1   Equivalence Testing and the Second Generation P-Value

2   Daniël Lakens[1] & Marie Delacre[2]

3   [1] Eindhoven University of Technology, Eindhoven, The Netherlands
4   [2] Service of Analysis of the Data, Université Libre de Bruxelles, Belgium

5   Author Note

10   Correspondence concerning this article should be addressed to Daniël Lakens, Den

11   Dolech 1, IPO 1.33, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

12   Abstract

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

14   developed where the observed data are compared against a range of values that are

15   equivalent to the absence of a meaningful effect. Specifying a range of values around zero

16   allows researchers to statistically reject the presence of effects large enough to matter, and

17   prevents practically insignificant effects from being interpreted as a statistically significant

18   difference. We compare the behavior of the recently proposed second generation p-value

19   (Blume, D'Agostino McGowan, Dupont, & Greevy, 2018) with the more established Two

20   One-Sided Tests (TOST) equivalence testing procedure (Schuirmann, 1987). We show that

21   the two approaches yield almost identical results under optimal conditions. Under

22   suboptimal conditions (e.g., when the confidence interval is wider than the equivalence range,

23   or when confidence intervals are asymmetric) the second generation p-value becomes difficult

24   to interpret. The second generation p-value is interpretable in a dichotomous manner (i.e.,

25   when the SGPV equals 0 or 1 because the confidence intervals lies completely within or

26   outside of the equivalence range), but this dichotomous interpretation does not require

27   calculations. We conclude that equivalence tests yield more consistent p-values, distinguish

28   between datasets that yield the same second generation p-value, and allow for easier control

29   of Type I and Type II error rates.

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

31   statistical inference

³²   Equivalence Testing and the Second Generation *P*-Value

³³   To test predictions researchers predominantly rely on null-hypothesis tests. This

³⁴ statistical approach can be used to examine whether observed data are sufficiently surprising

³⁵ under the null hypothesis to reject an effect that equals exactly zero. Null-hypothesis tests

³⁶ have an important limitation, in that this procedure can only reject the hypothesis that

³⁷ there is no effect, while scientists should also be able to provide statistical support for

³⁸ *equivalence*. When testing for equivalence researchers aim to examine whether an observed

³⁹ effect is too small to be considered meaningful, and therefore is practically equivalent to zero.

⁴⁰ By specifying a range around the null hypothesis of values that are deemed practically

⁴¹ equivalent to the absence of an effect (i.e., 0 ± 0.3) the observed data can be compared

⁴² against an *equivalence range* and researchers can test if a meaningful effect is absent (Hauck

⁴³ & Anderson, 1984; Kruschke, 2018; Rogers, Howard, & Vessey, 1993; Serlin & Lapsley, 1985;

⁴⁴ Spiegelhalter, Freedman, & Parmar, 1994; Wellek, 2010; Westlake, 1972).

⁴⁵   Second generation *p*-values (SGPV) were recently proposed as a statistic that

⁴⁶ represents "the proportion of data-supported hypotheses that are also null hypotheses"

⁴⁷ (Blume et al., 2018). The researcher specifies an equivalence range around a null hypothesis

⁴⁸ of values that are considered practically equivalent to the null hypothesis. The SGPV

⁴⁹ measures the degree to which a set of data-supported parameter values falls within the

⁵⁰ interval null hypothesis. If the estimation interval falls completely within the equivalence

⁵¹ range, the SGPV is 1. If the confidence interval falls completely outside of the equivalence

⁵² range, the SGPV is 0. Otherwise the SGPV is a value between 0 and 1 that expresses the

⁵³ overlap of data-supported hypotheses and the equivalence range. When calculating the

⁵⁴ SGPV the set of data-supported parameter values can be represented by a confidence

⁵⁵ interval (CI), although one could also choose to use credible intervals or Likelihood support

⁵⁶ intervals (SI). When a confidence interval is used, the SGPV and equivalence tests such as

⁵⁷ the Two One-Sided Tests (TOST) procedure (Lakens, 2017; Meyners, 2012; Quertemont,

58 2011; Schuirmann, 1987) appear to have close ties, because both tests compare a confidence

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

60 differences between the TOST procedure and the SGPV. We limit our analysis to continuous

61 data sampled from a bivariate normal distribution.

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

63 TOST procedure the data are tested against the lower equivalence bound in the first

64 one-sided test, and against the upper equivalence bound in the second one-sided test

65 (Lakens, Scheel, & Isager, 2018). For an excellent discussion of the strengths and weaknesses

66 of different frequentist equivalence tests, including alternatives to the TOST procedure, see

67 Meyners (2012). If both tests statistically reject an effect as extreme or more extreme than

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

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

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

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

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

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

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

75 contribution of the SGPV to the statistical toolbox.

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

77 **intervals are symmetrical**

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

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

79 where I is the interval based on the data (e.g., a 95% confidence interval) and H$_0$ is the

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

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

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

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

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

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

86 Replicating the example by Blume et al. (2018), in Figure 1 p-values are plotted for the

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

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

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

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

91 1 the SGPV and the TOST p-value is calculated for a hypothetical study with a sample size

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

93 for the equivalence test is 1, and the SGPV is 0.

94   Our conclusions about the relationship between TOST p-values and SGPV hold for

95 second generation p-values calculated from confidence intervals, and assuming data is

96 sampled from a bivariate normal distribution. Readers can explore the relationship between

97 TOST p-values and SGPV for themselves in an online Shiny app:

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

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

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

101 of comparison we can plot 1-SGPV (see Figure 2) to make the values more easily comparable.

102 We see that the p-value from the TOST procedure and the SGPV follow each other closely.

103 When we discuss the relationship between the p-values from TOST and the SGPV, we focus

104 on their correspondence at three values, namely where the TOST $p = 0.025$ and SGPV is 1,

105 where the TOST $p = 0.5$ and SGPV $= 0.5$, and where the TOST $p = 0.975$ and SGPV $= 1$.

