RESEARCH ARTICLE

Conditional equivalence testing: An alternative remedy for publication bias

Harlan Campbell*, **Paul Gustafson**

Department of Statistics, University of British Columbia, Vancouver, Canada

* harlan.campbell@stat.ubc.ca

Abstract

We introduce a publication policy that incorporates "conditional equivalence testing" (CET), a two-stage testing scheme in which standard NHST is followed conditionally by testing for equivalence. The idea of CET is carefully considered as it has the potential to address recent concerns about reproducibility and the limited publication of null results. In this paper we detail the implementation of CET, investigate similarities with a Bayesian testing scheme, and outline the basis for how a scientific journal could proceed to reduce publication bias while remaining relevant.

Introduction

Poor reliability, within many scientific fields, is a major concern for researchers, scientific journals and the public at large. In a highly cited essay, Ioannidis (2005) [1] uses Bayes theorem to claim that more than half of published research findings are false. While not all agree with the extent of this conclusion, e.g. [2, 3], the argument raises concerns about the trustworthiness of science, amplified by a disturbing prevalence of scientific misconduct [4]. That a result may be substantially less reliable than the p-value suggests has been "underappreciated" (Goodman & Greenland, 2007) [2], to say the least. To address this problem, some journal editors have taken radical measures [5], while the reputations of researchers are tarnished [6]. The merits of null hypothesis significance testing (NHST) [7–10] and the p-value are now vigorously debated [11–13].

The limited publication of null results is certainly one of the most substantial factors contributing to low reliability. Whether due to a reluctance of journals to publish null results or to a reluctance of investigators to submit their null research [14, 15], the consequence is severe publication bias [16, 17]. Despite repeated warnings, publication bias persists and, to a certain degree, this is understandable. Accepting a null result can be difficult, owing to the well-known fact that "absence of evidence is not evidence of absence" [18, 19]. As Greenwald (1975) [20] writes: "it is inadvisable to place confidence in results that support a null hypothesis because there are too many ways (including incompetence of the researcher), other than the null hypothesis being true, for obtaining a null result." Indeed, this is the foremost critique of NHST, that it cannot provide evidence in favour of the null hypothesis. The commonly held

belief that for a non-significant result to show high "retrospective power" [21] implies support in favour of the null, is problematic [22, 23].

In order to address publication bias, it is often suggested [24–27] that publication decisions should be made without regards to the statistical significance of results, i.e. "result-blind peer review" [28]. In fact, a growing number of psychology and neuroscience journals are requiring pre-registration [29] by means of adopting the publication policy of "Registered Reports" (RR) [30], in which authors "pre-register their hypotheses and planned analyses before the collection of data" [31]. If the rationale and methods are sound, the RR journal agrees (before any data are collected) to publish the study regardless of the eventual data and outcome obtained. Among many potential pitfalls with result-blind peer review [32], a genuine and substantial concern is that, if a journal were to adopt such a policy, it might quickly become a "dumping ground" [28] for null and ambiguous findings that do little to contribute to the advancement of science. To address these concerns, RR journals require that, for any manuscript to be accepted, authors must provide a-priori (before any data are collected) sample size calculations that show statistical power of at least 90% (in some cases 80%). This is a reasonable remedy to a difficult problem. Still, it is problematic for two reasons.

First, it is acknowledged that this policy will disadvantage researchers who work with "expensive techniques or who have limited resources" (Chambers et al., 2014) [31]. While not ideal, small studies can provide definitive value and potential for learning [33]. Some go as far as arguing against any requirement for a-priori sample size calculations (i.e. needing to show sufficient power as a requisite for publication). For example, Bacchetti (2002) [34] writes: "If a study finds important information by blind luck instead of good planning, I still want to know the results"; see also Aycaguer and Galbán (2013) [35]. While unfortunate, the loss of potentially valuable "blind luck" results and small sample studies [36] appears to be a necessary price to pay for keeping a "result-blind peer review" journal relevant. Is this too high a price? Based on simulations, Borm et al. (2009) [37] conclude that the negative impact of publication bias does not warrant the exclusion of studies with low power.

Second, a-priori power calculations are often flawed, due to the unfortunate "sample size samba" [38]: the practice of retrofitting the anticipated effect size in order to obtain an affordable sample size. Even under ideal circumstances, a-priori power estimation is often "wildly optimistic" [39] and heavily biased due to "follow-up bias" (see Albers and Lakens, 2018 [40]) and the "illusion of power" [41]. This "illusion" (also known as "optimism bias") occurs when the estimated effect size is based on a literature filled with overestimates (to be expected in many fields due, somewhat ironically, to publication bias). Djulbegovic et al. (2011) [42] conduct a retrospective analysis of phase III randomized controlled trials (RCTs) and conclude that optimism bias significantly contributes to inconclusive results; see also Chalmers and Matthews (2006) [43]. What's more, oftentimes due to unanticipated difficulties with enrolment, the actual sample size achieved is substantially lower than the target set out a-priori [44]. In these situations, RR requires that either the study is rejected/withdrawn for publication or a certain leeway is given under special circumstances (Chambers (2017), personal communication). Neither option is ideal. Given these difficulties with a-priori power calculations, it remains to be seen to what extent the 90% power requirement will reduce the number of underpowered publications that could lead a journal to be a dreaded "dumping ground".

