields the conjugate prior, but it does not if the variance is unknown.)

The Bayes factor based on the intrinsic prior, of the alternative over the null, becomes

$$B^{\mathrm{I}}_{1,0}(m) = \sqrt{\left(\frac{m}{2n+m}\right)} \exp\left(t_0^2 \frac{n}{2n+m}\right). \tag{5}$$

The conventional value of m is the 'minimal' training sample size, which, if we are to keep the correspondence between theoretical and real samples, in this example is $m = 1$. In the example the sample sizes are not

explicit, so we assume (consistent with the information given) that $\sigma_0^2 = 4$ and $n_0 = 150$. The Bayes factor based on the conventional intrinsic prior with $m = 1$ yields for the original study $B^{\mathrm{I}}_{1,0}(m = 1) = 24$, and for the replication 0.42. In probabilities, with equal model prior probabilities $\pi_0 = \frac{1}{2}$, we obtain $P(H_1|m = 1)$ equal to 0.96 and 0.3 for the replicate, which overall may be approximated (there are other possibilities) as the product 0.28.

We now define the 'handicapped' (a reversed Bayes concept) probabilities values based on equation (5). The strong handicap is calculated as the fractional training sample m^{**} for which the original experiment yields $B_{1,0}(m^{**}) = 1$, and then m^{**} is kept to obtain the Bayes factor under the replicated experiment. Similarly, the intermediate handicap m^* is the fractional training sample for which $B_{1,0}(m^*) = 3.2$ (mild evidence), and then for the replicate calculate $B_{1,0}(y|m^*)$, where y is the replicated data. It turns out that $m^{**} = 0.0017$ and $m^* = 0.017$, and $B_{1,0}(m^{**}) = 0.02$ and $B_{1.0}(m^*) = 0.05$. So the handicapped result yields a very low probability of a replication result between 0.02 and 0.05.

Kit C. B. Roes (*Radboud University Medical Center, Nijmegen*)

The importance of replication of experiments for science cannot be overstated, and Professor Held adds a valuable armamentarium to set standards for replication. In the standard drug regulatory setting, replicated evidence from at least two independent clinical trials is required before a licensing decision is made. These trials are typically run in parallel and cannot be divided into an original and a replication study. They can be quite different in design and an independent interpretation of results is considered a fundamental aspect. This is lost with sceptical p-values that depend on an earlier study. This may make the proposed standard not directly applicable. Simplifying the regulatory standard for the sake of this discussion, it is assumed that the standard of evidence corresponds to a p-value smaller than 0.0025. The standard proposed for replication studies may be relevant in case of conditional marketing approval: a temporary licence is granted based on an initial positive clinical trial (typically at a p-value less than 0.05, but not necessarily less than 0.0025), which can be either made definite or be revoked on the basis of evidence from a second clinical trial as post-marketing commitment. This is often presented as a 'continuum of evidence generation and decision making', but that is wrong: with the temporary licence the world has changed; the treatment is available to patients and withdrawing it represents very different decision making compared with initial licensing. The key question becomes whether the results of the second study are sufficiently close to those of the original study to justify the original decision for conditional approval and sufficiently convincing to meet the regulatory standard of evidence. In a frequentist framework it can be shown that replicating the original study with the same sample size, hoping for a 'positive' study at the standard 0.05-level, does not have adequate power to 'weed' out false positive results. Also other alternatives, such as directly evaluating the study by treatment interaction or planning the preconditional and post-conditional approval studies as a one-group sequential trial at a level of 0.0025 with sufficient overall power, directly lead to larger sample sizes for the post-approval study, very much in line with the findings for the sceptical p-value and power for replication studies. In this conditional licensing setting, it is justified and logical to interpret the data from the replication study in light of the results of the original study. The standard proposed is definitely interesting to explore in this increasingly relevant setting.

Stephen Senn (*Edinburgh*)

Leo Held's provision of a new and intriguing way to look at data is welcome. However, the on-going debate about p-values shows that it can take decades and even centuries (Shafer, 2019) for statisticians to get a feel for how statistical tools behave, so I draw attention to some puzzling features. In what follows I consider a *sceptic*, *originator* and *replicator*.

I start with the sceptic's limits as proposed by Matthews (2001, 2018), which are expression (1) of Held. This requirement can be re-expressed as the implied ratio of the sceptic's sample size to that provided by originator, namely

$$n_{\mathrm{ratio}} = \left(\frac{Z_P}{Z_\alpha}\right)^2 - 1, \qquad |Z_P| \geqslant |Z_\alpha|, \tag{6}$$

where Z_P is the ratio of the mean to the standard error and hence the Z-value for the p-value and Z_α is the value corresponding to the claimed confidence or significance level. Fig. 1 shows that, for the example of Fig. 1 of Held, the sceptic claims to know more than twice what the originator does. This formula has an implication that Bayesians of another persuasion would find objectionable. The sceptic claims to have seen n_{ratio} times as many subjects as the originator, *however many subjects the originator has seen*.

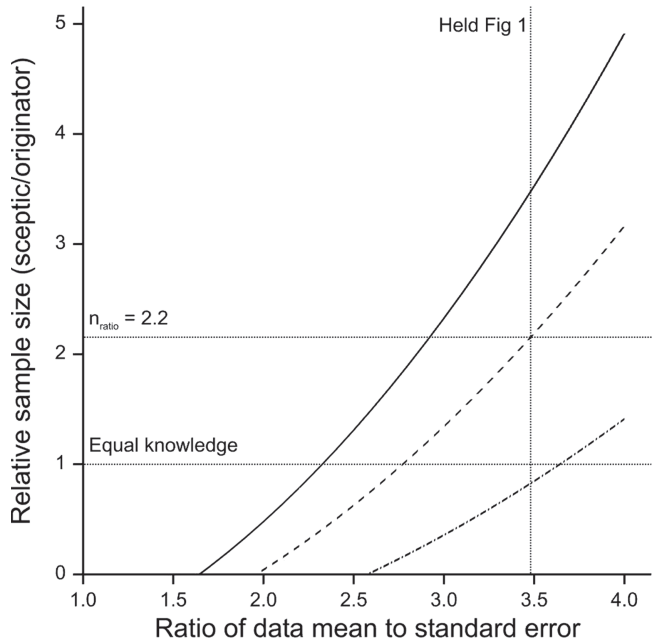

Fig. 1. The information in implicit relative sample size for the sceptic compared with the originator as a function of the ratio of mean to standard error, Z_P, at three levels of confidence (———, 90%; – – – –, 95%; · – · – ·, 99%): the value of Z_P for Held's example is about 3.5 and the implicit relative sample size is about 2.2

However, Matthews's approach is otherwise of a meta-analytic and hence Bayesian form: 'what does it take for current non-significance to overturn a new significant result?'. It is a pooling question analogous to the second of three questions in Senn (2002).

