<sup>1</sup> 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. We compare the behavior of the recently proposed

14   second generation $p$-value (Blume, McGowan, Dupont, & Greevy, 2018) with the

15   Two-One-Sided Tests equivalence testing procedure (Schuirmann, 1987). We show that the

16   two approaches yield almost identical results under optimal circumstances, but under

17   suboptimal circumstances (e.g., when the confidence interval is wider than the equivalence

18   range, or when confidence intervals are asymmetric) the second generation $p$-value becomes

19   difficult to interpret as a descriptive statistic. The second generation is interpretable in a

20   dichotomous manner when it is 0 or 1 (and lies completely within or outside of the

21   equivalence range), but these cases can be observed visually and do not require calculation.

22   Equivalence tests yield more consistent $p$-values, distinguish between datsets that yield the

23   same second generation $p$-value, and allow for easier control of Type I and Type II error

24   rates.

25          *Keywords:* equivalence testing, second generation $p$-values, $p$-values, hypothesis testing,

26   statistics

<sub>27</sub>                                    Equivalence Testing and the Second Generation P-Value

<sub>28</sub>        To test predictions most researchers predominantly rely on null-hypothesis tests. This

<sub>29</sub> statistical approach can be used to examine whether observed data is sufficiently surprising

<sub>30</sub> under the null hypothesis to reject an effect size of zero. Null-hypothesis tests have an

<sub>31</sub> important limitation, in that this procedure can only reject the hypothesis that there is no

<sub>32</sub> effect, while scientists should also be able to provide statistical support for *equivalence.*

<sub>33</sub> When testing for equivalence researchers aim to examine whether the effect is practically

<sub>34</sub> zero, or that there is no meaningful effect. By specifying a range around the null hypothesis

<sub>35</sub> of values that are deemed practically equivalent to the absence of an effect (i.e., $0 \pm 0.3$) the

<sub>36</sub> observed data can be compared against an *equivalence range* and researchers can test if a

<sub>37</sub> meaningful effect is absent (Hauck & Anderson, 1984; Kruschke, 2018; Rogers, Howard, &

<sub>38</sub> Vessey, 1993; Serlin & Lapsley, 1985; Spiegelhalter, Freedman, & Parmar, 1994; Wellek, 2010;

<sub>39</sub> Westlake, 1972).

<sub>40</sub>        Second generation $p$-values (SGPV) were recently proposed to as a descriptive statistic

<sub>41</sub> that represents "the proportion of data-supported hypotheses that are also null hypotheses"

<sub>42</sub> (Blume et al., 2018). The researcher specifies an equivalence range around a classical simple

<sub>43</sub> null hypothesis of parameter values that are considered practically equivalent to the null

<sub>44</sub> hypothesis. The SGPV measures the degree to which a set of data-supported parameter

<sub>45</sub> values falls within the interval null hypothesis. If the estimation interval falls completely

<sub>46</sub> within the equivalence range, the SGPV is 1. If the confidence interval falls completely

<sub>47</sub> outside of the equivalence range, the SGPV is 0. Otherwise the SGPV is a value between 0

<sub>48</sub> and 1 that expresses the overlap of data-supported hypotheses and the equivalence range.

<sub>49</sub> When calculating the SGPV the set of data-supported parameter values can be represented

<sub>50</sub> by a confidence interval (CI) or credible intervals or Likelihood support intervals (SI). When

<sub>51</sub> a confidence interval is used, the SGPV and equivalence tests such as the Two One-Sided

<sub>52</sub> Tests (TOST) procedure (Lakens, 2017; Meyners, 2012; Schuirmann, 1987) appear to have

<sub>53</sub> close ties. Here, we aim to examine the similarities and differences between the TOST

54   procedure and the SGPV.

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

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

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

58   & Isager, 2018). If both tests allow you to reject an effect as extreme or more extreme than

59   the equivalence bound, you can reject the presence of an effect large enough to be

60   meaningful, and conclude the observed effect is practically equivalent to zero. Because

61   one-sided tests are performed, one can also conclude equivalence by checking whether the

62   1-2*alpha confidence interval (e.g., when the alpha level is 0.05, a 90% CI) falls completely

63   within the equivalence bounds. Because both equivalence tests as the SGPV are based on

64   whether and how much a confidence interval overlaps with equivalence bounds, it seems

65   logical to compare the behavior of the newly proposed SGPV to equivalence tests.

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

67                         **intervals are symmetrical**

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

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

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

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

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

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

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

74   than twice as wide as the equivalence range.

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

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

77   procedure and the SGPV. The statistics are calculated for hypothetical one-sample $t$-tests

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

79 145 ± 2 (i.e., an equivalence range from 143 to 147), the observed standard deviation is

80 assumed to be 2, and the sample size is 100. For example, for the left-most point in 1 the

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

82 100, an observed standard deviation of 2, and an observed mean of 140, where the *p*-value for

