

University of Groningen

Practical Implications of Equating Equivalence Tests

Linde, Maximilian; Tendeiro, Jorge; Wagenmakers, Eric-Jan; van Ravenzwaaij, Don

Published in:
 Psychological Methods

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
 Early version, also known as pre-print

Publication date:
 2022

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Linde, M., Tendeiro, J., Wagenmakers, E.-J., & van Ravenzwaaij, D. (Accepted/In press). Practical Implications of Equating Equivalence Tests: Reply to Campbell and Gustafson (2022). *Psychological Methods*.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

**Practical Implications of Equating Equivalence Tests: Reply to Campbell and
Gustafson (2022)**


Maximilian Linde¹, Jorge N. Tendeiro², Eric-Jan Wagenmakers³, and
Don van Ravenzwaaij¹


¹Unit of Psychometrics and Statistics, Department of Psychology, Faculty of Behavioural
and Social Sciences, University of Groningen, Groningen, The Netherlands


²Office of Research and Academia-Government-Community Collaboration, Education and
Research Center for Artificial Intelligence and Data Innovation, Hiroshima University,
Hiroshima, Japan


³Unit of Psychological Methods, Department of Psychology, Faculty of Social and
Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands

Author Note

Maximilian Linde  <https://orcid.org/0000-0001-8421-090X>

Jorge N. Tendeiro  <https://orcid.org/0000-0003-1660-3642>

Eric-Jan Wagenmakers  <https://orcid.org/0000-0003-1596-1034>

Don van Ravenzwaaij  <https://orcid.org/0000-0002-5030-4091>

This research was supported by a Dutch scientific organization VIDI fellowship grant awarded to Don van Ravenzwaaij (016.Vidi.188.001), a Dutch scientific organization VICI fellowship grant awarded to Eric-Jan Wagenmakers (016.Vici.170.083), and a Japanese JSPS KAKENHI grant awarded to Jorge N. Tendeiro (21K20211). Correspondence concerning this article should be addressed to: Maximilian Linde, University of Groningen, Department of Psychology, Grote Kruisstraat 2/1, Heymans Building, room 217, 9712 TS Groningen, The Netherlands, Phone: (+31) 50 363 2702, E-mail: m.linde@rug.nl

The analysis code is available at <https://osf.io/rvs9g/>.

Abstract

Linde et al. (2021) compared the “two one-sided tests”, the “highest density interval – region of practical equivalence”, and the “interval Bayes factor” approaches to establishing equivalence in terms of power and Type I error rate using typical decision thresholds. They found that the interval Bayes factor approach exhibited a higher power but also a higher Type I error rate than the other approaches. In response, Campbell and Gustafson (2022) showed that the performances of the three approaches can approximate one another when they are calibrated to have the same Type I error rate. In this article, we argue that these results have little bearing on how these approaches are used in practice; a concrete example is used to highlight this important point.

Keywords: two one-sided tests, highest density interval, region of practical equivalence, interval Bayes factor, optimal test, calibration

Practical Implications of Equating Equivalence Tests: Reply to Campbell and Gustafson (2022)

In a recent simulation study (Linde et al., 2021, henceforth LTSWvR), we compared the performances of three approaches to establishing equivalence between two groups. Specifically, we examined the statistical power and Type I error rate of the following methods: (1) the “two one-sided tests” (TOST; Hodges & Lehmann, 1954; Schuirmann, 1987; Westlake, 1976); (2) the “highest density interval – region of practical equivalence” (HDI-ROPE; Kruschke, 2011, 2015, 2018); and (3) the “interval Bayes factor” (BF; Morey & Rouder, 2011; van Ravenzwaaij et al., 2019; see also Linde & van Ravenzwaaij, 2019). LTSWvR found that when using typical parameters (i.e., $\alpha = .05$ for TOST, 95% HDI for HDI-ROPE, and BF thresholds of 3 and 10 for BF) the BF approach exhibited a higher statistical power to detect equivalence but also a higher Type I error rate than the TOST and HDI-ROPE approaches. This difference in operating characteristics was most evident for relatively small sample sizes (i.e., $n = 50$ and $n = 100$ per group) and narrow equivalence intervals (i.e., $m = .1$ and $m = .2$); in these cases, the TOST and HDI-ROPE approaches were unable to conclude equivalence at all, and application of these procedures would therefore result in a foregone conclusion.

