Equivalence Testing and the Second Generation P-Value

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

₁₀   Abstract

₁₁ To move beyond the limitations of null-hypothesis tests, statistical approaches have been

₁₂ developed where the observed data is compared against a range of values that are equivalent

₁₃ to the absence of a meaningful effect. Specifying a range of values around zero allows

₁₄ researchers to statistically reject the presence of effects large enough to matter, and prevents

₁₅ practically insignificant effects from being interepreted as a statistically significant difference.

₁₆ We compare the behavior of the recently proposed second generation p-value (Blume,

₁₇ McGowan, Dupont, & Greevy, 2018) with the more established Two One-Sided Tests

₁₈ (TOST) equivalence testing procedure (Schuirmann, 1987). We show that the two

₁₉ approaches yield almost identical results under optimal conditions. Under suboptimal

₂₀ conditions (e.g., when the confidence interval is wider than the equivalence range, or when

₂₁ confidence intervals are asymmetric) the second generation p-value becomes difficult to

₂₂ interpret as a descriptive statistic. The second generation p-value is interpretable in a

₂₃ dichotomous manner (i.e., when the SGPV equals 0 or 1 because the confidence intervals lies

₂₄ completely within or outside of the equivalence range), but this dichotomous interpretation

₂₅ does not require calculations. We conclude that equivalence tests yield more consistent

₂₆ p-values, distinguish between datsets that yield the same second generation p-value, and

₂₇ allow for easier control of Type I and Type II error rates.

₂₈   *Keywords:* equivalence testing, second generation p-values, hypothesis testing, TOST,

₂₉ statistical inference

³⁰ <center>Equivalence Testing and the Second Generation P-Value</center>

³¹   To test predictions researchers predominantly rely on null-hypothesis tests. This

³² statistical approach can be used to examine whether observed data is sufficiently surprising

³³ under the null hypothesis to reject an effect that equals exactly zero. Null-hypothesis tests

³⁴ have an important limitation, in that this procedure can only reject the hypothesis that

³⁵ there is no effect, while scientists should also be able to provide statistical support for

³⁶ *equivalence.* When testing for equivalence researchers aim to examine whether an observed

³⁷ effect is too small to be considered meaningful, and therefore is practically equivalent to zero.

³⁸ By specifying a range around the null hypothesis of values that are deemed practically

³⁹ equivalent to the absence of an effect (i.e., 0 ± 0.3) the observed data can be compared

⁴⁰ against an *equivalence range* and researchers can test if a meaningful effect is absent (Hauck

⁴¹ & Anderson, 1984; Kruschke, 2018; Rogers, Howard, & Vessey, 1993; Serlin & Lapsley, 1985;

⁴² Spiegelhalter, Freedman, & Parmar, 1994; Wellek, 2010; Westlake, 1972).

⁴³   Second generation *p*-values (SGPV) were recently proposed to as a descriptive statistic

⁴⁴ that represents "the proportion of data-supported hypotheses that are also null hypotheses"

⁴⁵ (Blume et al., 2018). The researcher specifies an equivalence range around a null hypothesis

⁴⁶ of values that are considered practically equivalent to the null hypothesis. The SGPV

⁴⁷ measures the degree to which a set of data-supported parameter values falls within the

⁴⁸ interval null hypothesis. If the estimation interval falls completely within the equivalence

⁴⁹ range, the SGPV is 1. If the confidence interval falls completely outside of the equivalence

⁵⁰ range, the SGPV is 0. Otherwise the SGPV is a value between 0 and 1 that expresses the

⁵¹ overlap of data-supported hypotheses and the equivalence range. When calculating the

⁵² SGPV the set of data-supported parameter values can be represented by a confidence

⁵³ interval (CI), although other one could also choose to use credible intervals or Likelihood

⁵⁴ support intervals (SI). When a confidence interval is used, the SGPV and equivalence tests

⁵⁵ such as the Two One-Sided Tests (TOST) procedure (Lakens, 2017; Meyners, 2012;

⁵⁶ Schuirmann, 1987) appear to have close ties, because both tests conpare a confidence interval

against an equivalence range. Here, we aim to examine the similarities and differences

between the TOST procedure and the SGPV.

The TOST procedure also relies on the confidence interval around the effect. In the

TOST procedure the data is tested against the lower equivalence bound in the first one-sided

test, and against the upper equivalence bound in the second one-sided test (Lakens, Scheel,

& Isager, 2018). If both tests statistically reject an effect as extreme or more extreme than

the equivalence bound, you can conclude the observed effect is practically equivalent to zero

from a Neyman-Pearson approach to statistical inferences. Because one-sided tests are

performed, one can also conclude equivalence by checking whether the 1-2*alpha confidence

interval (e.g., when the alpha level is 0.05, a 90% CI) falls completely within the equivalence

bounds. Because both equivalence tests as the SGPV are based on whether and how much a

confidence interval overlaps with equivalence bounds, it seems worthwhile to compare the

behavior of the newly proposed SGPV to equivalence tests to examine the unique

contribution of the SGPV to the statistical toolbox.

The relationship between p-values from TOST and SGPV when confidence intervals are symmetrical

The second generation p-value (SGPV) is calculated as:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2\,|H_0|}, 1\right\}$$

where I is the interval based on the data (e.g., a 95% confidence interval) and H_0 is the

equivalence range. The first term of this formula implies that the second generation p-value

is the width of the confidence interval that overlaps with the equivalence range, divided by

the total width of the confidence interval. The second term is a "small sample correction"

(which will be discussed later) that comes into play whenever the confidence interval is more

than twice as wide as the equivalence range.

To examine the relation between the TOST p-value and the SGPV we can calculate

both statistics across a range of observed effect sizes. In Figure 1 p-values are plotted for the

TOST procedure and the SGPV. The statistics are calculated for hypothetical one-sample

t-tests for observed means ranging from 140 to 150 (on the x-axis). The equivalence range is

set to 145 ± 2 (i.e., an equivalence range from 143 to 147), the observed standard deviation

is assumed to be 2, and the sample size is 30. For example, for the left-most point in Figure

1 the SGPV and the TOST p-value is calculated for a hypothetical study with a sample size

of 30, an observed standard deviation of 2, and an observed mean of 140, where the p-value

for the equivalence test is 1, and the SGPV is 0. Our conclusions about the relationship

between TOST p-values and SGPV in this article are not dependent upon any specific

example, as readers can explore for themselves in an online Shiny app:

http://shiny.ieis.tue.nl/TOST_vs_SGPV/.

The SGPV treats the equivalence range as the null-hypothesis, while the TOST

procedure treats the values outside of the equivalence range as the null-hypothesis. For ease

of comparison we can reverse the SGPV (by calculating 1-SGPV in Figure 2) to make the

values more easily comparable. We see that the p-value from the TOST procedure and the

SGPV follow each other closely.

When the observed sample mean is 145, the sample size is 30, and the standard

deviation is 2, and we are testing against equivalence bounds of 143 and 147 using the TOST

procedure for a one-sample t-test, the equivalence test is significant, $t(29) = 5.48$, $p < .001$.

Because the 95% CI falls completely within the equivalence bounds, the SGPV is 1 (see

Figure 1).