An alternative proposal to address publication bias and the related issues surrounding low reliability is for researchers to adopt Bayesian testing schemes; e.g. [45–47]. It has been suggested that with Bayesian methods, publication bias will be mitigated "because the evidence can be measured to be just as strong either way" [48]. Bayesian methods may also provide for a better understanding of the strength of evidence [49]. However, sample sizes will typically need to be larger than with equivalent frequentist testing in situations when there is little prior

information incorporated [50]. Furthermore, researchers in many fields remain uncomfortable with the need to define (subjective) priors and are concerned that Bayesian methods may increase "researcher degrees of freedom" [51]. These issues may be less of a concern if studies are pre-registered, and indeed, a number of RR journals allow for a Bayesian option. At registration (before any data are collected), rather than committing to a specific sample size, researchers commit to attaining a certain Bayes Factor (BF). For example, the journals *Comprehensive Results in Social Psychology* (CRSP) and *NFS Journal* (the official journal of the Society of Nutrition and Food Science) require that one pledges to collect data until the BF is more than 3 (or less than 1/3) [52]. The journals *BMC Biology* and the *Journal of Cognition* require, as a requisite for publication, a BF of at least 6 (or less than 1/6) [53, 54].

In this paper, we propose an alternate option made possible by adapting NHST to conditionally incorporate equivalence testing. While equivalence testing is by no means a novel idea, previous attempts to introduce equivalence testing have "largely failed" [55] (although recently, there has been renewed interest). In Section 2, in order to further the understanding of conditional equivalence testing, we provide a brief overview including how to establish appropriate equivalence margins.

One reason conditional equivalence testing (CET) is an appealing approach is that it shares many of the properties that make Bayesian testing schemes so attractive. As such, the publication policy we put forward is somewhat similar to the RR "Bayesian option". With CET, evidence can be measured in favour of either the alternative or the null (at least in some pragmatic sense), and as such is "compatible with a Bayesian point of view" [56]. In Section 3, we conduct a simple simulation study to demonstrate how one will often arrive at the same conclusion whether using CET or a Bayesian testing scheme. In Section 4, we outline how a publication policy, similar to RR, could be framed around CET to encourage the publication of null results. Finally, Section 5 concludes with suggestions for future research.

Conditional equivalence testing overview

Standard equivalence testing is essentially NHST with the hypotheses reversed. For example, for a two-sample study of means, the equivalence testing null hypothesis would be a difference (beyond a given margin) in population means, and the alternative hypothesis would be equal (within a given margin) population means. *Conditional* equivalence testing (CET) is the practice of standard NHST followed conditionally (if one fails to reject the null) by equivalence testing. CET is not an altogether new way of testing. Rather it is the usage of established testing methods in a way that permits better interpretation of results. In this regard, it is similar to other proposals such as the "three-way testing" scheme proposed by Goeman et al. (2010) [57], and Zhao (2016) [58]'s proposal for incorporating both statistical and clinical significance into one's testing. To illustrate, we briefly outline CET for a two-sample test of equal means (assuming equal variance). Afterwards, we discuss defining the equivalence margin. In Appendix 1, we provide further details on CET including a look at operating characteristics and sample size calculations.

Let x_{i1}, for $i = 1, \ldots, n_1$ and x_{i2}, for $i = 1, \ldots, n_2$ be independent random samples from two normally distributed populations of interest with μ_1, the true mean of population 1; μ_2, the true mean of population 2; and σ^2, the true common population variance. Let $n = n_1 + n_2$ and define sample means and sample variances as follows: $\bar{x}_g = \sum_{i=1}^{n_g} x_{gi}/n_g$, and $s_g^2 = \sum_{i=1}^{n_g} (x_{gi} - \bar{x}_g)^2/(n_g - 1)$, for $g = 1, 2$. Also, let $s_p = \sqrt{((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)}$. The true difference in population means, $\mu_d = \mu_1 - \mu_2$, under the standard null hypothesis, H_0, is equal to zero. Under the standard alternative, H_1, we have that $\mu_d \neq 0$.

The term equivalence is not used in the strict sense that $\mu_1 = \mu_2$. Instead, equivalence in this context refers to the notion that the two means are "close enough", i.e. their difference is within the *equivalence margin*, $\delta = [-\Delta, \Delta]$, chosen to define a range of values considered equivalent. The value of Δ is ideally chosen to be the "minimum clinically meaningful difference" [59, 60] or as the "smallest effect size of interest" [61].

Let $F_{T_{df}}()$ be the cumulative distribution function (cdf) of the t distribution with df degrees of freedom and define the following critical t values: $t^*_{\alpha_1/2} = F^{-1}_{T_{n-2}}(1 - 0.5\alpha_1)$ (i.e. the upper $100 \cdot \frac{\alpha_1}{2}$-th percentile of the t-distribution with $n-2$ degrees of freedom) and $t^*_{\alpha_2} = F^{-1}_{T_{n-2}}(1 - \alpha_2)$, (i.e. the upper $100 \cdot \alpha_2$-th percentile of the t-distribution with $n-2$ degrees of freedom). As such, α_1 is the maximum allowable type I error (e.g. $\alpha_1 = 0.05$) and α_2 is the maximum allowable "type E" error (erroneously concluding equivalence), possibly set equal to α_1. If a type E error is deemed less costly than a type I error, α_2 may be set higher (e.g. $\alpha_2 = 0.10$). CET is the following conditional procedure consisting of five steps:

Step 1- A two-sided, two-sample *t*-test for a difference of means.

Calculate the t-statistic, $T = (\bar{x}_1 - \bar{x}_2)/(s_p\sqrt{1/n_1 + 1/n_2})$ and associated p-value, $p_1 = 2 \cdot F_{T_{n-2}}(-|T|)$.

Step 2- If $|T| > t^*_{\alpha_1/2}$, then declare a *positive result*. There is evidence of a statistically significant difference, p-value = p_1.

Otherwise, if $|T| \leq t^*_{\alpha_1/2}$, proceed to Step 3.