A problem with Matthews's approach is that if confidence distributions or *p*-values are used it cannot be applied. Held draws attention to this difficulty but addresses a different question: the replicator's result is not just pooled analogously to some meta-analytic updating. I expect much debate, not just directly on this paper, but in years to come, about whether this is a good idea or not. Frequentists see things differently and in the context of drug development there has been much discussion of the Food and Drug Administration's two trials rule (Darken and Ho, 2004; Fisher, 1999; Maca *et al.*, 2002; Rosenkranz, 2002; Senn, 1997).

Finally, I draw attention to the recalibration problem. Bayesians tend to assume that frequentist rules *ought to* be Bayesian but, when changing the system, retaining frequentist probability levels (e.g. 5% and 95%) is unlikely to be sensible. The analogy would be to say that a drug that could be safely given to anyone over 10 years of age would be safe to use for anyone weighing more than 10 kg (Senn, 2001).

I thank Leo Held for helpful comments.

Eric-Jan Wagenmakers (*University of Amsterdam*) **and Alexander Ly** (*University of Amsterdam and Centrum Wiskunde und Informatica, Amsterdam*)

The paper by Held provides a fresh perspective on the assessment of replicability. We are impressed by Held's work based on reverse engineering a sceptic's prior, but we are confused about the setting with $\hat{\theta}_j \sim \mathcal{N}(\theta, \kappa^2/n_j)$ where $j \equiv \mathrm{o}, \mathrm{r}$ refer to the original and replication study respectively, and $n_\mathrm{r} \to \infty$. In this case, $t_j = \sqrt{n_j}\hat{\theta}_j/\kappa$ and $c = n_\mathrm{r}/n_\mathrm{o}$ and consequently

$$z_\mathrm{S}^2 = \frac{n_\mathrm{o}}{2\kappa^2}\{|\hat{\theta}_\mathrm{r}|\sqrt{(4\hat{\theta}_\mathrm{o}^2 - \theta_\mathrm{r}^2)} - \hat{\theta}_\mathrm{r}^2\} + \mathcal{O}(n_\mathrm{r}^{-1}) \qquad \text{as } n_\mathrm{r} \to \infty. \tag{7}$$

For instance, if $\kappa = 4$, $\hat{\theta}_\mathrm{o} = 2$, $n_\mathrm{o} = 16$ (i.e. $t_\mathrm{o} = 2$; one-sided $\tilde{p}_\mathrm{o} = 0.02$), $\hat{\theta}_\mathrm{r} = 2$ (i.e. $t_\mathrm{r} \uparrow \infty$ and $\tilde{p}_\mathrm{r} \downarrow 0$), then $z_\mathrm{S} \leqslant 1.572$, and thus, $\tilde{p}_\mathrm{S} \geqslant 0.0579$ for all n_r. In particular, for $n_\mathrm{r} = 1000$ we obtain $t_\mathrm{r} = 15.8$ and $\tilde{p}_\mathrm{S} = 0.0582$, whereas the replication Bayes factor (e.g. Ly *et al.* (2019)) yields $\mathrm{BF}_{+0}(\hat{\theta}_\mathrm{r}|\hat{\theta}_0) = 2.54 \times 10^{53}$, indicating

overwhelming evidence against $\mathcal{H}_0 : \theta = 0$. Moreover, $\mathrm{BF}_{+0}(\hat{\theta}_\mathrm{r}|\hat{\theta}_\mathrm{o}) \to \infty$ as $n_\mathrm{r} \to \infty$, which we believe is the desired behaviour.

The upper bound from equation (7) suggests that Held's sceptic is a stubborn learner: someone who remains unconvinced by overwhelmingly informative data against the null acquired in the replication study, because this sceptic cannot let go of the original result, which he initially disputed.

The **authors** replied later, in writing, as follows.

Kenneth Rice, Chloe Krakauer and Tyler Bonnett

We thank the Royal Statistical Society for the opportunity to present our work, and to all the discussants for their contributions, which we hope will stimulate more research in this area.

Professor Hutton lists six requirements for good research. We agree with these but hasten to point out that there are open problems for statisticians to think about in all of them. Our work focuses on the first item in the list: the need for clear aims. Here again we suspect that we agree with Professor Hutton that the question should not be 'to p or not to p' but instead 'what do you need to determine?' As we have shown, if one's aim is knowing the sign of a parameter, using p in some way will, in many situations, be a large part of analyses addressing that aim. (We might also want to do model checking, in particular, of which more below.) But of course knowing a sign may not be the aim—in which case what to do becomes considerably murkier. To our knowledge no general syntax is available for translating scientific aims into statistical procedures—no language for stating what we want to know, how accurately we aim to know it or how willing we are to trade off veracity about what we want to know in exchange for simplicity in how we say it. Skilful statisticians do of course address these questions as they choose models and methods, but rarely in ways that can be stated directly in terms of the underlying science. Being more direct about how scientific aims inform choice of statistical methods should make that process simpler to understand, and easier to generalize to new situations.