In response, Campbell and Gustafson (2022, henceforth CG) demonstrated that the HDI-ROPE, the BF, and the so-called “optimal test” (OT; Möllenhoff et al., 2022; Romano, 2005) procedures for establishing equivalence “can be reverse-engineered from the other (at least approximately)” so that they exhibit almost identical performance characteristics (CG agree with LTSWvR where TOST is concerned). However, the calibration results of CG often required the adoption of extreme and highly unorthodox values, such as HDIs with coverage close to 0% instead of the typical 95% or BF thresholds higher than 196. In our assessment, no researcher would ever apply these procedures with such threshold values, and therefore we argue that the results of CG have limited impact on statistical practice. Although we believe that CG provide a valuable contribution and

important qualification to the findings of LTSWvR, we are worried that the readership is left with the impression that it does not matter in practice which approach to establishing equivalence is chosen.

To demonstrate why the choice of methodology for establishing equivalence matters *in practice*, we reanalyzed a series of real studies (Galak et al., 2012) that attempted to replicate experiments 8 and 9 of Bem (2011). In his original experiments, Bem (2011) claimed to have found consistent evidence for the existence of precognition, the idea that future events can retroactively influence current experiences and processes. We specifically chose the replications of Galak et al. (2012) for our reanalyses because (1) we can safely assume that the null hypothesis, claiming the absence of an effect, is in fact true (unless we are willing to turn over major parts of the current scientific framework), which is necessary since we want to determine whether the approaches to establishing equivalence make correct decisions; and (2) Galak et al. (2012) provide summary statistics in their Table 1 that we can use for the reanalyses.¹ We reanalyzed experiments 1–5, 6 (test-before practice), and 7; we did not reanalyze the original experiments of Bem (2011) and experiment 6 (practice-before-test) because in the latter case an effect was expected. The relevant summary statistics of Table 1 in Galak et al. (2012) are reproduced in Table 1.

Some alterations to the procedures described in LTSWvR were necessary: Firstly, in order to conform to the study designs in Galak et al. (2012), we adjusted the TOST, HDI-ROPE, and BF procedures to work for one-sample designs. Secondly, CG agree with LTSWvR that TOST is less suitable for relatively small sample sizes and therefore suggest to use OT instead.² This led us to include OT in addition to the TOST, HDI-ROPE, and

¹ Table 1 of Galak et al. (2012) claims to report standard deviations but the reported t -values suggest that the values that are indicated as standard deviations are actually standard errors.

² It is noteworthy that when the sample mean (in case of the one-sample design) or the difference in sample means (in case of the two-sample design) is exactly 0, OT provides a p -value of exactly 0, irrespective of the sample size and the equivalence margin used. This is undesirable as the p -value should not be so extreme in case we only have, say, $N = 5$ cases. Mathematically speaking, a sample mean or a

Table 1

*Summary statistics from Table 1
of Galak et al. (2012).*

Study	N	M	SE
1	112	-1.21	1.01
2	158	0.00	0.77
3	124	1.17	0.92
4	109	1.59	0.90
5	211	-0.49	0.69
6 (TBP ^a)	106	-0.29	0.88
7	2469	-0.05	0.22

Note. ^aTBP = test-before-practice.

BF approaches. For OT, we adapted the code in the Appendix of CG. Similar to LTSWvR, we chose $\alpha = .05$ for TOST and OT, 95% HDI for HDI-ROPE, and $BF_{thr} = 10$ for BF. Moreover, we used a standardized equivalence margin of $m = .1$ and a Cauchy prior scale of $r = 1/\sqrt{2}$ for HDI-ROPE and BF. All code for our analyses can be examined at <https://osf.io/rvs9g/>.

All-or-none Decisions

Table 2 shows the results of our reanalyses. Both TOST and HDI-ROPE only make the correct decision of concluding equivalence in study 7. OT declares equivalence in studies 2 and 7. Lastly, BF concludes equivalence in studies 2, 5, 6, and 7. Thus, BF makes the most correct decisions.

Following the logic of CG, we can calibrate the four approaches, so that they make sample mean difference of exactly 0 should never occur, but in practice researchers round their numbers to finite decimals and as such this test may well produce extremely low p -values even for very little data.

the same decisions. Specifically, if we would have used a Bayes factor threshold between 32.528 (cf. study 5) and 39.911 (cf. study 2), BF would have made the same two equivalence conclusions as OT. Similarly, if we would have used a significance level anywhere between $\alpha = .214$ (cf. study 5) and $\alpha = .544$ (cf. study 3), OT would have made the same four equivalence decisions as BF. However, equating TOST, OT, HDI-ROPE, and BF through this kind of calibration is cumbersome (it requires sifting through all the analysis results and making post-hoc corrections), ad-hoc (different calibration levels for each data set), and defeats the purpose of conducting statistical inference in the first place (thresholds should be set based on substantive reasons or convention).