Step 3- Two one-sided tests (TOST) for equivalence of means [62, 63].

Calculate two t-statistics: $T_1 = (\bar{x}_1 - \bar{x}_2 + \Delta)/(s_p\sqrt{1/n_1 + 1/n_2})$, and $T_2 = (\bar{x}_1 - \bar{x}_2 - \Delta)/(s_p\sqrt{1/n_1 + 1/n_2})$. Calculate an associated p-value, $p_2 = max(F_{T_{n-2}}(-T_1), F_{T_{n-2}}(T_2))$. Note: p_2 is a *marginal* p-value, in that it is not calculated under the assumption that $p_1 > \alpha_1$.

Step 4- If $T_1 > t^*_{\alpha_2}$ and $T_2 < (-t^*_{\alpha_2})$, declare a *negative result*. There is evidence of a statistically significant equivalence (with margin $\delta = [-\Delta, \Delta]$), p-value = p_2.

Otherwise, proceed to Step 5.

Step 5- Declare an *inconclusive result*. There is insufficient evidence to support any conclusion.

For ease of explanation, let us define $p_{CET} = p_1$ if the result is positive, and $p_{CET} = 1 - p_2$ if the result is negative or inconclusive. Thus, a small value of p_{CET} suggests evidence in favour of a positive result, whereas a large value of p_{CET} suggests evidence in favour of a negative result. As is noted above, it is important to acknowledge that, despite the above procedure being dubbed "conditional", p_2 is a *marginal* p-value, i.e. it is not calculated under the assumption that $p_1 > \alpha_1$. The interpretation of p_2 is taken to be the same regardless of whether it was obtained following Steps 1 and 2 or was obtained "on its own" via standard equivalence testing.

Standard NHST involves the same first and second steps and ends with an alternative Step 3 which states that if $p_1 > \alpha_1$, one declares an inconclusive result ('there is insufficient evidence

to reject the null.'). Similar to two-sided CET, one-sided CET is straightforward, making use of non-inferiority testing in Step 3.

CET is a procedure applicable to any type of outcome. Let θ be the parameter of interest in a more general testing setting. CET can be described in terms of calculating confidence intervals around $\hat{\theta}$, the statistic of interest. For example, θ may be defined as the the difference in proportions, the hazard ratio, the risk ratio, or the slope of a linear regression model, etc. [64–70]. Note that in these other cases, the equivalence margin, δ, may not necessarily be centred at zero, but will be a symmetric interval around θ_0, the value of θ under the standard null hypothesis. In general, let $\delta = [\delta_L, \delta_U]$, with Δ equal to half the length of the interval. Consider, more generally, CET as the following procedure:

Step 1- Calculate a $(1 - \alpha_1)$% Confidence Interval for θ.

Step 2- If this C.I. excludes θ_0, then declare a *positive result*.

Otherwise, if θ_0 is within the C.I., proceed to Step 3.

Step 3- Calculate a $(1 - 2\alpha_2)$% Confidence Interval for θ.

Step 4- If this C.I. is entirely within δ, declare a *negative result*.

Otherwise, proceed to Step 5.

Step 5- Declare an *inconclusive result*.

There is insufficient evidence to support any conclusion.

Fig 1 illustrates the three conclusions with their associated confidence intervals. We once again consider two-sample testing of normally distributed data as before, with $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$. Situation "a", in which the $(1 - \alpha_1)$% C.I. is entirely outside the equivalence margin is what Guyatt et al. (1995) [71] call a "definitive-positive" result. The lower confidence limit of the parameter is not only larger than zero (the null value, θ_0), implying a "positive" result, but also is above the Δ threshold. Situations "b" and "c", in which the $(1 - \alpha_1)$% C.I. excludes zero and the $(1 - 2\alpha_2)$% C.I. is within $[-\Delta, \Delta]$ are considered "positive" results but require some additional interpretation. One could describe the effect in these cases as "significant yet not meaningful" or conclude that there is evidence of a significant effect, yet the effect is likely "not minimally important". A "positive" result such as "d", with a wider confidence interval, represents a significant, albeit imprecisely estimated, effect. One could conclude that additional studies, with larger sample sizes, are required to determine if the effect is of a meaningful magnitude.