Dr Kumar and Dr Chai would prefer that we give some examples of the method. Following Rice (2010), we shall consider testing Hardy–Weinberg equilibrium (HWE): a parametric constraint that is encountered in categorical genetic data. For the common case of a biallelic marker, the data from n unrelated subjects can be denoted

$$\begin{array}{ll} \text{genotype} & \mathrm{AA} \quad \mathrm{Aa} \quad \mathrm{aa} \\ \text{count} & n_{\mathrm{AA}} \quad n_{\mathrm{Aa}} \quad n_{\mathrm{aa}} \end{array},$$

and the data modelled as a simple random sample of size n from a multinomial distribution, with cell probabilities $(p_{\mathrm{AA}}, p_{\mathrm{Aa}}, p_{\mathrm{aa}})$ in the unit simplex. Under HWE the two alleles are assigned independently, and the cell probabilities lie on a curve parameterized as

$$(p_{\mathrm{AA}}, p_{\mathrm{Aa}}, p_{\mathrm{aa}}) = (p_\mathrm{A}^2, 2p_\mathrm{A}(1-p_\mathrm{A}), (1-p_\mathrm{A})^2)$$

where p_A denotes the probability that an allele is of type A. Departures from HWE can be explained by several phenomena (Weir, 1996); we shall focus on potential inbreeding, which could violate the independent assignment. To quantify this dependence, we define the inbreeding coefficient

$$\theta = \frac{2p_{\mathrm{aa}}(1+p_{\mathrm{AA}}) - p_{\mathrm{aa}}^2 - p_{\mathrm{AA}}^2}{1-(p_{\mathrm{aa}}-p_{\mathrm{AA}})^2}.$$

Negative and positive values of θ correspond respectively to deficient and excessive numbers of Aa observations, relative to HWE (Weir, 1996). We take the null value to be $\theta_0 = 0$, which accords with exactly no inbreeding. For simplicity of exposition, we shall use a Dirchlet(1, 1, 1) prior on $(p_{\mathrm{AA}}, p_{\mathrm{Aa}}, p_{\mathrm{aa}})$: this gives uniform support to all values of $(p_{\mathrm{AA}}, p_{\mathrm{Aa}}, p_{\mathrm{aa}})$, and support over the range of $-1 < \theta < 1$. Formally, it gives exactly zero support to $\theta = \theta_0 = 0$, ruling out states of nature that are perfectly free from inbreeding—an assumption that would also be typical of an elicited, substantive prior.

The calculations of relevant posterior tail areas are straightforward. In Fig. 2 we give results for all possible data sets with $n = 200$, applying the symmetric version of two-sided loss (3) using $\alpha = 0.05$, which we here assume captures the relative costs of misstatements about the two forms of inbreeding. For comparison we also show results from a standard exact test (Graffelman, 2015), again using $\alpha = 0.05$. In line with theory, the two tests behave extremely similarly, giving different results in only a handful of situations, most notably when at least one of the cell counts is small. (For priors with even closer frequentist agreement see Altham (1971).) We hope that, from Fig. 2, it seems reasonable to accept that our decision theoretic approach provides a Bayesian motivation to use something *very much like* a two-sided frequentist test.

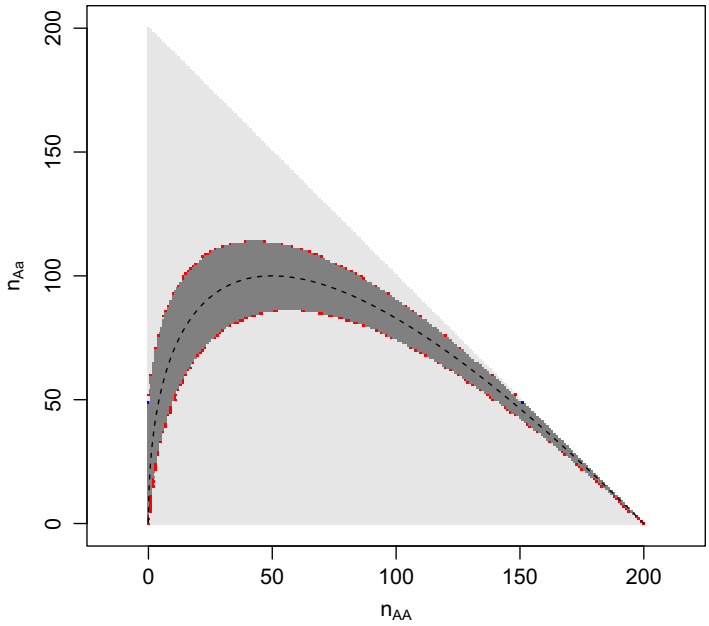

Fig. 2. Bayes rule *versus* exact test of HWE, $n = 200$, $\alpha = 0.05$: ——, reject, both; ——, do not reject, both; ——, reject, Bayes only; ——, reject, exact only; - - - - -, perfect HWE

In the example we have used the familiar $\alpha = 0.05$ for simplicity. But we agree with Dr Ferguson and Dr Fitz-Simon that, without careful thought about the consequences, using $\alpha = 0.05$ or any other default level is arbitrary and hence unsatisfactory. This is true for decision theoretic approaches and also more familiar frequentist arguments. Our thoughts here align closely with Lakens *et al.* (2018)—that users, when testing, should justify their α. We hope that our decision theoretic approach's underlying meaning of the chosen α (as relative loss) may help some investigators to choose meaningful values.

Professor Longford's comments are similar; we apologize to him if it was not clear that we also advocate collaborating with scientific colleagues when setting up loss functions that are relevant to analysis of their data. These colleagues are who we expect the papers' 'non-experts' will be, and they are the ones who will primarily determine what the scientific questions of interest actually are. Collaboration of this sort should also set α or other elements in the loss function, and we do not advocate research-area-wide defaults of, say, 0.05 or 2.9×10^{-7}. We also agree with him that careful evaluation of loss functions may, in many cases, result in decisions that look nothing like standard tests. As we wrote in Section 3, in our view our approach's simplicity is a strength, facilitating discussions about whether its loss functions are relevant or not. Based on our applied experience, in high throughput genetic association studies where signals are weak but valuable, just determining their sign is a realistic scientific goal and can set up useful next steps such as replication and functional follow-up. Where tests are relevant it would seem perverse for them to be locked up in Professor Longford's museum.

