Reply to Reviews

Daniël Lakens[1] & Marie Delacre[2]

[1] Eindhoven University of Technology, Eindhoven, The Netherlands

[2] Service of Analysis of the Data, Université Libre de Bruxelles, Belgium

Author Note

Correspondence concerning this article should be addressed to Daniël Lakens, Den Dolech 1, IPO 1.33, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

Abstract

We respond to all the comments raised by the two reviewers and the editors. We would like to thank the editor and reviewers for their suggestions for improvements, which we believe made our main argument much clearer, focussed, and formally correct. We feel we were able to accomodate all suggestions, and we hope you think our submission is ready to be accepted for publication in Meta-Psychology.

Reply to Reviews

Dear editor,

Thank you for your comments on our submitted manuscript. We have incorporated your three comments:

1. We moved the first sentence on page 3 into the next paragraph (well spotted).
2. The code for the Shiny app was on GitHub, and the GitHub repository is linked to the OSF project page, but we have added a component to the OSF page and also directly added the Shiny App code there to make it easier to find: https://osf.io/7f8uq/
3. We now immediately refer to Figure 4 in the beginning of the paragraph in which the Figure is described.

We will respond to the points raised by both reviewers in turn.

Reply to Reviewer 1

1. We have changed the part of the sentence "Our conclusions about the relationship between TOST *p*-values and SGPV in this article are not dependent upon any specific example" into "Our conclusions about the relationship between TOST *p*-values and SGPV hold for an SGPV calculated from confidence intervals, and assuming continuous normally distributed data.". We mainly intended to refer to our conclusion that: "When confidence intervals are symmetric we can think of the SGPV as a straight line that is directly related to the *p*-value from an equivalence test for three values" namely p = 0.5 for TOST when the SGPV will be 0.5, and the *p*-values for TOST of 0.025 and 0.975 (corresponding to the largest observed mean where the SGPV is 0 and the smallest observed mean where the SGPV is 1). We have now

made it clearer that when we talk about "correspondence between TOST p-values and the SGPV", we mean that there is a direct relation between three points.

2. We appreciate correcting our use of the term "unrelated" and teaching us the word "bijective". We used the word "unrelated" somewhat naievely, looking at the Figures and wanting to express that across the three identified situations, a different TOST p-value does not imply a different SGPV - the SGPV across these ranges of means is aways the same (either 0, 0.5, or 1) while the TOST p-value differs across this range of means. We now just directly say what we originally meant, and no longer use the word "unrelated". For example, "There are three situations where p-values for TOST differentiate between observed results, while the SGPV does not differentiate." and "While the SGPV is 1 as long as the confidence interval falls completely within the equivalence bounds, the p-value for the TOST continues to differentiate between results as a function of how far the confidence interval lies within the equivalence bounds (the further the confidence interval is from both bounds, the lower the p-value)." and similarly "A third situation in which the SGPV remains stable across a range of observed effects, while the TOST p-value continues to differentiate, is whenever the CI is wider than the equivalence range, and the CI overlaps with the upper *and* lower equivalence bound."

3. Please do not feel like you ever need to apologize for having formal training we lack, that allows you to point out whenever we make formally incorrect statements. This is exactly why we submit our work for peer review - to be able to improve it based on feedback from experts. We now preface the sentence with "As long as data is normally distributed,". We follow your recommendation throughout to not generalize beyond the scenarios we study. In the second paragraph of the manuscript we now clearly state: "We limit our analysis to normally distributed continuous data." See also point 5 below.

4. After making the change at point 2 above, we now think the sentence “The conclusion of this section on asymmetric confidence intervals is that a SGPV of 1 or 0 can still be interpreted as a p < 0:025 or p > 0:975 in an equivalence test, since the SGPV and *p*-value for the TOST procedure are always directly related at these values.” should no longer be confusing. To make the sentence clearer, we have changed “these values” into “at the values p = 0.025 and p = 0.975” It should now be clear the sections are not contradictory - TOST differentiates below p < 0.025, when the SGPV is 1 for all p < 0.025.