106 These three values are important for the SGPV because they indicate the values at which
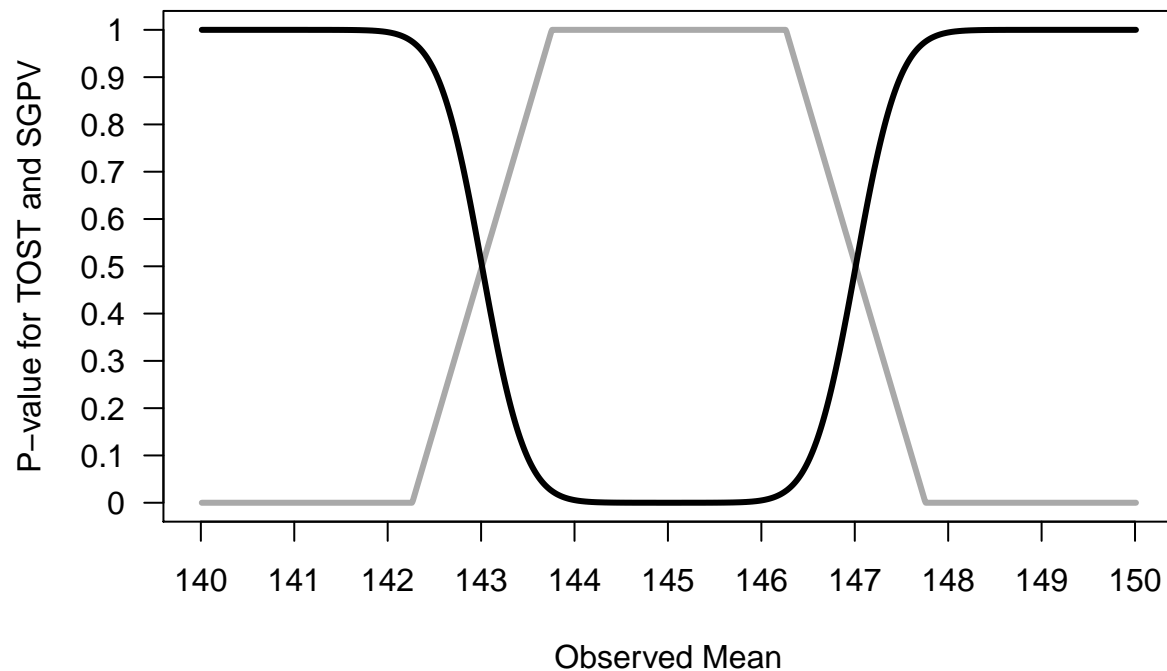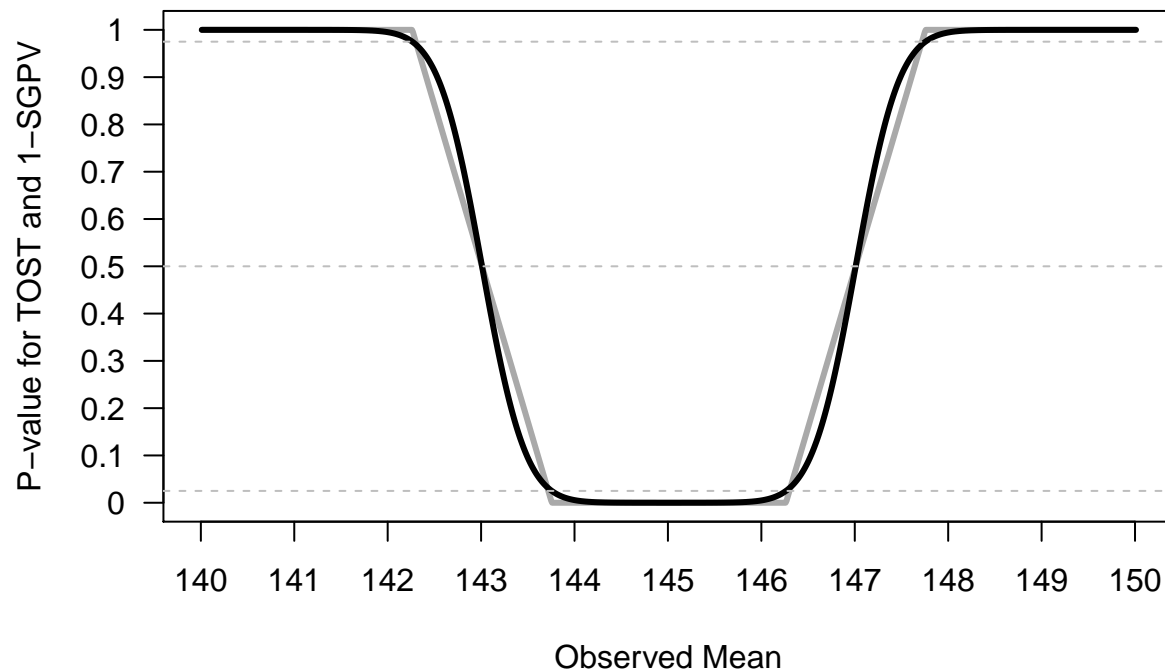
Figure 1. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2, illustrating that when the TOST *p*-value = 0.5, the SGPV = 0.5, when the TOST *p*-value is 0.975, 1-SGPV = 1, and when the TOST *p*-value = 0.025, 1-SGPV = 0.

107   the SGPV indicates the data should be interpreted as compatible with the null hypothesis

108   (SGPV = 1), or with the alternative hypothesis (SGPV = 0), or when the data are strictly

109   inconclusive (SGPV = 0.5). These three points of overlap are indicated by the horizontal

110   dotted lines in Figure 2 at TOST *p*-values of 0.975, 0.5, and 0.025.

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

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

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

Figure 2. Comparison of *p*-values from TOST (black line) and 1-SGPV (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.

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

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

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

117 -8.22, $p = 1$ (or more accurately, $p > .999$ as *p*-values are bounded between 0 and 1). Because

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

119 **SGPV as a uniform measure of overlap**

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

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

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

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

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

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

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

127 0, the *t*-value for the equivalence test is also 0, and thus the *p*-value is 0.5 (situation A,

128 Figure 3).
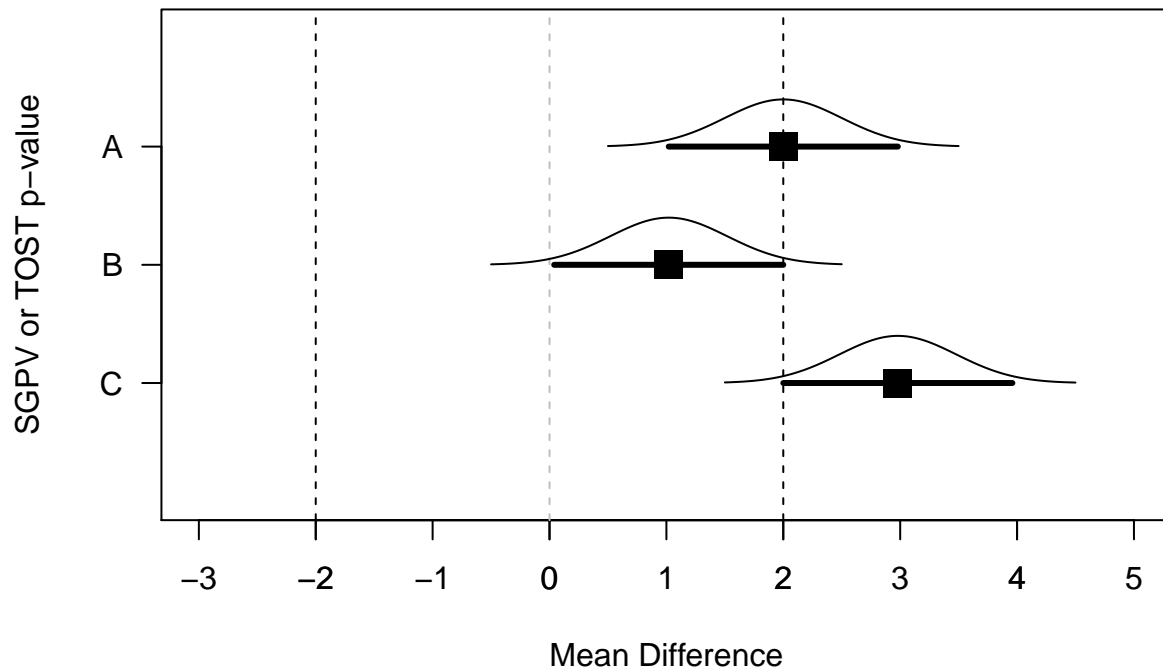
Figure 3. Means, normal distribution, and 95% CI for three example datasets that illustrate
the relationship between *p*-values from TOST and SGPV.