Professor Hennig and Dr Ly seem to prefer approaches with support for point nulls, in some form. For Dr Ly the goal is confirmation of point nulls, which is not strictly part of the significance testing framework—in our motivation of it or anyone else's—but which can be done by including point masses at the null. In the HWE example and many others in complex biological systems with continuous valued θ, these point nulls are simply not realistic statements of prior belief, so we find it difficult to accept Dr Ly's assertion that whether they can be considered true is a 'universal scientific question'. The Bayes factor's unrealistic foundations for these problems seems reason alone to consider alternatives, but we also note that in practice Bayes factors with point nulls are well known to be 'highly' (Gelman *et al.* (2013), page 184) or indeed 'hideously' (Draper (2005), page 246) sensitive to small and arbitrary details in how the non-null part is specified. As we hope Section 2.1 illustrates, we believe that for some problems Bayes factors can be worth thinking hard about. For testing point nulls, however, we remain sceptical.

Professor Hennig's concern is more about clear thinking about models, and whether they are true or

simply posited for the sake of arguments. Regarding the p-value as a statement about replicates under a hypothetical point null is of course not wrong, but its relevance in practice, particularly to non-statisticians, is difficult to grasp or use. The problem is 'cognitive, not statistical' (Chow and Greenland, 2019). At the admitted price of some approximation—though often very little—our motivation for using p removes the need to consider data under the point null, or to give probabilities that the point null is true. Professor Hennig also likens arguments relying on point nulls to those relying on parametric models, and how neither is a realistic assumption. We note with interest then that, in the large sample setting at least for some standard forms of regression (Szpiro *et al.*, 2010), the reliance on parametric models can be essentially removed, with a careful statement of the estimand, while keeping the familiar regression output. In such settings our Bayes rule for testing would give a large sample approximation of a robust Wald test.

Professor Hennig alludes to Box's famous space of models that are wrong but possibly useful. Giving less parametric interpretations of regression output and removing support for point nulls can be viewed as restricting attention to models that are *less* wrong while omitting some that are *rarely* useful.

We are grateful to Dr Kumar for pointing out Killeen's (2006) call for use of decision theory. This seems, in spirit, very close to our paper—and also recent work by Esteves *et al.* (2016), Manski (2019), Manski and Tetenov (2019) and Professor Longford. We particularly appreciate Killeen's labels of 'act' *versus* 'balk' for our active ($d \equiv$ above or below) *versus* null ($d = N$) decisions. Similarly to the losses that were considered by Professor Longford, Killeen introduced a dependence of decision on the magnitude as well as the sign. We have no dispute with using the magnitude, where it reflects real scientific concerns, and it need not even rule out the use of p-values in some form; see Rice (2010) for a connection to Wald tests. For continuously valued θ it would, however, appear to rule out being able to consider the set of all possible loss functions, that can be done when we only use θ through its sign.

Professor Dowe's interest lies in the range of values—we presume parameter values—for which $d = N$. At a basic level and assuming that the same α is used for every null value θ_0, this range is just the credible interval of Section 2.2, And, although its credibility (or its confidence coverage over repeated applications, up to large sample approximations) does not change over its range—as it is a feature of the whole interval—we interpret Professor Dowe's enquiry as asking how such decisions can be evaluated. This is an interesting and largely open question: in our Appendix B we briefly discussed how to summarize potential tests, and how this ends up motivating a quantity not unlike severity: a frequentist measure of test evaluation. Evaluating intervals could presumably be done by extending much the same idea. But the literature about tests warns of minefields ahead: Hoenig and Heisey (2001), for example, showed that *post hoc* power evaluation (of a particular type) tells us strictly nothing that the p-value does not already capture. Evaluation of the way that inferences rely on preliminary steps is also fundamentally difficult; there are proven impossibility theorems on the availability of inference in this situation (Leeb and Pötscher, 2005). Useful progress could be made here in two directions. Perhaps we can formulate (and then use) a general framework of how both to test hypotheses and to assess tests or intervals simultaneously, and their sensitivity to *a priori* assumptions or particular patterns in the data. Or perhaps we can prove that in general no such system is available, in which case statisticians might aim to help non-experts to understand when their results rely on assumptions that are supported only through contextual knowledge.

A final note is the issue of prior choice, raised by Dr Ly. This large topic, discussed at length in the literature, is well beyond the scope of our paper. Our focus is on situations like that of Fig. 2, where inference is fairly stable with respect to the prior (see for example Edwards *et al.* (1963)) and where we aim to give a Bayesian interpretation of classic procedures. Where those classic procedures, such as the z-test, accord with Bayesian results under relevant priors and losses we have no desire to improve on them.

Leonhard Held

I thank all the discussants for their useful comments on my paper. After rereading the discussion of Box (1980), I can only confirm

> 'how happy I am at the reception afforded my paper which I was particularly anxious to present here because of the unique vitality of this Society and its well known willingness to entertain and criticize ideas'.

The same spirit was present nearly 40 years later here in Belfast and I would like specifically to thank the Discussion Meetings Committee of the Royal Statistical Society for making this possible.

I have used the traditional two-sided 0.05-threshold for replication success throughout my paper but have already noted at the end of Section 3 that this choice is remarkably strict with $\Pr(p_S \leqslant 0.05 | H_0) \approx 0.0001$ for relative sample size $c = 1$. I am therefore grateful to Professor Senn who asks for a recalibration of

p_S. This can be done through type I error control at some suitable level and the two-study paradigm (Senn (1997), section 12.2.7) suggests a value of $0.025^2 = 1/1600$ (one sided) if the two studies are treated as exchangeable. The sceptical p-value treats the original and replication study as exchangeable only for $c = 1$. I could recently show that the (one-sided) threshold 0.065 for $\tilde{p}_S$ will control the type I error rate at 1/1600 (Held, 2019c). This confirms Professor Ioannidis's conjecture that the traditional (one-sided) 0.025-threshold may be too stringent for scientific fields with rigorous designs. I also agree with Professor Ioannidis that a more stringent threshold may be required in other disciplines unless they follow the high standards in drug and medical device regulation, mentioned by Professor Hutton. In the application on the replicability of psychological science reported in Section 5 of my paper, the one-sided 0.065-threshold turns out to increase the replication success rate to $22/73 = 30\%$, but it may be too loose in the light of the sobering result of the quality checks reported by Professor Hutton.

