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

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

11   Abstract

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

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

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

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

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

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

18   D'Agostino McGowan, Dupont, & Greevy, 2018) with the more established Two One-Sided

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

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

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

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

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

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

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

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

27   p-values, distinguish between datasets that yield the same second generation p-value, and

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

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

30   statistical inference

₃₁   Equivalence Testing and the Second Generation *P*-Value

₃₂   To test predictions researchers predominantly rely on null-hypothesis tests.

₃₃   This statistical approach can be used to examine whether observed data is sufficiently

₃₄ surprising under the null hypothesis to reject an effect that equals exactly zero.

₃₅ Null-hypothesis tests have an important limitation, in that this procedure can only reject the

₃₆ hypothesis that there is no effect, while scientists should also be able to provide statistical

₃₇ support for *equivalence.* When testing for equivalence researchers aim to examine whether an

₃₈ observed effect is too small to be considered meaningful, and therefore is practically

₃₉ equivalent to zero. By specifying a range around the null hypothesis of values that are

₄₀ deemed practically equivalent to the absence of an effect (i.e., 0 ± 0.3) the observed data can

₄₁ be compared against an *equivalence range* and researchers can test if a meaningful effect is

₄₂ absent (Hauck & Anderson, 1984; Kruschke, 2018; Rogers, Howard, & Vessey, 1993; Serlin &

₄₃ Lapsley, 1985; Spiegelhalter, Freedman, & Parmar, 1994; Wellek, 2010; Westlake, 1972).

₄₄   Second generation *p*-values (SGPV) were recently proposed to as a descriptive statistic

₄₅ that represents "the proportion of data-supported hypotheses that are also null hypotheses"

₄₆ (Blume et al., 2018). The researcher specifies an equivalence range around a null hypothesis

₄₇ of values that are considered practically equivalent to the null hypothesis. The SGPV

₄₈ measures the degree to which a set of data-supported parameter values falls within the

₄₉ interval null hypothesis. If the estimation interval falls completely within the equivalence

₅₀ range, the SGPV is 1. If the confidence interval falls completely outside of the equivalence

₅₁ range, the SGPV is 0. Otherwise the SGPV is a value between 0 and 1 that expresses the

₅₂ overlap of data-supported hypotheses and the equivalence range. When calculating the

₅₃ SGPV the set of data-supported parameter values can be represented by a confidence

₅₄ interval (CI), although one could also choose to use credible intervals or Likelihood support

₅₅ intervals (SI). When a confidence interval is used, the SGPV and equivalence tests such as

₅₆ the Two One-Sided Tests (TOST) procedure (Lakens, 2017; Meyners, 2012; Quertemont,

₅₇ 2011; Schuirmann, 1987) appear to have close ties, because both tests compare a confidence

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

59 differences between the TOST procedure and the SGPV.

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

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

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

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

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

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

66 performed, one can also conclude equivalence by checking whether the 1-2×α confidence

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

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

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

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

71 contribution of the SGPV to the statistical toolbox.

72 **The relationship between p-values from TOST and SGPV when confidence**

73 **intervals are symmetrical**

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

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

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

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

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

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

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

80 than twice as wide as the equivalence range. To examine the relation between the TOST

81 p-value and the SGPV we can calculate both statistics across a range of observed effect sizes.

82 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

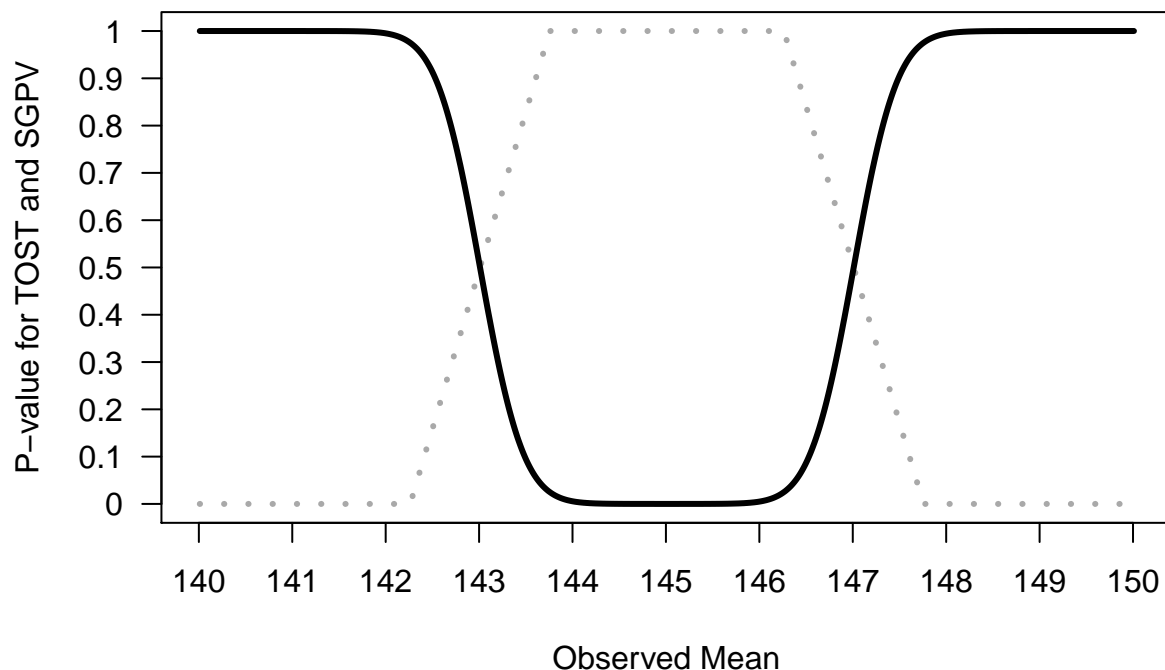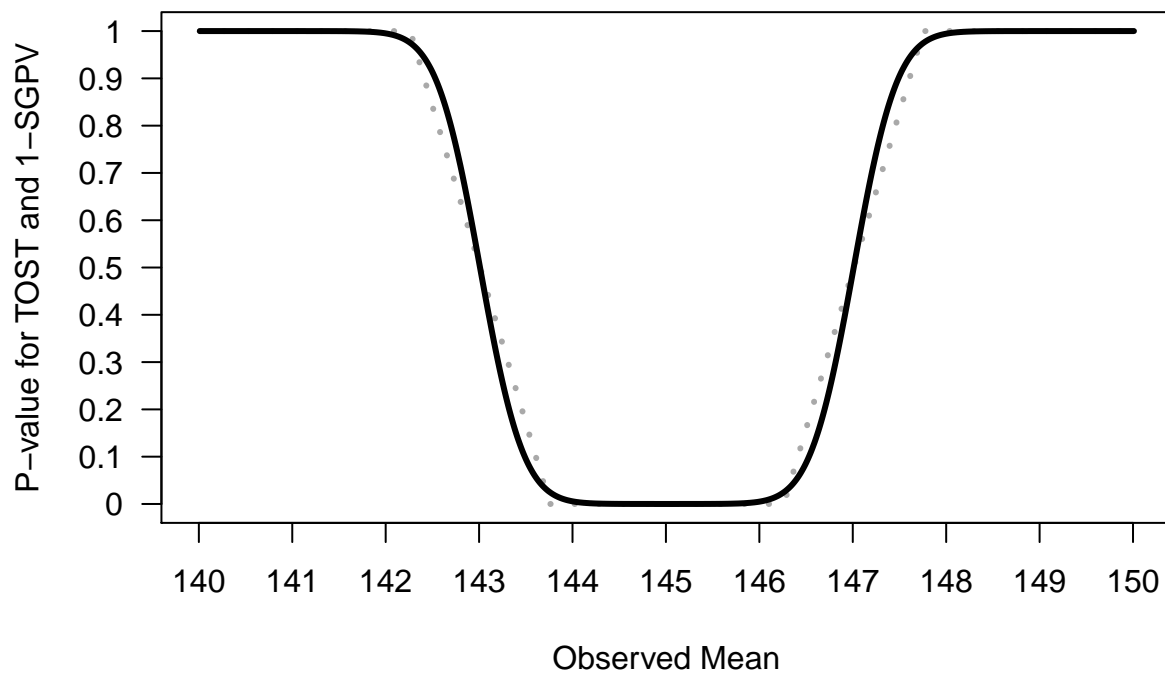Shiny app: http://shiny.ieis.tue.nl/TOST_vs_SGPV/.