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

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

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

132 one-sided test) is always 0.025, and the SGPV is 1. Note that when sample sizes are small or

133 equivalence bounds are narrow, small *p*-values for the TOST or a SGPV = 1 might not be

134 observed in practice if too few observations are collected. The third point where the SGPV

135 and the *p*-value from the TOST procedure should overlap is where the 95% CI falls

136 completely outside of the equivalence range, but one endpoint of the confidence interval is

137 equal to the equivalence bound (see situation C in Figure 3), when the *p*-value will always be

138 0.975, and the SGPV is 0. Note that this situation is in essence a minimum-effect test

139 (Murphy, Myors, & Wolach, 2014). The goal of a minimum-effect is not just to reject a

140 difference of zero, but to reject the smallest effect size of interest (i.e., the equivalence

141 bounds). An equivalence test and minimum effect test against the same equivalence bound

142 are complementary, and when a TOST *p*-value is larger than 0.975, the *p*-value for the

143 minimum effect test is smaller than 0.05 (and therefore the minimum effect test provides no

144 additional information that can not be derived from the *p*-value from the equivalence test).

145 The SGPV summarizes the information from an equivalence test (and the complementary

146 minimum-effect test). These can be two relevant questions to ask, although it often makes

147 sense to combine an equivalence test and a null-hypothesis test instead (Lakens et al., 2018).

148   For example, in Figure 4 we have plotted four SGPV's. From A to D the SGPV is 0.76,

149 0.81, 0.86, and 0.91. The difference in the percentage of overlap between A and B (-0.05) is

150 identical to the difference in the percentage of overlap between C and D as the mean gets 0.1

151 closer to the test value (-0.05). As the observed mean in a one-sample *t*-test lies closer to the

152 test value, from situation A to D, the difference in the overlap changes uniformly. As we

153 move the observed mean closer to the test value in steps of 0.1 across A to D the *p*-value

154 calculated for normally distributed data are not uniformly distributed. The probability of

155 observing data more extreme than the upper bound of 2 is (from A to D) 0.16, 0.12, 0.08,

156 and 0.05. As we can see, the difference between A and B (0.04) is not the same as the

157 difference between C and D (0.03). Indeed, the difference in *p*-values is the largest as you

158 start at $p = 0.5$ (when the observed mean falls on the test value), which is why the line in
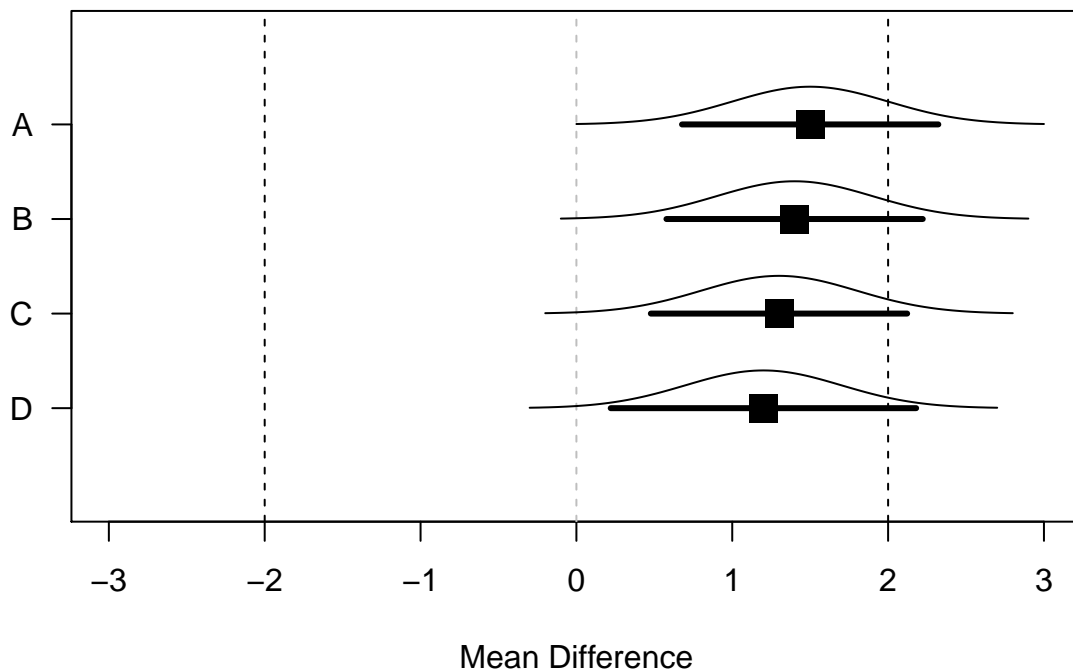
Figure 4. Means, normal distribution, and 95% CI for samples where the observed population mean is 1.5, 1.4, 1.3, and 1.2.

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

160 closely approximate 0 and 1, but never reach these values.

When different p-values for equivalence tests yield the same SGPV

162   There are three situations where p-values for TOST differentiate between observed

163 results, while the SGPV does not differentiate. The first two situations were discussed before

164 and can be seen in Figure 1. When the SGPV is either 0 or 1, p-values from the equivalence

165 test fall between 0.975 and 1 or between 0 and 0.025. Where the SGPV is 1 as long as the

166 confidence interval falls completely within the equivalence bounds, the p-value for the TOST

167 continues to differentiate between results as a function of how far the confidence interval lies

168 within the equivalence bounds (the further the confidence interval is from both bounds, the

169 lower the *p*-value). The easiest way to see this is by plotting the SGPV against the *p*-value

170 from the TOST procedure. The situations where the *p*-values from the TOST procedure

171 continue to differentiate based on how extreme the results are, but the SGPV is a fixed value

