Journal of Psychiatric Research 169 (2024) 38–41

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/jpsychires

Redefining effect size interpretations for psychotherapy RCTs in depression

Anders Nordahl-Hansen [a], Hugo Cogo-Moreira [a], Sareh Panjeh [b], Daniel S. Quintana [c,d,e,f,*]

[a] Department of Education, ICT and Learning, Østfold University College, Halden, Norway
[b] Department of Psychiatry and Medical Psychology, Federal University of São Paulo, São Paulo, Brazil
[c] Department of Psychology, University of Oslo, Oslo, Norway
[d] NevSom, Department of Rare Disorders, Oslo University Hospital, Oslo, Norway
[e] KG Jebsen Centre for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway
[f] NORMENT Centre for Psychosis Research, Division of Mental Health and Addiction, University of Oslo and Oslo University Hospital, Oslo, Norway

ARTICLE INFO

Keywords:
Effect size distribution
Treatment
Depression
Psychotherapy

ABSTRACT

Introduction: Effect sizes are often used to interpret the magnitude of a result and in power calculations when planning research studies. However, as effect size interpretations are context-dependent, Jacob Cohen's suggested guidelines for what represents a small, medium, and large effect are unlikely to be suitable for a diverse range of research populations and interventions. Our objective here is to determine empirically-derived effect size thresholds associated with psychotherapy randomized controlled trials (RCTs) in depression by calculating the effect size distribution.

Methods: We extracted effect sizes from 366 RCTs provided by the systematic review of Cuijpers and colleagues (2020) on psychotherapy for depressive disorders across all age groups. The 50th percentile effect size, as this represents a medium effect size, and the 25th (small) and 75th (large) percentile effect sizes were calculated to determine empirically-derived effect size thresholds.

Results: After adjusting for publication bias, 0.27, 0.53, and 0.86 represent small, medium, and large effect sizes, respectively, for psychotherapy treatment for depressive disorders.

Discussion: The effect size distribution for psychotherapy treatment of depression indicates that observed effect size thresholds are larger than Cohen's suggested effect size thresholds (0.2, 0.5, and 0.8). These results have implications for the interpretation of study effects and the planning of future studies via power analyses, which often use effect size thresholds.

1. Introduction

Cohen's *d* effect sizes are typically used to indicate the magnitude of group differences in clinical research, which are often classified as small, medium, or large effects (Wasserstein and Lazar, 2016). Jacob Cohen proposed that a medium-sized effect should represent the average effect size within a field, with small and large effects to be equidistant from this medium-sized effect (Cohen, 2013). Accordingly, small, medium, and large effects have been associated with the 25th, 50th, and 75th percentile of effect sizes for a research field. Cohen suggested that *d* values of 0.2, 0.5, and 0.8 can be used to represent small, medium, and large effects, respectively. But despite the wide use of these effect size thresholds, Cohen's intention was for them to serve as a fallback option when effect size percentiles are unknown (Cohen, 2013; Glass et al., 1981; Thompson, 2009) and he subsequently expressed regret suggesting these thresholds in the first place (Funder and Ozer, 2019) . As the distribution of effect sizes vary from field-to-field, relying on Cohen's thresholds risks under- or over-estimating an effect size distribution (ESD) that represents the published literature for a given field (e.g., Quintana, 2017; Cherubini and MacDonald, 2021; Panjeh et al., 2023a; Panjeh et al., 2023b; Szucs and Ioannidis, 2017).

In terms of study planning, using Cohen's thresholds instead of a *larger* empirically-derived effect size threshold would lead to sample sizes that are larger than required, which can subject more people than necessary to study intervention risks and can lead to a misuse of resources that could be directed elsewhere. For example, a recent ESD analysis in the field of endothelial function reported a small effect size threshold of $d = 0.28$ (Cherubini and MacDonald, 2021). If one were to plan a future independent samples study using a *d* of 0.28 in a power analysis, a total of 404 participants would be required (assuming the

* Corresponding author. Department of Psychology, University of Oslo, Norway.
E-mail address: daniel.quintana@psykologi.uio.no (D.S. Quintana).