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

84 TOST *p*-values and SGPV in this article are not dependent upon any specific example.
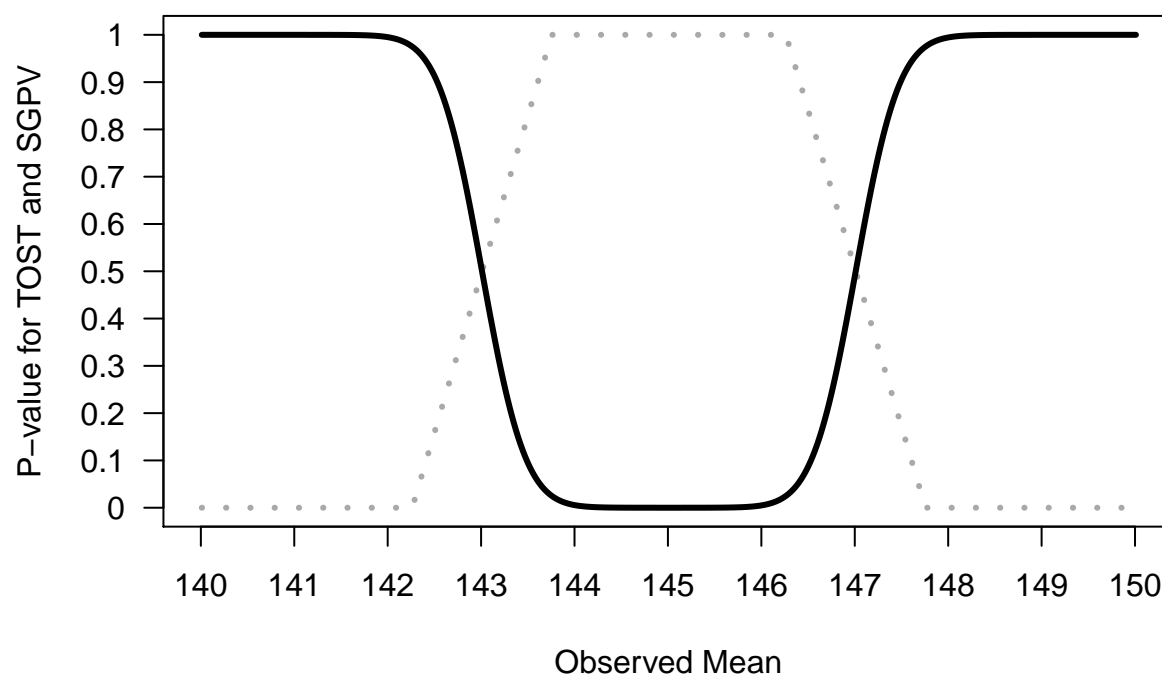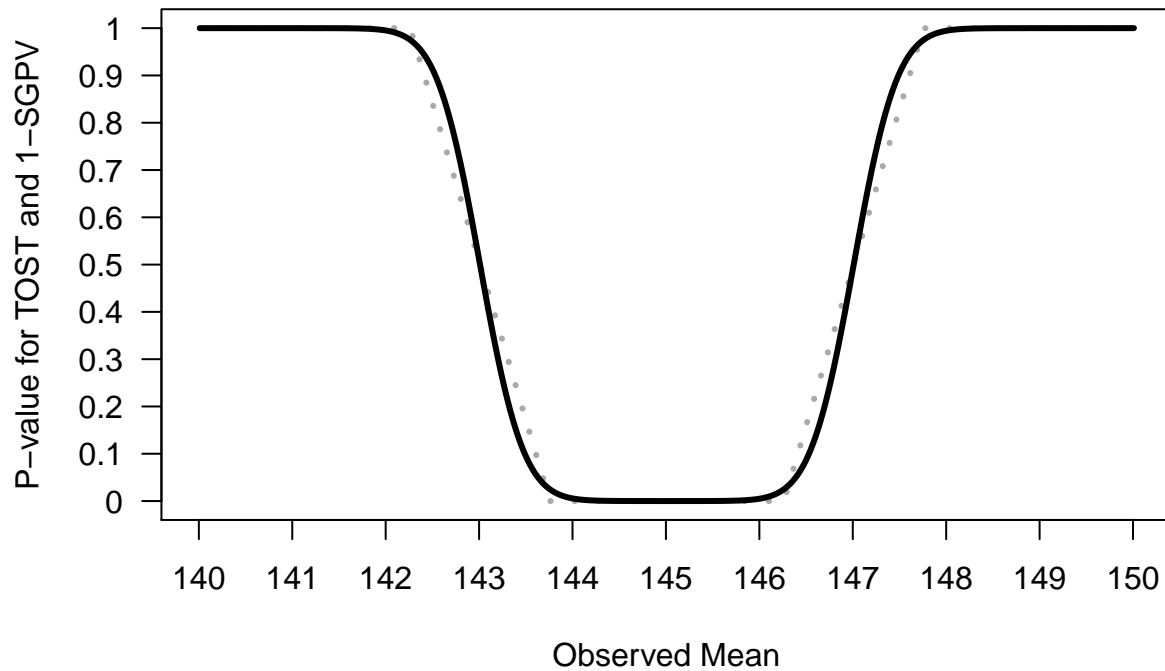


*Figure 1*. Comparison of *p*-values from TOST (black line) and SGPV (dotted grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2.

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

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

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

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

89  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.

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

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

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

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

94  Figure 1).

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

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

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

98  falls completely outside the equivalence bounds, the SGPV is 0 (see Figure 1).

**SGPV as a uniform measure of overlap**

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

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

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

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

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

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

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

107  0, the $t$-value for the equivalence test is also 0, and thus the $p$-value is 0.5 (situation A, 3).

108  Two other points always have to overlap. When the 95% CI falls completely (but only

109  just) inside the equivalence region, the TOST (which relies on a one-sided test) should be

110  significant at an alpha level of 0.025. When the SGPV changes from <1 to 1 the 95% CI

111  touches one of the equivalence bounds (see situation B in 3, where the 95% CI falls

112  completely inside the equivalence bounds) and the TOST $p$-value is always 0.025. The third

113  point where the SGPV and the $p$-value from the TOST procedure should overlap is where

114  the SGPV changes from a positive value to 0 (i.e., when the 95% CI falls completely outside

115  of the equivalence range, but touches one of the equivalence bounds, see situation C in 3).

116  When the 95% CI touches the outside of the equivalence bound and the TOST $p$-value will

117  be 0.975.

118  The confidence interval width is uniformly distributed across the mean differences. As

119  the observed mean in a one-sample $t$-test gets closer to the test value (in the plot below,

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

121  overlap is stable.

122  For example, the SGPV from A to D is 0.76, 0.81, 0.86, and 0.91. The difference in the
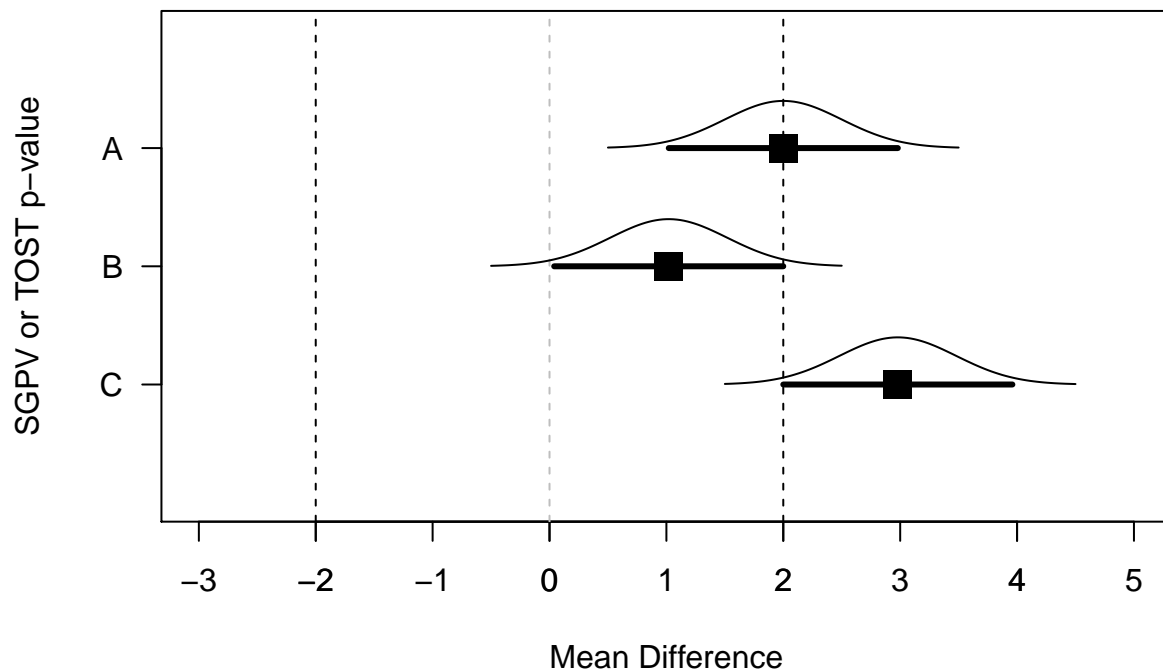
*Figure 3*. Means, normal distribution, and 95% CI for three example datasets that illustrate the relationship between *p*-values from TOST and SGPV.