Evidently, in some cases, the three categories are not sufficient on their own for adequate interpretation. For example, without any additional information the "positive" vs. "negative" distinction between cases "b" and "f" is incomplete and potentially misleading. These two results appear very similar: in both cases a substantial (i.e. greater than Δ-sized) effect can be ruled out. Indeed, positive result "b" appears much more like the negative result "f" than positive result "a". Additional language and careful attention to the estimated effect size is required for correct interpretation. While case "k" has a similar point estimate to "b" and "f", the wider C.I. (Δ is within $(1 - 2\alpha_2)$% C.I.) means that one cannot rule out the possibility of a meaningful effect, and so it is rightly categorized as "inconclusive".
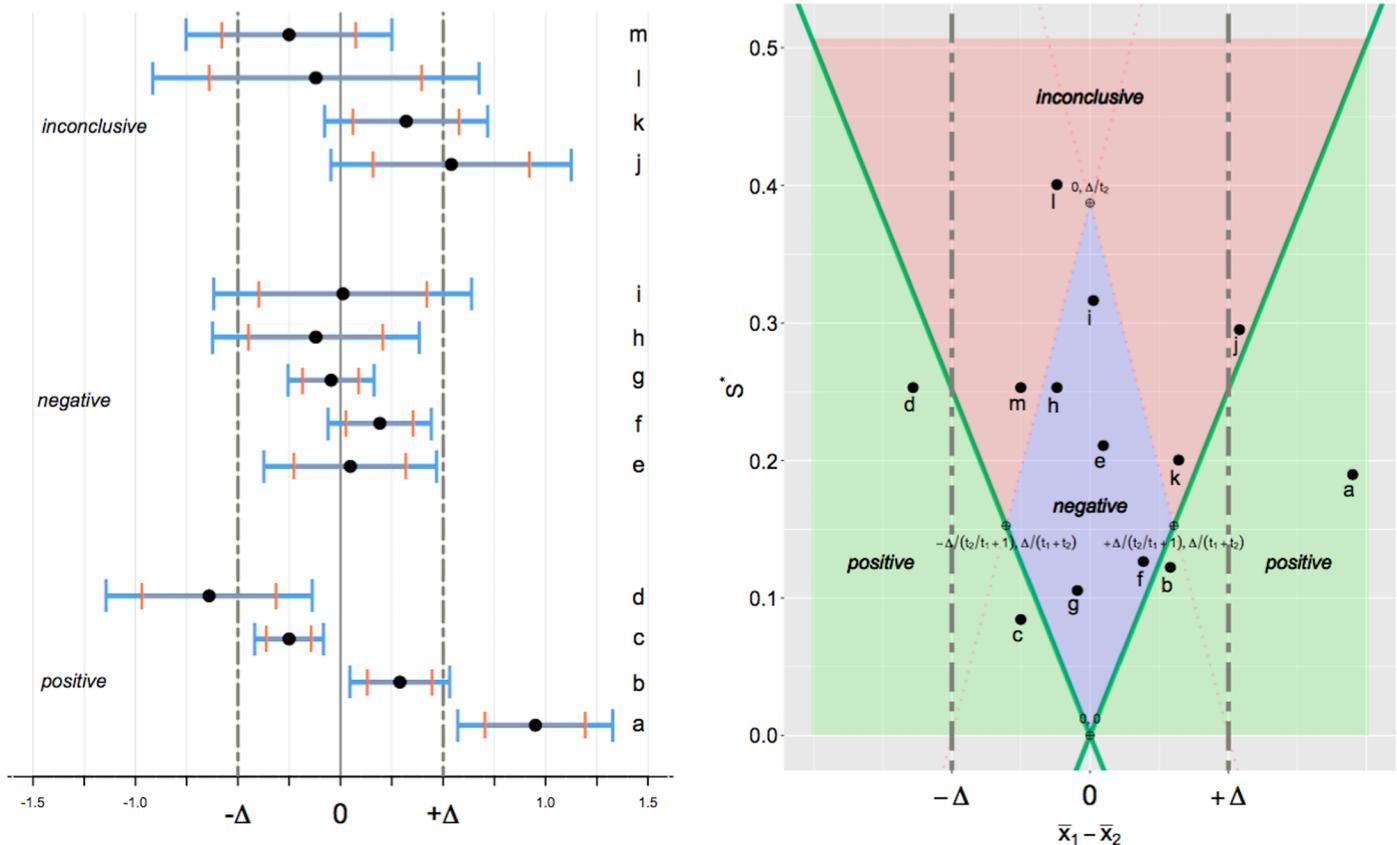
Fig 1. Left panel—CET point estimates and confidence intervals. Point estimates and confidence intervals of thirteen possible results from two-sample testing of normally distributed data are presented alongside their corresponding conclusions. Let $n = 90$, $\Delta = 0.5$, $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$. Black points indicate point estimates; blue lines (wider intervals) represent $(1-\alpha_1)\%$ confidence intervals; and orange lines (shorter intervals) represent $(1-2\alpha_2)\%$ confidence intervals. **Right panel- Rejection regions.** Let $t_1 = t^*_{\alpha_1/2}$, $t_2 = t^*_{\alpha_2}$. The values of $\hat{\mu}_d = \bar{x}_1 - \bar{x}_2$ (on the x-axis) and $s^* = s_p\sqrt{(1/n_1 + 1/n_2)}$ (on the y-axis) are obtained from the data. The three conclusions, positive, negative, inconclusive correspond to the three areas shaded in green, blue and red respectively. The black lettered points correspond to the thirteen scenarios on the left panel.

It may be argued that CET is simply a recalibration of the standard confidence interval, like converting Fahrenheit to Celsius. This is valid commentary and in response, it should be noted that our suggestion to adopt CET (in the place of standard confidence intervals) is not unlike suggestions that confidence intervals should replace p-values [72–74]. One advantage of CET over confidence intervals is that it may improve the interpretation of null results [75, 76]. By clearly distinguishing between what is a *negative* versus an *inconclusive* result, CET serves to simplify the long "series of searching questions" necessary to evaluate a "failed outcome" [77]. However, as can be seen with the examples of Fig 1, the use of CET should not rule out the complementary use of confidence intervals. Indeed, the best interpretation of a result will be when using both tools together.