https://doi.org/10.1016/j.jpsychires.2023.11.009
Received 6 January 2023; Received in revised form 27 October 2023; Accepted 15 November 2023
Available online 18 November 2023

input parameters of a two-tailed *t*-test, alpha of 0.05, power of 0.8, and a group sample size ratio of 1:1). This sample size is considerably smaller than the 788 participants required when using Cohen's 0.2 threshold for a small effect size using the same input parameters described above (to reproduce these example power analyses, see https://osf.io/e7yt5/). Conversely, using Cohen's threshold instead of a *smaller* empirically-derived effect size threshold could increase the risk of false positive results, as such study designs would not include enough participants to reliably detect the desired effect size threshold of interest (Ioannidis, 2005). Of course, empirically-derived ESD thresholds for some fields may match or be roughly equivalent to Cohen's suggestions. But an ESD analysis is first required to make this determination.

Several meta-analyses have concluded that psychotherapy can have positive effects on depressive symptoms (Barth et al., 2016; Cuijpers et al., 2008, 2020), however, research is yet to characterize the ESD of psychotherapy RCTs. Extracting data from a recent meta-analysis by Cuijpers et al. (2020), we here calculate the ESD of 453 RCTs on psychotherapy treatments for depression. Other scholars in the field can use these empirically derived ESDs as an approach for study planning to determine effect sizes of interest for power calculations (Lakens, 2022) and to convey more precise effect size magnitudes for completed studies in this especially active and resource intensive research area.

2. Materials and methods

To calculate an ESD, we extracted data from a recent systematic review by Cuijpers et al. (2020) that included RCTs of psychotherapy for depressive disorders. A total of 453 effect sizes in three different age-groups were included [n = 43 in children and adolescents (up to 18 years); n = 331 in young adults and adults (18–55 years); and n = 79 in older adults (55 years and above)]. These effect sizes were derived from 366 trials, which included a total of 36 702 individuals (17158 in the control conditions and 19 544 in the treatment conditions). Next, we calculated the 50th percentile effect size (medium effect), and the 25th (small effect) and 75th percentile (large effect) effects, as they are equidistant from the average effect size (Cohen, 1992), as also conducted by Quintana (2017). Note we use Hedges' *g* effect size, which is an unbiased estimate of effect size, and Cohen's *d* effect size has a negligible difference when the sample size of the RCT is greater than 20 (see Lakens, 2013). In the present sample, less than 9% of the RCTs had a sample size <20. Empirically derived cutoffs were calculated considering all the included depression studies (n = 453) and stratifying studies by a) those that did or did not include participants with comorbidities, and b) age group (i.e., children/adolescents, young adults/adults, older adults). Empirically derived cutoffs were also calculated for the six psychotherapy intervention subtypes with at least 20 reported effects: Third wave therapy, behavioral activation therapy, cognitive behavioral therapy, interpersonal psychotherapy, problem-solving therapy, and non-directive supportive therapy.

To identify studies that may significantly impact the overall ESD, we also created a graphical display of study heterogeneity (GOSH) plot (also called a combinatorial meta-analysis), which involves conducting a series of meta-analyses using all potential combination of studies. A GOSH plot is useful for identifying whether a single study or distinct subgroup of studies has an impact on the summary effect size estimate. We specified 20 000 random subset models out of all possible models.

Sample sizes needed to reliably identify effect size values at the 25th, 50th, and 75th percentiles were computed, assuming 80% power, a significance level of 0.05, and a two-tailed test. To identify potential publication bias that could lead to inflated effect sizes, a selection model was fitted using the approach outlined by Vevea and Hedges (1995). A selection model assumes that studies with non-significant *p*-values are less likely to be published, thus studies associated with non-significant *p*-values contribute more weight in this model. Weight-selection models operate under the assumption of effect size independence. Thus, for studies presenting multiple effect sizes, we opted conservatively to choose the largest effect size. A likelihood ratio test was used to determine whether there was a significant difference between the unadjusted model and the model that was adjusted for publication bias. A threshold of $p = 0.1$ was applied, as suggested by Begg and Mazumdar (1994). The R script and data to reproduce analyses are available at https://osf.io/e7yt5/.