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

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

125         As we move the observed mean closer to the test value in steps of 0.1 across A to D the

126  *p*-value calculated for normally distributed data is not uniformly distributed. The

127  probability of observing data more extreme than the upper bound of 2 is (from A to D) 0.16,

128  0.12, 0.08, and 0.06. As we can see, the difference between A and B (0.04) is not the same as

129  the difference between C And D (0.03). Indeed, the difference in *p*-values is the largest as

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

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

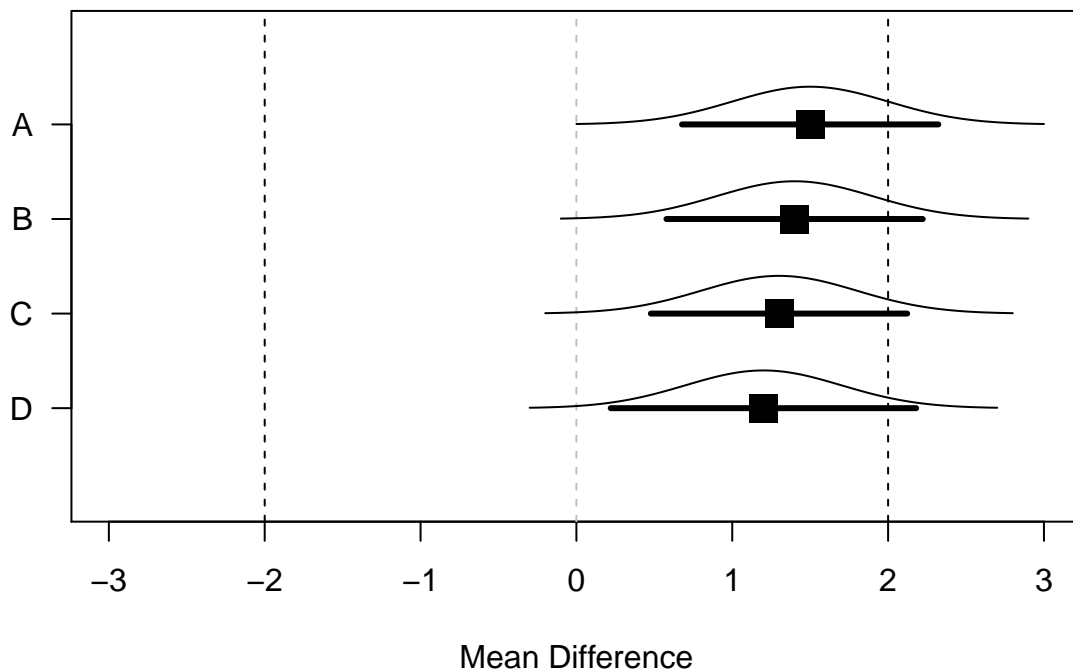132  approximate 0 and 1, but never reach these values. ##When are the SGPV and Equivalence

*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.

Test Unrelated? There are 4 situations where $p$-values from TOST and SGPV are unrelated. The first two situations were discussed earlier, and can be seen in 1. When the SGPV is either 0 or 1 $p$-values from the equivalence test fall between 0.975 and 1 or between 0 and 0.025. Because $p$-values approach 0 or 1, but are never exactly 0 or 1, while the SGPV is exactly 0 or 1, the two statistics are completely unrelated. The easiest way to see this is by plotting the SGPV against the $p$-value from the TOST procedure. The situations where the SPGV and $p$-values from the TOST procedure are unrelated are indicated by the parts of the curve where there are vertical lines at SGPV of 0 and 1.

A third situation in which the SGPV deviates strongly from the TOST $p$-value is whenever the CI is more than twice as wide as the equivalence range, and the CI overlaps
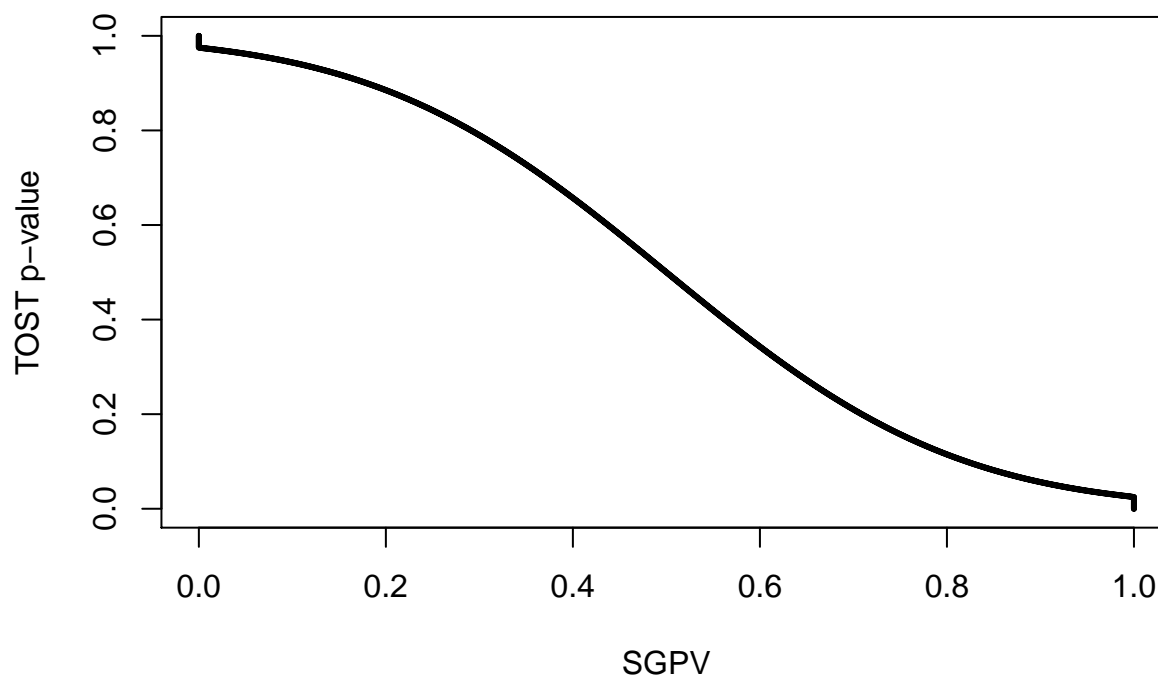
*Figure 5*. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 1.