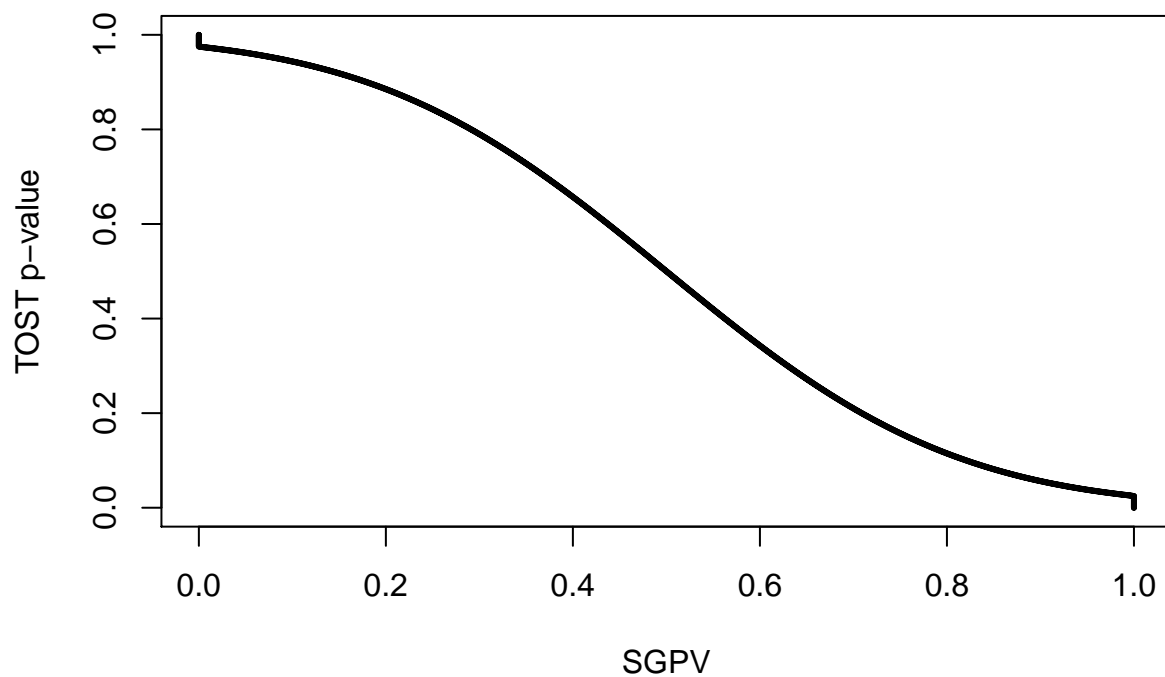172 are indicated by the parts of the curve where there are vertical straight lines at second

173 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.

174   A third situation in which the SGPV remains stable across a range of observed effects,

175 while the TOST *p*-value continues to differentiate, is whenever the CI is wider than the

176 equivalence range, and the CI overlaps with the upper *and* lower equivalence bound. When

177 the confidence interval is more than twice as wide as the equivalence range the SGPV is set

178 to 0.5. Blume et al. (2018) call this the "small sample correction factor". However, it is not a

179 correction in the typical sense of the word, since the SGPV is not adjusted to any "correct"

180 value. When the normal calculation would be "misleading" (i.e., the SGPV would be small,

181 which normally would suggest support for the alternative hypothesis, but at the same time

182 all values in the equivalence range are supported), the SGPV is set to 0.5 which according to

183 Blume and colleagues signals that the SGPV is "uninformative". Note that the CI can be

184 twice as wide as the equivalence range whenever the sample size is small (and the confidence

185 interval width is large) *or* when then equivalence range is narrow. It is therefore not so much

186 a "small sample correction" as it is an exception to the typical calculation of the SGPV

187 whenever the ratio of the confidence interval width to the equivalence range exceeds 2:1 and

188 the CI overlaps with the upper and lower bounds.

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

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

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

192 statistics are plotted against each other we can see where the SGPV is the same while the

193 TOST *p*-value still differentiates different observed means (indicated by straight lines in the

194 curve, see Figure 7). We see the SGPV is 0.5 for a range of observed means where the

195 *p*-value from the equivalence test still varies. It should be noted that in these calculations

196 the *p*-values for the TOST procedure are *never* smaller than 0.05 (i.e., they do not get below

197 0.05 on the y-axis). In other words, we cannot conclude equivalence based on any of the

198 observed means. This happens because the we are examining a scenario where the 90% CI is

199 so wide that it never falls completely within the two equivalence bounds. As Lakens (2017)

200 notes: "in small samples (where CIs are wide), a study might have no statistical power (i.e.,

201 the CI will always be so wide that it is necessarily wider than the equivalence bounds)."

202 None of the *p*-values based on the TOST procedure are below 0.05, and thus, in the long run
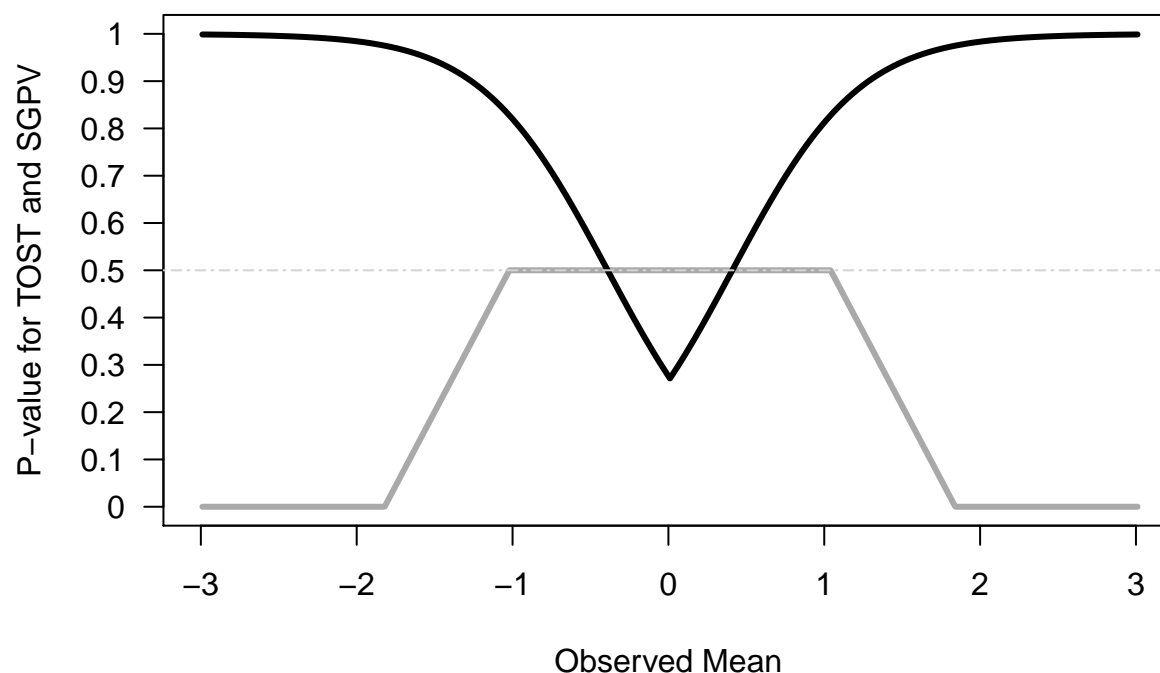
203 we have 0% power.

Figure 6. Comparison of *p*-values from TOST (black line) and SGPV (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.

204   The *p*-value from the TOST procedure still differentiates observed means, while the
205 SGPV does not, when the CI is wider than the equivalence range (so the precision is low)
206 and overlaps with the upper and lower equivalence bound, but the CI is *not* twice as wide as
207 the equivalence range. In the example below, we see that the CI is only 1.79 times as wide as
208 the equivalence bounds, but the CI overlaps with the lower and upper equivalence bounds
209 (Figure 8). This means the SGPV is not set to 0.5, but it is constant across a range of
210 observed means, while the TOST *p*-value is not constant across this range.