3. Results

For the 453 extracted effect sizes, the 25th (small effect), 50th (medium effect), and 75th (large effect) percentiles corresponded to Hedges' *g* values of 0.32, 0.62, and 1.00, respectively (Fig. 1). The 25th (small effect), 50th (medium effect), and 75th (large effect) percentile effect sizes in studies including participants with and without comorbidities and also across age groups are presented in Table 1. Effect size thresholds across six different psychotherapy subtypes, whose medium-sized effects ranged from 0.55 (non-directive supportive therapy) to 0.93 (behavioral activation therapy), are presented in Supplementary Table 1. The GOSH plot did not reveal any apparent outlier cluster, suggesting that no individual study or group of studies had an overt influence (Fig. 2). Removal of 11 potential outliers only had a negligible effect on results (see Supplementary Materials). Table 2 shows the simulation results of the sample sizes required to reliably detect a range of effect sizes when specifying statistical power at 0.8 (alpha = 0.05, a two-tailed test).

A selection model was fitted including 366 effects sizes, after the exclusion of 87 dependent effect sizes, suggesting that the adjusted publication bias model had an estimate of 0.614 (SE = 0.05), which in relation to the estimate of the unadjusted model (intercept = 0.717, SE = 0.03), represents a 14.29% attenuation due to publication bias. Based on the statistical test of the goodness-of-fit between two models, this difference was statistically significant ($\chi^2_{(1)} = 7.000$, p-value <0.001), which is consistent with the presence of publication bias. Given the attenuation effect, we reduced the small, medium, and large effects by 14.29%, yielding publication bias adjusted ESDs of 0.274, 0.531, and 0.857 respectively (see Table 1 with the publication bias adjusted estimates in parentheses). The same procedure was used for the subset of non-comorbidity studies (excluding 66 studies). This analysis suggested that the adjusted publication bias model has an intercept of 0.552 (SE = 0.074), which in relation to the intercept of the unadjusted model (intercept = 0.759 (SE = 0.046), ($\chi^2_{(1)} = 14.53$, p-value <0.001), represented an attenuation of 29.48%.

4. Discussion

Our analysis of 366 effect sizes from RCTs of psychotherapy treatment for depression indicates that effect sizes of 0.27, 0.53, and 0.86 correspond to small, medium, and large magnitudes, respectively. We provide empirically derived cutoffs for study designs that include participants with and without comorbidities, across different age groups, and for six psychotherapy subtypes. Cohen's effect size thresholds for group differences are commonly used across the psychological sciences, despite the limitations associated this approach. Our results suggest that using Cohen's threshold set would underestimate effect sizes compared to reported effect sizes associated with psychotherapy treatment for depression. For instance, an interpretation using Cohen's thresholds would categorise an effect size of 0.2 as small. However, our empirically established effect size criterion indicates that a small effect is associated with an effect size of 0.27. Prior ESD analyses have also demonstrated that using Cohen's defaults are smaller than empirically-derived effect sizes in diverse fields to a similar degree as the present analysis, such as heart rate variability (Quintana, 2017; small effect = 0.26), endothelial function (Cherubini and MacDonald, 2021; small effect = 0.28), and psychology (Szucs and Ioannidis, 2017; small effect = 0.29).

It is important to note that our publication bias adjusted threshold set was based on an inflation estimate derived from a fitted selection model.