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.

₉₂   The SGPV treats the equivalence range as the null-hypothesis, while the TOST

₉₃   procedure treats the values outside of the equivalence range as the null-hypothesis. For ease

₉₄   of comparison we can reverse the SGPV (by calculating 1-SGPV in Figure 2) to make the

₉₅   values more easily comparable. We see that the *p*-value from the TOST procedure and the

₉₆   SGPV follow each other closely.

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.

₉₇   When the observed sample mean is 145, the sample size is 30, and the standard

₉₈   deviation is 2, and we are testing against equivalence bounds of 143 and 147 using the TOST

₉₉   procedure for a one-sample *t*-test, the equivalence test is significant, $t(29) = 5.48$, $p < .001$.

₁₀₀   Because the 95% CI falls completely within the equivalence bounds, the SGPV is 1 (see

101 Figure 1). On the other hand, when the observed mean is 140, the equivalence test is not

102 significant (the observed mean is far outside the equivalence range of 143 to 147), $t(29) =$

103 -8.22, $p = 1$ (or more accurately, $p > .999$ as p-values are bounded between 0 and 1). Because

104 the 95% CI falls completely outside the equivalence bounds, the SGPV is 0 (see Figure 1).

SGPV as a uniform measure of overlap

106   It is clear the SGPV and the p-value from TOST are closely related. When confidence

107 intervals are symmetric we can think of the SGPV as a straight line that is directly related

108 to the p-value from an equivalence test for three values. When the TOST p-value is 0.5, the

109 SGPV is also 0.5 (note that the reverse is not true). The SGPV is 50% when the observed

110 mean falls exactly on the lower or upper equivalence bound, because 50% of the symmetrical

111 confidence interval overlaps with the equivalence range. When the observed mean equals the

112 equivalence bound, the difference between the mean in the data and the equivalence bound is

113 0, the t-value for the equivalence test is also 0, and thus the p-value is 0.5 (situation A,

114 Figure 3).

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

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

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

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

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

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

121 bound (see situation C in Figure 3), when the p-value will always be 0.975. Note that this

122 situation is in essence a minimum-effect test (Murphy, Myors, & Wolach, 2014). The goal of

123 a minimum-effect is not just to reject a difference of zero, but to reject the smallest effect

124 size of interest (i.e., the equivalence bounds). The SGPV summarizes the information from

125 an equivalence test and a minimum-effect test. These can be two relevant questions to ask,

126 although it often makes sense to combine an equivalence test and a null-hypothesis test
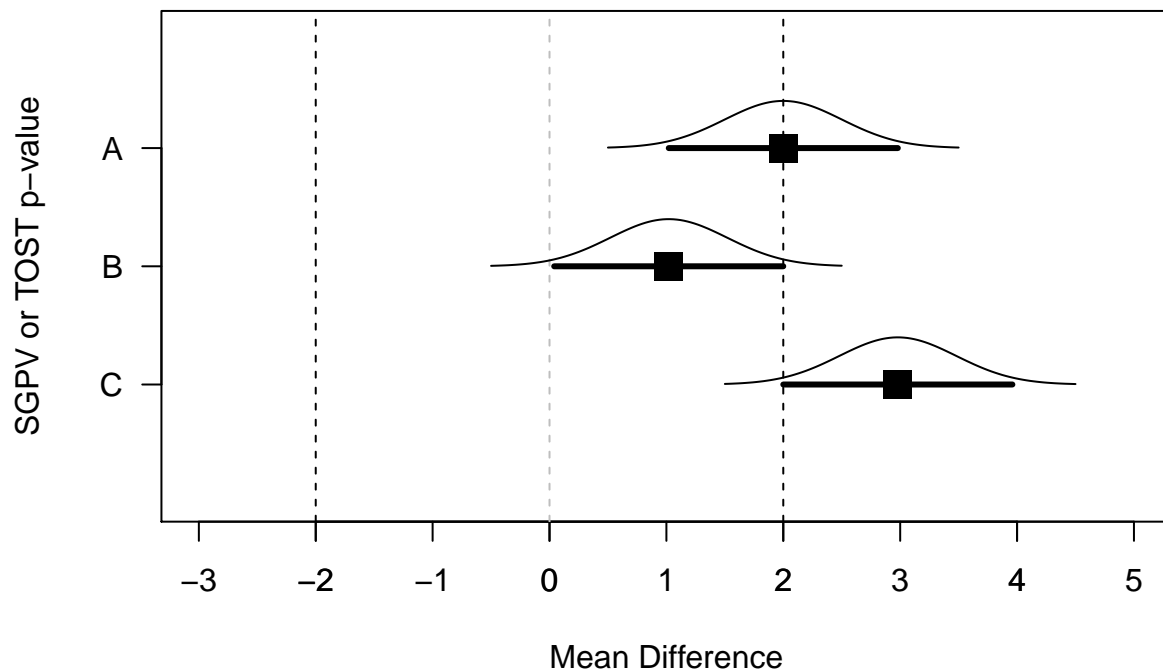
Figure 3. Means, normal distribution, and 95% CI for three example datasets that illustrate the relationship between *p*-values from TOST and SGPV.