On the other hand, if the observed mean is 140, the equivalence test is not significant

(the observed mean is far outside the equivalence range of 143 to 147), $t(29) = -8.22$, $p = 1$

(or more accurately, $p > .999$ as p-values are bounded between 0 and 1). Because the 95% CI

falls completely outside the equivalence bounds, the SGPV is 0 (see Figure 1).
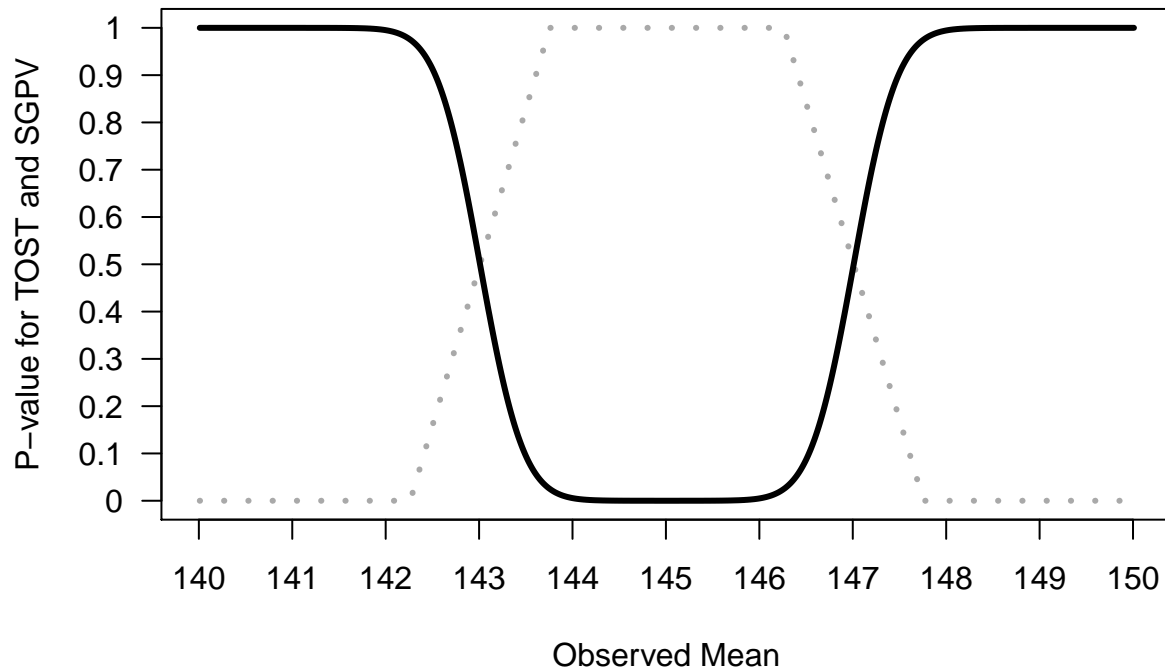
Figure 1. Comparison of *p*-values from TOST (black line) and SGPV (dotted grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2.

SGPV as a uniform measure of overlap

It is clear the SGPV and the *p*-value from TOST are closely related. 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. When the TOST *p*-value is 0.5, the SGPV is also 0.5 (note that the reverse is not true). 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,
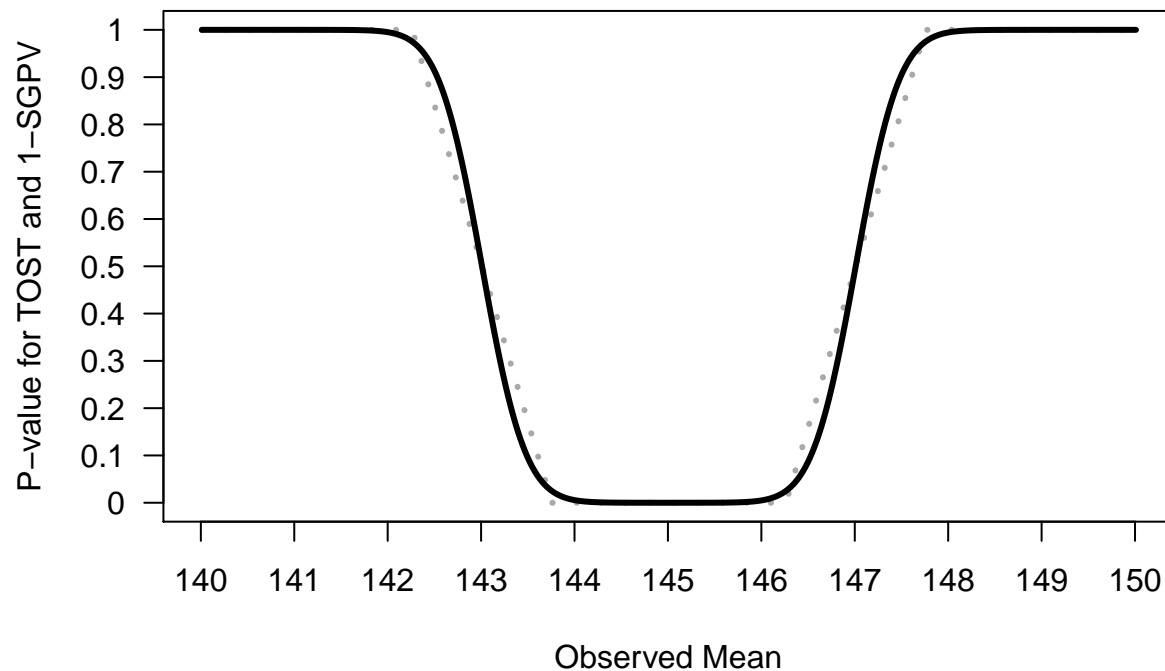
Figure 2. Comparison of *p*-values from TOST (black line) and 1-SGPV (dotted grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2.

115 Figure 3).

116   Two other points always have to overlap. When the 95% CI falls completely inside the

117 equivalence region, and one endpoint of the confidence interval is exactly equal to one of the

118 equivalence bounds (see situation B in Figure 3) the TOST *p*-value (which relies on a

119 one-sided test) is always 0.025. The third point where the SGPV and the *p*-value from the

120 TOST procedure should overlap is where the 95% CI falls completely outside of the

121 equivalence range, but one endpoint of the confidence interval is equal to the equivalence

122 bound (see situation C in Figure 3), when the *p*-value will always be 0.975.

123   As the observed mean in a one-sample *t*-test lies closer to the test value (in the plot
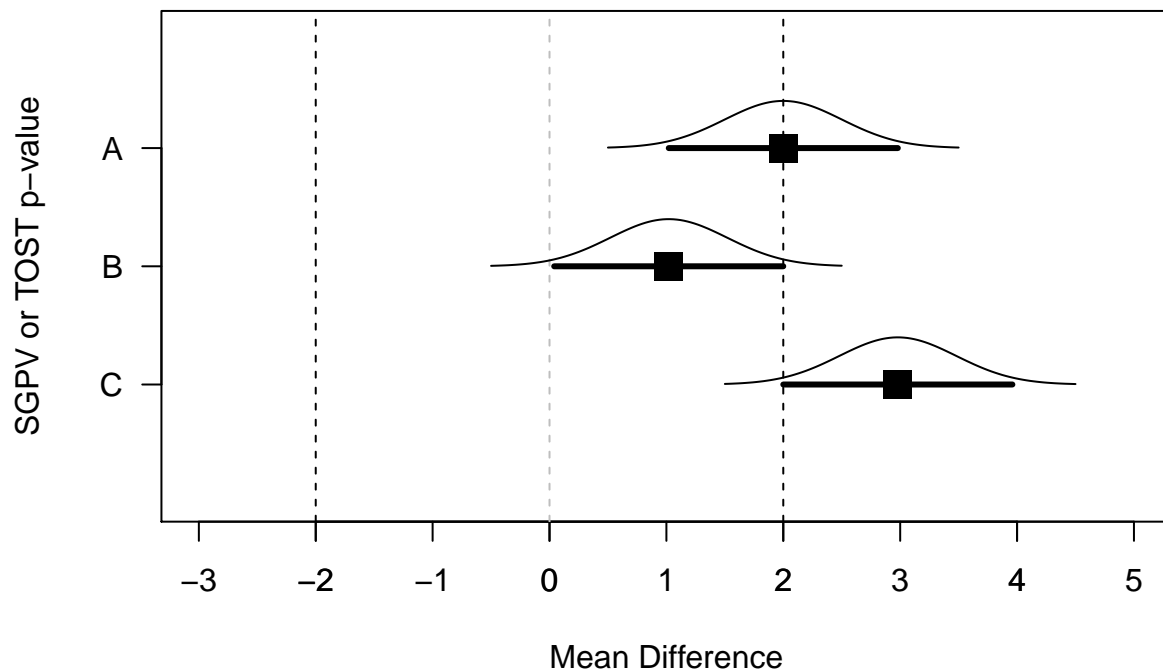
Figure 3. Means, normal distribution, and 95% CI for three example datasets that illustrate the relationship between *p*-values from TOST and SGPV.

below, from situation A to D, the mean gets closer to the test value by 0.1) the difference in the overlap changes uniformly.