143 with the upper *and* lower equivalence bound. In this situation the normal calculation of the

144 proportion of overlap is skipped, and the SGPV is set to 0.5 instead. Blume et al. (2018)

145 call this the "small sample correction factor". However, it is not a correction in the typical

146 sense of the word, since the SGPV is not adjusted to any "correct" value. When the normal

147 calculation would be "misleading" (i.e., the SGPV would be small, which normally would

148 suggest support for the alternative hypothesis, when all values in the equivalence range are

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

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

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

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

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

154 interval width to the equivalence range exceeds 2:1 and the CI overlaps with the upper and

155 lower bounds.

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

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

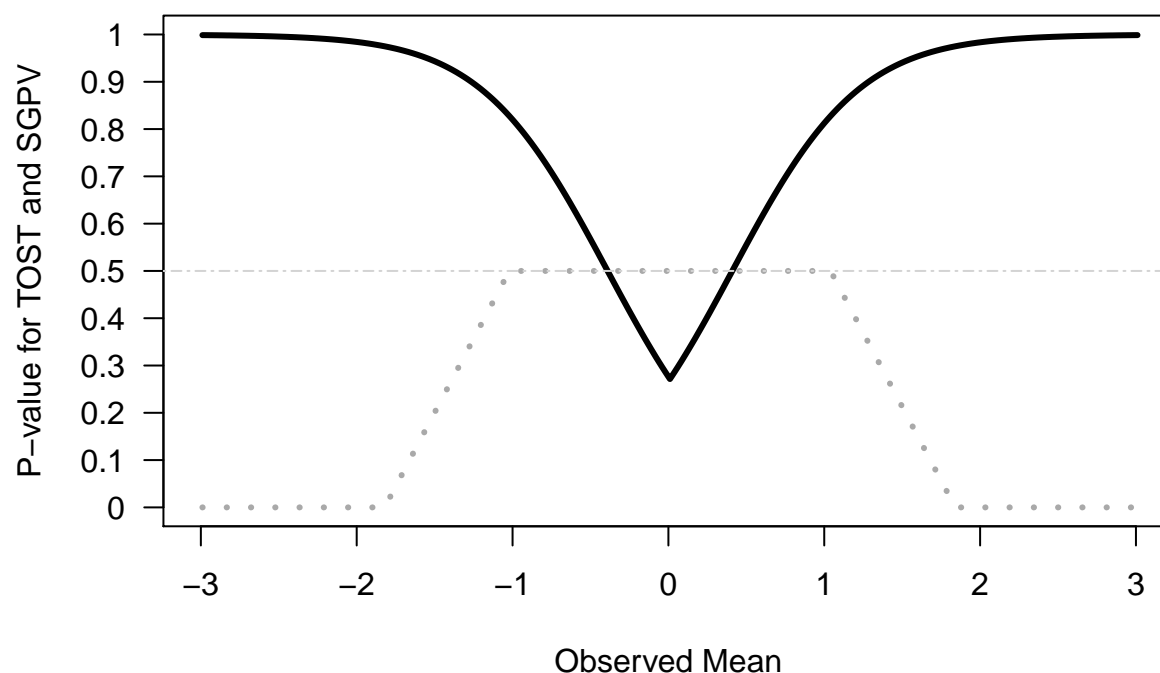158 CI is more than twice as large as the equivalence range.



*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.

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

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

observed means where the $p$-value from the equivalence test still varies. It should be noted that in these calculations the $p$-values for the TOST procedure are *never* smaller than 0.05 (i.e., they do not get below 0.05 on the y-axis). In other words, we cannot conclude equivalence based on any of the observed means. How is this possible? Remember that the TOST procedure consists of two one-sided tests against the upper and lower equivalence bound. The TOST $p$-value is smaller than 0.05 if the 90% CI falls completely between the upper and lower equivalence bounds. However, we are examining a scenario where the 90% CI is so wide that it never falls completely within the two equivalence bounds. As Lakens (2017) notes: "in small samples (where CIs are wide), a study might have no statistical power (i.e., the CI will always be so wide that it is necessarily wider than the equivalence bounds)." None of the $p$-values based on the TOST procedure are below 0.05, and thus, in the long run we have 0% power.

There is one last situation where the $p$-value from the TOST procedure and the SGPV are unrelated. This is when the CI is wider than the equivalence range (so the precision is low) and overlaps with the upper and lower equivalence bound, but the CI is *not* twice as wide as the equivalence range. This fourth category exists because of the decision by Blume and colleagues to set the SGPV to 0.5 whenever the CI is twice as wide as the equivalence range and the CI overlaps with both equivalence bounds. This means that there are situations where the CI interval overlaps with both equivalence bounds, while the CI is less than twice as large as the equivalence bound. For example, in the example below, we see that the CI is only 1.79 times as wide as the equivalence bounds, but the CI overlaps with the lower and upper equivalence bounds.