127   instead (Lakens et al., 2018).

128   For example, the SGPV from A to D is 0.76, 0.81, 0.86, and 0.91. The difference in the

129   percentage of overlap between A and B (-0.05) is identical to the difference in the percentage

130   of overlap between C and D as the mean gets 0.1 closer to the test value (-0.05). As the

131   observed mean in a one-sample *t*-test lies closer to the test value (in Figure 4), from

132   situation A to D, the mean gets closer to the test value by 0.1) the difference in the overlap

133   changes uniformly. As we move the observed mean closer to the test value in steps of 0.1

134   across A to D the *p*-value calculated for normally distributed data is not uniformly

135   distributed. The probability of observing data more extreme than the upper bound of 2 is

136   (from A to D) 0.16, 0.12, 0.08, and 0.06. As we can see, the difference between A and B
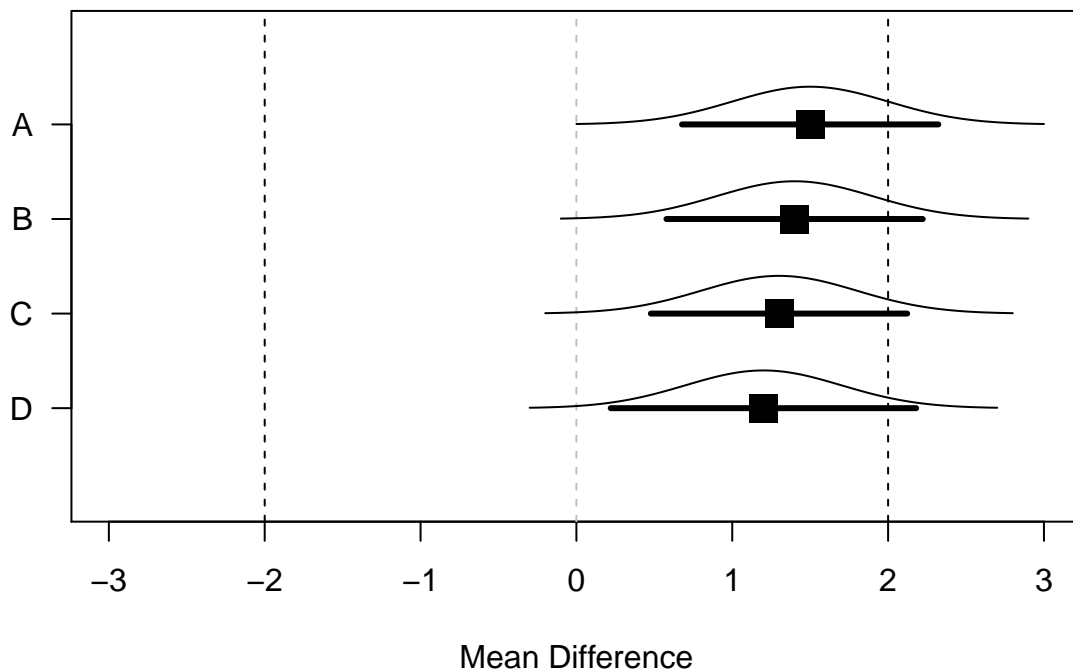
Figure 4. Means, normal distribution, and 95% CI for samples where the observed population mean is 1.5, 1.4, 1.3, and 1.2.

137 (0.04) is not the same as the difference between C and D (0.03). Indeed, the difference in

138 p-values is the largest as you start at $p = 0.5$ (when the observed mean falls on the test

139 value), which is why the line in Figure 1 is the steepest at $p = 0.5$. Note that where the

140 SGPV reaches 1 or 0, p-values closely approximate 0 and 1, but never reach these values.

141 **When are the SGPV and Equivalence Test Unrelated?**

142   There are three situations where p-values from TOST and SGPV are unrelated. The

143 first two situations were discussed before and can be seen in Figure 1. When the SGPV is

144 either 0 or 1 p-values from the equivalence test fall between 0.975 and 1 or between 0 and

145 0.025. Because p-values approach 0 or 1, but are never exactly 0 or 1, and the SGPV is

146 exactly 0 or 1, the two statistics are completely unrelated within these ranges. The easiest

147 way to see this is by plotting the SGPV against the *p*-value from the TOST procedure. The

148 situations where the SPGV and *p*-values from the TOST procedure are unrelated are