5. We agree that is unclear whether the original statement by Blume et al that “the degree of overlap conveys how compatible the data are with the null premise” holds in general. We have now amended the sentence to read: “Although a SGPV of 1 or 0 can be directly interpreted, a SGPV between 0 and 1 is not interpretable as ‘compatibility with the null hypothesis’ under the assumption of normally distributed data, and the generalizability of this statement needs to be examined beyond normally distributed data requires further examination.”

6. We have removed our earlier statements that the SGPV only works as a descriptive statistic. We deleted the sentence “the SGPV is a descriptive statistic (unlike the *p*-value that is calculated for an equivalence test, which is an inferential statistic) that provides the proportion of overlap of the confidence interval and the equivalence range.” and deleted “The novelty of the SGPV lies in its use as a descriptive statistic” from the conclusion. Indeed, we acknowledge that “Although a SGPV of 1 or 0 can be directly interpreted, a SGPV between 0 and 1 is not interpretable as “compatibility with the null hypothesis”." So the SGPV is an inferential statistic when interpreted dichotomously. We will leave our uncertainty of how specific SGPV values (e.g., SGPV = 0.3) should be interpreted to future researchers, and we do not make any statements about whether such values are descriptive or inferential.

7. We agree that using TOST for small samples (e.g., N = 4) will not be useful. Indeed, an a-priori power analysis for equivalence tests would prevent anyone from planning such a study, since the result will be that such a study has 0 power - the CI will never fit withing the equivalence bounds.

8. With our recommendations, we have limited our conclusions to the specific cases we have examined, and noted the possibility for future research to examine a wider set of assumptions. An additional change not mentioned above is that we renamed the section on asymmetric confidence intervals to the more specific "The relation between equivalence tests and SGPV for asymmetrical confidence intervals around correlations" instead of the more general "The relation between equivalence tests and SGPV when confidence intervals are not symmetrical". We appreciate the very positive evaluation of our manuscript, but even more so the criticism which allowed us to improve the manuscript.

Reply to Reviewer 2

Before listing major points, the reviewer raises the point that we compare SGPV with *p*-values from TOST throughout our paper, and in all figures, but then at the end, in page 20, we write:

"Equivalence tests combined with null-hypothesis significance tests also allow researchers to conclude an effect is significant and equivalent (i.e., statistically different from zero, but also too small to be considered meaningful). Thus, the SGPV is used to classify results into one of three possible categories (with the data falling inside or outside the equivalence range, or being inconclusive), while equivalence tests combined with null-hypothesis tests classify results into four possible categories."

The reviewer says that "the authors suddenly switch the premise; now the SGPV should be replaced by two *p*-values, one from an equivalence test and one from a NHST".

However, we do not suggest to replace the SGPV with an equivalence test and a NHST at all. In our manuscript, we already pointed out how the SGPV is basically the same as an equivalence test and a minimum effect test (on page 7):

"The third point where the SGPV and the *p*-value from the TOST procedure should overlap is where the 95% CI falls completely outside of the equivalence range, but one endpoint of the confidence interval is equal to the equivalence bound (see situation C in Figure 3), when the *p*-value will always be 0.975. Note that this situation is in essence a minimum-effect test (Murphy, Myors, & Wolach, 2014). The goal of a minimum-effect is not just to reject a difference of zero, but to reject the smallest effect size of interest (i.e., the equivalence bounds). The SGPV summarizes the information from an equivalence test and a minimum-effect test."

One might argue that we suggest to replace the SGPV with two tests. However, the overlap of the TOST *p*-value at 0.975 when SGPV = 0 gives away that the equivalence test can be interpreted as a minimum effect test against the equivalence bound. When the TOST *p*-value is 0.975, the 95% CI touches either the upper or lower equivalence bound, and fall complately outside of the equivalence range. This is mathematically identical to a statistically significant minimum effect test at $p < 0.05$. We didn't make this explicit, but we clearly should have. We have added the sentence (on page 7):

"An equivalence test and minimum effect test against the same equivalence bound are complementary, and when a TOST *p*-value is larger than 0.975, the *p*-value for the minimum effect test is smaller than 0.05, and the minimum effect test provides no additional information that can not be derived from the *p*-value from the equivalence test."