As Dienes and Mclatchie (2017) [45] clearly explain, with standard NHST, one is unable to make the "three-way distinction" between the positive, inconclusive and negative ("evidence for H_1", "no evidence to speak of", and "evidence for H_0"). Fig 2 shows the distribution of standard two-tailed p-values under NHST and corresponding p_{CET} values under CET for three different sample sizes. These are the result of two-sample testing of normally distributed data with $n_1 = n_2$.
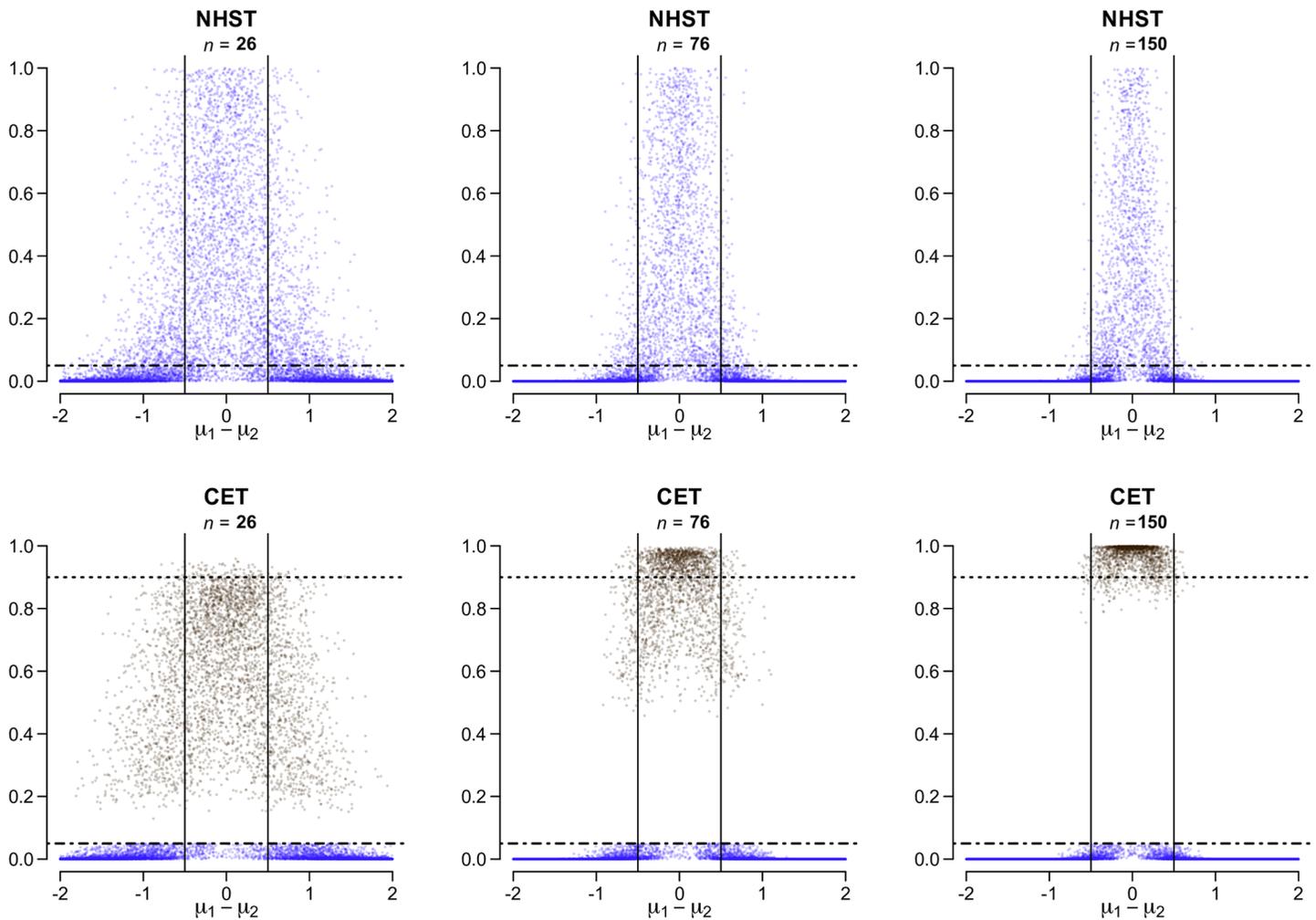
Fig 2. NHST p-values vs. p_{CET} values. Distribution of two-tailed p-values from NHST and p_{CET} values from CET, with varying n (total sample size) and $\delta = [-0.5, 0.5]$. Data are the results from 10,000 Monte Carlo simulations of two-sample normally distributed data with equal variance, $\sigma^2 = 1$. The true difference in means, $\mu_d = \mu_1 - \mu_2$, is drawn from a uniform distribution between -2 and 2. Black points ($= 1 - p_2$) might indicate evidence in favour of equivalence ("negative result"), whereas blue points ($= p_1$) indicate evidence in favour of non-equivalence ("positive result"). An "inconclusive result" would occur for all black points falling in between the two dashed horizontal lines at $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$, (i.e for $p_2 > \alpha_2$). The format of this plot is based on Fig 1, "Distribution of two-tailed p-values from Student's t-test", from Lew (2013) [11].

Under the standard null ($\mu_d = 0$) with NHST, one is just as likely to obtain a large p-value as one is to obtain a small p-value. Under the standard null with CET, a small p_2 value (a large p_{CET} value) will indicate evidence in favour of equivalence given sufficient data. In Fig 2, note that the blue points that fall below the $\alpha_1 = 0.05$ threshold in the upper panels (NHST) remain unchanged in the lower panels (CET). However, blue points that fall above the $\alpha_1 = 0.05$ threshold in the upper panels (NHST) are no longer present in the lower panels (CET). They have been replaced by black points ($= p_{CET} = 1 - p_2$) which, if near the top, suggest evidence in favour of equivalence. This treatment of the larger p-values is more conducive to the interpretation of null results, bringing to mind the thoughts of Amrhein et al. (2017) [78] who write "[a]s long as we treat our larger p-values as unwanted children, they will continue disappearing in our file drawers, causing publication bias, which has been identified as the possibly most prevalent threat to reliability and replicability."

Defining the equivalence margin

As with standard equivalence and non-inferiority testing, defining the equivalence margin will be one of the "most difficult issues" [79] for CET. If the margin is too large, then a claim of equivalence is meaningless. If the margin is too small, then the probability of declaring equivalence will be substantially reduced [80]. As stated earlier, the margin is ideally chosen as a boundary to exclude the smallest effect size of interest [61]. However, these can be difficult to define, and there is generally no clear consensus among stakeholders [81]. Furthermore, previously agreed-upon meaningful differences may be difficult to ascertain as they are rarely specified in protocols and published results [42].