149 indicated by the parts of the curve where there are vertical straight lines at second
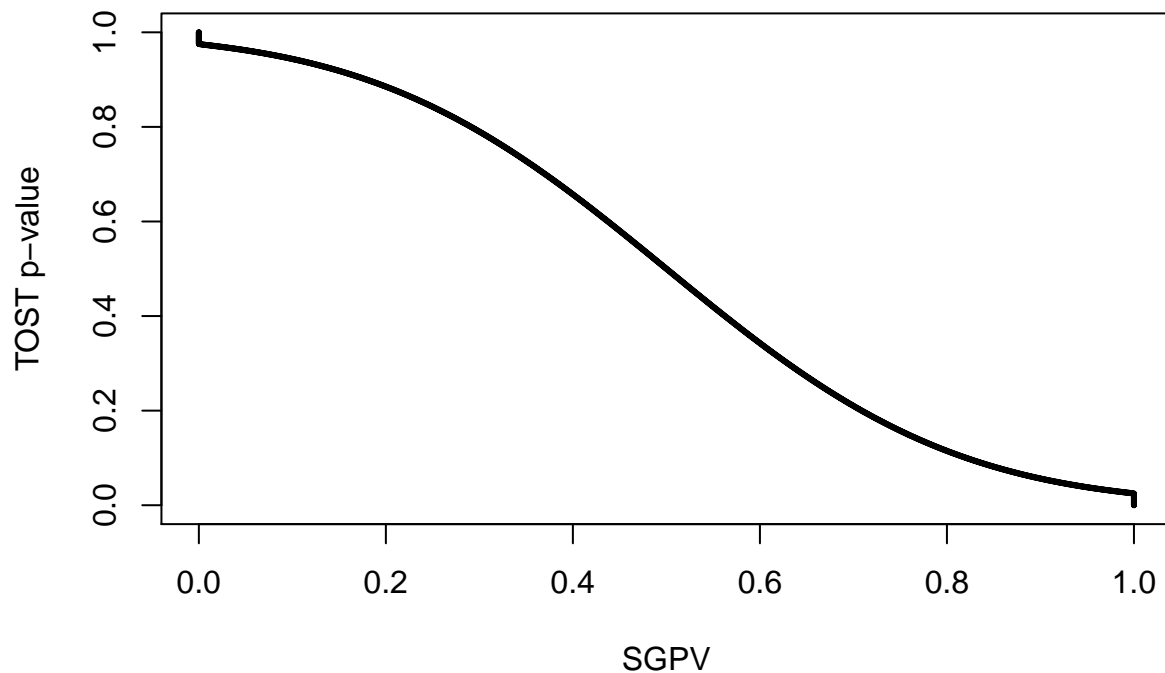
150 generation *p*-values of 0 and 1.

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

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

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

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

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

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

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

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

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

159 supported), the SGPV is set to 0.5 which according to Blume and colleagues signals that the

160 SGPV is "uninformative". Note that the CI can be twice as wide as the equivalence range

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

162 equivalence range is narrow. It is therefore not so much a "small sample correction" as it is

163 an exception to the typical calculation of the SGPV whenever the ratio of the confidence

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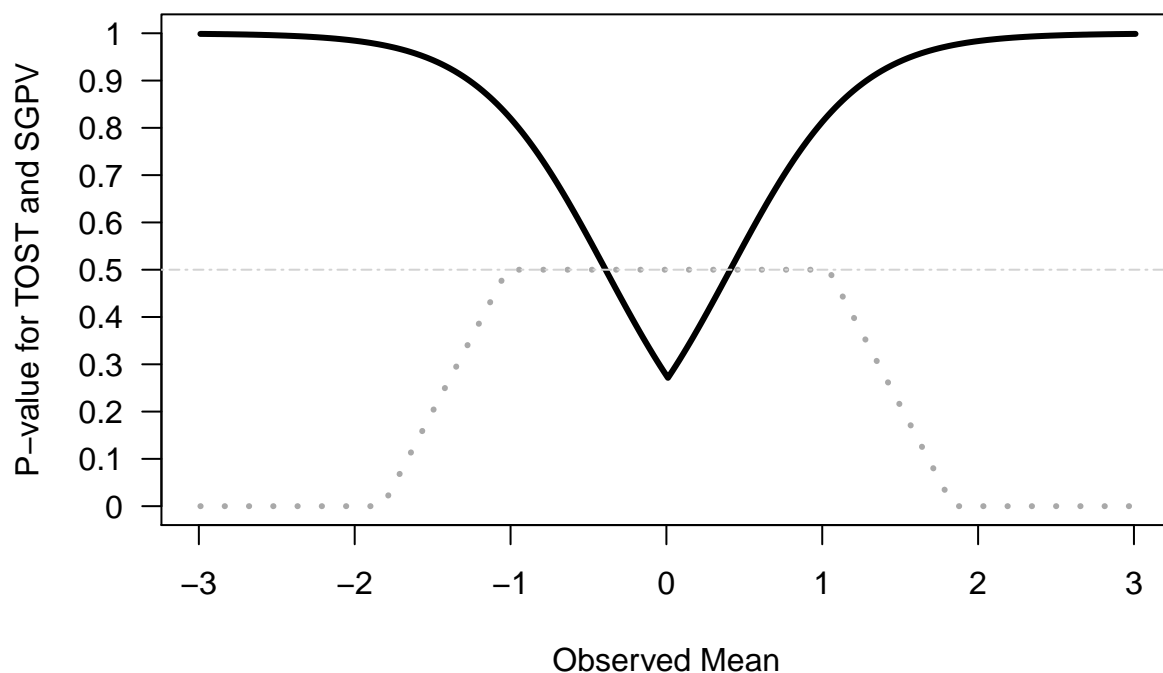
165 lower bounds.

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.

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

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

168 CI is more than twice as large as the equivalence range (see Figure 6). When the two

169 statistics are plotted against each other we can see where they are unrelated (indicated by

170 straight lines in the curve, see Figure 7). We see the SGPV is 0.5 for a range of observed

171 means where the p-value from the equivalence test still varies. It should be noted that in

172 these calculations the p-values for the TOST procedure are *never* smaller than 0.05 (i.e.,

173 they do not get below 0.05 on the y-axis). In other words, we cannot conclude equivalence

174 based on any of the observed means. This happens because the TOST p-value is smaller

175 than 0.05 only when the 90% CI falls completely between the upper and lower equivalence

176 bounds. However, we are examining a scenario where the 90% CI is so wide that it never

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

178 significant. As Lakens (2017) notes: "in small samples (where CIs are wide), a study might

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

180 the equivalence bounds)." None of the p-values based on the TOST procedure are below 0.05,

181 and thus, in the long run we have 0% power.