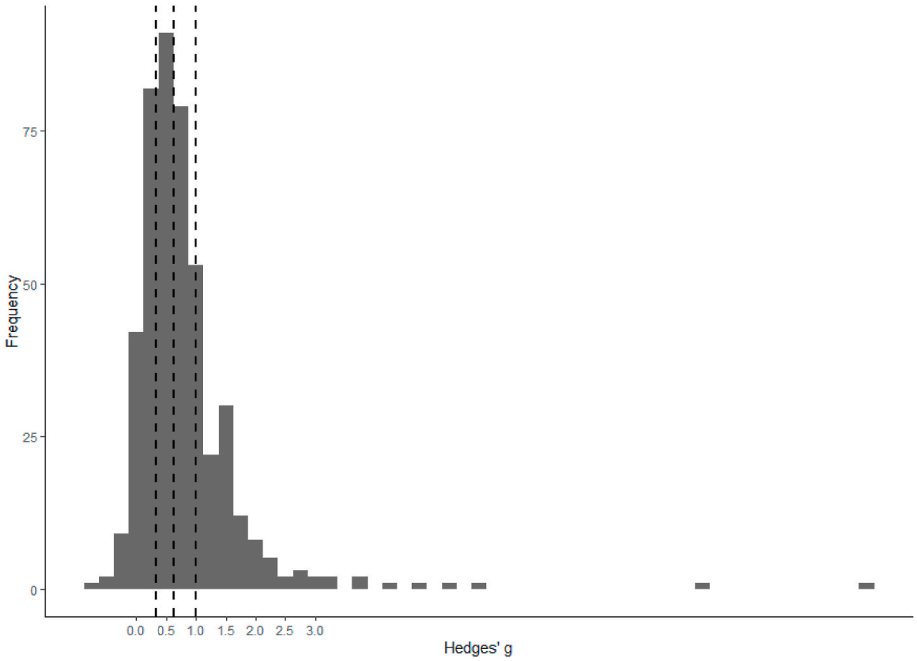

Fig. 1. The effect size distribution of 453 raw effect sizes (without attenuation) from studies evaluating the effect of psychotherapy on depression. The 25th, 50th, and 75th percentiles (dashed lines) represent the calculated thresholds for small (0.32), medium (0.62), and large (1.00) effects.

Table 1
Effect size percentiles for all studies (n = 453) per age and sub-group analyses for RCTs where patients without comorbidities were included (n = 280). Percentiles that consider the publication bias attenuation effect are also presented.
* = the attenuated effect was calculated considering 366 studies, ‡ the attenuated effect was calculated considering 214 studies.

	N = 453 (All studies)			N = 280 (No comorbidity studies)		
	25%	50%	75%	25%	50%	75%
Cohen's suggested guidelines	0.2	0.5	0.8	0.2	0.5	0.8
All ages	0.320 (0.274)*	0.620 (0.531)*	1.00 (0.857)*	0.357 (0.256)‡	0.660 (0.474)‡	1.07 (0.7690)‡
Children and adolescents	0.215	0.340	0.685	0.220	0.340	0.720
Young adults and adults	0.355	0.660	1.01	0.410	0.710	1.06
Older adults	0.300	0.510	1.11	0.390	0.700	1.350

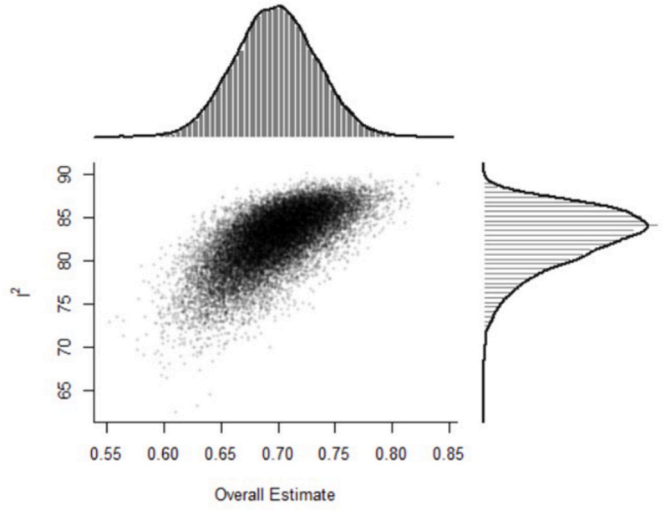

Fig. 2. A Graphical Display of Heterogeneity (GOSH) plot illustrating the summary effect sizes and heterogeneity (I^2) of 20 000 different combinations of studies from the original meta-analysis. There were no distinct clusters, suggesting that no single study or group of studies had an influential effect on the summary effect size.

For a less conservative approach that does not rely on a selection model estimate, the pre-adjusted threshold set can also be used (i.e., small = 0.32, medium = 0.62, and large = 1.00). However, users should be aware that the unadjusted set likely represents modestly inflated effect sizes. Moreover, it is worth mentioning that this ESD approach is particularly well suited for active and established research areas, whereas it less suited for emerging fields with only a small number of studies available (e.g., <20).