211   If the observed mean would be somewhat closer to 0, or further away from 0, the
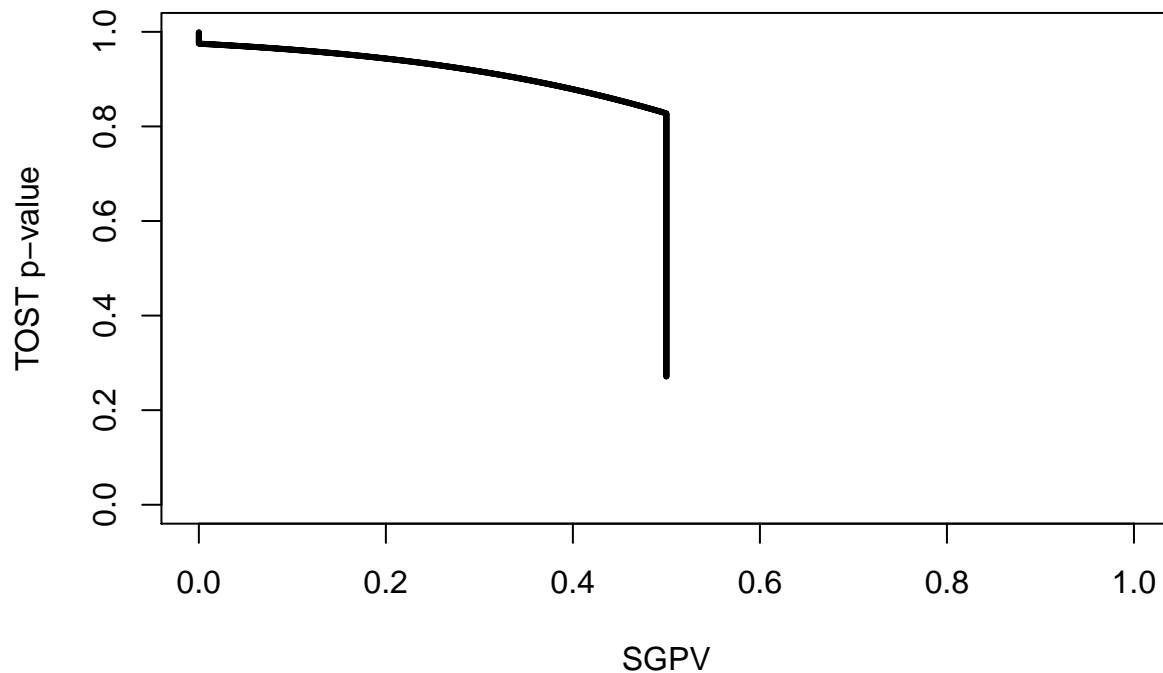
Figure 7. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 6.

²¹² SGPV remains constant (the CI width does not change, and it completely overlaps with the ²¹³ equivalence range) while the *p*-value for the TOST procedure does vary. We can see this in ²¹⁴ Figure 9 below. The SGPV is not set to 0.5, but is slightly higher than 0.5 across a range of ²¹⁵ means. How high the SGPV will be for a CI that is not twice as wide as the equivalence ²¹⁶ range, but overlaps with the lower and upper equivalence bounds, depends on the width of ²¹⁷ the CI and the equivalence range.

²¹⁸   If we once more plot the two statistics against each other we see the SGPV is 0.56 for a ²¹⁹ range of observed means where the *p*-value from the equivalence test still varies, as indicated ²²⁰ by the straight section of the line (Figure 10).
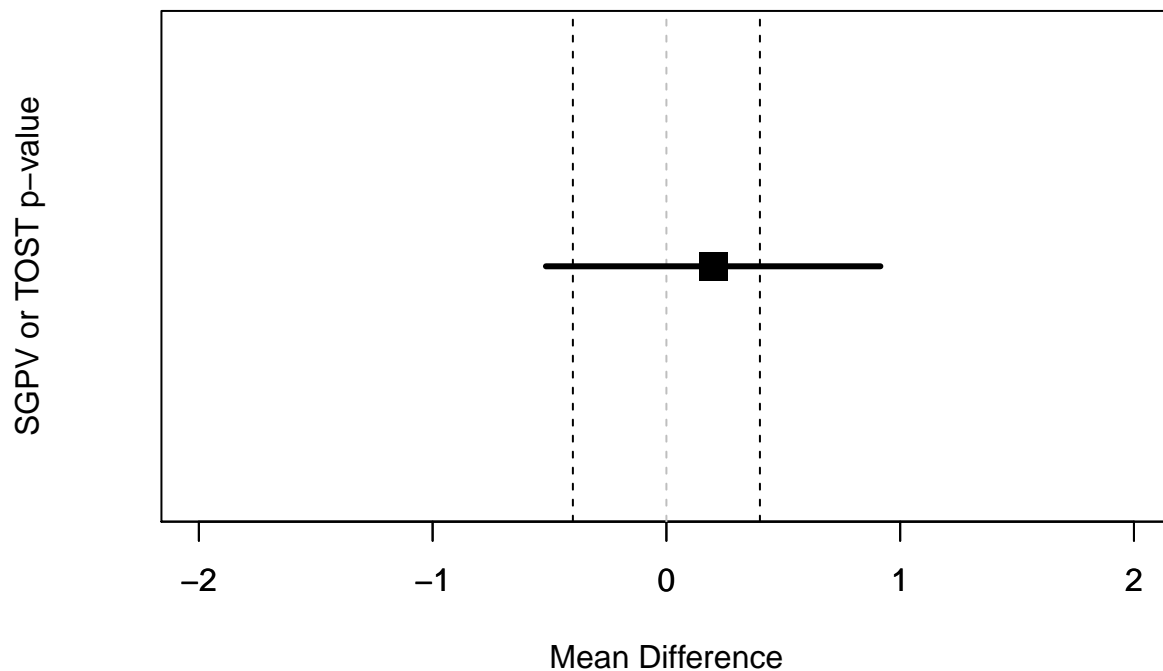
Figure 8. Example of a 95% CI that overlaps with the lower and upper equivalence bound (indicated by the vertical dotted lines).

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

222 procedure continues to differentiate, while the SGPV does not. Therefore, interpreted as a

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

The relation between equivalence tests and SGPV for asymmetrical confidence
intervals around correlations

226   So far we have only looked at the relation between equivalence tests and the SGPV

227 when confidence intervals are symmetric (e.g., for confidence intervals around mean

228 differences). For correlations, which are bound between -1 and 1, confidence intervals are