One might wonder why then, on page 20, why we introduce the idea to perform both a null-hypothesis significance test and an equivalence test. We were not clear enough. And indeed, the sentence "This interpretation of a SGPV as allowing researchers to reject the null, reject the presence of a meaningful effect, or remaining inconclusive is very similar to

the Neyman-Pearson interpretation of combining a null-hypothesis test and an equivalence test" was simply confusing. We have deleted this sentence and replaced it with "As already mentioned, this interpretation of a SGPV as is very similar to the Neyman-Pearson interpretation of an equivalence test and a minimum effect tests (which are complementary)."

Furthermore, to prevent the apparent confusion about why we suggested to perform a null-hypothesis test and equivalence test. Part of this confusion was our original sentence comparing the SGPV and a combination of NHST and an equivalence test which was not needed, and we understand it made it seem as if SGPV needs to be replaced by a combination of other tests. But that is not what we propose. Werewrote the paragraph:

"Some researchers might be interested in combining an equivalence tests with a null-hypothesis significance test. This allows a researcher to ask whether there is an effect that is statistically different from zero, and whether effect sizes that are considered meaningful can be rejected. Equivalence tests combined with null-hypothesis tests classify results into four possible categories, and for example allow researchers to conclude an effect is significant and equivalent (i.e., statistically different from zero, but also too small to be considered meaningful, see (Lakens et al., 2018))."

We hope this makes it clear we are not pulling a bait and switch. The SGPV is strongly related to an equivalence test (or it's complement, a minimum effect test). Describing the relation between the SGPV and TOST is our main contribution. We just wanted to mention that researchers might want to combine an equivalence test with a NHST, because in the literature, this is a question we see many psychologists ask.

We will now discuss the major points in turn.

1. In line with the comments by reviewer 1, in the second paragraph of the manuscript we now clearly state: "We limit our analysis to normally distributed continuous data." After all, p-values and confidence are directly related, but Bayesian credible

intervals would not necessarily be, depending on the prior. Blume et al (2018) limit their examples to confidence intervals, but explicitly acknowledge other possibilities (such as Bayesian intervals). We look forward to future papers that illustrate the SGPV with other interval estimates and are happy to add the limitation to continuous normally distributed data in our article. Note that we did not add that the correspondence only holds for situations where the SE's are small compared to the width of the equivalence range, as the reviewer suggests.

2. The reviewer repeats that the correspondence only holds when SE's are small compared to the width of the equivalence range. The correspondence is not in the shape of the curve (which clearly differs across Figures 2, 6, and 9). We clearly state that: "When confidence intervals are symmetric we can think of the SGPV as a straight line that is directly related to the *p*-value from an equivalence test for three values." This correspondence is not dependent upon the SE's. It will always hold for continuous normal data. The reviewer says that "The paper does not provide theoretical justification for this connection nor any proofs." We are unsure which proof is needed for the overlap between the three points, since it is rather trivial and we believe the explanation in the paper is very clear.

For the relation between TOST p = 0.5 and SGPV = 0.5, TOST p = 0.025 and SGPV = 1, and TOST p = 0.975 and SGPV = 1: "The SGPV is 50% when the observed mean falls exactly on the lower or upper equivalence bound, because 50% of the symmetrical confidence interval overlaps with the equivalence range. When the observed mean equals the equivalence bound, the difference between the mean in the data and the equivalence bound is 0, the t-value for the equivalence test is also 0, and thus the *p*-value is 0.5 (situation A, Figure 3). Two other points always have to overlap. When the 95% CI falls completely inside the equivalence region, and one endpoint of the confidence interval is exactly equal to one of the equivalence bounds (see situation B in Figure 3) the TOST

192 p-value (which relies on a one-sided test) is always 0.025, and the SGPV is 1. The third
193 point where the SGPV and the p-value from the TOST procedure should overlap is where
194 the 95% CI falls completely outside of the equivalence range, but one endpoint of the
195 confidence interval is equal to the equivalence bound (see situation C in Figure 3), when
196 the p-value will always be 0.975, and the SGPV is 0.