Bayes Factors Quantify Evidence

A Bayes factor does not *require* a decision threshold. That is, the Bayes factor quantifies evidence; it compares the probability of the data under \mathcal{H}_0 to the probability of the data under \mathcal{H}_1 . In other words, it quantifies the relative predictive performance of the rival hypotheses, and thereby yields the extent to which researchers should update their beliefs after seeing the data. In contrast, a p -value resulting from TOST and OT does not relate to the probability of the data under two hypotheses, but instead focuses solely on the probability of the observed data (or data more extreme) under \mathcal{H}_0 .

In our reanalyses of the Galak et al. (2012) replication experiments, the obtained Bayes factors always provide evidence towards \mathcal{H}_0 (i.e., all $\text{BF}_{01} > 3$). For example, for studies 1 and 3 we obtained $\text{BF}_{01} = 8.205$ and $\text{BF}_{01} = 8.062$, respectively. Even though we cannot conclude equivalence in an all-or-none decision for these Bayes factors (because the evidence falls below the stipulated threshold of 10), they still provide moderate evidence in favor of equivalence. On the other hand, when the result of TOST or OT is a non-significant p -value, we cannot conclude anything because it is impossible to disentangle whether the non-significant p -value resulted from (1) equivalence at the population level combined with insufficient statistical power; (2) non-equivalence at the population level (Bakan, 1966; Keyzers et al., 2020).

Conclusion

We agree with CG that OT, HDI-ROPE, and BF can be calibrated to have the same operating characteristics in terms of statistical power and Type I error rate. However, this calibration requires the use of extreme, highly unorthodox settings; consequently, we do not believe the calibration to be helpful for pragmatic researchers. Our reanalysis of the Galak et al. (2012) replication experiments confirms the practical advantages of the BF approach for equivalence testing.

Table 2

Comparison of TOST, OT, HDI-ROPE, and BF for 7 replication experiments from Galak et al. (2012).

Study	Method			
	TOST (p -value)	OT (p -value)	HDI-ROPE ($\max \text{HDI} $)	BF (BF_{01})
1	.555	.544	0.293	8.205
2	.105	< .001	0.154	39.911
3	.563	.554	0.285	8.062
4	.764	.763	0.351	3.348
5	.229	.214	0.182	32.528
6 (TBP ^a)	.243	.155	0.218	21.664
7	< .001	< .001	0.044	> 1000

Note. Cells with green background indicate equivalence decisions. For TOST and OT, equivalence is concluded if $p < \alpha = .05$; for HDI-ROPE, equivalence is concluded if $\max |\text{HDI}| < m = 0.1$; for BF, equivalence is concluded if $\text{BF}_{01} > \text{BF}_{thr} = 10$.

^aTBP = test-before-practice.

References

- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. <https://doi.org/10.1037/h0020412>
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. <https://doi.org/10.1037/a0021524>
- Campbell, H., & Gustafson, P. (2022). re:Linde et al. (2021): The Bayes factor, HDI-ROPE and frequentist equivalence tests can all be reverse engineered – almost exactly – from one another.
- Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*(6), 933–948. <https://doi.org/10.1037/a0029709>
- Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, *16*(2), 261–268.
- Keyesers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*, 788–799. <https://doi.org/10.1038/s41593-020-0660-4>
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. <https://doi.org/10.1177/1745691611406925>
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd). Academic Press.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. <https://doi.org/10.1177/2515245918771304>

- Linde, M., Tendeiro, J. N., Selker, R., Wagenmakers, E.-J., & van Ravenzwaaij, D. (2021). Decisions about equivalence: A comparison of tost, hdi-rope, and the Bayes factor. *Psychological Methods*, Advance online publication. <https://doi.org/10.1037/met0000402>
- Linde, M., & van Ravenzwaaij, D. (2019). baymedr: An R package for the calculation of Bayes factors for equivalence, non-inferiority, and superiority designs.
- Möllenhoff, K., Loingeville, F., Bertrand, J., Nguyen, T. T., Sharan, S., Zhao, L., Fang, L., Sun, G., Grosser, S., Mentré, F., & Dette, H. (2022). Efficient model-based bioequivalence testing. *Biostatistics*, *23*(1), 314–327. <https://doi.org/10.1093/biostatistics/kxaa026>
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. <https://doi.org/10.1037/a0024377>
- Romano, J. P. (2005). Optimal testing of equivalence hypotheses. *The Annals of Statistics*, *33*(3), 1036–1047. <https://doi.org/10.1214/0090536050000000048>
- Schuurmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. <https://doi.org/10.1007/BF01068419>
- van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(1), 71. <https://doi.org/10.1186/s12874-019-0699-7>
- Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*(4), 741–744. <https://doi.org/10.2307/2529259>