Meta-Psychology, 2020, vol 4, MP.2018.933
https://doi.org/10.15626/MP.2018.933
Article type: Original Article
Published under the CC-BY4.0 license

Open data: N/A
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: N/A

Edited by: S. R. Martin
Reviewed by: J. D. Blume, O. L. Olvera Astivia
Analysis reproduced by: André Kalmendal
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/ZP3KF

Equivalence Testing and the Second Generation P-Value.

Daniël Lakens
Eindhoven University of Technology, The Netherlands

Marie Delacre
Université Libre de Bruxelles, Belgium

Abstract

To move beyond the limitations of null-hypothesis tests, statistical approaches have been developed where the observed data are compared against a range of values that are equivalent to the absence of a meaningful effect. Specifying a range of values around zero allows researchers to statistically reject the presence of effects large enough to matter, and prevents practically insignificant effects from being interpreted as a statistically significant difference. We compare the behavior of the recently proposed second generation *p*-value (Blume, D'Agostino McGowan, Dupont, & Greevy, 2018) with the more established Two One-Sided Tests (TOST) equivalence testing procedure (Schuirmann, 1987). We show that the two approaches yield almost identical results under optimal conditions. Under suboptimal conditions (e.g., when the confidence interval is wider than the equivalence range, or when confidence intervals are asymmetric) the second generation *p*-value becomes difficult to interpret. The second generation *p*-value is interpretable in a dichotomous manner (i.e., when the SGPV equals 0 or 1 because the confidence intervals lies completely within or outside of the equivalence range), but this dichotomous interpretation does not require calculations. We conclude that equivalence tests yield more consistent *p*-values, distinguish between datasets that yield the same second generation *p*-value, and allow for easier control of Type I and Type II error rates.

Keywords: equivalence testing, second generation p-values, hypothesis testing, TOST, statistical inference

To test predictions researchers predominantly rely on null-hypothesis tests. This statistical approach can be used to examine whether observed data are sufficiently surprising under the null hypothesis to reject an effect that equals exactly zero. Null-hypothesis tests have an important limitation, in that this procedure can only reject the hypothesis that there is no effect, while scientists should also be able to provide statistical support for *equivalence*. When testing for equivalence researchers aim to examine whether an observed effect is too small to be considered meaningful, and therefore is practically equivalent to zero. By specifying a range around the null hypothesis of values that are deemed practically equivalent to the absence of an effect (i.e., 0 +- 0.3) the observed data can be compared against an *equivalence range* and researchers can test if a meaningful effect is absent (Hauck & Anderson, 1984; Kruschke, 2018; Rogers, Howard, & Vessey, 1993; Serlin & Lapsley, 1985; Spiegelhalter, Freedman, & Parmar, 1994; Wellek, 2010; Westlake, 1972).

Second generation *p*-values (SGPV) were recently proposed as a statistic that represents "the proportion of