For example, the SGPV from A to D is 0.76, 0.81, 0.86, and 0.91. The difference in the percentage of overlap between A and B (-0.05) is identical to the difference in the percentage of overlap between C and D as the mean gets 0.1 closer to the test value (-0.05).

As we move the observed mean closer to the test value in steps of 0.1 across A to D the *p*-value calculated for normally distributed data is not uniformly distributed. The probability of observing data more extreme than the upper bound of 2 is (from A to D) 0.16, 0.12, 0.08, and 0.06. As we can see, the difference between A and B (0.04) is not the same as the difference between C And D (0.03). Indeed, the difference in *p*-values is the largest as
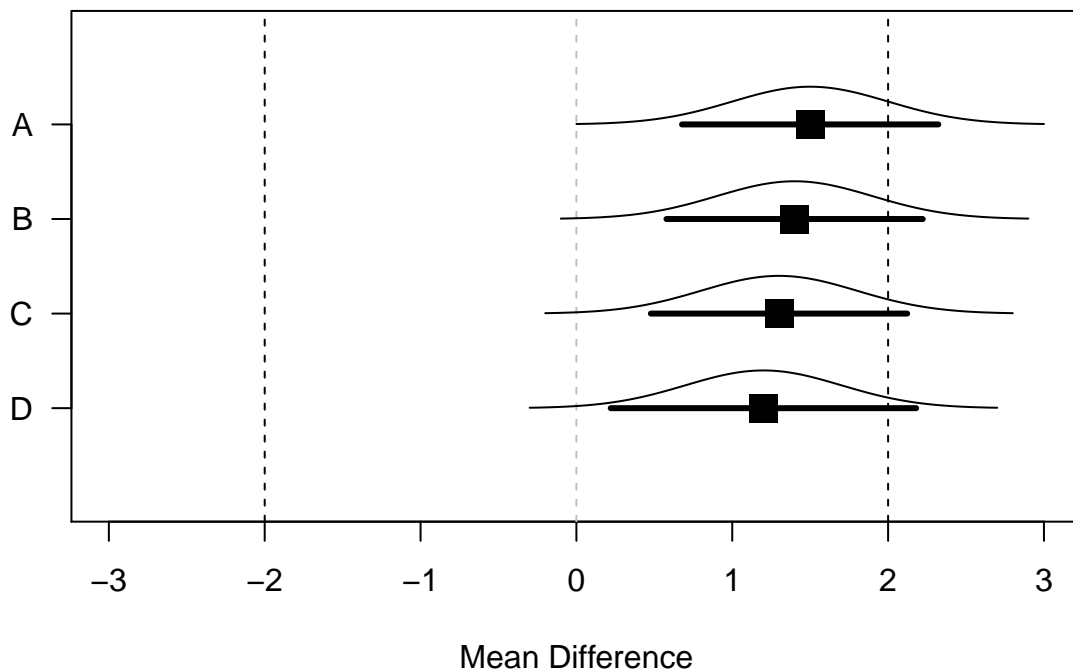
Figure 4. Means, normal distribution, and 95% CI for samples where the observed population mean is 1.5, 1.4, 1.3, and 1.2.

134 you start at $p = 0.5$ (when the observed mean falls on the test value), which is why the line

135 in Figure 1 is the steepest at $p = 0.5$. Note that where the SGPV reaches 1 or 0, p-values

136 closely approximate 0 and 1, but never reach these values. ##When are the SGPV and

137 Equivalence Test Unrelated? There are 4 situations where p-values from TOST and SGPV

138 are unrelated. The first two situations were discussed before and can be seen in Figure 1.

139 When the SGPV is either 0 or 1 p-values from the equivalence test fall between 0.975 and 1

140 or between 0 and 0.025. Because p-values approach 0 or 1, but are never exactly 0 or 1, and

141 the SGPV is exactly 0 or 1, the two statistics are completely unrelated within these ranges.

142 The easiest way to see this is by plotting the SGPV against the p-value from the TOST

143 procedure. The situations where the SPGV and p-values from the TOST procedure are

144 unrelated are indicated by the parts of the curve where there are vertical straight lines at

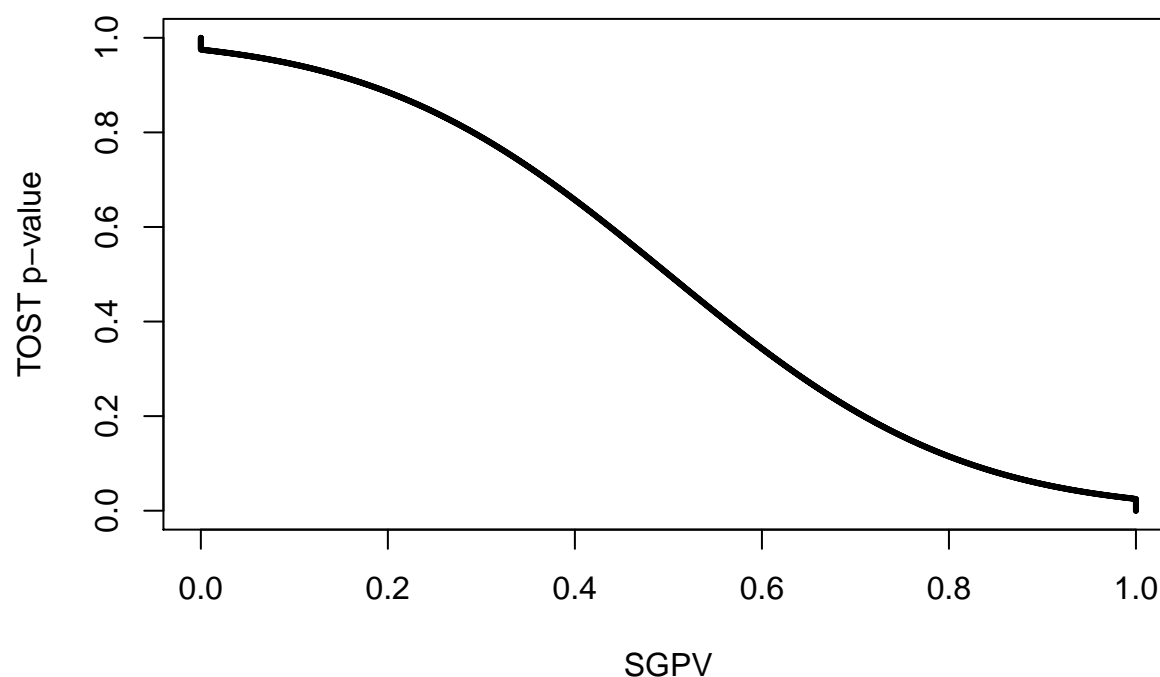145 second generation p-values of 0 and 1.