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

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

184 lower equivalence bound, but the CI is *not* twice as wide as the equivalence range. In the

185 example below, we see that the CI is only 1.79 times as wide as the equivalence bounds, but

186 the CI overlaps with the lower and upper equivalence bounds (Figure 8). This means the

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

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

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

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

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

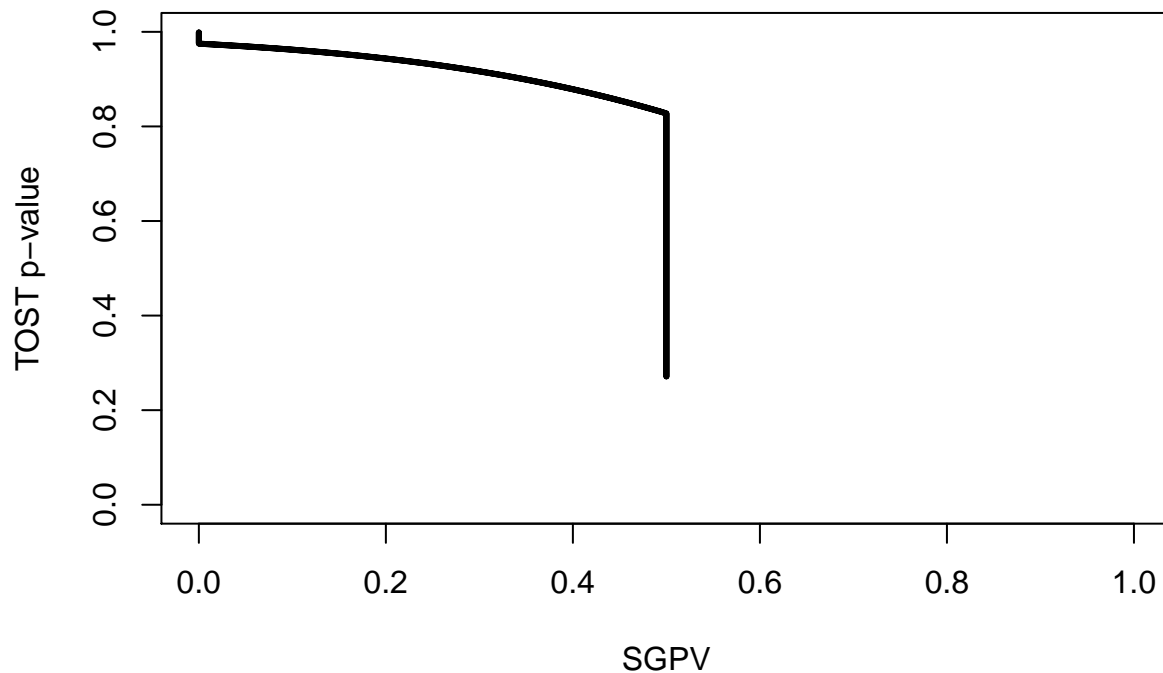192 means. How high the SGPV will be for a CI that is not twice as wide as the equivalence

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

₁₉₃ range, but overlaps with the lower and upper equivalence bounds, depends on the width of

₁₉₄ the CI and the equivalence range.

₁₉₅   If we once more plot the two statistics against each other to see where they are

₁₉₆ unrelated (indicated by straight lines in the curve), we see the SGPV is 0.56 for a range of

₁₉₇ observed means where the *p*-value from the equivalence test still varies (Figure 10).

₁₉₈   To conclude this section, there are three situations where the *p*-value from the TOST

₁₉₉ procedure is unrelated to the SGPV. In all these situations the *p*-value for the equivalence

₂₀₀ test differentiates tests with different means, but the SGPV does not. Therefore, as a purely

₂₀₁ continuous descriptive statistic, the SGPV is more limited than the *p*-value from the TOST
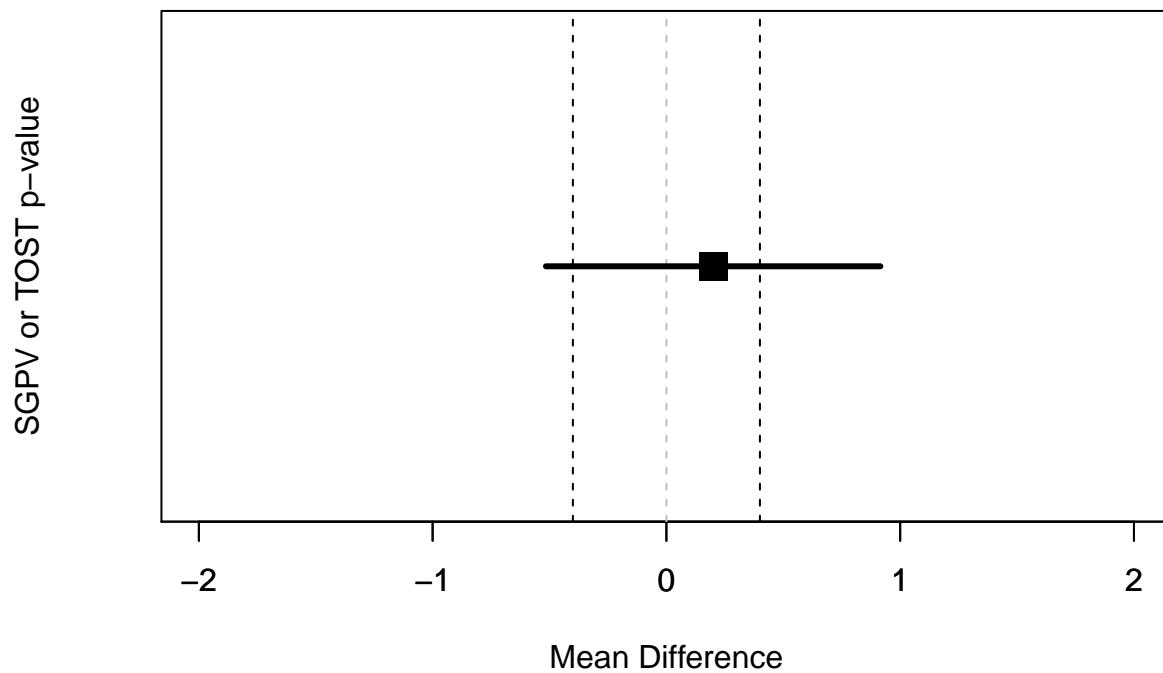
₂₀₂ procedure.

