Decision Letter: Equivalence Testing and the Second Generation P-value

Stephen R. Martin

02/05/2019

Two reviewers (Jeffrey Blume, Oscar Olvera) provided useful feedback, concerns, and questions about the manuscript. The reviews are available in full below. Given their comments, I believe a revision will improve the manuscript, and I invite you to read and address their comments.

Editorially, my comments are minor:

- Page 3: Move first sentence into the paragraph below it.
- Page 5: Is the code for the shiny app available? If so, it would be ideal to include this as supplementary material in the OSF page for this manuscript.
- Page 8: Move the parenthetical reference to figure 4 to earlier in the paragraph, so that it is clear from the beginning what you are referring to.

Dr. Olvera points out several areas where the claims need clarification, correction, or tempering. I agree with these points, and would like them addressed. Dr. Blume has several major concerns, some which I agree with, and some which were shared by Dr. Olvera. In particular:

- The SGPV simultaneously describes agreement and disagreement with the null region. For TOST to provide equivalent information, the researcher must conduct the TOST *and* a null test. However, the plots only seemingly show one of these; it may be useful to add additional curves to the present figures for the null p-values (since in practice, one would need both TOST and NHST to make a claim).
- The frequency claims about the SGPV should be better substantiated.
- When discussing the asymmetric CIs, ensure that the comparisons are equivalent. Namely, that if the CIs are constructed in a transformed space wherein they are symmetric, the criticism may not be applicable *on that transformed space*, but limited to when the endpoints are inverse-transformed.
- The SGPV is not limited to being a descriptive statistic.

Both reviewers provided many useful comments, including points of possible confusion and potential improvements. I look forward to reading a revision of your manuscript, with reviewer comments addressed.

1 Review 1: Dr. Olvera

Review for *Equivalence Testing and the Second Generation P-Value*

October 3, 2018

Hello. My name is Oscar L. Olvera Astivia. I am a post-doctoral researcher in the School of Population and Public Health (SPPH) in the University of British Columbia (UBC) in Vancouver, Canada. I would like to thank Dr. Stephen Martin, for the opportunity to be a reviewer for this manuscript. I sincerely hope you'll find some of my comments useful to improve it.

First, I'd like to praise Dr.Lakens and (future) Dr.Delacre for this manuscript. It offers a very clear overview of both the Two One-Sided Test (TOST) and Second Generation P Value (SGPV) so that researchers unfamiliar with the work can become better acquainted with them. Given the abundance of misconceptions regarding proper ways to show evidence for equivalence or lack of differences, it is encouraging to find a manuscript that is both principled yet accessible to researchers. It is definitely an excellent contribution to the scientific literature.

Even though I could not find anything particularly objectionable within the manuscript, there are a few areas where I must say I was left either slightly confused or where I believe the conclusions need to be tempered. Perhaps either elaborating further (or providing a reference) would help readers like myself understand what the point of certain paragraphs is. Also, some of the comments that I make in this review are unnecessarily pedantic and I apologize in advance. This comes as part of my formal technical training so using words or conveying the sense that something"always" or "never" happens immediately triggers in me the need to either see a formal proof for or a counter-example. For the sake of clarity, every time I make a point and reference the n^{th} line in a page, it is always the n^{th} line counting from the top of the page to the bottom of the page.

- p#5–8th line: **Our conclusions about the relationship between TOST p-values and SGPV in this article are not dependent upon any specific example**.

  I can see where the need to make a general statement like this is coming from but I am not sure if one can claim this. Consider the example in Figure 6 where there is no "overlap" between the TOST and SGPV curves in the same way there is one in Figure 1. To get the confidence interval (CI) to overlap with the equivalence bounds, one needs the sample size to be low so that the uncertainty around the estimates is large.

Indeed, for n=10 in Figure 6 this holds true. However, if one uses the shiny calculator kindly provided by the authors (an excellent resource in itself) and set n=1000, we can see the new plot looks a lot more like the Plot on Figure 1. So, in my mind, many of the statements made throughout the manuscript *do* rely on specific examples, namely, cases with small sample sizes.