Figure 5. The relationship between p-values from the TOST procedure and the SGPV for the same scenario as in Figure 1.

146   A third situation in which the SGPV deviates from the TOST p-value is whenever the

147 CI is wider than the equivalence range, and the CI overlaps with the upper *and* lower

148 equivalence bound. When the confidence interval is more than twice as wide as the

149 equivalence range the SGPV is set to 0.5. Blume et al. (2018) call this the "small sample

150 correction factor". However, it is not a correction in the typical sense of the word, since the

151 SGPV is not adjusted to any "correct" value. When the normal calculation would be

152 "misleading" (i.e., the SGPV would be small, which normally would suggest support for the

153 alternative hypothesis, but at the same time all values in the equivalence range are

154 supported), the SGPV is set to 0.5 which according to Blume and colleagues signal the SGPV

155 is "uninformative". Note that the CI can be twice as wide as the equivalence range whenever

156 the sample size is small (and the confidence interval width is large) *or* when then equivalence

157 range is narrow. It is therefore not so much a "small sample correction" as it is an exception

158 to the typical calculation of the SGPV whenever the ratio of the confidence interval width to

159 the equivalence range exceeds 2:1 and the CI overlaps with the upper and lower bounds.

160   We can examine this situation by calculating the SGPV and performing the TOST for

161 a situation where sample sizes are small and the equivalence range is narrow, such that the

162 CI is more than twice as large as the equivalence range.

163   We can again plot the two statistics against each other to see where they are unrelated

164 (indicated by straight lines in the curve, see Figure 7). We see the SGPV is 0.5 for a range of

165 observed means where the *p*-value from the equivalence test still varies. It should be noted

166 that in these calculations the *p*-values for the TOST procedure are *never* smaller than 0.05

167 (i.e., they do not get below 0.05 on the y-axis). In other words, we cannot conclude

168 equivalence based on any of the observed means. This happens because the TOST *p*-value is

169 smaller than 0.05 only when the 90% CI falls completely between the upper and lower

170 equivalence bounds. However, we are examining a scenario where the 90% CI is so wide that

171 it never falls completely within the two equivalence bounds, and thus the equivalence test is

172 never significant. As Lakens (2017) notes: "in small samples (where CIs are wide), a study

173 might have no statistical power (i.e., the CI will always be so wide that it is necessarily wider

174 than the equivalence bounds)." None of the *p*-values based on the TOST procedure are

175 below 0.05, and thus, in the long run we have 0% power.

176   The *p*-value from the TOST procedure and the SGPV are also unrelated when the CI

177 is wider than the equivalence range (so the precision is low) and overlaps with the upper and

178 lower equivalence bound, but the CI is *not* twice as wide as the equivalence range. For

179 example, in the example below, we see that the CI is only 1.79 times as wide as the

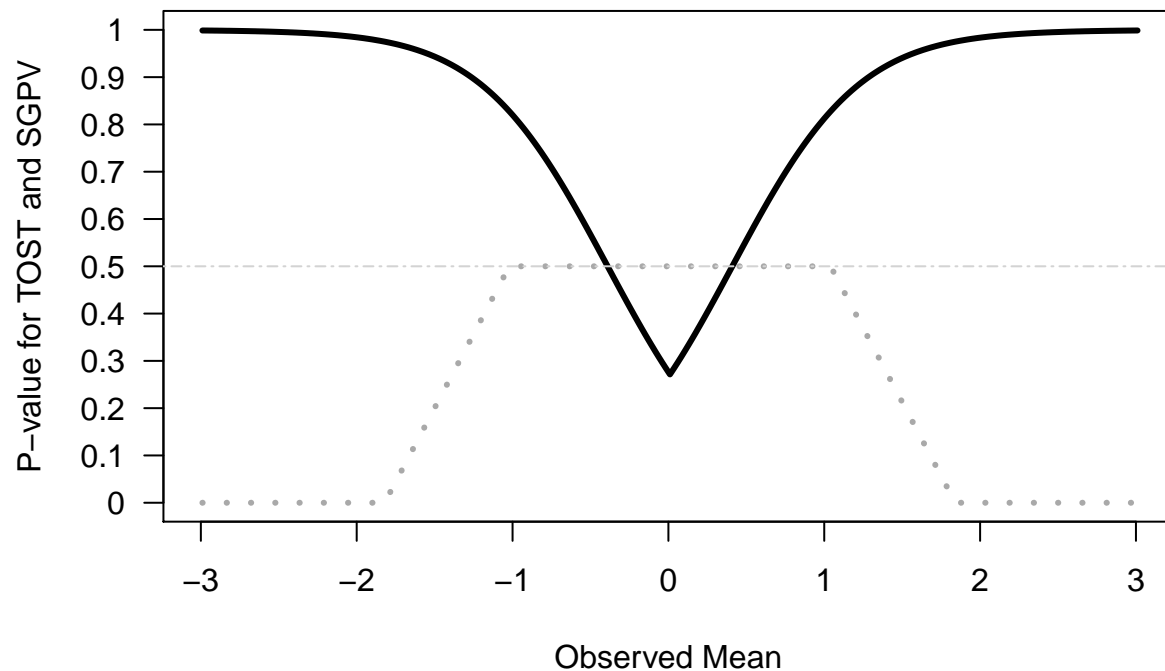180 equivalence bounds, but the CI overlaps with the lower and upper equivalence bounds. This

Figure 6. Comparison of *p*-values from TOST (black line) and SGPV (dotted grey line) across a range of observed sample means (x-axis). Because the sample size is small (n = 10) and the CI is more than twice as wide as the equivalence range (set to -0.4 to 0.4), the SGPV is set to 0.5 (horizontal lightgrey line) across a range of observed means.

181 means the SGPV is not set to 0.5, but it is constant across a range of observed means.

182   If the observed mean would be somewhat closer to 0, or further away from 0, the

183 SGPV remains constant (the CI width does not change, and it completely overlaps with the

184 equivalence range) while the *p*-value for the TOST procedure does vary. We can see this in

185 Figure 9 below. The SGPV is not set to 0.5, but is slightly higher than 0.5 across a range of

186 means. How high the SGPV will be for a CI that is not twice as wide as the equivalence

187 range, but overlaps with the lower and upper equivalence bounds, depends on the width of
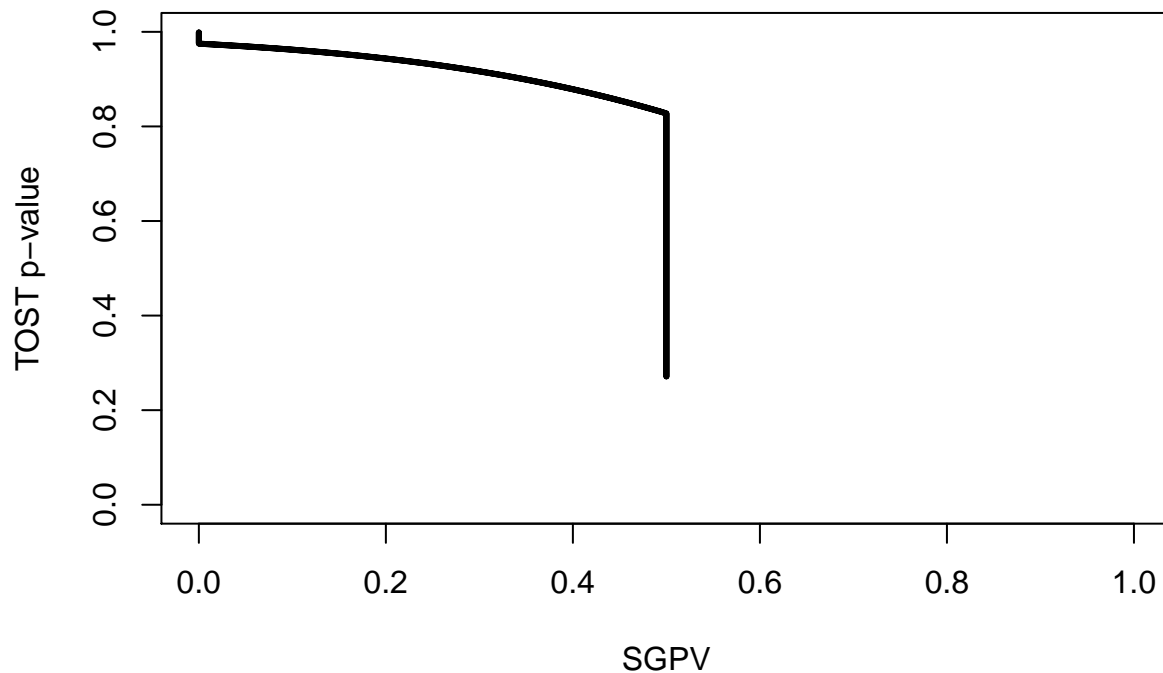
188 the CI and the equivalence range.