A recalibration of p_S may also address the concerns by Professor Diggle, Dr Ferguson and Dr Fitz-Simon regarding the 'peculiar' property $p_S \geqslant \max\{p_o, p_r\}$. Specifically, with the threshold 0.065 the first two examples by Dr Neuenschwander and Professor Zwahlen now both lead to replication success. Their third example is particularly interesting, as the 0.065-threshold is still not met ($\tilde{p}_S = 0.098$). However, the replication sample size is five times larger than in the original study and so the replication effect estimate (which was not listed by Neuenschwander and Zwahlen) is just $1/\sqrt{5} = 0.45$ times as large as the original effect estimate. Only because of the larger sample size, the shrunken effect estimate still achieved significance. The sceptical p-value takes this into account and penalizes this shrinkage accordingly.

Dr Ferguson and Dr Fitz-Simon as well as Professor Wagenmakers and Dr Ly question the utility of p_S as a measure of evidence. Both consider a scenario where the original study is just borderline significant whereas the evidence of an effect in the replication study is overwhelming. The degree of replication success, as quantified by the sceptical p-value, will then be surprisingly low. These concerns are understandable but deserve some additional comments. The proposed reverse Bayes assessment of replication success has much in common with the two-study paradigm, with requires

> 'at least two adequate and well-controlled studies, each convincing in its own, to establish effectiveness'

(Food and Drug Administration (1998), page 3). The method is hence more stringent than a standard meta-analysis and it is not surprising that the sceptical p-value remains 'stubborn' if the evidence from the original study is only weak. Replication success represents a substantiation of a claim of a new discovery but there is nothing to substantiate if the original study was not convincing on its own. I believe that this stringency of p_S is a good thing, in contrast with more established evidence synthesis methods (such as the popular method by Fisher (1958) to combine p-values) which can produce a significant overall result even if one of the studies was negative, perhaps even significant in the opposite direction. This is related to the replication paradox, to which the modified replication Bayes factor (Ly *et al.* (2018), appendix C) and the one-sided sceptical p-value are not prone. The sceptical p-value additionally treats the original and replication study in an asymmetric way, taking into account relative effect and sample sizes. It would be interesting to investigate whether the replication Bayes factor has a similar feature.

Intrinsic credibility is a special case of the proposed framework in the absence of a replication study and leads to the double-the-variance rule; see equation (12). A referee pointed out that this is precisely the rule-of-thumb that was proposed by Copas and Eguchi (2005) for dealing with locally misspecified models. Professor Matthews notes that the p-value for intrinsic credibility (Held, 2019b) is easy to interpret in terms of the probability of replicating an effect (Killeen, 2005), whereas the sceptical p-value is not. It is therefore of interest to develop more direct probability measures of replication success within the Bayesian framework. Professor Pericchi points out that the reverse Bayes approach can also be combined with Bayesian hypothesis testing, where his 'handicap' Bayes factor provides an alternative and perhaps more intuitive measure for replication success.

Professor Diggle asks about the possibility of a compatible 'sceptical confidence interval' and I am pleased to report that this can indeed be defined on the basis of inversion of the proposed assessment of replication success. The trick is to extend the approach to a non-zero sceptical prior mean μ (a proposal also made by Dr Mathur and Professor VanderWeele) which defines the sceptical confidence interval at level $1 - \alpha$ as the set of all values of μ which do not lead to replication success at level α. This is achieved by using the more general test statistics $t_o = (\hat{\theta}_o - \mu)/\sigma_o$ and $t_r = (\hat{\theta}_r - \mu)/\sigma_r$ to calculate the sceptical p-value with equation (9) of the paper. Fig. 3 displays the sceptical confidence interval for the introductory example from the paper. Shown is the sceptical confidence interval for the nominal 95% level and for 87% $(1 - 2 \times 0.065)$ level, calibrated to the two-study paradigm as described above. Both sceptical confidence intervals are regular in this example but it is worth noting that, if there is more conflict between the original

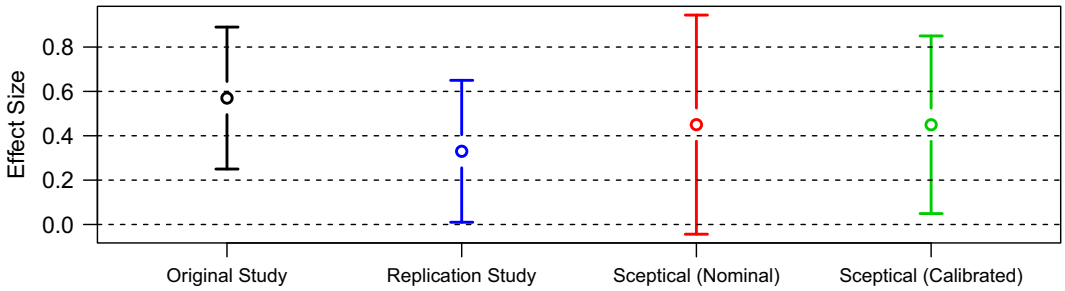

Fig. 3. Revisiting the example from Fig. 1 in my paper: the sceptical confidence interval at nominal level 95% has limits −0.04 and 0.94; the calibrated sceptical confidence interval at level 87% has limits 0.05 and 0.85

and replication study, the sceptical confidence interval may actually be a sceptical confidence region defined as the union of two non-intersecting intervals. The sceptical confidence interval or region thus behaves like a mixture of two normal distributions rather than a single normal distribution as the confidence interval for the combined effect obtained from a meta-analysis.

There are other potential applications of a non-zero sceptical prior mean. The clinically relevant difference mentioned by Professor Diggle is a natural choice. It may also help to apply the proposed method to the problem described by Professor Grieve. Here the national authority (the sceptic) postulated a (non-zero) mean LD50 of 200 mg kg^{-1}, challenging earlier results which suggested substantially larger values. The new experiment has been conducted to convince the national authority that a mean LD50 of 200 mg kg^{-1} is unrealistic. The proposed assessment of replication success may thus be applicable to this setting, perhaps after suitable transformation of LD50 to achieve approximate normality of the fiducial interval.

