# A New Standard for the Analysis and Design of Replication Studies

## Response to discussion

I would like to thank all the discussants for their useful comments on my paper. After re-reading 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 in Belfast and I would like to specifically thank the Discussion Meetings Committee of the 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 $\Pr(p_S \leq 0.05 \,|\, H_0) \approx 0.0001$ for relative sample size $c = 1$. I am therefore grateful to Professor Senn who asks for a re-calibration of $p_S$. This can be done through Type-I error control at some suitable level and the two-study paradigm (Senn, 2007, Section 12.2.8) suggests a value of $0.025^2 = 1/1600$ (one-sided) if the two studies are treated as exchangeable. The sceptical $p$-value treats original and replication study as exchangeable only for $c = 1$. I was recently able to show that the (one-sided) threshold 0.065 for $\tilde{p}_S$ will control the Type-I error at 1/1600 (Held, 2019b). This confirms Professor Ioannidis 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 chapter 5 of my paper, the one-sided 0.065 threshold turns out to increase the replication success rate to 22/73=30%, but may be too loose in the light of the sobering results of the quality checks reported by Professor Hutton.

A re-calibration of $p_S$ may also address the concerns by Professor Diggle, Dr Ferguson and Dr Fitz-Simon regarding the "peculiar" property $p_S \geq \max\{p_o, p_r\}$. 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 (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 comprehensible, but deserve

some additional comments. The proposed reverse-Bayes assessment of replication success has much in common with the two-study paradigm, which requires "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness." (FDA, 1998, p. 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 to 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 to. 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, cp. equation (12). A referee has pointed out that this is precisely the rule-of-thumb proposed by Copas and Eguchi (2005) for dealing with locally misspecified models. Professor Matthews notes that the $p$-value for intrinsic credibility (Held, 2019a) 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 Perrichi 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 based on inversion of the proposed assessment of replication success. The trick is to extend the approach to a non-zero sceptical prior mean $\mu$ (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 $\mu$ which do not lead to replication success at level $\alpha$. 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. Figure 1 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% $(1 - 2 \cdot 0.065)$ level, calibrated to the two-study paradigm as described above. Both sceptical confidence intervals are regular in this example but it is worth to note that if there is more conflict between the original 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/region thus behaves like a mixture of two normals rather than a single normal as the confidence interval for the combined effect obtained from a meta-analysis.

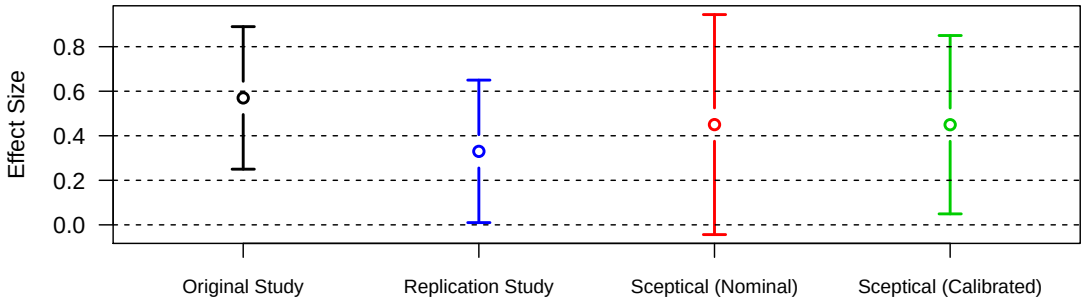


Figure 1: Revisiting the example from Figure 1. The sceptical confidence interval at nominal level 95% has limits −0.04 and 0.94. The calibrated sceptical confidence interval at level 87% has limits 0.05 and 0.85.

There are other 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, 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 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 if 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, 2007, Section 12.2.8). 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 reverse-Bayes.

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 half 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

empirical Bayes (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 $\Delta$ (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 $\Delta$ 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 $\Delta$:

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

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 $2\,\Delta$ and $\Delta$ 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 to 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 remaning 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 authors have asked how to analyse multiple replication studies (Mateu, Dowe, Mathur and VanderWeele) or even multiple original studies (Chai). The sceptical confidence interval may then be used to summarize 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 a lot 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 proposed methodology. 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 in order to combat the reproducibility problems we are currently facing.

**Acknowledgments** Support by the Swiss National Science Foundation (Project # 189295) is gratefully acknowledged.

## References

Bayarri, M. J. and Mayoral, M. (2002). Bayesian design of "successful" replications. *The American Statistician*, 56:207–214. https://www.doi.org/10.1198/000313002155.

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

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430. https://www.jstor.org/stable/2982063.

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. *Nature Human Behaviour*, 2(9):637–644. https://doi.org/10.1038/s41562-018-0399-z.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):459–513.

Copas, J. B. (1983). Regression, prediction and shrinkage. *J. Roy. Statist. Soc. Ser. B*, 45(3):311–354.

FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products. Technical report, US Food and Drug Administration. www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 13th ed.(rev.) edition.

Held, L. (2019a). The assessment of intrinsic credibility and a new argument for $p < 0.005$. *Royal Society Open Science*. https://doi.org/10.1098/rsos.181534.

Held, L. (2019b). The harmonic mean $\chi^2$ test to substantiate scientific findings. Technical report, University of Zurich.

Ioannidis, J. P. A. (2019). The importance of predefined rules and prespecified statistical analyses: Do not abandon significance. *JAMA*, 321(21):2067–2068.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16(5):345–353. https://doi.org/10.1111/j.0956-7976.2005.01538.x.

Kirkwood, B. R. and Sterne, J. A. C. (2003). *Essential Medical Statistics*. Blackwell Science, Malden.

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1092-x.

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. *Perspectives on Psychological Science*, 11(4):539–544. https://doi.org/10.1177/1745691616646366.

Pawel, S. and Held, L. (2019). Probabilistic forecasting of replication studies. Technical report, University of Zurich.

Senn, S. (2007). *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, U.K., second edition.