Figure 7. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 6.

If we once more plot the two statistics against each other to see where they are unrelated (indicated by straight lines in the curve), we see the SGPV is 0.56 for a range of observed means where the *p*-value from the equivalence test still varies.

To conclude this section, there are four situations where the *p*-value from the TOST procedure is unrelated to the SGPV. In all these situations the *p*-value for the equivalence test differentiates tests with different means, but the SGPV does not. Therefore, as a purely continuous descriptive statistic, the SGPV is more limited than the *p*-value from the TOST procedure.
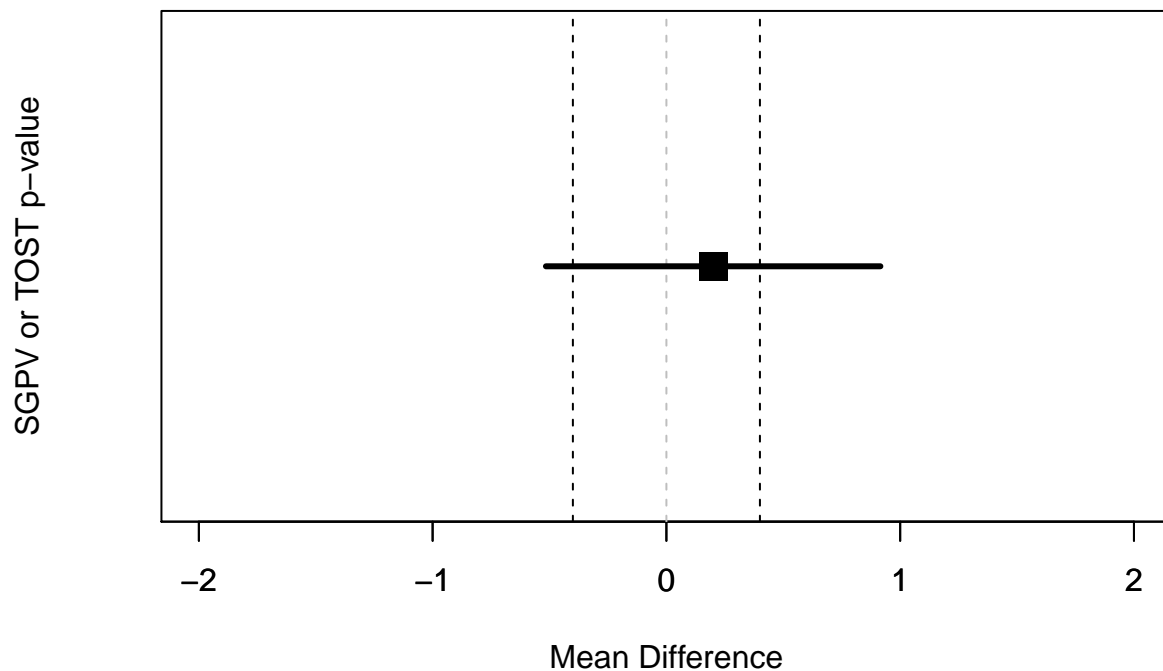
Figure 8. Example of a 95% CI that overlaps with the lower and upper equivalence bound (indicated by the vertical dotted lines).

The relation between equivalence tests and SGPV when confidence intervals are not symmetrical

So far we have only looked at the relation between equivalence tests and the SGPV when confidence intervals are symmetric (e.g., for confidence intervals around mean differences). For correlations, which are bound between -1 and 1, confidence intervals are only symmetric for a correlation of exactly 0. The confidence interval becomes increasingly asymmetric as the observed correlation nears -1 or 1. For example, with ten observations, an observed correlation of 0 has a symmetric 95% confidence interval ranging from -0.629 to 0.629, while and observed correlation of 0.7 has an asymmetric 95% confidence interval ranging from 0.126 to 0.992.
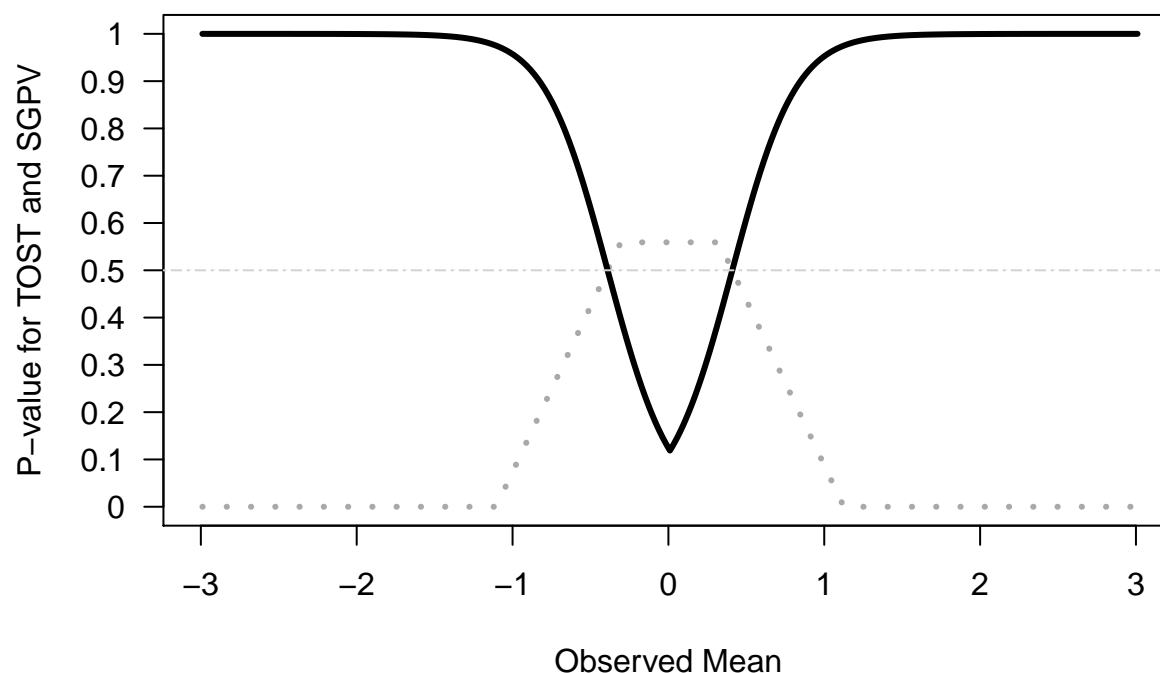
Figure 9. Comparison of *p*-values from TOST (black line) and SGPV (dotted grey line) across a range of observed sample means (x-axis). The sample size is small (n = 10), but because the sd is half as big as in Figure 7 (1 instead of 2) the CI is less than twice as wide as the equivalence range (set to -0.4 to 0.4). The SGPV is not set to 0.5 (horizontal lightgrey line) but reaches a maximum slightly above 0.5 across a range of observed means.