229 only symmetric for a correlation of exactly 0. The confidence interval for a correlation
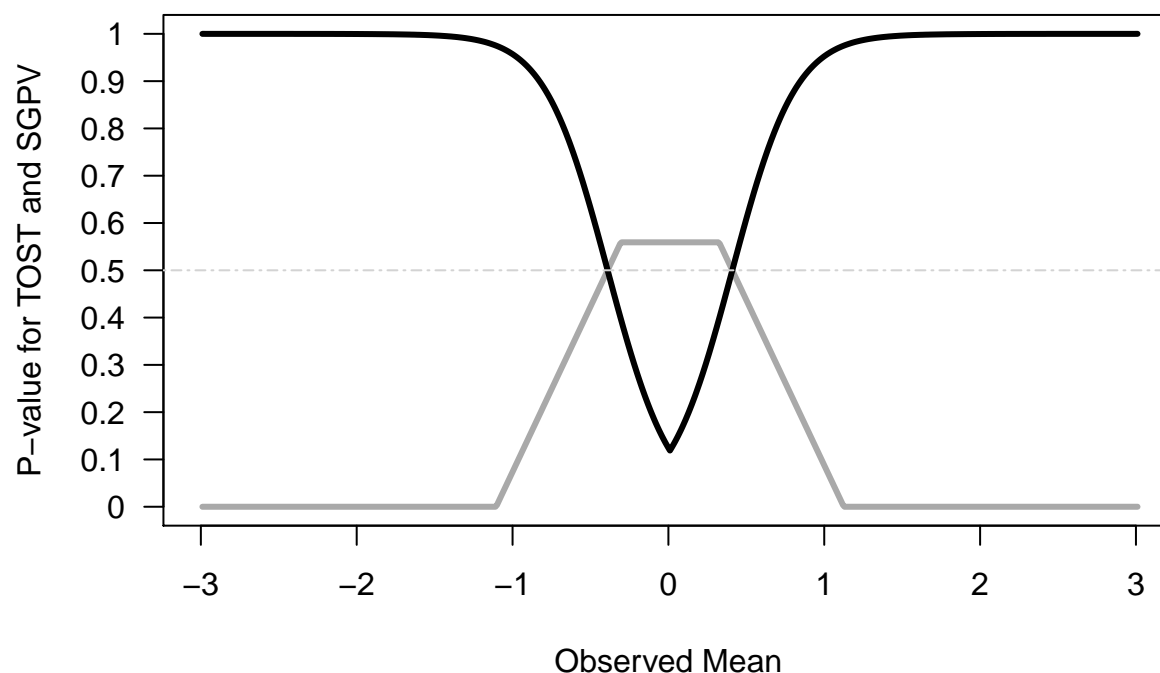
Figure 9. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis). The sample size is small (n = 10), but because the sd is half as big as in Figure 7 (1 instead of 2) the CI is less than twice as wide as the equivalence range (set to -0.4 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.

230 becomes increasingly asymmetric as the observed correlation nears -1 or 1. For example,

231 with ten observations, an observed correlation of 0 has a symmetric 95% confidence interval

232 ranging from -0.63 to 0.63, while and observed correlation of 0.7 has an asymmetric 95%

233 confidence interval ranging from 0.13 to 0.92. Note that calculating confidence intervals for a

234 correlation involves a Fisher's z-transformation, which transforms values such that they are

235 approximately normally z-distributed, which allows one to compute symmetric confidence

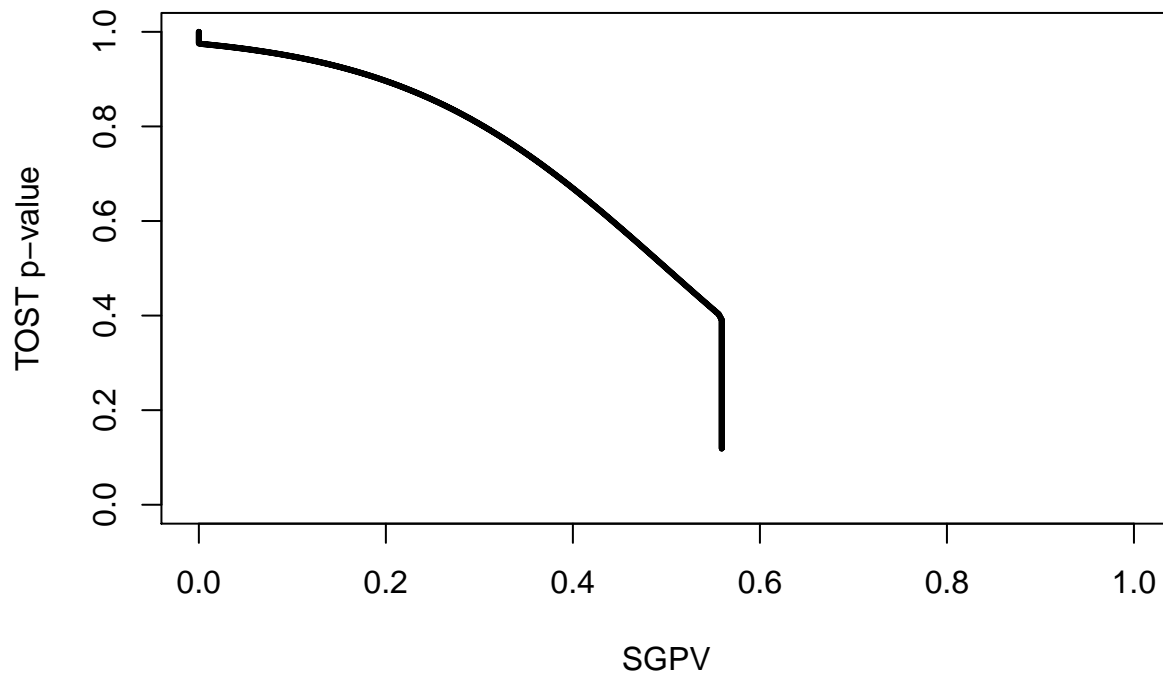236 intervals. These confidence intervals are then retransformed into a correlation, where the

Figure 10. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 9.

₂₃₇ confidence intervals are asymmetric if the correlation is not exactly zero.

₂₃₈   The effect of asymmetric confidence intervals around correlations is most noticeable at ₂₃₉ smaller sample sizes. In Figure 11 we plot the *p*-values from equivalence tests and the SGPV ₂₄₀ (again plotted as 1-SGPV for ease of comparison) for correlations. The sample size is 30 ₂₄₁ pairs of observations, and the lower and upper equivalence bounds are set to -0.45 and 0.45, ₂₄₂ with an alpha of 0.05. As the observed correlation in the sample moves from -1 to 0 the ₂₄₃ *p*-value from the equivalence test becomes smaller, as does 1-SGPV. The pattern is quite ₂₄₄ similar to that in Figure 2. The *p*-value for the TOST procedure and 1-SGPV are still ₂₄₅ related as discussed above, with TOST *p*-values of 0.975 and 0.025 corresponding to a ₂₄₆ 1-SGPV of 1 and 0, respectively. There are two important differences, however. First of all,

247 the SGPV is no longer a straight line, but a curve, due to the asymmetry in the 95% CI.