In some fields, there are some generally accepted norms. For example, in bioavailability studies, equivalence is routinely defined (and listed by regulatory authorities) as a difference of less than 20%. In oncology trials, a small effect size has been defined as odds ratio or hazard ratio of 1.3 or less [82]. In ecology, a proposed equivalence region for trends in population size (the log-linear population per year regression slope) is $\delta = [-0.0346, 0.0346]$ [68].

In cases when a specified equivalence margin may not be as clear-cut, less conventional options have been put forth (e.g. the "equivalence curve" of Hauck and Anderson (1986) [76], and the least equivalent allowable difference (LEAD) of Meyners (2007) [83]). One important choice a researcher must make in defining the equivalence margin is whether the margin should be defined on a raw or standardized scale; for further details, see [55, 61]. For example, in our two-sample normal case, one could define equivalence to be a difference within half the estimated standard deviation, i.e. defining $\Delta = q s_p$, with pre-specified $q = 0.5$. Note, that the probability of obtaining a negative result may be zero (or negligible) for certain combinations of values of α_1, α_2, n_1, n_2 and q. For example, with $q = 0.5$, $\alpha_1 = 0.05$, and $\alpha_2 = 0.10$, $Pr(negative)$ = 0 for all $n \leq 26$ (with $n_1 = n_2$). As such, q must be chosen with additional practical considerations (see additional details Appendix 1). For binary and time-to-event outcomes, there is an even greater number of different ways one can define the margin [65, 84–86].

Since the choice of margin is often difficult in the best of circumstances, a retrospective choice is not ideal as there will be ample room for bias in one's choice, regardless of how well intentioned one may be. For this reason, for equivalence and non-inferiority RCTs, it is generally expected (and we recommend, when possible with CET) that margins are to be pre-specified [87].

A comparison with Bayesian testing

Recently, Bayesian statistics have been advocated for as a "possible solution to publication bias" [88]. In particular, there have been many Bayesian testing schemes proposed in the psychology literature [89, 90]. What's more, publication policies based on Bayesian testing schemes are currently in use by a small number of journals and are the preferred approach for some [45]. In response to these developments, we will compare, with regards to their operating characteristics, CET and a Bayesian testing scheme. This brings to mind Dienes (2014) [91] who compares testing with Bayes Factors (BF) to testing with "interval methods" and notes that with interval methods, a study result is a "reflection of the data", whereas with BFs the result reflects the "evidence of one theory over another". What follows is a brief overview of one Bayesian scheme and an investigation of how it compares to CET. (We recognize that this scheme is based on reference priors while some advocate for weakly informative priors or elicitation of subjective priors.)

The Bayes Factor is a valuable tool for determining the relative evidence of a null model compared to an alternative model [92]. Consider, for the two-sample testing of normally distributed data, a Bayes Factor testing scheme in which we take the JZS (Jeffreys-Zellner-Siow)
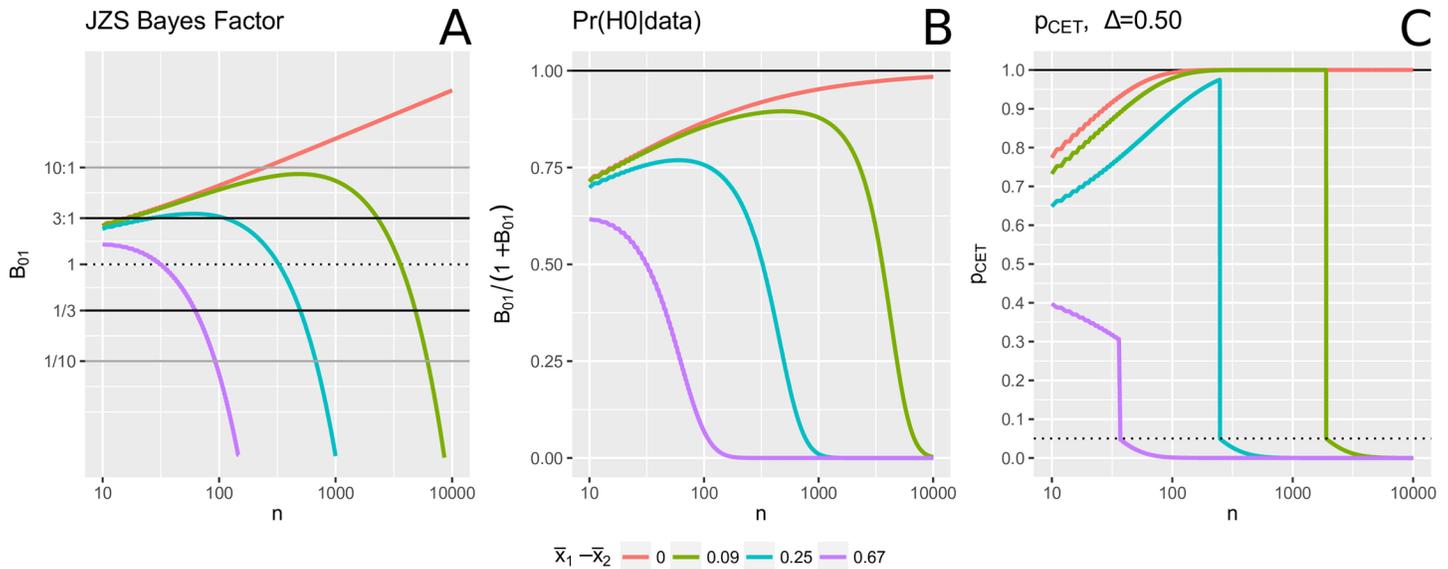
Fig 3. JZS Bayes Factor vs. p_{CET}. Based partially on Fig 5 from Rouder et al. (2009) [93]. Left (middle; right) panel shows how the JZS Bayes Factor (the posterior probability of H_0; p_{CET}) changes with sample size for four different observed mean differences, $\hat{\mu}_d = \bar{x}_1 - \bar{x}_2$. The observed variance is constant, $s_p^2 = 1$.

prior (recommended as a reasonable "objective prior") for the alternative hypothesis; see Appendix 2 and Rouder et al. (2009) [93] for details.