207   The effect of assymetric confidence intervals is most easily noticable at smaller sample
208 sizes, therefore in Figure 11 below we plot the *p*-values from equivalence tests and the SGPV
209 (again plotted as 1-SGPV for ease of comparison) for correlations. The sample size is 30 pairs
210 of observations, and the lower and upper equivalence bounds are set to -0.45 and 0.45, with
211 an alpha of 0.05. As the observed correlation in the sample moves from -1 to 0 the *p*-value
212 from the equivalence test becomes smaller, as does 1-SGPV. The pattern is quite similar to
213 that in Figure 2. The *p*-value for the TOST procedure and 1-SGPV are still identical when
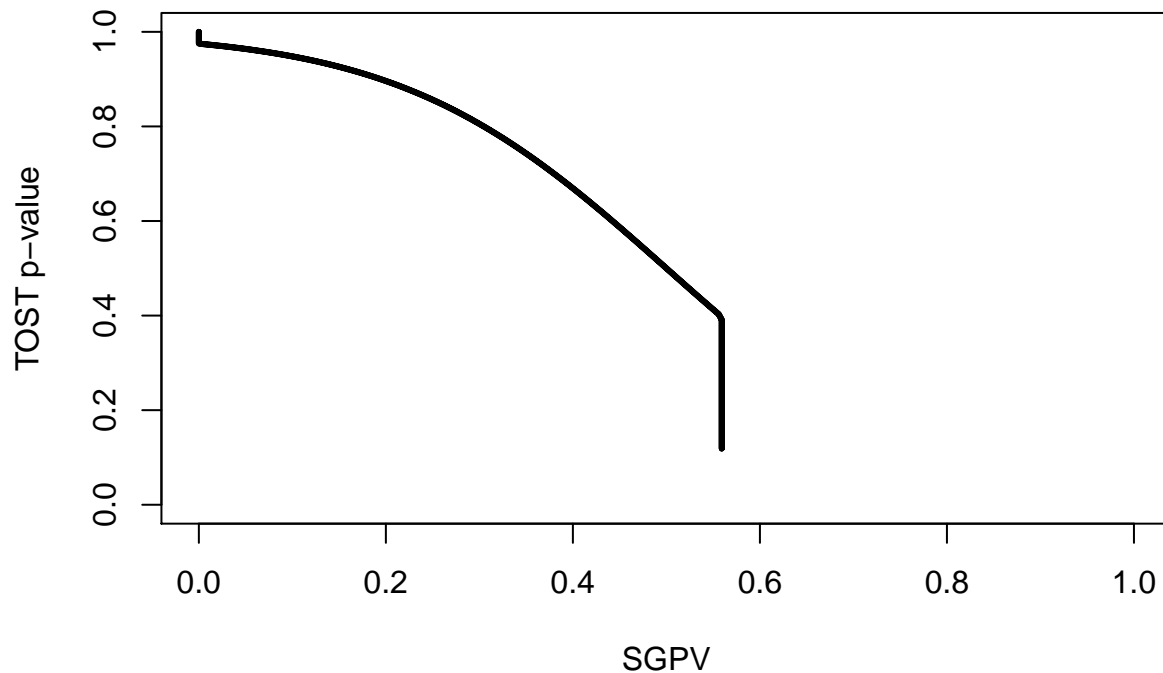
Figure 10. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 9.

214 *p*-values are 0.975 and 0.025 (indicated by the upper and lower horizontal dotted lines).

215 There are two important differences, however. First of all, the SGPV is no longer a straight

216 line, but a curve, due to the asymmetry in the 95% CI. Second, and most importantly, the

217 *p*-value for the equivalence test and the SGPV do no longer overlap at $p = 0.5$.

218   The reason that the equivalence test and SGPV no longer overlap is also because of

219 asymmetric confidence intervals. If the observed correlation falls exactly on the equivalence

220 bound the *p*-value for the equivalence test indicates that the probability of observing the

221 observed or more extreme data, assuming the equivalence bound is the true effect size, is

222 50%. In other words, if the true effect size is the same as the equivalence bound, it is equally

223 likely to find an effect more extreme than the equivalence bound, as it is to observe an effect
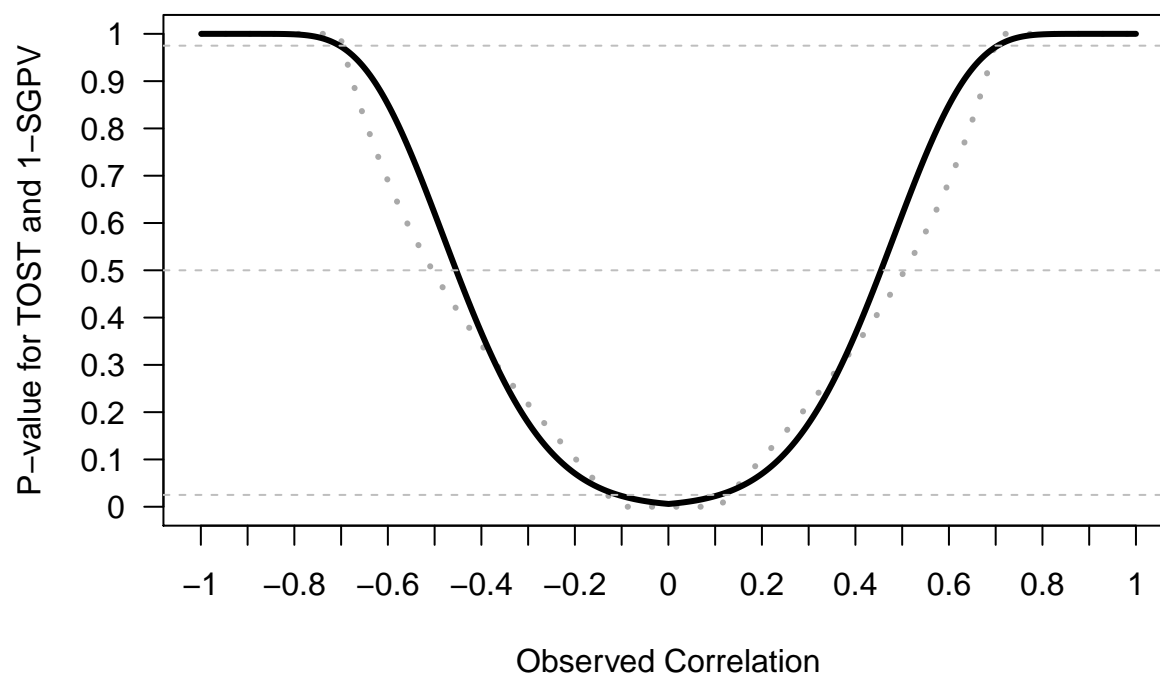
Figure 11. Comparison of *p*-values from TOST (black line) and 1-SGPV (dotted grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = -0.45 and r = 0.45 with n = 30 and an alpha of 0.05.

224 that is less extreme than the equivalence bound. However, as can be seen in Figure 12, the

225 two second generation *p*-values associated with the observed correlations at r = -0.45 and r

226 = 0.45 are 0.58. Because the confidence intervals are asymmetric around the observed effect

227 size of 0.45 (ranging from 0.11 to 0.70) according to Blume et al. (2018) 58.10% of the

228 data-supported hypotheses are null hypotheses, and therefore 58.10% of the data-supported

229 hypotheses are compatible with the null premise.