5. Conclusion

Based on power calculations using empirically derived thresholds, we suggest sample sizes which researchers can use to more suitably power future study designs in psychotherapy for depression. However, there are some limitations to the present research that are worth noting. First, these empirically derived thresholds only faithfully represent the sample of the studies included in the meta-analysis by Cuijpers et al. (2020), who applied their own specific inclusion and exclusion criteria for study eligibility. However, if researchers would like to calculate their own empirically derived thresholds for psychotherapy interventions in depression that include a different set of studies, the provided analysis code can be used. Second, for our publication bias adjusted values we assumed that the attenuation level of small, medium, and large effects are equivalent, which may not necessarily be the case. Future research can explore approaches that adjust effect sizes for potential publication

Table 2
Simulation for RCTs with a target power of 0.8, alpha of 0.05, and an allocation ratio of 1:1.

Target Power	Actual Power	N1	N2	N	*d* Effect size	Alpha
0.80	0.8003	340	340	680	0.22	0.050
0.80	0.8007	274	274	548	0.24	0.050
0.80	**0.8009**	**225**	**225**	**450**	**0.27**	**0.050**
0.80	0.8008	188	188	376	0.29	0.050
0.80	0.8021	160	160	320	0.32	0.050
0.80	0.8007	137	137	274	0.34	0.050
0.80	0.8007	119	119	238	0.37	0.050
0.80	0.8031	105	105	210	0.39	0.050
0.80	0.8038	93	93	186	0.42	0.050
0.80	0.8045	83	83	166	0.44	0.050
0.80	0.8023	74	74	148	0.47	0.050
0.80	0.8038	67	67	134	0.49	0.050
0.80	0.8055	61	61	122	0.52	0.050
0.80	**0.8014**	**55**	**55**	**110**	**0.54**	**0.050**
0.80	0.8067	51	51	102	0.57	0.050
0.80	0.8079	47	47	94	0.59	0.050
0.80	0.8048	43	43	86	0.62	0.050
0.80	0.8070	40	40	80	0.64	0.050
0.80	0.8056	37	37	74	0.67	0.050
0.80	0.8005	34	34	68	0.69	0.050
0.80	0.8039	32	32	64	0.72	0.050
0.80	0.8046	30	30	60	0.74	0.050
0.80	0.8027	28	28	56	0.77	0.050
0.80	0.8129	27	27	54	0.79	0.050
0.80	0.8060	25	25	50	0.82	0.050
0.80	0.8129	24	24	48	0.84	0.050
0.80	**0.8003**	**22**	**22**	**44**	**0.87**	**0.050**
0.80	0.8035	21	21	42	0.89	0.050
0.80	0.8050	20	20	40	0.92	0.050
0.80	0.8049	19	19	38	0.94	0.050
0.80	0.8030	18	18	36	0.97	0.050
0.80	0.8226	18	18	36	0.99	0.050
0.80	0.8184	17	17	34	1.02	0.050
0.80	0.8123	16	16	32	1.04	0.050
0.80	0.8038	15	15	30	1.07	0.050
0.80	0.8216	15	15	30	1.09	0.050
0.80	0.8105	14	14	28	1.12	0.050

Note: N1 and N2 are the number of participants per group. N is the total sample size. Actual Power is the achieved power. Because N1 and N2 are integers, this value is often (slightly) larger than the target power. In bold are the sample sizes for small, medium, and large effect sizes using the redefined cutoffs.

bias at the individual study level, such as 'limit' meta-analysis (Rücker et al., 2011).

In summary, we have demonstrated that the use of default effect size thresholds can lead to inaccurate effect size interpretations, which can have implications for study planning and the interpretation of results. We also provide the means for researchers to calculate effect size thresholds and determine effect sizes of interest that may better represent their field or research question of interest.

Statement of ethics

Ethical approval was not required as this study extracted data from a systematic review by Cuijpers et al. (2020).