If the observed mean would be somewhat closer to 0, or further away from 0, the SGPV would remain constant (the CI width does not change, it completely overlaps with the equivalence range) while the $p$-value for the TOST procedure can vary between 0 and .025. We can see this in 9 below. The SGPV is not set to 0.5, but is slightly higher than 0.5 across a range of means. How high the SGPV will be when the CI overlaps with the lower and
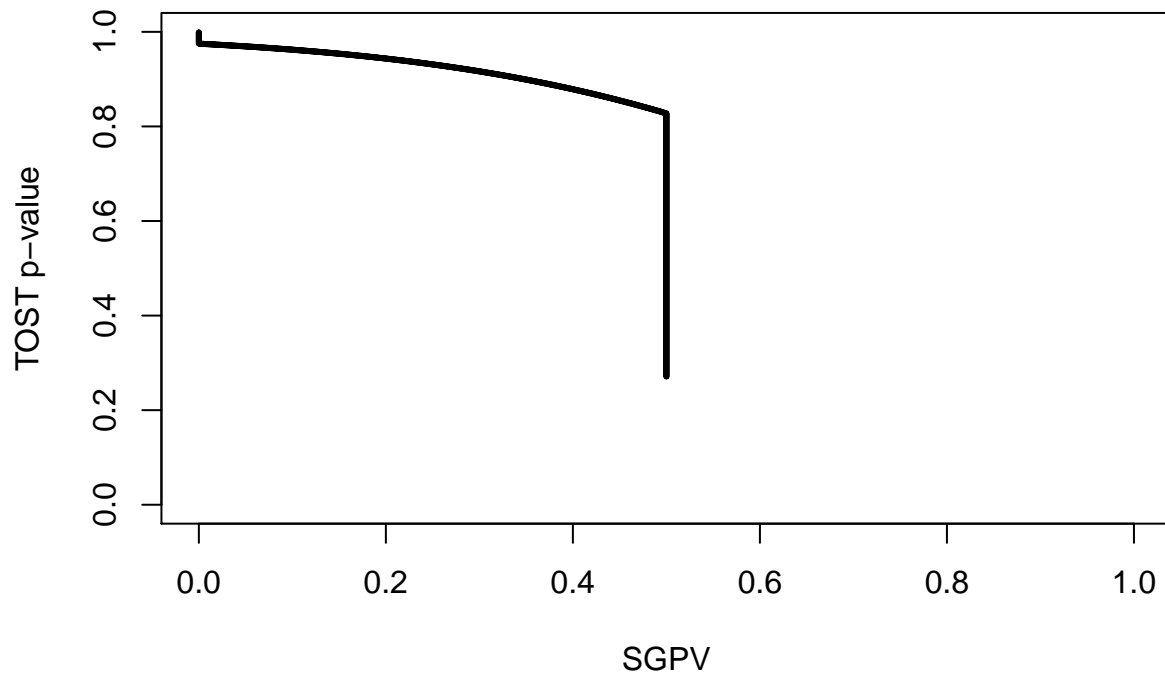
*Figure 7.* The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 6.

188 upper equivalence bounds, but the CI is not twice as large as the equivalence range, depends

189 on the width of the CI and the equivalence range.

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

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

192 observed means where the *p*-value from the equivalence test still varies.

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

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

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

196 descriptive statistic, the SGPV is more limited than the value from the TOST procedure.

197 The proportion of overlap can be the same value when the observed mean is 0 or when the
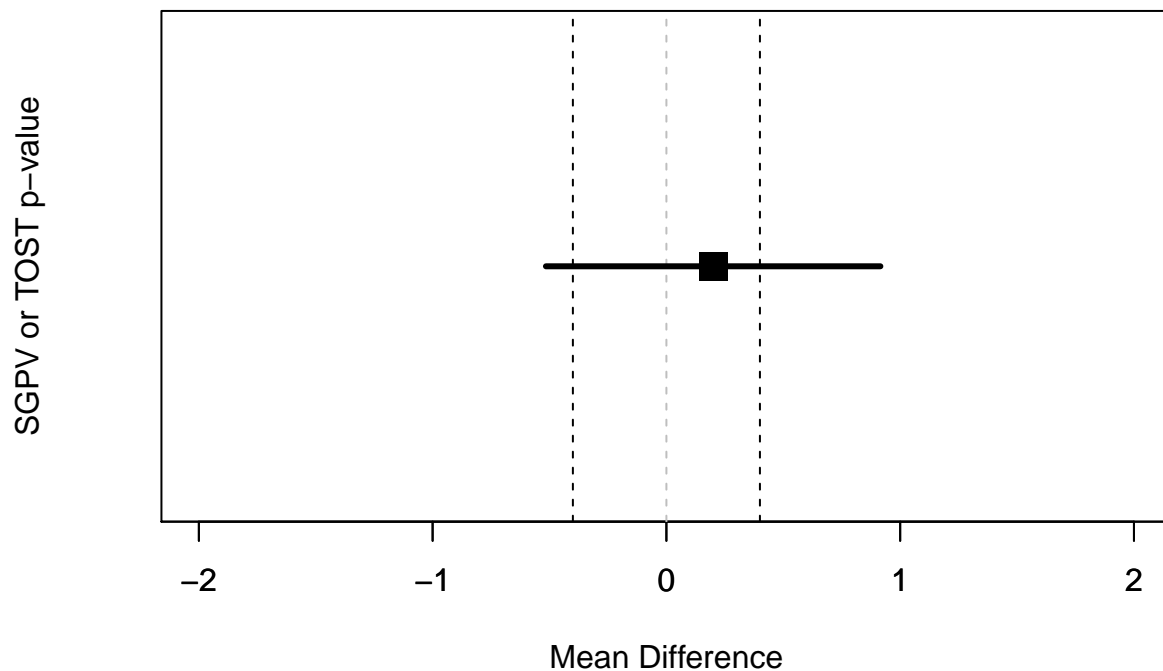
*Figure 8*. Example of a 95% CI that overlaps with the lower and upper equivalence bound (indicated by the vertical dotted lines).

198   observed mean falls just inside the equivalence bound, and additional information (e.g., the

199   95% CI) is required to differentiate these situations. One way to mitigate this limitation of

200   the SGPV would be to set the SGPV to 0.5 whenever the CI overlaps with both the upper

201   and lower equivalence bound (irrespective of the width of the CI).

202   **The relation between equivalence tests and SGPV when confidence intervals**

203   **are not symmetrical**

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

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

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

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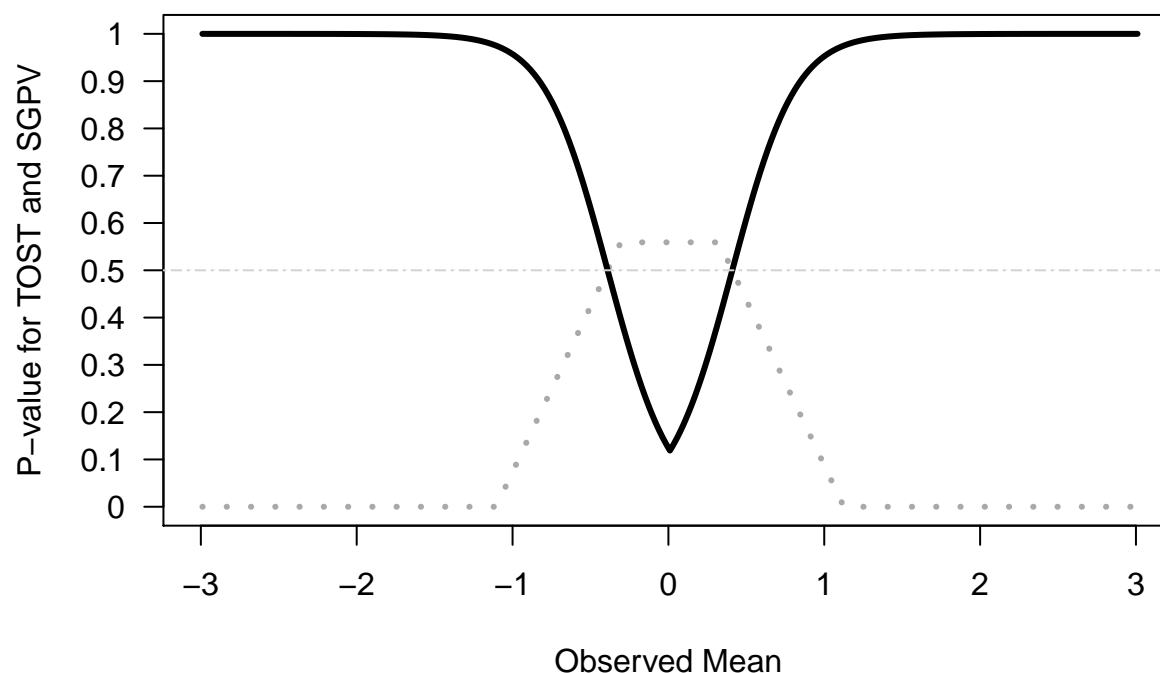
*Figure 9*. Comparison of *p*-values from TOST (black line) and SGPV (dotted grey line) across a range of observed sample means (x-axis). The sample size is small (n = 10), but because the sd is half as big as in Figure 7 (1 instead of 2) the CI is less than twice as wide as the equivalence range (set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