230   This example illustrates the difference between a proportion and a probability. There

231 is always a 50% probability of observing a correlation smaller or larger than the true

232 correlation, but the SGPV for this situation depends on how far away the observed
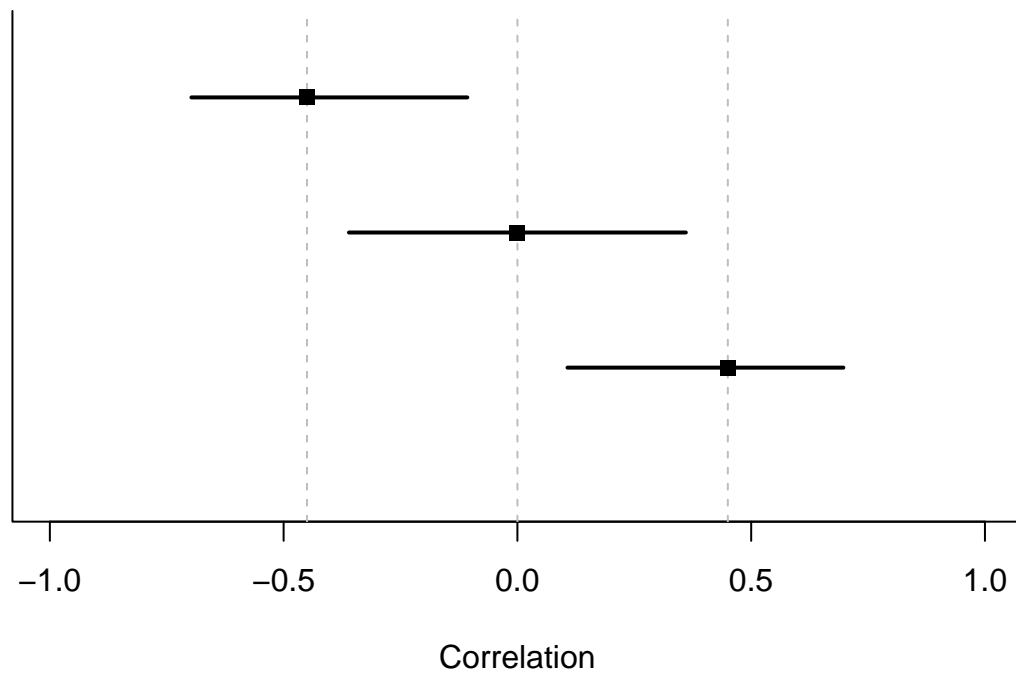
Figure 12. Three 95% confidence intervals for observed effect sizes of r = -0.45, r = 0, and r = 0.45 for n = 30. Only the confidence interval for r = 0 is symmetric.

233 correlation is from 0. The further away from 0, the larger the SGPV when the observed mean

234 falls on the equivalence bound. The SGPV is the proportion of values in a 95% confidence

235 interval that overlap with the equivalence range, but not the probability that these values

236 will be observed. In the most extreme case (i.e., a sample size of 4, and equivalence bounds

237 set to r = -0.99 and 0.99, with an observed correlation of 0.99) 97.60% of the confidence

238 interval overlaps with the equivalence range, even though in the long run only 50% of the

239 correlations observed in the future will fall in this range. It should be noted that in larger

240 sample sizes the SGPV is closer to 0.5 whenever the observed correlation falls on the

241 equivalence bound, but this extreme example nevertheless clearly illustrates the difference

242 between question the SGPV answer, and the question a *p*-value answers. The conclusion of

243 this section on asymmetric confidence intervals is that a SGPV of 1 or 0 can still be

244 interpreted as a $p < 0.025$ or $p > 0.975$ in an equivalence test, since the SGPV and p-value

245 for the TOST procedure are always directly related at these values. Although Blume et al.

246 (2018) state that "the degree of overlap conveys how compatible the data are with the null

247 premise" this definition of what the SGPV provides does not hold for asymmetric confidence

248 intervals. Although a SGPV of 1 or 0 can be directly interpreted, a SGPV between 0 and 1

249 is not interpretable as "compatibility with the null hypothesis". Indeed, Blume and

250 colleagues write in the supplemental material that "The magnitude of an inconclusive

251 second-generation p-value can vary slightly when the effect size scale is transformed.

252 However definitive findings, i.e. a p-value of 0 or 1 are *not* affected by the scale changes."

What are the Relative Strengths and Weaknesses of Equivalence Testing and the SGPV?

255 When introducing a new statistical method, it is important to compare it to existing

256 approaches and specify it's relative strengths and weaknesses. First of all, the SGPV is a

257 descriptive statistic (unlike the p-value that is calculated for an equivalence test, which is an

258 inferential statistic) that provides the proportion of overlap of the confidence interval and

259 the equivalence range.

260 Even though a SGPV of 1 or 0 has a clear interpretation (we can reject effects outside

261 or inside the equivalence range), intermediate values are not as easy to interpret (especially

262 for effects that have asymmetric confidence intervals). In one sense, they are what they are

263 (the proportion of overlap), but in any specific case it can be unclear what this number tells

264 us about the data we have collected. This is not too problematic, since the main use of the

265 SGPV (e.g., in all examples provided by Blume and colleagues) is to examine whether the

266 SGPV is 0, 1, or inconclusive. This interpretation of a SGPV as allowing researchers to

267 reject the null, reject the presence of a meaningful effect, or remaining inconclusive is very

268 similar to the Neyman-Pearson interpretation of combining a null-hypothesis test and an

equivalence test (Lakens et al. (2018)). The difference is that where a SGPV of 1 can be

interpreted as $p < .025$, equivalence tests provide exact p-values, and they continue to

differentiate between for example p = 0.048 and p = 0.002. Whether this is desireable

depends on the perspective that is used. 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. At the same time, it is recommended to

report exact p-values (American Psychological Association, 2010), and exact p-values might

provide information of interest to readers about how precisely how surprising the data is

under the null model. 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 cateogies (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

An important issue when calculating the SGPV is its reliance on the "small sample

correction", where the SGPV is set to 0.5 whenever the ratio of the confidence interval width

to the equivalence range exceeds 2:1 and the CI overlaps with the upper and lower bounds.

This exception to the normal calculation of the SGPV is introduced to prevent misleading

values. Without this correction it is possible that a confidence interval is extremely wide,

and an equivalence range is extremely narrow, which without the correction would lead to a

very low value for the SGPV. Blume et al. (2018) suggest that under such a scenario "the

data favor alternative hypotheses", even when a better interpretation would be that there is

not enough data to accurately estimate the true effect compared to the width of the

equivalence range. Although it is necessary to set the SGPV to 0.5 whenever the ratio of the

confidence interval width to the equivalence range exceeds 2:1, it leads to a range of

²⁹⁶ situations where the SGPV is set to 0.5, while the *p*-value from the TOST procedure

²⁹⁷ continues to differentiate (see for example Figure 6). An important benefit of equivalence

²⁹⁸ tests is that is does not need such a correction to prevent misleading results.

²⁹⁹   As a more extreme example of the peculiar behavior of the "small sample correction"

³⁰⁰ as currently implemented in the calculation of the SGPV see Figure 13 below. In this figure

³⁰¹ observed correlations (from a sample size of 10) from -1 to 1 are tested against an

³⁰² equivalence range from r = 0.4 to r = 0.8. We can see the SGPV has a peculiar shape

³⁰³ because it is set to 0.5 for certain observed correlations, even though there is no risk of a

³⁰⁴ meaningless SGPV in this range. This example suggests that the current implementation of

³⁰⁵ the "small sample correction" could be improved. If, on the other hand, the SGPV is mainly

³⁰⁶ meant to be interpreted when it is 0 or 1, it might be preferable to simply never apply the

³⁰⁷ "small sample correction".