- p#9–5th line: **There are three situations where p-values from TOST and SGPV are unrelated.**

  Saying that two quantities are unrelated to me implies that there is no relationship between them. But here there clearly is! When the SGPV is 0, e.g., the TOST p-value is between 0 and 0.025, and vice-versa. The ranges are intimately linked. Critically, they both give the same evidentiary information here. Perhaps the authors were aiming for something like there is no (bijective) function that defines a one-to-one correspondence between the p-value of the TOST procedure and the SGPV? Statements about the lack of relationship between the TOST p-value and the SGPV need to be tempered a little more, in my opinion. Similar claims are made on p#12 and p#13. Curiously, on p#10 the claim reads: *A third situation in which the SGPV deviates from the TOST p-value* which rings more closely to what I believe the authors are trying to convey.

- p#14–3th line: **For correlations, which are bound between -1 and 1, confidence intervals are only symmetric for a correlation of exactly 0.**

  True, but only under the assumption of normal theory confidence intervals. I realize it seems like minutia but in the next point it will become apparent why I make this distinction.

- p#17–7th line: **There is always a 50 % probability of observing a correlation smaller or larger than the true correlation.**

  This is not true. Again, this is me being pedantic but I have come to realize over the years that our reliance on normal theory results tends to sway our intuition towards generalizations that, when examined mathematically, simply do not hold. This statement is only true if the data are being sampled from a bivariate normal distribution. I will present a counter-example using binary data (so the phi coefficient) because it is more straightforward to show. I could build a similar argument for continuous, non-normal data, though, but that would distract from the main point. Consider the following example of two jointly-distributed Bernoulli random variables with parameters ($p_1 = 0.5, p_2 = 0.75$) and population Pearson correlation $\rho = 0.55$. A quick simulation with a large sample of $N = 10,000$ using the `bindata` R package shows:

```
library(bindata)
set.seed(123)

reps<-1000
```

```
r <- double(reps)

for (i in 1:reps) {
rho <- 0.55
m <- matrix(c(1,rho,rho,1), ncol=2)
Q <- rmvbin(10000, margprob = c(0.5, 0.75), bincorr = m)
r[i] <- cor(Q)[1,2]}

sum(r<rho)/reps
[1] 0.23
sum(r>rho)/reps
[1] 0.77
```

So correlations *greater* than the population parameter are more common than those less than it. This is simply using the well known fact that, if $p_1 \neq p_2 \neq 0.5$ the phi coefficient does not span the full $[-1, +1]$ range that the Pearson correlation should do.

- p#19–1st-3rd line: **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.**

  The same argument was using on p#9 (or the 2nd point in this list) to say that the TOST p-values and the SGPV are *un*related. And I already expressed my thoughts about this issue of the relationship between the TOST procedure and SGPV. So, are they related or unrelated based on this argument of whether 1 or 0 maps on to $p < 0.025$ or $p > 0.975$? As it stands I feel the sections are self-contradictory.

- p#19–3rd-6th line: **Although Blume et al. (2018) state that "the degree of overlap conveys how compatible the data are with the null premise" this definition of what the SGPV provides does not hold for asymmetric confidence intervals.**

  Given what was shown before regarding the issue of 50 % probability of an observed, sample correlation being larger or smaller than the true, population correlation, I am no longer sure if this holds true in general. Yes, this is problematic under *normal theory* CIs but with a cleverly-constructed sampling distribution (like that of the phi coefficient simulated before) I am wondering whether or not one can create example where Blume et al. (2018) is right. If anything, I would ask the authors to temper the claim and say that this is not known and a more theoretical investigation of the properties of the SGPV under more general conditions is needed.

- p#19–12th line: **First of all, the SGPV is a descriptive statistic (unlike the p-value that is calculated for an equivalence test, which is an inferential statistic)**

  It is difficult for me to conceptualize the SGPV as a purely *descriptive* statistic given that the definition provided by the authors (which I'm assuming is taken from Blume et.al. (2018)) depends on the calculation of a confidence interval, which is most definitely an inferential statistic. Then again it is unclear to me from the literature review whether Blume et.al. (2018) themselves consider the SGPV as a purely *descriptive* statistic, in which case I guess one should stay truthful to what the creators intended. But, if it is not, I would need to see more evidence that this is exclusively descriptive. Particularly for cases when the inferences are the same as those of the TOST procedure.