208  asymmetric as the observed correlation nears -1 or 1. For example, with ten observations, an

209  observed correlation of 0 has a symmetric 95% confidence interval ranging from -0.629 to

210  0.629, while and observed correlation of 0.7 has an asymmetric 95% confidence interval

211  ranging from 0.126 to 0.992.

212        The effect of assymetric confidence intervals is most easily noticable at smaller sample

213  sizes, therefore in 11 below we plot the *p*-values from equivalence tests and the SGPV (again

214  plotted as 1-SGPV for ease of comparison) for correlations. The sample size is 30 pairs of
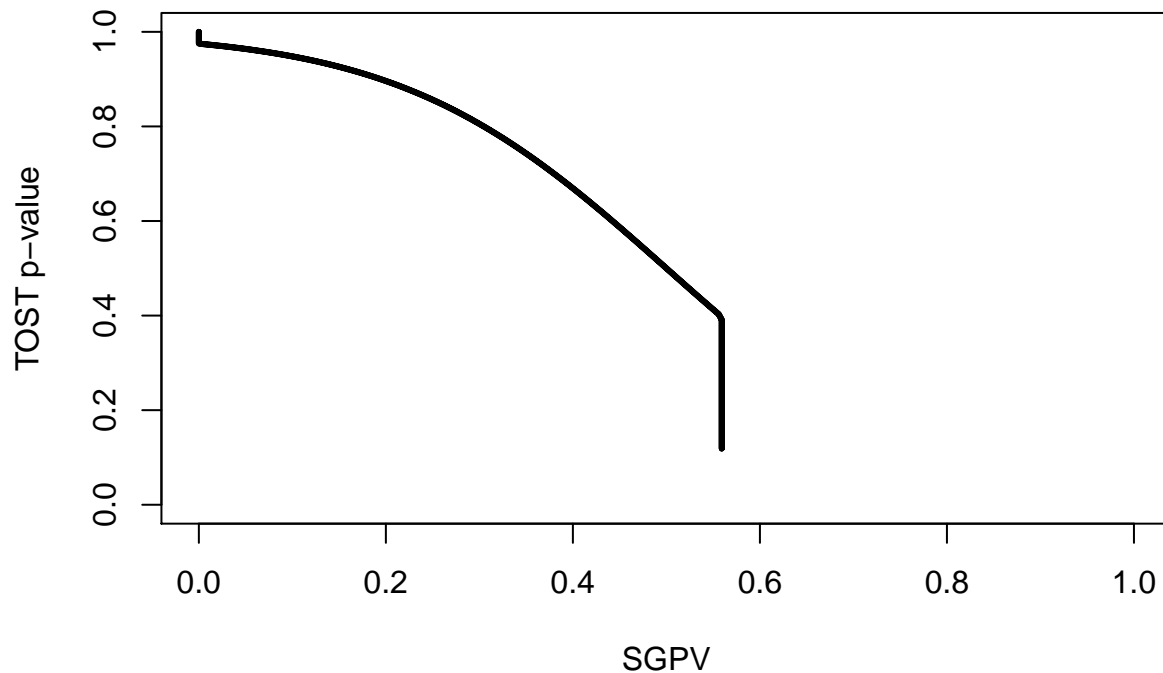
*Figure 10*. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 9.

215  observations, and the lower and upper equivalence bounds are set to -0.45 and 0.45, with an

216  alpha of 0.05. As the observed correlation in the sample moves from -1 to 0 the *p*-value from

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

218  in 2. The *p*-value for the TOST procedure and 1-SGPV are still identical when *p*-values are

219  0.975 and 0.025 (indicated by the upper and lower horizontal dotted lines). There are two

220  important differences, however. First of all, the SGPV is no longer a straight line, but a

221  curve, due to the asymmetry in the 95% CI. Second, and most importantly, the *p*-value for

222  the equivalence test and the SGPV do no longer overlap at $p = 0.5$.

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

224  asymmetric confidence intervals. If the observed correlation falls exactly on the equivalence
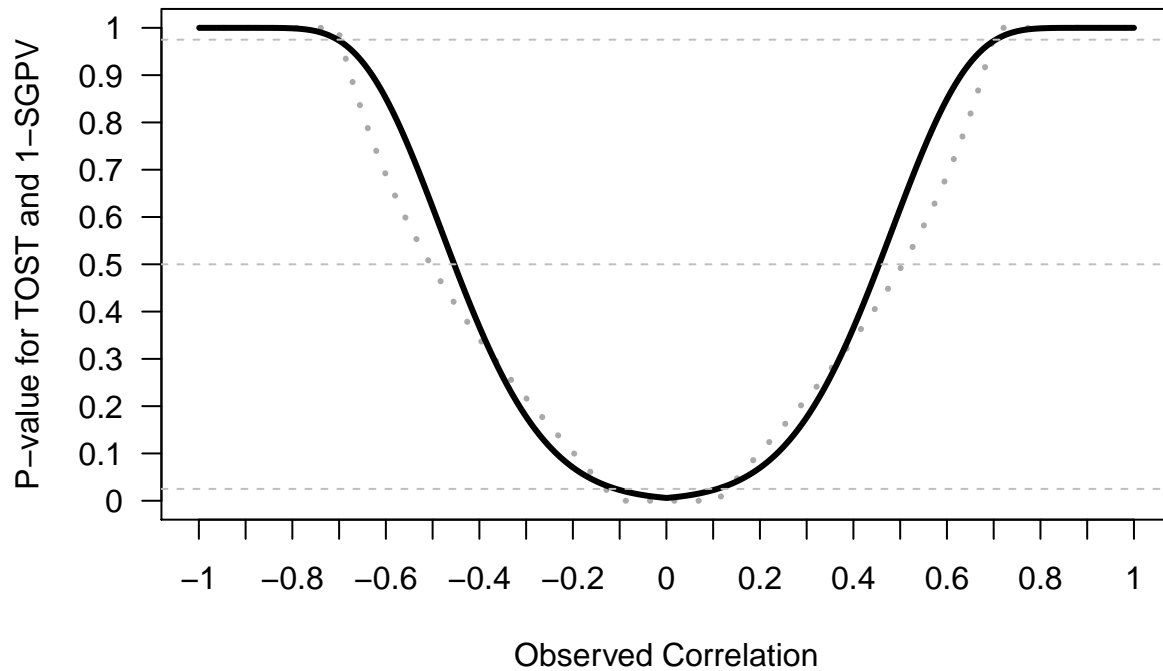
*Figure 11*. Comparison of *p*-values from TOST (black line) and 1-SGPV (dotted grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = -0.45 and r = 0.45 with n = 30 and an alpha of 0.05.