197 We do not think we can provide any other proof for this. A confidence interval and a
198 p-value are directly related. The TOST procedure is a test using the confidence interval,
199 and the SGPV (now that we clearly limited our discussion to a SGPV based on a
200 confidence interval calculated from continuous normal data) compared against the same
201 equivalence bounds will be related as we describe. But we believe our article might not
202 have been clear enough about that we limit our discussion about the correspondence of the
203 curves to these 3 points. Indeed, in some figures, the curves are less similar (e.g., Figure 6,
204 9, 13). Nevertheless, in these graphs, we still have the correspondence between these 3
205 points (or at least the 2 points if the CI is too wide compared to the equivalence bounds to
206 ever observe a $p = 0.025$). To make it clearer that we talk about the overlap in these 3
207 critical points, we have added:

208 "When we discuss the relationship between the p-values from TOST and the SGPV,
209 we focus on their correspondence at three values, namely where the TOST $p = 0.025$ and
210 SGPV is 1, where the TOST $p = 0.5$ and SGPV = 0.5, and where the TOST $p = 0.975$
211 and SGPV = 1. These three values are important for the SGPV because they indicate the
212 values at which the SGPV indicates the data should be interpreted as compatible with the
213 null hypothesis (SGPV = 1), or with the alternative hypothesis (SGPV = 0), or when the
214 data are strictly inconclusive (SGPV = 0.5). These three points of overlap are indicated by
215 the horizontal dotted lines in Figure 2."

216 We believe this satisfactorily addresses the point by the reviewer that "Because of
217 this lack of discussion, the reader can easily miss that the close correspondence only occurs

when the sample size is large enough that the standard error is very small relative to the width of the equivalence zone." Indeed, the correspondence between the TOST and SGPV at these 3 points should be obvious to anyone who understands the relationship between a p-value and a confidence interval, so we do not provide any further proof, but we hope that clarifying we talk about the correspondence of these 3 points, and not the entire curve, is sufficient to address point 2.

3. We are aware of the Beger and Hsu (1996) article. But if the reviewer reads more closely, it should be clear that we never say the TOST is UMP. We say that TOST is a test procedure that aims to control error rates - our use of the term "Neyman-Pearson perspective" refers to the classic difference between significance tests (Fisher) and hypothesis tests (Neyman-Pearson). We do not imply that any test that fixes an error rate is a NP test. Because many psychologists are not trained in the difference between a Fisherian approach and a Neyman Pearson approach to the use of p-values, we are simply adding a "Neyman-Pearson" approach in the text to explain that "From a Neyman-Pearson perspective on statistical inferences the main conclusion is based on whether or not $p < \alpha$, and thus an equivalence test and SGPV can be performed by simply checking whether the confidence interval falls within the equivalence range, just as a null-hypothesis test can be performed by checking whether the confidence interval contains zero or not." We believe this is a correct use of the term "Neyman-Pearson perspective on statistical inferences".

4. We tried to find the passage the reviewer refers to when he writes "In light of this, the assertion without proof that classical p-values provide more or better assessments of the "evidence" appear out of place." but we could not. We do not use the word "evidence" in the text (because neither the p-value from TOST nor the SGPV are measures of "evidence" - we fully agree with BLume et al., 2018, in this respect). We also do not say the p-value is "naturally informative". Indeed, related to point 3

above, we tried to be consistent in interpreting the TOST procedure from a Neyman-Pearson perspective, and focus on error control. So we regrettably do not know how to incorporate this point.

5. We fully agree with the reviewer on point 5 that "The maximum Type I Error rate for the SGPV over the null interval is 5%, exactly like the TOST test." Indeed, given the direct overlap between the TOST p-value and the SGPV at $p = 0.025$ and 0.0975 and SGPV of 1 and 0, any error rates (when concluding support for the null or alternative when the SGPV is either 1 or 0) must be identical. This is why we were so surprised when Blume et al (2018) write "They control the Type I error naturally, forcing it to zero as the sample size grows. This, in turn, offsets Type I Error inflation that results from multiple comparisons or multiple examinations of accumulating data. Findings identified by second-generation p-values are less likely to be false discoveries than findings identified by classical p-values. Consequently, second-generation p-values do not require ad-hoc adjustments to provide strict error control and this improves power in studies with massive multiple comparisons." Clearly, it can not both be true that the error rate of SGPV and TOST is identical, TOST requires adjusting for multiple comparisons, but the SGPV does not.