As a quick summary of my observations (and for most applied, practical purposes) I found that the TOST procedure and the SGPV perform much more similarly than I expected them to. Two concerns were raised regarding whether or not they offer different interpretations with either wide, symmetric CIs or asymmetric CIs. For the wide, symmetric CI case I feel the cases presented would make the use of the TOST procedure *and* the SGPV questionable. Performing inference in sample sizes of 4 and 10 would raise a red flag (I hope) for any reader looking at a published study. But if the differences between SGPV and the TOST procedure only arise in those cases, then, if anything, this becomes a cautionary note on the dangers of using small samples when working with anything that involves p-values, not an inherent shortcoming of the SGPV itself.

The asymmetric CI case is very interesting, but (at least from what was described in the example) it hinges on very specific properties of the sampling distribution of the statistic in question (the Pearson correlation sampled from a bivariate normal distribution) that are not true in general (as shown in my simulation). I am not requesting the authors to provide a simulation to evaluate which procedure is more robust to assumptions under which conditions, but I would like to strongly encourage the authors to restrict their conclusion regarding differences to the one case that they are presenting. If a more general case can be derived under more general conditions, it would be a welcome addition to the manuscript.

Overall, this is an excellent manuscript, it does a very well job in showing the cases where both procedures diverge, but it would help the reader if the conclusions regarding their differences were stated *exclusively* for the conditions described and not implying they work in more general settings.

Yours truly,

Oscar L. Olvera Astivia

2 Review 2: Dr. Blume

Lakens and Delacre
Equivalence testing and the second generation P-value

Summary
This paper examines the relationship between the newly proposed second-generation p-value (SGPV) and the (classical) p-value from an equivalence test (classical p-value). The paper presents several examples in which there is an apparent correspondence between the classical p-value and the SGPV. The paper then claims because of this correspondence that a classical p-value from an equivalence test, combined with a classical p-value from a NHST, is preferable to a SGPV. The paper makes a number of (mistaken) claims about the frequency properties of the SGPV. In more than one place, the paper fails to recognize that the SGPV and the TOST p-value have quite similar frequency properties.

Comments
I found that

- the paper oversells equivalence tests (to its own detriment; a more objective narrative would hold the reader's attention longer)
- that the paper lacks proper discussions about the generalizability of the examples presented (which are not support by any theoretical arguments)
- that the paper fails to mention any of the well know theoretical issues of TOST equivalence tests by Berger and Hsu (1996 in statistical science) and Pearlman and Wu (1999 statistical science). The former calls for the outright ban of TOST tests.
- that the paper contains multiple technical misstatements that impact the paper's claims

Most concerning, however, was that the paper seemed to pull a bait and switch on the reader. The abstract, introduction and conclusion imply that (1) the SGPV should be compared to classical p-values from an equivalence test, and (2) the classical p-value from an equivalence test is all that is needed. Yet on page 20 (!) 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, despite many examples of the first comparison, there is not a single example of the second. So it is impossible for the reader to assess this (very strongly worded) claim.

The authors are correct that the numerical comparison between the SGPV and the classical p-value is flawed. When the SGPV is 1, it indicates support for the null. But when the classical p-value is 1, it indicates that data are inconclusive. So when they are numerically equal near 1, they are also inferentially opposed. This nuance is completely lost in the first 19 pages and this is a major flaw of the paper.

In my opinion, the paper should be re-worked to compare like with like. Since the SGPV has three regions of inference, the authors need a comparator metric that has three regions. They can either argue against 100+ years of accepted statistical practice that large p-values from an hypothesis test can be interpreted as evidence for the null hypothesis. Or they can map their proposed two p-value metric to the three inference regions and compare. They should also be upfront about this, stating their intentions in the abstract, introduction, and conclusions as well as outlining their approach and the mapping to the three regions early on.

Major concerns

1. The paper needs to more clearly explain when the supposed correspondence between SGPV and classical equivalence testing p-values holds. As far as I can tell, this holds only for cases

when (1) the SEs are very small relative to the width of indifference zone, (2) the sampling distribution is continuous, (3) the sampling distribution is approximately normal, and (3) nominal CIs and p-values are computed. The concern is that this is highly specific, especially the required conditions of such a large sample size, which are rarely met in practice.