The other interesting aspect of Professor Grieve's story is that the regulatory authority had used mice instead of rats. This relates to the comment made by Professor Friede and Dr Röver that additional between-study heterogeneity will raise the bar for replication success even further. Such adjustments for heterogeneity are commonly done in meta-analysis where the goal is to compute an overall effect estimate but I am not sure whether they are required in the assessment of replication success. For example, the two-study paradigm is the preferred approach for drug approval exactly because the two trials may have been performed in different settings (Senn (1997), section 12.2.7). If the requirement of the two-study paradigm is met, the results are considered more robust than significance of a pooled effect estimate obtained through a meta-analysis. Possible between-study heterogeneity is implicitly incorporated in the requirement of two independent studies, just as replication studies try to confirm original results in independent investigations. Additional explicit incorporation of heterogeneity in the analysis of replication success does not seem to match the idea to challenge the original finding through the reverse Bayes approach.

However, if the goal is to calculate the required replication sample size to achieve replication success, heterogeneity could be incorporated in the design prior, as well as a possible exaggeration of original effect estimates. Regarding the latter Professor Bird mentions the popular rule of thumb to halve the original effect size in sample size calculations, as used in the recent social science replication project (Camerer *et al.*, 2018). A more principled approach is to estimate the shrinkage prior variance with the empirical Bayes method (Pawel and Held, 2019), which has connections to shrinkage methods in regression for optimal prediction (Copas, 1983).

Professor Diggle reminds us of the problem how to specify the clinically relevant difference Δ (see also Bland (2009)), but this difficulty does not seem to apply in a replication setting, where the original (possibly shrunken) study effect estimate is a natural choice for Δ and the associated uncertainty can also be incorporated. It also seems worth mentioning that the precision of the estimate as quantified by the width w of the 95% confidence interval for the parameter of interest is directly related to Δ:

$$w = 2\Delta \frac{1.96}{1.96 + u}$$

where $u = \Phi^{-1}(\text{power})$ depends on the conditional power. This relationship is easy to derive from the standard formula mentioned by Professor Bird and the corresponding formula for sample size calculation based on precision (Kirkwood and Sterne (2003), Table 35.1(b)). It shows that the width w is between 2Δ and Δ for studies with power between 50% and 97.5%. So, from a purely technical perspective, sample size

calculation based on precision rather than power is just the other side of the same coin. Whether estimation should be preferred over testing is a more general issue and currently the subject of an intensive debate (e.g. Ioannidis (2019)). The investigation of consistency between original and replication study, as mentioned by Dr Mathur and Professor VanderWeele, is more in the spirit of estimation and perhaps best described through probabilistic forecasting of the replication result based on the original result (Bayarri and Mayoral, 2002; Patil *et al.*, 2016; Pawel and Held, 2019). In rare cases it may then happen that replication success is declared although the replication effect size is inconsistently large compared with the original effect size: a scenario mentioned by Professor Bird. This is the price to be paid for the otherwise attractive property of the sceptical p-value to react to shrinkage of the replication effect estimate.

I thank the remaining contributors for their many constructive remarks. I agree that the role of the sceptical limit could be investigated in more detail, as suggested by Professor Senn and Professor Mansmann. Several of the contributors have asked how to analyse multiple replication studies (Mateu, Dowe, and Mathur and VanderWeele) or even multiple original studies (Chai). The sceptical confidence interval may then be used to summarize the original and replication study and could be used to assess the success of a second replication study. I am also grateful for the list of possible interesting applications outside the standard replication setting where the conditional marketing setting in drug development, as outlined by Professor Roes, seems particularly promising.

Needless to say that plenty of work remains to be done to address the current helplessness in the assessment of replication studies, mentioned by Professor Mansmann. I am most grateful for the comments being made that have already initiated further developments of the methodology proposed. I hope that my method will contribute to a joint effort of statisticians and researchers, scientific journals and funding agencies to establish a replication culture in science to combat the reproducibility problems that we are currently facing.

References in the discussion

Altham, P. M. E. (1971) Exact Bayesian analysis of an intraclass 2×2 table. *Biometrika*, **58**, 679–680.

Altman, D. G. (1994) The scandal of poor medical research. *Br. Med. J.*, **308**, 283–284.

Bayarri, M. J. and Mayoral, M. (2002) Bayesian design of "successful" replications. *Am. Statistn*, **56**, 207–214.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Holland Jones J., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafò, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. and Johnson, V. E. (2018) Redefine statistical significance. *Nat. Hum. Behav.*, **2**, 6–10.

Berger, J. O. and Pericchi, L. R. (2004) Training samples in objective Bayesian model selection. *Ann. Statist.*, **32**, 841–869.

Bird, S. M. and Hutchinson, S. J. (2003) Male drugs-related deaths in the fortnight after release from prison: Scotland, 1996–1999. *Addiction*, **98**, 185–190.

Bland, J. M. (2009) The tyranny of power: is there a better way to calculate sample size? *Br. Med. J.*, **339**, article b3985.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J. and Wu, H. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.*, **2**, 637–644.

Chow, Z. R. and Greenland, S. (2019) Semantic and cognitive tools to aid statistical inference: replace confidence and significance by compatibility and surprise. *Preprint*. (Available from `https://arxiv.org/abs/1909.08579`.)

Copas, J. B. (1983) Regression, prediction and shrinkage (with discussion). *J. R. Statist. Soc.* B, **45**, 311–354.

Copas, J. and Eguchi, S. (2005) Local model uncertainty and incomplete-data bias (with discussion). *J. R. Statist. Soc.* B, **67**, 459–513.

Cox, D. R. (1982) Statistical significance tests. *Br. J. Clin. Pharmcol.*, **14**, 325–331.

Darken, P. F. and Ho, S.-Y. (2004) A note on sample size savings with the use of a single well-controlled clinical trial to support the efficacy of a new drug. *Pharm. Statist.*, **3**, 61–63.