The editor asks us to better substantiate the frequency claims for the comparison of TOST and SGPV, but since the SGPV and TOST p-value are directly related (a TOST $p < 0.025$ = a SGPV of 1) we do not think any additional substantiation is possible, especially since the reviewer agrees that (as long as confidence intervals are used to calculate the SGPV, which we now clearly limit our conclusions to) "The maximum Type I Error rate for the SGPV over the null interval is 5%, exactly like the TOST test".

Nevertheless, we agree this discussion was a bit too extensive. Given my own fondness of error control in statistical inferences, I get worried if researchers suggest we can ignore it, especially when making multiple comparisons. Imagine people start to use the SGPV

instead of the TOST procedure whenever they perform 20 tests, and then argue they do not need to correct for multiple comparison - I think we can all agree this is undesireable! We have reduced the two paragraphs to one hopefully clear statement about the need to control error rates, also when using the SGPV. We hope this addresses the concerns by the editor as well as the reviewer, while still making an important point in the literature about the need to control error rates (if one cares about drawing incorrect conclusions).

Blume, D'Agostino McGowan, Dupont, and Greevy (2018) claim that "Adjustments for multiple comparisons are obviated" (p. 15). However, this is not correct. Given the direct relationship between TOST and SGPV highlighted in this manuscript (where to TOST $p = 0.025$ equals SGPV = 1, as long as the SGPV is calculated based on confidence intervals, and assuming continuous and normally distributed data), not correcting for multiple comparisons will inflate the probability of concluding the absence of a meaningful effect based on the SGPV in exactly the same way as for equivalence tests. Whenever statistical tests are interpreted as support for a hypothesis (e.g., SPGV = 0 or SGPV =1), it is possible do so erroneously, and if researchers want to control error rates, they need to correct for multiple comparisons.

6. We have removed this section while addressing point 5 above, and thus incorporated the reviewers'suggestion to remove these statements.

7. We have not removed the paragraph about multiple comparisons, as the reviewer asks, but we have simplified our treatment of it. We do not need any other argument to make our claims about multiple comparisons that fact, noted by the reviewer, that "The maximum Type I Error rate for the SGPV over the null interval is 5%, exactly like the TOST test." This is true, and it logically follows (hence, no mathematical proof is needed) that error rates need to be controlled when using SGPV to conclude support for the null-hypothesis. We believe error control is important enough to include this paragraph in the paper, as we believe that lack of error control was one

of the causes of the replication crisis in psychology.

8. We thank the reviewer for this comment. Indeed, to calculate confidence intervals around a correlation, the Fisher's z-transformation is used, which is approximately normally distributed, and symmatric CI are calculated. When transforming these confidence intervals back to correlations, they become asymmetric. We have added to the text: "Note that calculating confidence intervals for a correlation involves a Fisher's z-transformation, which transforms values such that they are approximately normally distributed, which allows one to compute symmetric confidence intervals, which are then retransformed into a correlation, where the confidence intervals are asymmetric if the correlation is not exactly zero." This is simply one example of a general effect mentioned in Blume et al (2018), which we already cited in the main text: "The magnitude of an inconclusive second-generation p-value can vary slightly when the effect size scale is transformed. However definitive findings, i.e. a p-value of 0 or 1 are *not* affected by the scale changes."

Minor Concerns:

1. We added ",or more extreme data,"

2. This text was deleted (see major point 5 above)

3. We see how "reverse the SGPV" might have been confusing. We did not mean to imply to reverse the null and alternative. We now just say that we plot 1-SGPV (which is what we incorrectly tried to express with "reverse").

4. The issue of the SGPV needing a "correction" by setting it to 0.5 is not related to the fact that large p-values are not evidence for the null. The p-value is behaving as it is intended to, without any correction. A high p-value is not a "misleading result" - although it can be misinterpreted (which is exactly what we try to prevent with our work on equivalence testing). We therefore did not incorporate this point.