2. The paper does not provide theoretical justification for this connection nor any proofs. The paper also does not discuss the extent to which these connections generalize. Because of 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. Most applications of statistical tests will never meet this criterion.

3. The authors appear unware that the TOST is not a Neyman-Pearson (NP) hypothesis test. The paper implies that any test that fixes its Type I Error rate is a NP test, but this is mistaken. A NP test is very specific test. It rejects the null hypothesis when the likelihood ratio is large and only when the likelihood ratio is large. Neyman and Pearson showed that tests of this type are uniformly most powerful (but that UMP tests are not unique). However, TOST tests are not NP tests; they are intersection-union test (Berger and Hsu, 1996 in statistical science). This is partly why they have severely criticized in the literature (Berger and Hsu, 1996 in statistical science) and (Pearlman and Wu, 1999 statistical science).

4. The paper claims in many places that classical p-values are naturally informative. Most statisticians would not agree, especially since a single p-value can map back to many different CIs with varying lengths. It is not too hard to do this for equivalence tests. In light of this, the assertion without proof that classical p-values provide more or better assessments of the "evidence" appear out of place.

5. The paragraphs about frequency properties (page 22) are largely off base and should be stricken. For example, the two approaches have very similar frequency properties when it comes to the Type I error rate. The maximum Type I Error rate for the SGPV over the null interval is 5%, exactly like the TOST test. And the observed TOST Type I Error rate is often much less than 5%, exactly like SGPVs.

6. Page 22: The statement "The TOST procedure always has higher power to declare equivalence than the SGPV…" should be stricken; it is false. This is too general to be true and I was easily able to generate a counter example for a one-sample exact test for correlation. Claims like this should be supported by a proof. Also note that only tests with equal Type I Error rates can be compared on the basis of their power alone.

7. Page 22: Paragraph on multiple comparisons should be removed. This paragraph is too vague and non-specific to disprove anything; and it is distracting. Blume (2018) appears to provide a careful mathematical argument and an example. Yet this paper just claims that Blume (2018) without argument, example or proof. Moreover, it is not clear how the premise to use two different p-values would not be doubly impacted in cases of many multiple comparisons.

8. Page 21: Discussion of asymmetric CI. It seems as if the paper is relying on a type of normal approximation to the sampling distribution of a Pearson correlation, which means the CIs ARE symmetric, but in a different space. Please clarify or correct.

Minor concerns

1. Page 20: the sentence "…exact p-values might provide information of interest to readers about how precisely how surprising the data is under the null model". Is technically incorrect. It is 'data or data more extreme'.

2. Page 22: "However, the SGPV has a lower error rate than a null-hypothesis test, not a more accurate error rate". And subsequent sentences. What does "more accurate" mean here? These are fixed parameters, not estimates. Any wouldn't a lower error rate imply more accuracy? This is just confusing.

3. 1-SGPV is not equal to the SGPV for the flipped hypotheses. However, 1-pvalue is at least interpretable. Consider correcting this on the graphs or make it clear to the reader what the issue is.

4. Page 21: "An important benefit of equivalence tests is that is [sic] does not need such a correction to prevent misleading results." This statement ignores the fact that large p-values are inconclusive and cannot indicate evidence for the null (unlike SGPVs). Remove or revise.

5. Page 19: SGPVs would appear to be "inferential" statistics in the same way the p-values are. Define these two statistics or remove this sentence.

6. The discussion about the problem with asymmetric CIs is fraught with difficulty. Asymmetric CIs would imply asymmetric p-value computations and this would maintain the key relationships. But also, CIs can be defined in any number of ways (one-sided, with holes in the middle, for example). This message was very hard for the reader to unpack and it was not clear that it was correct or useful.

7. Page 12: "None of the p-values based on the TOST procedure are below 0.05, and thus, in the long run we have 0% power." This statement can't be correct. In the long run, the test has zero power? Why use the TOST test then?

8. Page 10. Delete: "However, it is not a correction in the typical sense of the word, since the SGPV is not adjusted to any "correct" value." Correction factors in statistics don't replace an estimate with a more "correct" one. They shrink the estimate towards some value.