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 -0.630 to
0.630, while and observed correlation of 0.7 has an asymmetric 95% confidence interval
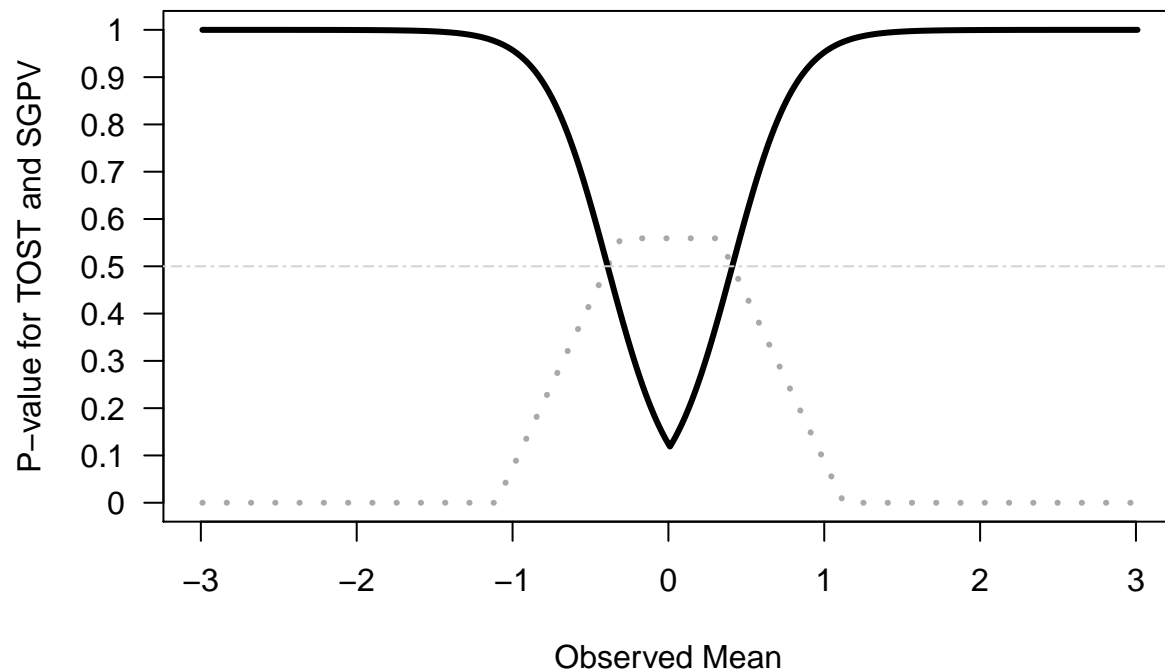ranging from 0.126 to 0.993.

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 light grey line) but reaches a maximum slightly above 0.5 across a range of observed means.

²¹³ The effect of asymmetric confidence intervals is most noticeable at smaller sample sizes.

²¹⁴ In Figure 11 we plot the *p*-values from equivalence tests and the SGPV (again plotted as

²¹⁵ 1-SGPV for ease of comparison) for correlations. The sample size is 30 pairs of observations,

²¹⁶ and the lower and upper equivalence bounds are set to -0.45 and 0.45, with an alpha of 0.05.

²¹⁷ As the observed correlation in the sample moves from -1 to 0 the *p*-value from the

²¹⁸ equivalence test becomes smaller, as does 1-SGPV. The pattern is quite similar to that in

²¹⁹ Figure 2. The *p*-value for the TOST procedure and 1-SGPV are still identical when *p*-values
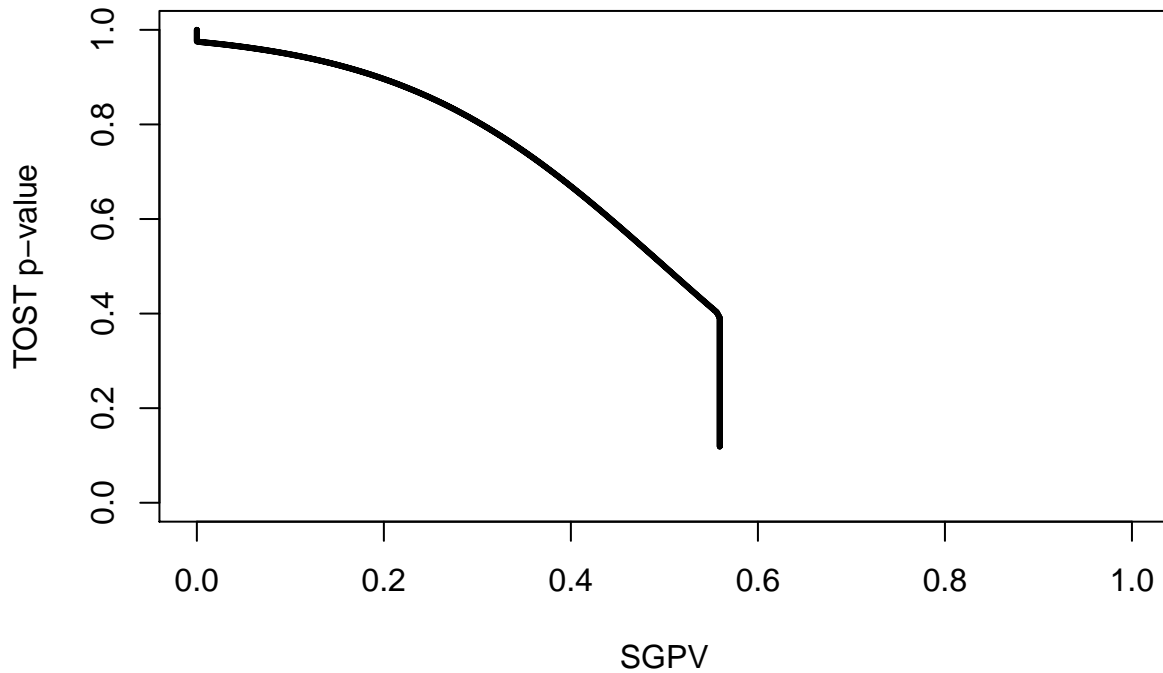
Figure 10. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 9.

are 0.975 and 0.025 (indicated by the upper and lower horizontal dotted lines). There are two important differences, however. First of all, the SGPV is no longer a straight line, but a curve, due to the asymmetry in the 95% CI. Second, and most importantly, the *p*-value for the equivalence test and the SGPV do no longer overlap at $p = 0.5$.