Fig 3(A) shows how the JZS Bayes Factor changes with sample size for four different values of the observed difference in means, $\hat{\mu}_d = \bar{x}_1 - \bar{x}_2$; the observed variance is held constant at $s_p^2 = 1$. When the observed mean difference is exactly 0, the BF increases logarithmically with n. For small to moderate $\hat{\mu}_d$, the BF supports the null for small values of n. However, as n becomes larger, the BF yields less support for the null and eventually favours the alternative. The horizontal lines mark the 3:1 and 1/3 thresholds ("moderate evidence") as well as the 10:1 and 1/10 thresholds ("strong evidence").

Fig 3(C) shows p_{CET}-values for the same four values of $\hat{\mu}_d$, constant $s_p^2 = 1$, and the equivalence margin of $[-0.50, 0.50]$. The curves suggest that CET possesses similar "ideal behaviour" [93] as is observed with the BF. When the observed mean difference is exactly zero, CET provides increasing evidence in favour of equivalence with increasing n. For small to moderate $\hat{\mu}_d$, CET supports the null at first and then, as n becomes larger, at a certain point favours the alternative. The sharp change-point represents border cases. Consider case "f" in Fig 1: if n increased, the confidence intervals would shrink and at a certain point, the $(1 - \alpha_1)\%$ C.I. would be exclude 0 (similar to case "b"). At that point, the result abruptly changes from "negative" to "positive". While this abrupt change may appear odd at first, it may in fact be more desirable than the smooth transition of the BF. Consider for example when $\hat{\mu}_d = 0.25$. Then for n between 112 and 496, the BF will be strictly above 1/3 and strictly below 3, and as such the result, by BF, is inconclusive. In contrast, for the same range in sample size, p_{CET} will be either above 0.90 or below 0.05. As such, with $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$, a conclusive result is obtained. While careful interpretation is required (e.g. "the effect is significant yet not of a meaningful magnitude"), this may be preferable in some settings to the BF's inconclusive result.

We can also consider the posterior probability of H_0 (i.e. $\mu_d = 0$) equal to $B_{01}/(1 + B_{01})$ (when the prior probabilities $Pr(H_0)$ and $Pr(H_1)$ are equal), plotted in Fig 3(B). The similarities and differences between p-values and posterior probabilities have been widely discussed previously [94–96]. Fig 3 suggests that the JZS-BF and CET testing may often result in similar conclusions. We investigate this further by means of a simple simulation study.
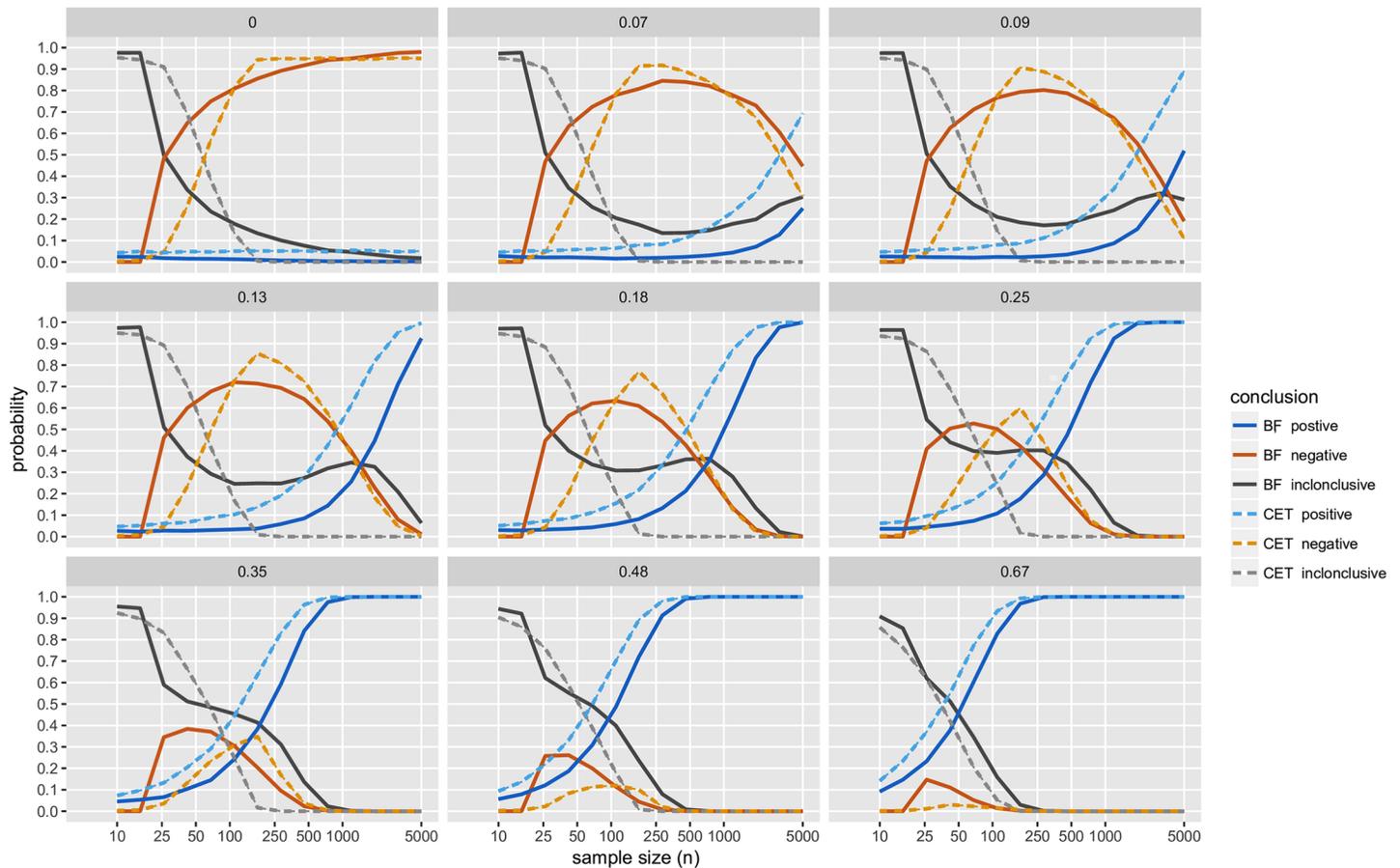
Simulation study