248 Second, and most importantly, the *p*-value for the equivalence test and the SGPV do no
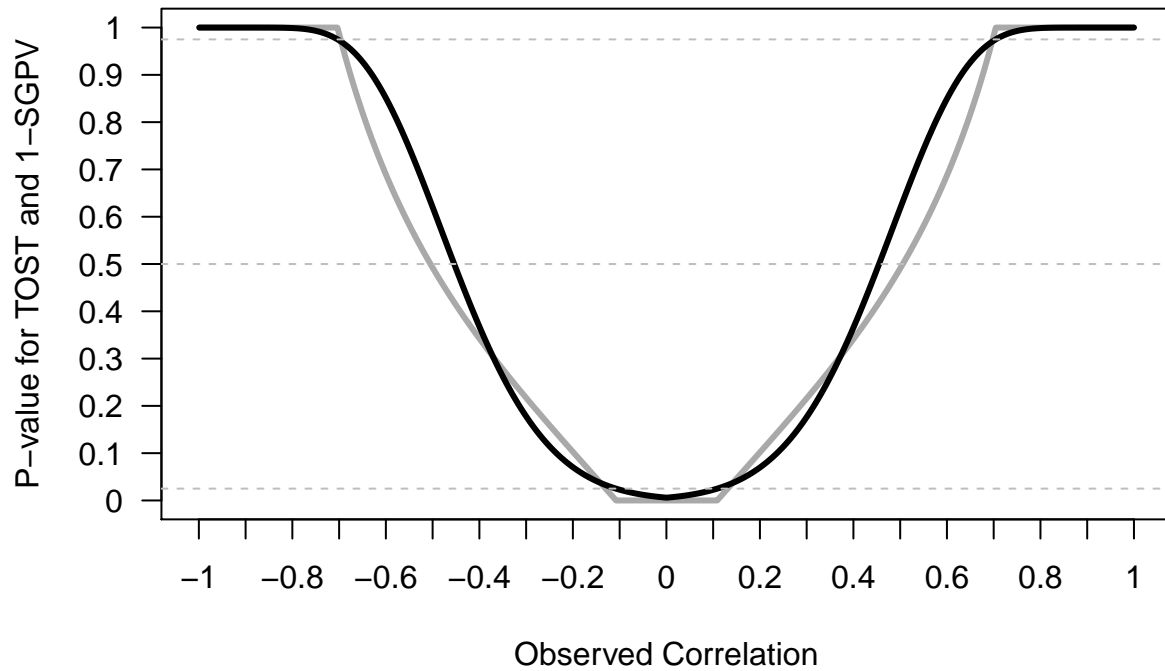
249 longer overlap at $p = 0.5$.

Figure 11. Comparison of *p*-values from TOST (black line) and 1-SGPV (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.

250   The reason that the equivalence test and SGPV no longer overlap is due to asymmetric

251 confidence intervals. If the observed correlation falls exactly on the equivalence bound the

252 *p*-value for the equivalence test is 0.5. In the equivalence test for correlations the *p* value is

253 computed based on a z-transformation which better controls error rates (Goertzen & Cribbie,

254 2010). This transformation is computed as follows, where r is the observed correlation and ρ

255 is the theoretical correlation under the null:

$$z = \frac{\frac{\log(\frac{1+r}{1-r})}{2} - \frac{\log(\frac{1+\rho}{1-\rho})}{2}}{\sqrt{\frac{1}{n-3}}}$$

[256] Because the z-distribution is symmetric, the probability of observing the observed or
[257] more extreme z-score, assuming the equivalence bound is the true effect size, is 50%.
[258] However, because the r distribution is not symmetric, this does not mean that there is always
[259] a 50% probability of observing a correlation smaller or larger than the true correlation. As
[260] can be seen in Figure 12, the two second generation p-values associated with the observed
[261] correlations at $r = $ -0.45 and $r = 0.45$ are larger than 50%. Because the confidence intervals
[262] are asymmetric around the observed effect size of 0.45 (ranging from 0.11 to 0.70) according
[263] to Blume et al. (2018) 58.11% of the data-supported hypotheses are null hypotheses, and
[264] therefore 58.11% of the data-supported hypotheses are compatible with the null premise.

[265] This example illustrates the difference between a proportion and a probability. As long
[266] as data are sampled from a bivariate normal distribution, there is always a 50% probability
[267] of observing a correlation smaller or larger than the true correlation, but the SGPV for this
[268] situation depends on how far away the observed correlation is from 0. The further away from
[269] 0, the larger the SGPV when the observed mean falls on the equivalence bound. The SGPV
[270] is the proportion of values in a 95% confidence interval that overlap with the equivalence
[271] range, but not the probability that these values will be observed. In the most extreme case
[272] (i.e., a sample size of 4, and equivalence bounds set to $r = $ -0.99 and 0.99, with a true
[273] correlation of 0.99) 97.60% of the confidence interval overlaps with the equivalence range,
[274] even though in the long run only 36% of the correlations observed in the future will fall in
[275] this range.

[276] It should be noted that in larger sample sizes the SGPV is closer to 0.5 whenever the
[277] observed correlation falls on the equivalence bound, but this extreme example nevertheless
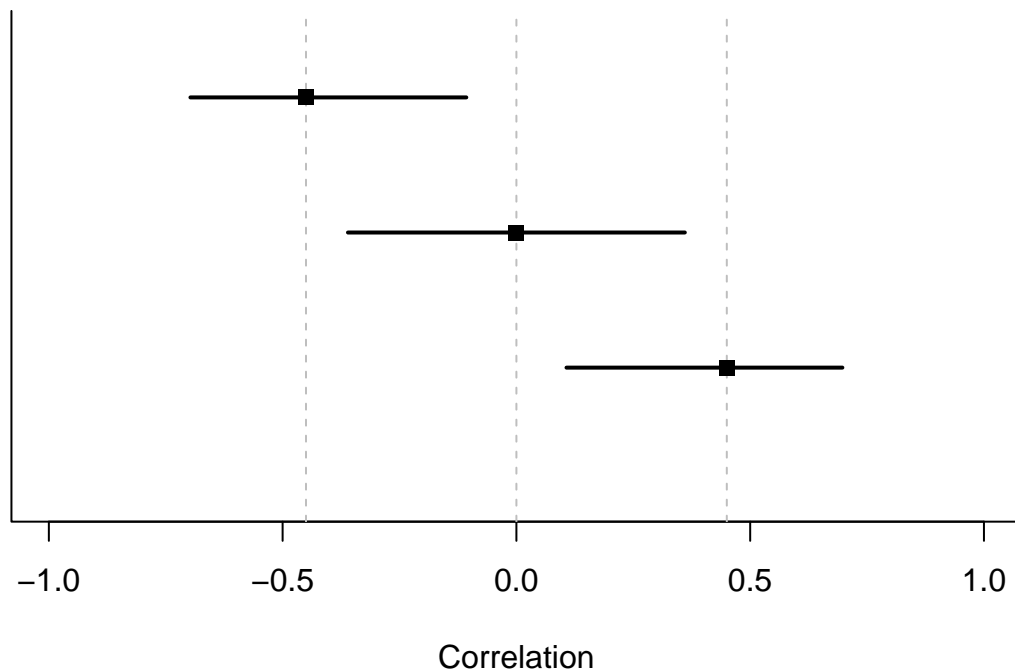[278] clearly illustrates the difference between question the SGPV answers, and the question a

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.

279   *p*-value answers. The conclusion of this section on asymmetric confidence intervals is that a

280   SGPV of 1 or 0 can still be interpreted as a $p < 0.025$ or $p > 0.975$ in an equivalence test,

281   since the SGPV and *p*-value for the TOST procedure are always directly related at the

282   values $p = 0.025$ and $p = 0.975$. Although Blume et al. (2018) state that "the degree of

283   overlap conveys how compatible the data are with the null premise" this definition of what

284   the SGPV provides does not hold for asymmetric confidence intervals. Although a SGPV of

285   1 or 0 can be directly interpreted, a SGPV between 0 and 1 is not interpretable as

286   "compatibility with the null hypothesis" under the assumption of a bivariate normal

287   distribution, and the generalizability of this statement needs to be examined beyond normal

288   bivariate distributions. Indeed, Blume and colleagues write in the supplemental material

₂₈₉ that "The magnitude of an inconclusive second-generation p-value can vary slightly when the

₂₉₀ effect size scale is transformed. However definitive findings, i.e. a p-value of 0 or 1 are *not*

₂₉₁ affected by the scale changes."