The reason that the equivalence test and SGPV no longer overlap is also because of asymmetric confidence intervals. If the observed correlation falls exactly on the equivalence bound the *p*-value for the equivalence test indicates that the probability of observing the observed or more extreme data, assuming the equivalence bound is the true effect size, is 50%. In other words, if the true effect size is the same as the equivalence bound, it is equally likely to find an effect more extreme than the equivalence bound, as it is to observe an effect
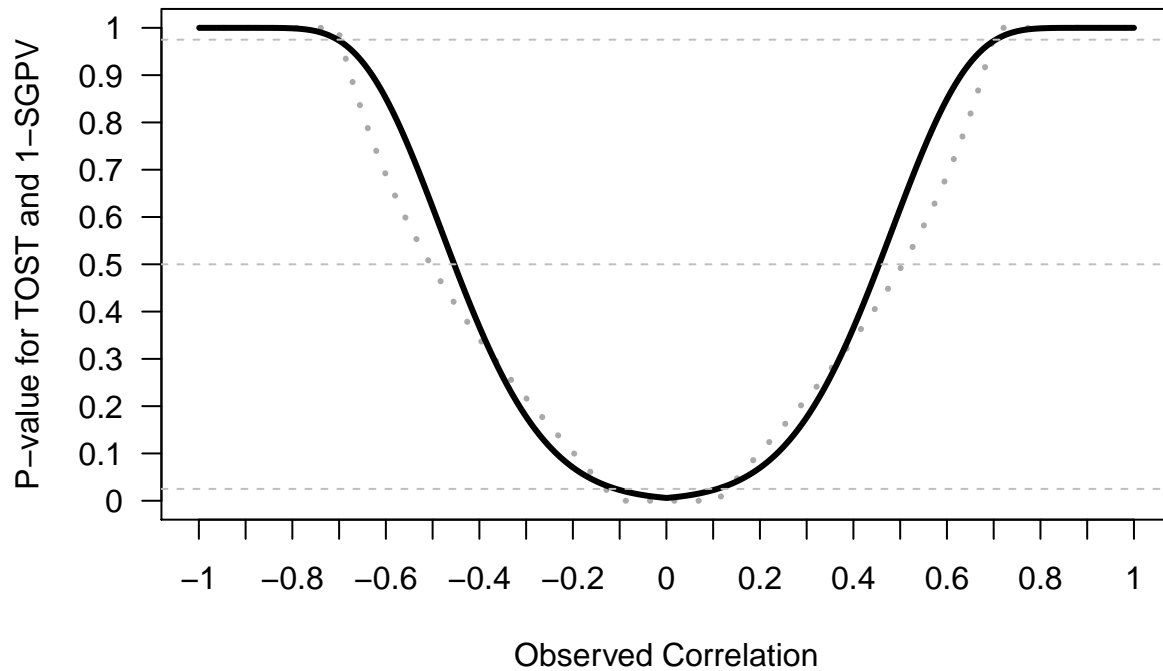
Figure 11. Comparison of *p*-values from TOST (black line) and 1-SGPV (dotted grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = -0.45 and r = 0.45 with n = 30 and an alpha of 0.05.

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

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

232 = 0.45 are larger than 50%. Because the confidence intervals are asymmetric around the

233 observed effect size of 0.45 (ranging from 0.11 to 0.70) according to Blume et al. (2018)

234 58.11% of the data-supported hypotheses are null hypotheses, and therefore 58.10% of the

235 data-supported hypotheses are compatible with the null premise.

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

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

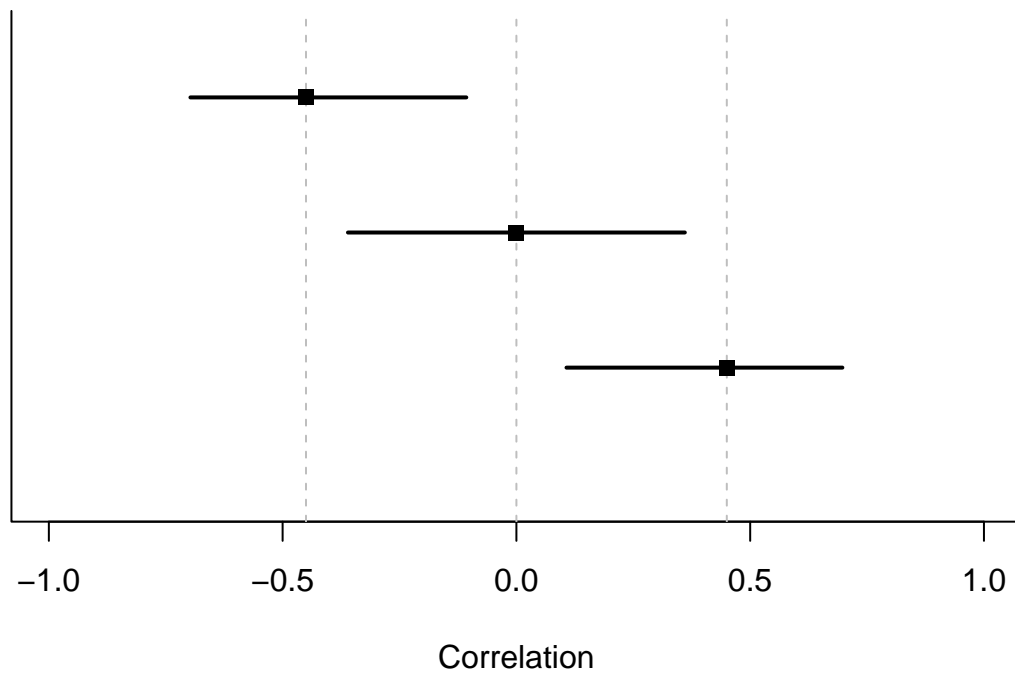238 correlation, but the SGPV for this situation depends on how far away the observed

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.

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

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

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

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

243 set to r = -0.99 and 0.99, with an observed correlation of 0.99) 58.10% of the confidence

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

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

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

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

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

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

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

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

252 (2018) state that "the degree of overlap conveys how compatible the data are with the null

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

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

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

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

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

258 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?

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

262 approaches and specify its relative strengths and weaknesses. First of all, the SGPV is a

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

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

265 the equivalence range. Even though a SGPV of 1 or 0 has a clear interpretation (we can

266 reject effects outside or inside the equivalence range), intermediate values are not as easy to

267 interpret (especially for effects that have asymmetric confidence intervals). In one sense, they

268 are what they are (the proportion of overlap), but it can be unclear what this number tells

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

270 SGPV (e.g., in all examples provided by Blume and colleagues) seems to be to examine

271 whether the SGPV is 0, 1, or inconclusive. This interpretation of a SGPV as allowing

272 researchers to reject the null, reject the presence of a meaningful effect, or remaining

273 inconclusive is very similar to the Neyman-Pearson interpretation of combining a

274 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 desirable 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 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.