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

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

227 50%. In other words, if the true effect size is the same as the equivalence bound, it is equally

228 likely to find an effect more extreme than the equivalence bound, as it is to observe an effect

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

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

231 0.45 are 0.58. Because the confidence intervals are asymmetric around the observed effect

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

233 data-supported hypotheses are null hypotheses, and therefore 58.10% of the data-supported
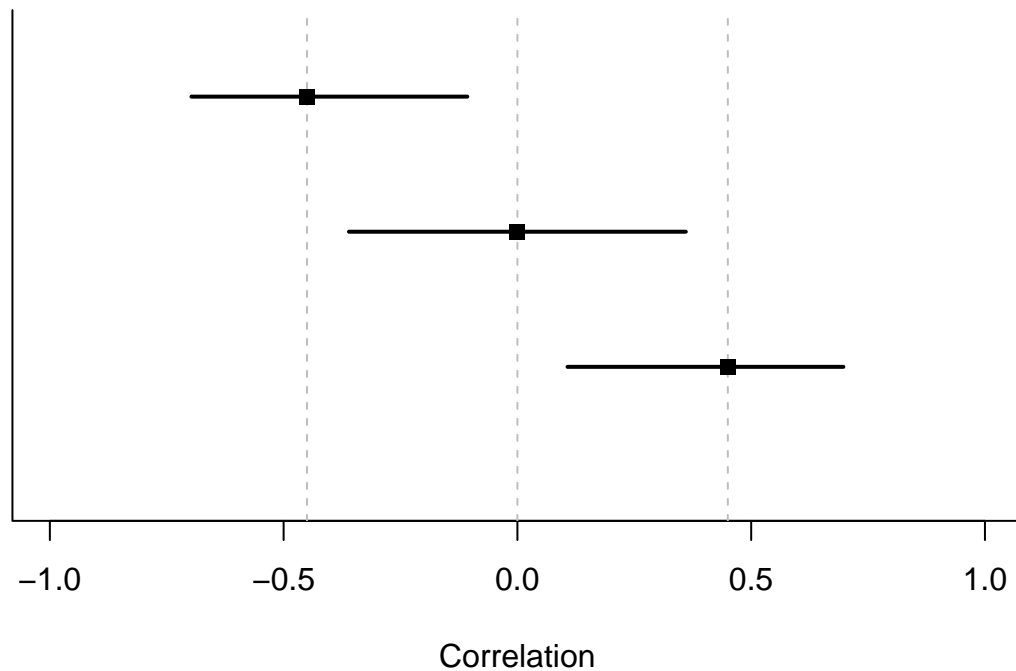
234   hypotheses are compatible with the null premise.



*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.

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

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

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

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

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

240   confidence interval that overlap with the equivalence range, but not the probability that

241   these values will be observed. In the most extreme case (i.e., a sample size of 4, and

242   equivalence bounds set to r = -0.99 and 0.99, with an observed correlation of 0.99) 97.60% of

243   the confidence interval overlaps with the equivalence range, even though in the long run only

50% of the correlations observed in the future will fall in this range. It should be noted that in larger sample sizes the SGPV is closer to 0.5 whenever the observed correlation falls on the equivalence bound, but this extreme example nevertheless clearly illustrates the difference between two different questions the SGPV and a *p*-value are answers to. The conclusion of this in depth look at asymmetric confidence intervals is that a SGPV of 1 or 0 can still be interpreted the same way as a *p*-value of 0.025 and 0.975 can be interpreted in an equivalence test, since the SGPV and *p*-value for the TOST procedure are always directly related at these values. Although Blume et al. (2018) state that "the degree of overlap conveys how compatible the data are with the null premise" this definition of what the SGPV provides does not hold for asymmetric confidence intervals. Although a SGPV of 1 or 0 can be directly interpreted, a SGPV between 0 and 1 is not interpretable as "compatibility with the null hypothesis". Indeed, Blume and colleagues write in the supplemental material that "The magnitude of an inconclusive second-generation *p*-value can vary slightly when the effect size scale is transformed. However definitive findings, i.e. a *p*-value of 0 or 1 are *not* affected by the scale changes."

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

Given the strong relationship between SGPV and equivalence testing, a logical question is to ask what the introduction of SGPV adds to the existing statistical approaches, including equivalence tests, and what the relative strengths and weaknesses of either approach are. First of all, SGPV is a descriptive statistic (unlike the *p*-value that is calculated for an equivalence test, which is an inferential statistic). It numerically summarizes the information that is visually present in a plot of the equivalence range and the confidence interval around the observed effect (for example, see 3).

A possible weakness of the SGPV is that even though a SGPV of 1 or 0 has a clear interpretation (we can reject effects outside or inside the equivalence range), intermediate

270 values are not as easy to interpret (especially for effects that have asymmetric confidence

271 intervals). This is not too problematic, since the main use of the SGPV (e.g., in all examples

272 provided by Blume and colleagues) is to examine whether the SGPV is 0 or 1, or whether

273 the SGPV is inconclusive. This interpretation of a SGPV as allowing researchers to reject

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

275 to the Neyman-Pearson interpretation of combining a null-hypothesis test and an equivalence

276 test (Lakens et al. (2018)). The difference is that where a SGPV of 1 can be interpreted as $p$

277 $< .05$, and equivalence test provides an exact $p$-value, and continues to differentiate between

278 for example p = 0.048 and p = 0.002. $P$-values continue to differentiate for situations where

279 the SGPV are identical. For example, different equivalence tests with $p = 0.048$ and $p =$

280 0.002 have a SGPV of 1. Whether this is desireable depends on the perspecitve that is used.