³⁰⁸   Blume et al. (2018) argue that the SGPV has improved error control, in that the

³⁰⁹ conclusion that the 95% confidence interval lies completely outside the equivalence range

³¹⁰ around zero will necessarily occur less frequently than a Type I error in a null-hypothesis

³¹¹ test (where the 95% confidence interval only needs to not overlap with zero). However, the

³¹² SGPV has a *lower* error rate than a null-hypothesis test, not a *more accurate* error rate. In a

³¹³ Neyman-Pearson perspective (which forms the basis of equivalence tests) the goal is not to

³¹⁴ end up with an error rate that is as low as possible (to achieve that, one simply adjusts the

³¹⁵ alpha level), but with a decision procedure that, when applied, yields a desired error rate

³¹⁶ with high accuracy. We show here that the SGPV and equivalence tests are directly related,

³¹⁷ and so are their error rates. The TOST procedure uses a 90% confidence interval because it

³¹⁸ is based on two *one-sided* tests, and thus the confidence interval is 1-2*α, and because both

³¹⁹ tests need to be significant to conclude equivalence, no adjustment for multiple comparisons

³²⁰ is needed. The TOST procedure always has higher power to declare equivalence than the

³²¹ SGPV, which relies on a 95% confidence interval (but note that there is no reason to always

³²² set the alpha level to 0.05).

³²³ Blume et al. (2018) claim that "Adjustments for multiple comparisons are obviated"

³²⁴ (p. 15) and that "second-generation *p*-values provide a proper scientific adjustment for

³²⁵ multiple comparisons". However, this is not correct. With a sufficient number of looks error

³²⁶ rates can rise above the nominal alpha level (both when compared to equivalence and

³²⁷ minimal effect tests, but even when compared to a null-hypothesis test). When directly

³²⁸ comparing equivalence tests and the SGPV, multiple comparisons inflate the probability that

³²⁹ one erroneously concludes there is *no* effect, where there *is* a true effect size that equals the

³³⁰ equivalence bound, just as quickly for the SGPV as for equivalence tests, because they are

³³¹ directly related. To conclude, the idea that the SGPV improves error rates does not hold up

³³² under closer scrutiny, and the recommendation to ignore adjustments for multiple

³³³ comparisons has the potential to increase false positives in the literature. Equivalence tests

³³⁴ provide and easier and more formal way to control both Type I error rates (by setting the

³³⁵ alpha level) and the Type II error rate (by performing an a-priori power analysis).

Conclusion

³³⁷ We believe that our explanation of the similarities between the TOST procedure and

³³⁸ the SGPV provides context to interpret the contribution of second generation *p*-values to the

³³⁹ statistical toolbox. The novelty of te SGPV lies in its use as a descriptive statistic, but this

³⁴⁰ use can be limited when confidence intervals are asymmetrical or wider than the equivalence

³⁴¹ range. There are strong similarities with *p*-values from the TOST procedure, and in all

³⁴² situations where the statistics yield different results, the behavior of the *p*-value from the

³⁴³ TOST procedure is more consistent and easier to interpret. We hope this overview of the

³⁴⁴ relationship between the SGPV and equivalence tests will help researchers to make an

³⁴⁵ informed decision about which statistical approach provides the best answer to their

³⁴⁶ question. Our comparisons shows that when proposing alternatives to null-hypothesis tests,

³⁴⁷ it is important to compare new proposals to already existing procedures. We believe

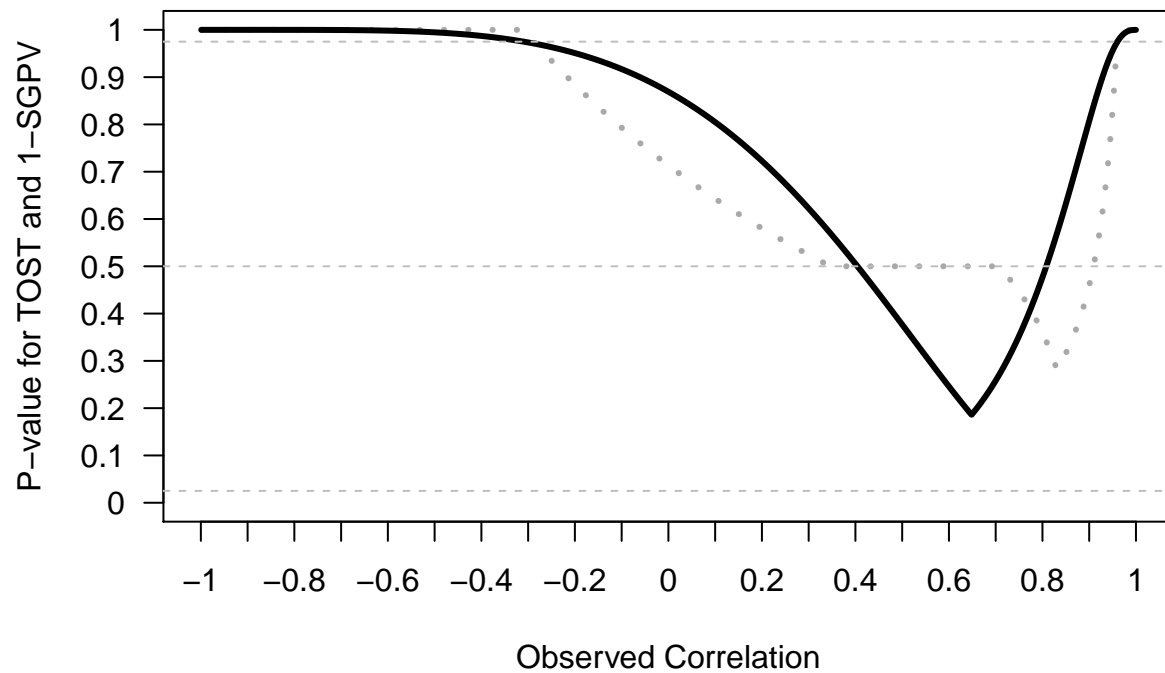³⁴⁸ equivalence tests achieve the goals of the second generation *p*-value while allowing users to

Figure 13. Comparison of *p*-values from TOST (black line) and 1-SGPV (dotted grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = 0.4 and r = 0.8 with n = 10 and an alpha of 0.05.

349 more easily control error rates, and while yielding more consistent statistical outcomes.

References

American Psychological Association (Ed.). (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Blume, J. D., McGowan, L. D., 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

Hauck, D. W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, *12*(1), 83–91. doi:10.1007/BF01063612

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 2515245918771304. doi:10.1177/2515245918771304

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918770963. doi:10.1177/2515245918770963

Meyners, M. (2012). Equivalence tests review. *Food Quality and Preference*, *26*(2), 231–245. doi:10.1016/j.foodqual.2012.05.003

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. doi:http://dx.doi.org/10.1037/0033-2909.113.3.553

Schuirmann, 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*

376   *Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680.

377   Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The

378   good-enough principle.

379   Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to

380   randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in*

381   *Society)*, 357–416. doi:10.2307/2983527

382   Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.).

383   Boca Raton: CRC Press.

384   Westlake, W. J. (1972). Use of Confidence Intervals in Analysis of Comparative

385   Bioavailability Trials. *Journal of Pharmaceutical Sciences*, *61*(8), 1340–1341.

386   doi:10.1002/JPS.2600610845