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

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

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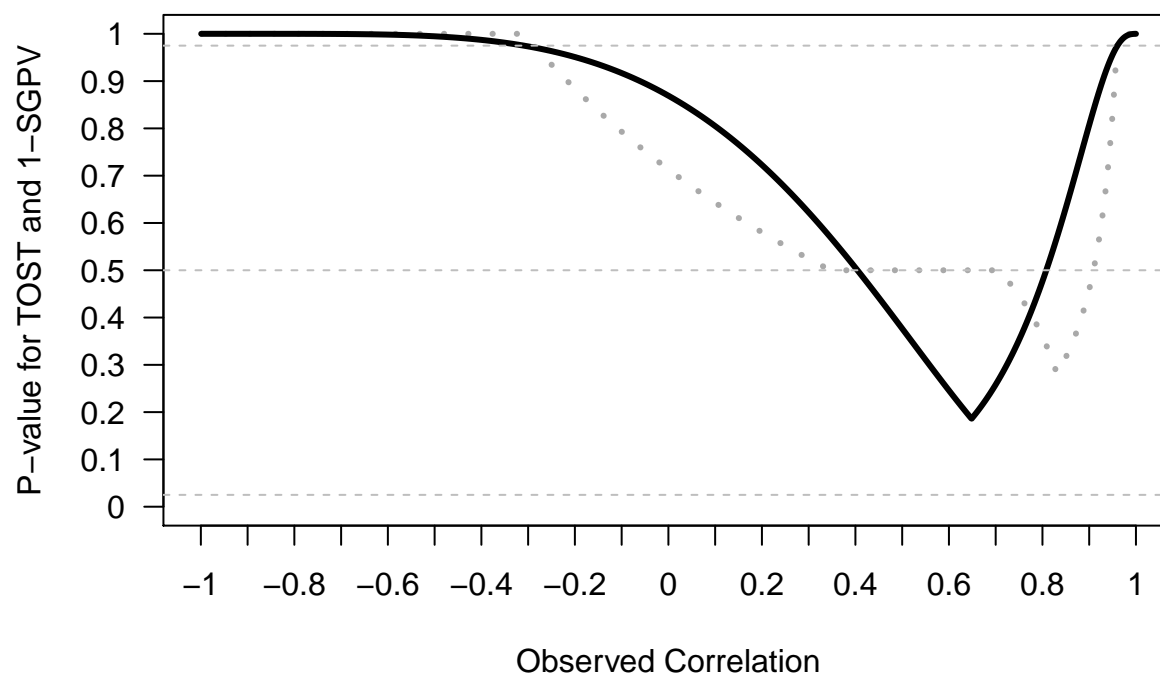

Figure 13. Comparison of *p*-values from TOST (black line) and 1-SGPV (dotted grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = 0.4 and r = 0.8 with n = 10 and an alpha of 0.05.

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

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

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

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

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

309 "misleading" SGPV in this range. This example suggests that the current implementation of

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

meant to be interpreted when it is 0 or 1, it might be preferable to simply never apply the "small sample correction".

Blume et al. (2018) argue that the SGPV has improved error control, in that the conclusion that the 95% confidence interval lies completely outside the equivalence range around zero will necessarily occur less frequently than a Type I error in a null-hypothesis test (where the 95% confidence interval only needs to not overlap with zero). However, the SGPV has a *lower* error rate than a null-hypothesis test, not a *more accurate* error rate. In a Neyman-Pearson perspective (which forms the basis of equivalence tests) the goal is not to end up with an error rate that is as low as possible (to achieve that, one simply adjusts the alpha level), but with a decision procedure that, when applied, yields a desired error rate with high accuracy. We show here that the SGPV and equivalence tests are directly related, and so are their error rates. The TOST procedure uses a 90% confidence interval because it is based on two *one-sided* tests, and thus the confidence interval is 1-2×α, and because both tests need to be significant to conclude equivalence, no adjustment for multiple comparisons is needed. The TOST procedure always has higher power to declare equivalence than the SGPV, which relies on a 95% confidence interval (but note that there is no reason to always set the alpha level to 0.05).

Blume et al. (2018) claim that "Adjustments for multiple comparisons are obviated" (p. 15) and that "second-generation *p*-values provide a proper scientific adjustment for multiple comparisons". However, this is not correct. With a sufficient number of looks error rates can rise above the nominal alpha level (both when compared to equivalence and minimal effect tests, but even when compared to a null-hypothesis test). When directly comparing equivalence tests and the SGPV, multiple comparisons inflate the probability that one erronously concludes there is *no* effect, where there *is* a true effect size that equals the equivalence bound, just as quickly for the SGPV as for equivalence tests, because they are directly related. To conclude, the idea that the SGPV improves error rates does not hold up under closer scrutiny, and the recommendation to ignore adjustments for multiple

comparisons has the potential to increase false positives in the literature. Equivalence tests provide an easier and more formal way to control both Type I error rates (by setting the alpha level) and the Type II error rate (by performing an a-priori power analysis).

Conclusion

We believe that our explanation of the similarities between the TOST procedure and the SGPV provides context to interpret the contribution of second generation p-values to the statistical toolbox. The novelty of the SGPV lies in its use as a descriptive statistic, but this use can be limited when confidence intervals are asymmetrical or wider than the equivalence range. There are strong similarities with p-values from the TOST procedure, and in all situations where the statistics yield different results, the behavior of the p-value from the TOST procedure is more consistent and easier to interpret. We hope this overview of the relationship between the SGPV and equivalence tests will help researchers to make an informed decision about which statistical approach provides the best answer to their question. Our comparisons show that when proposing alternatives to null-hypothesis tests, it is important to compare new proposals to already existing procedures. We believe equivalence tests achieve the goals of the second generation p-value while allowing users to 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., 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

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

Murphy, K. R., Myors, B., & Wolach, A. H. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (Fourth edition.). New York: Routledge, Taylor & Francis Group.

Quertemont, E. (2011). How to Statistically Show the Absence of an Effect. *Psychologica Belgica*, *51*(2), 109–127. doi:10.5334/pb-51-2-109

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

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 357–416. doi:10.2307/2983527

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton: CRC Press.

Westlake, W. J. (1972). Use of Confidence Intervals in Analysis of Comparative Bioavailability Trials. *Journal of Pharmaceutical Sciences*, *61*(8), 1340–1341. doi:10.1002/JPS.2600610845