5. In line with comments by reviewer 1, we removed statemetns about SGPV not being inferential, thereby solving the issue reviewer 2 raised.

6. In line with major point 8, we added information that asymmetric condifence intervals around a correlation are symmetric when Fisher z-transformed. This is just an example of what Blume et al (2018) refer to when they write "The magnitude of an inconclusive second-generation p-value can vary slightly when the effect size scale is transformed. However definitive findings, i.e. a p-value of 0 or 1 are *not* affected by the scale changes." We beieve this section, although the most detailed, is interesting, even only for more expert readers (indeed, Reviewer 1 notes that "The asymmetric CI case is very interesting"). We have made our point less general, now only talking about correlations, by changing the title of this section into "The relation between equivalence tests and SGPV for asymmetrical confidence intervals around correlations". The reviewer is correct confidence intervals can be created in many difference ways (e.g., with holes in the middle), but for this paper, we just focus on one specific case now, which we believe might be closest to what researchers would use in real life, the confidence interval around a correlation.

7. It is a well-known fact that the TOST can have 0 power (and indeed, this is one of the main issues in the Perlman and Wu (1999) paper the reviewer cited). As explained in Lakens (2017): "It is important to take statistical power into account when determining the equivalence bounds because, in small samples (where CIs are wide), a study might have no statistical power (i.e., the CI will always be so wide that it is necessarily wider than the equivalence bounds)." We agree with the reviewer it is not interesting to do a test if you have 0 power. People should design informative experiments by doing an a-priori power analysis. Any problem in this respect is not a problem with equivalence tests, but a problem with designing a bad study. Not that the exact same issue applies to the SGPV. It is also possible to have

0 power for the SGPV - so we do not understand why the reviewer is so surprised about this. The question "Why use the TOST test then?" equally applies to "Why use the SGPV then?" if you, from the outset, know that the "small sample correction" will lead to a SGPV of 0.5.

8. We do not see how setting the SGPV to 0.5 "shrinks" an estimate to some value. Regardless of how correction factors are applied in statistics, we are trying to educate our readers about what the "small sample correction" does. We believe our readers will associate "correction" with "correcting", and want to explain that the SGPV is not set to a "correct" value. We are not talking about statistical terms here - we are warning the reader to not interpret "correction" in terms of normal language use. This is similar to warning readers that "significant" does not have the normal language connotation of "important". We have therefore kept our original sentence.

As a final comment, altough not explicitly listed as a major point of minor point, we are happy to respond to the reviewers' suggestion that Perlman and Wu (1999) call for an "outright ban of TOST". The best summary of this discussion (of which Perlman & Wu is one contribution) can be found in Meyners (2012), which we cited. The discussion has different arguments, and Meyners draws the conclusion that there is no universially preferred approach. The power benefits of alternative approaches are typically modest, or only noticeable in situations where all alternatives are underpowered. The power issue is not that important - the main discussion is about whether equivalence tests should have a bounded rejection region, ike TOST, or an unbounded rejection region. It might be useful to note that TOST and SGPV are similar in this respect - both have bounded rejection regions, and if the reviewer agrees with Perlman and Wu (1999) in the criticism against TOST, the same criticism applies to the SGPV. If the reviewer reads Perlman and Wu (1999) as an argument to ban TOST, it is at the same time an argument to ban the SGPV. We agree with Meyners, who writes: "The debate has been quite heated in times; this

section intends to give a brief overview of the debate. We would like to emphasize again that the differences between the approaches are minor and to a large extent of academic interest only, as is this debate." and concludes: "To summarize, there is no right or wrong in most of the arguments; the two opposing parties rather apply different criteria to select the most appropriate test, one group focusing rather on mathematical/statistical criteria, the other one more on interpretability and reasonability." We have added a reference to this discussion for the interested reader, and added "For an excellent discussion of the strengths and weaknesses of different frequentist equivalence tests, including alternatives to the TOST procedure, see Meyners, 2012."

Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., & Greevy, R. A. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE*, *13*(3), e0188299. doi:10.1371/journal.pone.0188299