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The data that support the findings of this study are openly available at https://osf.io/e7yt5/

Author contributions

Contribution of authors is as follows: study conception and design: DSQ and ANH; analysis and interpretation of the results: SP, HCM, DSQ, ANH; draft manuscript preparation: ANH, DSQ, HCM, SP; reviewed the final version of the manuscript: ANH, HCM, SP & DSQ. All authors reviewed the results and approved the final version of the manuscript.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Acknowledgements

An earlier version of the manuscript was posted on the Open Science Framework Preprint server https://osf.io/erhmw/

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpsychires.2023.11.009.

References

Barth, J., Munder, T., Gerger, H., et al., 2016. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. Focus 14 (2), 229–243. https://doi.org/10.1176/appi.focus.140201.
Begg, C.B., Mazumdar, M., 1994. Operating characteristics of a rank correlation test for publication bias. Biometrics 50 (4), 1088–1101. https://doi.org/10.2307/2533446.
Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Routledge.
Cherubini, J.M., MacDonald, M.J., 2021. Statistical inferences using effect sizes in human endothelial function research. Artery Research 27 (4). https://doi.org/10.1007/s44200-021-00006-6. Article 4.
Cuijpers, P., Karyotaki, E., Eckshtain, D., Ng, M.Y., Corteselli, K.A., Noma, H., et al., 2020. Psychotherapy for depression across different age groups: a systematic review and meta-analysis. JAMA Psychiatr. 77 (7), 694–702. https://doi.org/10.1001/jamapsychiatry.2020.0164.
Cuijpers, P., van Straten, A., Andersson, G., van Oppen, P., 2008. Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. J. Consult. Clin. Psychol. 76 (6), 909–922. https://doi.org/10.1037/a0013075.
Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155.
Funder, D.C., Ozer, D.J., 2019. Evaluating effect size in psychological research: sense and nonsense. Advances in Methods and Practices in Psychological Science 2 (2), 156–168. https://doi.org/10.1177/2515245919847202.
Glass, G.V., McGaw, B., Smith, M.L., 1981. Meta-analysis in Social Research. Sage.
Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Med 2 (8), e124. https://doi.org/10.1371/journal.pmed.0020124.
Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 863. https://doi.org/10.3389/fpsyg.2013.00863.
Lakens, D., 2022. Sample size justification. Collabra: Psychology 8 (1). https://doi.org/10.1525/collabra.33267.
Panjeh, S., Nordahl-Hansen, A., Cogo-Moreira, H., 2023a. Establishing new cutoffs for Cohen's *d*: an application using known effect sizes from trials for improving sleep quality on composite mental health. Int. J. Methods Psychiatr. Res., e1969 https://doi.org/10.1002/mpr.1969.
Panjeh, S., Nordahl-Hansen, A., Cogo-Moreira, H., 2023b. Moving forward to a world beyond 0.2, 0.5, and 0.8 effects sizes: new cutoffs for school-based anti-bullying interventions. J. Interpers Violence 38 (11–12), 7843–7851. https://doi.org/10.1177/08862605221147065.
Quintana, D.S., 2017. Statistical considerations for reporting and planning heart rate variability case-control studies. Psychophysiology 54 (3), 344–349. https://doi.org/10.1111/psyp.12798.
Rücker, G., Schwarzer, G., Carpenter, J.R., Binder, H., Schumacher, M., 2011. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. Biostatistics 12 (1), 122–142. https://doi.org/10.1093/biostatistics/kxq046.
Szucs, D., Ioannidis, J.P., 2017. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 15 (3), e2000797 https://doi.org/10.1371/journal.pbio.2000797.
Thompson, B., 2009. A brief primer on effect sizes. J. Teach. Phys. Educ. 28 (3), 251–254. https://doi.org/10.1123/jtpe.28.3.251.
Vevea, J.L., Hedges, L.V., 1995. A general linear model for estimating effect size in the presence of publication bias. Psychometrika 60 (3), 419–435. https://doi.org/10.1007/BF02294384.
Wasserstein, R.L., Lazar, N.A., 2016. The ASA statement on p-values: context, process, and purpose. Am. Statistician 70 (2), 129–133. https://doi.org/10.1080/00031305.2016.1154108.