281 From a Neyman-Pearson perspective on statistical inferences the main conclusion is based on

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

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

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

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

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

287 readers about how surprising the data is under the null model. Equivalence tests combined

288 with null-hypothesis significance tests also allow researchers to conclude an effect is

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

290 considered meaningful). Thus, the SGPV is used to classify results into one of three possible

291 outcomes (with the data falling inside or outside the equivalence range, or being

292 inconclusive), while equivalence tests combined with null-hypothesis tests classify results into

293 four possible outcomes.

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

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

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

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

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

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

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

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

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

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

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

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

306 continues to differentiate (see for example 6). An important benefit of equivalence tests is

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

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

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

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

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

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

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

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

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

316 "small sample correction".

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

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

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

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

321 SGPV has a *lower* error rate, not a *more accurate* error rate. In a Neyman-Pearson

322 perspective (which forms the basis of equivalence tests) the goal is not to end up with an

323 error rate that is as low as possible (to achieve that, one simply adjusts the alpha level), but

<sup>324</sup> with a decision procedure that, when applied, yields a desired error rate with high accuracy.

<sup>325</sup> Traditional frequentist tests such as equivalence tests, null-hypothesis tests, and minimal

<sup>326</sup> effect tests (which have the same goal as the SGPV to test whether a confidence interval falls

<sup>327</sup> completely outside the equivalence range) make it straightforward to control the Type 1

<sup>328</sup> error rate, as well as the Type II error rate by performing an a-priori power analysis. Note

<sup>329</sup> that the TOST procedure uses a 90% confidence interval (because it is based on two

<sup>330</sup> one-sided tests, the confidence interval is 1-2*$\alpha$, but because both tests need to be significant

<sup>331</sup> to conclude equivalence, no adjustment for multiple comparisons is needed). Therefore, the

<sup>332</sup> TOST procedure always has higher power to declare equivalence than the SGPV, which

<sup>333</sup> relies on a 95% confidence interval.

<sup>334</sup>       Blume et al. (2018) claim that "Adjustments for multiple comparisons are obviated"

<sup>335</sup> (p. 15) and that "second-generation p-values provide a proper scientific adjustment for

<sup>336</sup> multiple comparisons". However, this is not correct. Even when for a single test the

<sup>337</sup> probability of concluding there is a meaningful effect (where the confidence interval falls

<sup>338</sup> outside of the equivalence range) is lower than the probability that there is any effect, with a

<sup>339</sup> sufficient number of looks and narrow confidence intervals error rates can rise above the

<sup>340</sup> nominal alpha level. More importantly, multiple comparisons inflate the probability that one

<sup>341</sup> erronously concludes there is *no* effect, where there *is* a true effect size that equals the

<sup>342</sup> equivalence bound, just as quickly for the SGPV as for null-hypothesis testing. To conclude,

<sup>343</sup> the idea that the SGPV improves error rates does not hold up under closer scrutiny, and the

<sup>344</sup> recommendation to ignore adjustments for multiple comparisons has the potential to increase

<sup>345</sup> false positives in the literature. Equivalence tests provide and easier and more formal way to

<sup>346</sup> control error rates.

<sup>347</sup>                                    **Conclusion**

<sup>348</sup>       We believe that our explanation of the similarities between the TOST procedure and

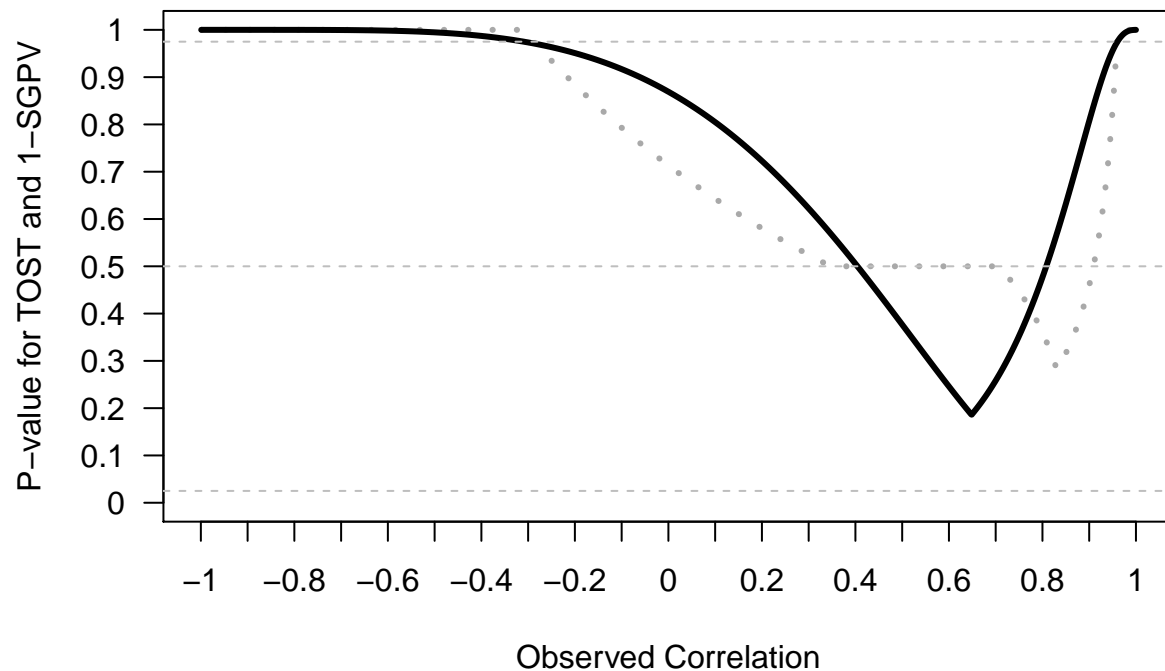<sup>349</sup> the SGPV provides some useful context to interpret the contribution of second generation

*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.

350  *p*-values to the statistical toolbox. The novelty lies in its use as a descriptive statistic, but

351  this use can be limited when confidence intervals are asymmetrical, or wider than the

352  equivalence range. There are strong similarities with *p*-values from the TOST procedure, and

353  in all situations where the statistics yield different results, the behavior of the *p*-value from

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

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

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

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

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

359   equivalence tests achieve the goals of the second generation $p$-value while allowing users to

360   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*

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

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

389      good-enough principle.

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

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

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

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

394      Boca Raton: CRC Press.

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

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

397      doi:10.1002/JPS.2600610845