Demany, L., Trost, W., Serman, M. and Semal, C. (2008) Auditory change detection: simple sounds are not memorized better than complex sounds. *Psychol. Sci.*, **19**, 85–89.
Dirnagl, U. (2019) Rethinking research reproducibility. *EMBO J.*, **38**, article e101117.
Dowe, D. L. (2008) Foreword re C. S. Wallace. *Comput. J.*, **51**, 523–560.
Dowe, D. L. (2011) MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In *Philosophy of Statistics* (eds P. S. Bandyopadhyay and M. R. Forster), pp. 901–982. New York: Elsevier.
Dowe, D. L. (2013) Introduction to Ray Solomonoff 85th memorial conference. In *Algorithmic Probability and Friends: Bayesian Prediction and Artificial Intelligence*, pp. 1–36. New York: Springer.
Draper, D. (2005) Bayesian modeling, inference and prediction. University of Southern California, Los Angeles. (Available from `https://users.soe.ucsc.edu/draper/draper-BMIP-dec2005.pdf`.)
Edwards, W., Lindman, H. and Savage, L. J. (1963) Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193.
Esteves, L. G., Izbicki, R., Stern, J. M. and Stern, R. B. (2016) The logical consistency of simultaneous agnostic hypothesis tests. *Entropy*, **18**, no. 7, article 256.
Farrell, T. M. (2008) Factors affecting beef enterprise profitability: experiences from a grazing group in north-west NSW. In *Proc. 23rd A. Conf. Grassland Society of New South Wales, Australia*, p. 15.
Fisher, L. D. (1999) One large, well-designed, multicenter study as an alternative to the usual FDA paradigm. *Drug Inform. J.*, **33**, 265–271.
Fisher, R. A. (1958) *Statistical Methods for Research Workers*, 13th edn. Edinburgh: Oliver and Boyd.
Food and Drug Administration (1998) Providing clinical evidence of effectiveness for human drug and biological products. *Technical Report*. US Food and Drug Administration, Rockville. (Available from `www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products`.)
Friede, T., Röver, C., Wandel, S. and Neuenschwander, B. (2017) Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometr. J.*, **59**, 658–67l.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*. Boca Raton: Chapman and Hall–CRC.
Gonzalez, J. A., Lagos, B. M. and Mateu, J. (2019) Factorial experiments of spatial point patterns in Minerals Engineering. To be published.
Good, I. J. (1950) *Probability and the Weighing of Evidence*. London: Griffin.
Goodman, S. N. (2016) Aligning statistical and scientific reasoning. *Science*, **352**, 1180–1181.
Graffelman, J. (2015) Exploring diallelic genetic markers: the HardyWeinberg package. *J. Statist. Softwr.*, **64**, no. 3, 1–23.
Greenland, S. (2006) Bayesian perspective for epidemiological research: I, Foundations and basic methods. *Int. J. Epidem.*, **35**, 765–775.
Grieve, A. P. (1992) Implementation of Bayesian methods in the pharmaceutical industry. *PhD Thesis*. University of Nottingham, Nottingham. (Available from `http://eprints.nottingham.ac.uk/14013/1/334866.pdf`.)
Grünwald, P., de Heide, R. and Koolen, W. (2019) Safe testing. *Preprint arXiv:1906.07801*. Centrum Wiskunde und Informatica, Amsterdam.
Hahn, S., Williamson, P. R., Hutton, J. L., Garner, P. and Flynn, V. (2000) Assessing the potental for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Statist. Med.*, **19**, 3325–3336.
Hardisty, D. J. and Weber, E. (2009) Discounting future green: money versus the environment. *J. Exptl Psych. Gen.*, **138**, 329–340.
Hedges, L. V. and Olkin, I. (1985) *Statistical Methods for Meta-analysis*. San Diego: Academic Press.
Held (2019a) A new standard for the analysis and design of replication studies. *Preprint arXiv:1811.10287v2*. University of Zurich, Zurich.
Held, L. (2019b) The assessment of intrinsic credibility and a new argument for $p < 0.005$. *R. Soc. Open Sci.*, **6**, no. 3.
Held, L. (2019c) The harmonic mean χ^2 test to substantiate scientific findings. *Technical Report*. University of Zurich, Zurich. (Available from `arxiv.org/abs/1911.10633`.)
Hendriksen, A., de Heide, R. and Grünwald, P. (2018) Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *Preprint arXiv:1807.09077*. Centrum Wiskunde und Informatica, Amsterdam.
Hill, A. B. (1965) *The Environment and Disease: Association or Causation?*, pp. 295–300. New York: Sage.
Hoenig, J. M. and Heisey, D. M. (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Statistn*, **55**, 9–24.
Hubbard, D. W. (2013) How to measure anything. *Professnl Safty*, **58**, 58–59.
Hutton, J. L. (1995) Statistics is essential for professional ethics. *J. Appl. Phil.*, **12**, 253–261.
Hutton, J. L. and Owens, R. G. (1993) Bayesian sample size calculations and prior beliefs about child sexual abuse. *Statistician*, **42**, 399–404.
Hutton, J. L. and Williamson, P. R. (2000) Bias in meta-analysis due to outcome variable selection within studies. *Appl. Statist.*, **49**, 359–370.

Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLOS Med.*, **2**, no. 8, 696–701.
Ioannidis, J. P. A. (2019) The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *J. Am. Med. Ass.*, **321**, 2067–2068.
Johnson, V. E., Payne, R. D., Wang, T., Asher, A. and Mandal, S. (2016) On the reproducibility of psychological science. *J. Am. Statist. Ass.*, **112**, 1–10.
Killeen, P. R. (2005) An alternative to null-hypothesis significance tests. *Psychol. Sci.*, **16**, 345–353.
Killeen, P. R. (2006) Beyond statistical inference: a decision theory for science. *Psychon. Bull. Rev.*, **13**, 549–562.
Kirkwood, B. R. and Sterne, J. A. C. (2003) *Essential Medical Statistics*. Oxford: Blackwell Science.
Kontopantelis, E., Springate, D. A. and Reeves, D. (2013) A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses. *PLOS One*, **8**, no. 7, article e69930.
Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E. *et al*. (2018) Justify your alpha. *Nat. Hum. Behav.*, **2**, no. 3, 168.
Lamb, E. (2012) 5 sigma what's that? *Scient. Am.,* blog. (Available from `https://blogs.scientificamerican.com/observations/five-sigmawhats-that/`.)
Leeb, H. and Pötscher, B. M. (2005) Model selection and inference: facts and fiction. *Econmetr. Theory*, **21**, 21–59.
Leek, J. T. and Peng, R. D. (2015) P values are just the tip of the iceberg. *Nature*, **520**, 612.
Lim, K. T. K. (2019) Statistical methods and reproducibiliiy in behavioural science. *PhD Thesis*. Department of Statistics, University of Warwick, Coventry.
Longford, N. T. (2013) *Statistical Decision Theory*. Heidelberg: Springer.
Longford, N. T. (2016) Comparing two treatments by decision theory. *Pharm. Statist.*, **15**, 387–395.
Ly, A., Etz, A., Marsman, M. and Wagenmakers, E.-J. (2018) Replication Bayes factors from evidence updating. *Behav. Res. Meth.* to be published.
Ly, A., Stefan, A., van Doorn, J., Dablander, P., van den Bergh, D., Sarafoglou, A., Kucharskỳ, Š., Derks, K., Gronau, Q. F., Raj, A., Boehm, U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman, M. and Wagenmakers, E.-J. (2019) The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p-value hypothesis test. *Computnl Brain Behav.*, to be published.
Maca, J., Gallo, P., Branson, M. and Maurer, W. (2002) Reconsidering some aspects of the two-trials paradigm. *J. Biopharm. Statist.*, **12**, 107–119.
Manski, C. F. (2019) Treatment choice with trial data: statistical decision theory should supplant hypothesis testing. *Am. Statistn*, **73**, suppl. 1, 296–304.
Manski, C. F. and Tetenov, A. (2019) Trial size for near-optimal choice between surveillance and aggressive treatment: reconsidering mslt-ii. *Am. Statistn*, **73**, suppl. 1, 305–311.
Mathur, M. B. and VanderWeele, T. J. (2019) New statistical metrics for multisite replication projects. *Preprint*. (Available from `https://osf.io/w89s5/`.)
Matthews, R. A. J. (2001) Methods for assessing the credibility of clinical trial outcomes. *Drug Inform. J.*, **35**, 1469–1478.
Matthews, R. A. J. (2018) Beyond 'significance': principles and practice of the analysis of credibility. *R. Soc. Open Sci.*, **5**, no. 1, article 171047.
Matthews, R. A. J. (2019) Moving towards the post $p < 0.05$ era via the analysis of credibility. *Am. Statistn*, **73**, suppl. 1, 202–212.
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, **349**, article aac47l6.
Patil, P., Peng, R. D. and Leek, J. T. (2016) What should researchers expect when they replicate studies?: A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.*, **11**, 539–544.
Pawel, S. and Held, L. (2019) Probabilistic forecasting of replication studies. *Technical Report*. University of Zurich, Zurich. (Available from `psyarxiv.com/fhwb7`.)
Pérez, M. E. and Pericchi, L. R. (2014) Changing statistical significance with the amount of information: the adaptive α significance level. *Statist. Probab. Lett.*, **85**, 20–24.
Pierce, M., Millar, T., Robertson, J. R. and Bird, S. M. (2018) Ageing opioid users increased risk of methadone-specific death in the UK. *Int. J. Drug Poly*, **55**, 121–127.
Piper, S. K., Grittner, U., Rex, A., Riedel, N., Fischer, F., Nadon, R., Siegerink, B. and Dirnagl, U. (2019) Exact replication: foundation of science or game of chance? *PLOS Biol.*, **17**, no. 4, article e3000188.
Rice, K. (2010) A decision-theoretic formulation of Fisher's approach to testing. *Am. Statistn*, **64**, 345–349.
Rosenkranz, G. (2002) Is it possible to claim efficacy if one of two trials is significant while the other just shows a trend? *Drug Inform. J.*, **36**, 875–879.
Seaman, S. R., Brettle, R. P. and Gore, S. M. (1998) Mortality from overdose among injecting drug users recently released from prison: database linkage study. *Br. Med. J.*, **316**, 426–428.
Senn, S. (1997) *Statistical Issues in Drug Development*. Chichester: Wiley.
Senn, S. J. (2001) Two cheers for P-values. *J. Epidem. Biostatist.*, **6**, 193–204.
Senn, S. J. (2002) A comment on replication, p-values and evidence. *Statist. Med.*, **21**, 2437–2444.
Shafer, G. (2019) On the nineteenth-century origins of significance testing and p-hacking. (Available from `http://probabilityandfinance.com/articles/55.pdf`.)

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004) *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. New York: Wiley.
Szpiro, A. A., Rice, K. M. and Lumley, T. (2010) Model-robust regression and a Bayesian "sandwich" estimator. *Ann. Appl. Statist.*, **4**, 2099–2113.
Turner, R. M., Bird, S. M. and Higgins, J. P. T. (2013) The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PLOS One*, **8**, article e59202.
Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G. and Higgins, J. P. T. (2015) Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statist. Med.*, **34**, 984–998.
Wallace, C. S. (2005) *Statistical and Inductive Inference by Minimum Message Length*. New York: Springer Science and Business Media.
Wallace, C. S. and Boulton, D. M. (1968) An information measure for classification. *Comput. J.*, **11**, 185–194.
Wallace, C. S. and Dowe, D. L. (1999) Minimum message length and Kolmogorov complexity. *Comput. J.*, **42**, 270–283.
Wallace, C. S. and Freeman, P. R. (1987) Estimation and inference by compact coding. *J. R. Statist. Soc.* B, **49**, 240–252.
Wasserstein, R. L. and Lazar, N. A. (2016) The ASA statement on *p*-values: context, process, and purpose. *Am. Statistn*, **70**, 129–133.
Weir, B. S. (1996) *Genetic Data Analysis II*. Sunderland: Sinauer.
Wilkinson, L. (1999) Statistical methods in psychology journals: guidelines and explanations. *Am. Psychol.*, **54**, 594–604.
Wrinch, D. and Jeffreys, H. (1919) On some aspects of the theory of probability. *Phil. Mag.*, **38**, 715–731.
Wrinch, D. and Jeffreys, H. (1921) On certain fundamental principles of scientific inquiry. *Phil. Mag.*, **42**, 369–390.
Wrinch, D. and Jeffreys, H. (1923) On certain fundamental principles of scientific inquiry. *Phil. Mag.*, **45**, 368–374.