₂₉₂ What are the Relative Strengths and Weaknesses of Equivalence Testing and
₂₉₃ the SGPV?

₂₉₄   When introducing a new statistical method, it is important to compare it to existing

₂₉₅ approaches and specify its relative strengths and weaknesses. Here, we aimed to compare the

₂₉₆ SGPV against equivalence tests based on the TOST procedure. First of all, even though a

₂₉₇ SGPV of 1 or 0 has a clear interpretation (we can reject effects outside or inside the

₂₉₈ equivalence range), intermediate values are not as easy to interpret (especially for effects that

₂₉₉ have asymmetric confidence intervals). In one sense, they are what they are (the proportion

₃₀₀ of overlap), but it can be unclear what this number tells us about the data we have collected.

₃₀₁ This is not too problematic, since the main use of the SGPV (e.g., in all examples provided

₃₀₂ by Blume and colleagues) seems to be to examine whether the SGPV is 0, 1, or inconclusive.

₃₀₃ As already mentioned, this interpretation of a SGPV is very similar to the Neyman-Pearson

₃₀₄ interpretation of an equivalence test and a minimum effect tests (which are complementary).

₃₀₅ The difference is that where a SGPV of 1 can be interpreted as $p < .025$, equivalence tests

₃₀₆ provide exact p-values, and they continue to differentiate between for example $p = 0.024$ and

₃₀₇ $p = 0.002$. Whether this is desirable depends on the perspective that is used. From a

₃₀₈ Neyman-Pearson perspective on statistical inferences the main conclusion is based on

₃₀₉ whether or not $p < \alpha$, and thus an equivalence test and SGPV can be performed by simply

₃₁₀ checking whether the confidence interval falls within the equivalence range, just as a

₃₁₁ null-hypothesis test can be performed by checking whether the confidence interval contains

₃₁₂ zero or not. At the same time, it is recommended to report exact p-values (American

₃₁₃ Psychological Association, 2010), and exact p-values might provide information of interest to

314 readers about how precisely how surprising the data, or more extreme data, is under the null

315 model. Some researchers might be interested in combining an equivalence tests with a

316 null-hypothesis significance test. This allows a researcher to ask whether there is an effect

317 that is statistically different from zero, and whether effect sizes that are considered

318 meaningful can be rejected. Equivalence tests combined with null-hypothesis tests classify

319 results into four possible categories, and for example allow researchers to conclude an effect

320 is significant *and* equivalent (i.e., statistically different from zero, but also too small to be

321 considered meaningful, see Lakens et al. (2018)).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

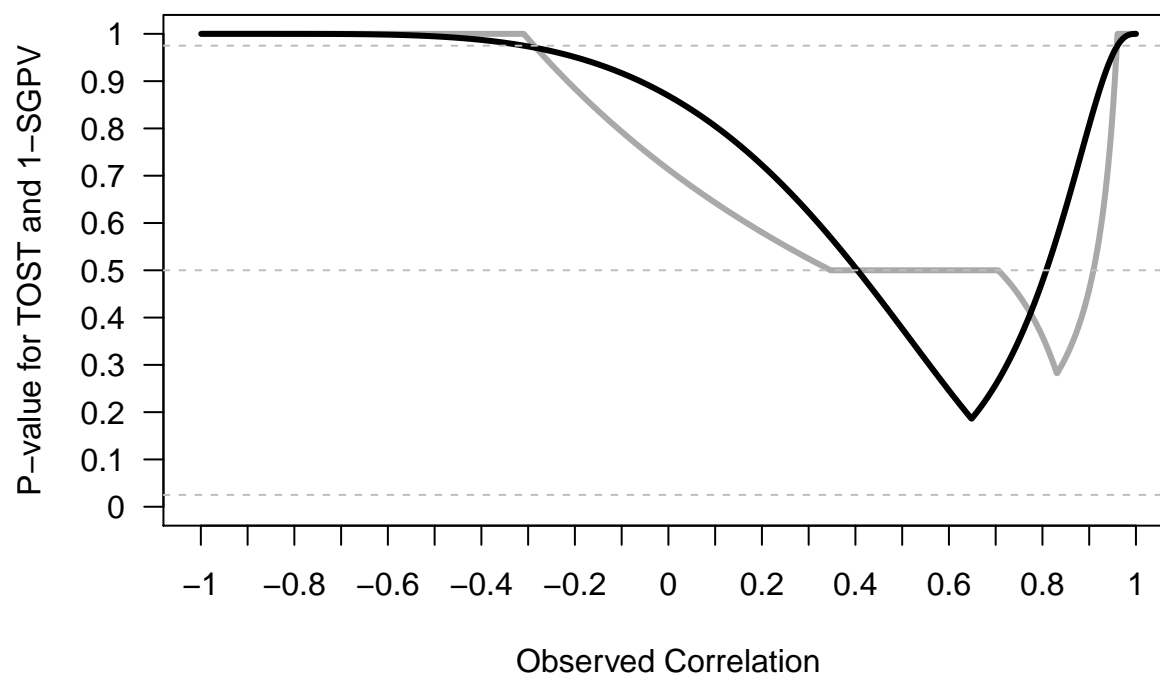339 equivalence range from r = 0.4 to r = 0.8. We can see the SGPV has a peculiar shape

Figure 13. Comparison of *p*-values from TOST (black line) and 1-SGPV (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.

because it is set to 0.5 for certain observed correlations, even though there is no risk of a "misleading" SGPV in this range. This example suggests that the current implementation of 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) claim that when using the SGPV "Adjustments for multiple comparisons are obviated" (p. 15). However, this is not correct. Given the direct relationship between TOST and SGPV highlighted in this manuscript (where a TOST $p = 0.025$ equals SGPV $= 1$, as long as the SGPV is calculated based on confidence intervals, and assuming

349 data are sampled from a continuous bivariate normal distribution), not correcting for

350 multiple comparisons will inflate the probability of concluding the absence of a meaningful

351 effect based on the SGPV in exactly the same way as it will for equivalence tests. Whenever

352 statistical tests are interpreted as support for a hypothesis (e.g., SPGV = 0 or SGPV = 1),

353 it is possible do so erroneously, and if researchers want to control error rates, they need to

354 correct for multiple comparisons.

Conclusion

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

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

358 statistical toolbox. The novelty of the SGPV can be limited when confidence intervals are

359 asymmetrical or wider than the equivalence range. There are strong similarities with

360 p-values from the TOST procedure, and in all situations where the statistics yield different

361 results, the behavior of the p-value from the TOST procedure is more consistent and easier

362 to interpret. We hope this overview of the relationship between the SGPV and equivalence

363 tests will help researchers to make an informed decision about which statistical approach

364 provides the best answer to their question. Our comparisons show that when proposing

365 alternatives to null-hypothesis tests, it is important to compare new proposals to already

366 existing procedures. We believe equivalence tests achieve the goals of the second generation

367 p-value while allowing users to easily control error rates, and while yielding more consistent

368 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

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*(3), 527–537. doi:10.1348/000711009X475853

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 A 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