We conducted a small simulation study to compare the operating characteristics of testing with the JZS-BF relative to with the CET approach. CET conclusions were based on setting $\Delta = 0.50$, $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$. JZS BF conclusions were based on a threshold of 3 or greater for evidence in favour of the a negative result and less than 1/3 for evidence in favour of a positive result. BFs in the 1/3–3 range correspond to an inconclusive result. A threshold of 3:1 can be considered "substantial evidence" [97]. We also conducted the simulations with a 6:1 threshold for comparison.

Note that one advantage of Bayesian methods, is that sample sizes need not be determined in advance [98]. With frequentist testing, interim analyses to re-evaluate one's sample size can be performed (while maintaining the Type 1 error). However, these must be carefully planned in advance [99, 100]. Schönbrodt and Wagenmakers (2016) [101] list three ways one might design the sample size for a study using the BF for testing. For the simulation study here we examine only the "fixed-n design".

For a range of μ_d (= 0, 0.07, 0.09, 0.13, 0.18, 0.25, 0.35, 0.48, 0.67) and 14 different sample sizes (n ranging from 10 to 5,000, with $n_1 = n_2$) we simulated normally distributed two-sample datasets (with $\sigma^2 = 1$). For each dataset, we obtained CET p-values, JZS BFs and declared the result to be positive, negative or inconclusive accordingly. Results are presented in Fig 4 (and Figs 5–7), based on 10,000 distinct simulated datasets per scenario.

Several findings merit comment (unless otherwise noted, comments refer to results displayed in Fig 4):

- In this simulation study, the JZS-BF admits a very low frequentist type I error, recorded at most ≈0.01, for a sample size of $n = 110$. As the sample size increases, the frequentist type I error diminishes to a negligible level.

- The JZS-BF requires less data to reach a negative conclusion than the CET. However, with moderate to large sample sizes ($n = 100$ to 5,000) and small true mean differences ($\mu_d = 0$ to 0.25), both methods are approximately equally likely to deliver a negative conclusion.

- While the JZS-BF requires less data to reach a conclusion when the true mean difference is small ($\mu_d = 0$ to 0.25) (see how solid black curve drops more rapidly than the dashed grey line), there are scenarios in which larger sample sizes will surprisingly reduce the likelihood of obtaining a conclusive result (see how the solid black curve drops abruptly then rises slightly as n increases for $\mu_d = 0.07$, 0.09, 0.13, and 0.18.)

- The JZS-BF is always less likely to deliver a positive conclusion (see how the dashed blue curve is always higher than the solid blue curve). In the scenarios like those considered, JZS-BF may require larger sample sizes for reaching a positive conclusion and may be considered "less powerful" in a traditional frequentist sense.

- Fig 5 shows the probability that, for a given dataset, both methods agree as to whether the result is "positive", "negative", or "inconclusive". Across the range of effect sizes and samples sizes considered, the probability of agreement is always higher than 50%. For the strictly dichotomous choice between "positive" and "not positive", the probability that both methods

Fig 4. Simulation study results: Conclusions; with BF threshold of 3:1. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 3:1) and CET ($\alpha_1 = 0.05$, $\alpha_2 = 0.10$). Each panel displays the results of simulations with true mean difference, $\mu_d = 0$, 0.07, 0.09, 0.13, 0.18, 0.25, 0.35, 0.48, and 0.67.

agree is above 70% for all cases considered except a few scenarios in which effect size is small ($\mu_d = 0.07$, 0.09, 0.13, and 0.18) and sample size is large ($n > 500$).

- With the 6:1 BF threshold, results are similar. The main difference with this more stringent threshold is that much larger sample sizes are required in order to a achieve high probabilities of obtaining a conclusive result, see Fig 6; and agreement with CET is substantially lower, see Fig 7.

The results of the simulation study suggest that, in many ways, the JZS-BF and CET operate very similarly. Think of JZS-BF and CET as two pragmatically similar, yet philosophically different, tools for making "trichotomous significance-testing decisions".

A CET publication policy

Many researchers have put forth ideas for new publication policies aimed at addressing the issue of publication bias. There is a wide range of opinions on how to incorporate more null results into the published literature [102–104]. RR is one of many proposed "two-step" manuscript review schemes in which acceptance for publication is granted prior to obtaining the
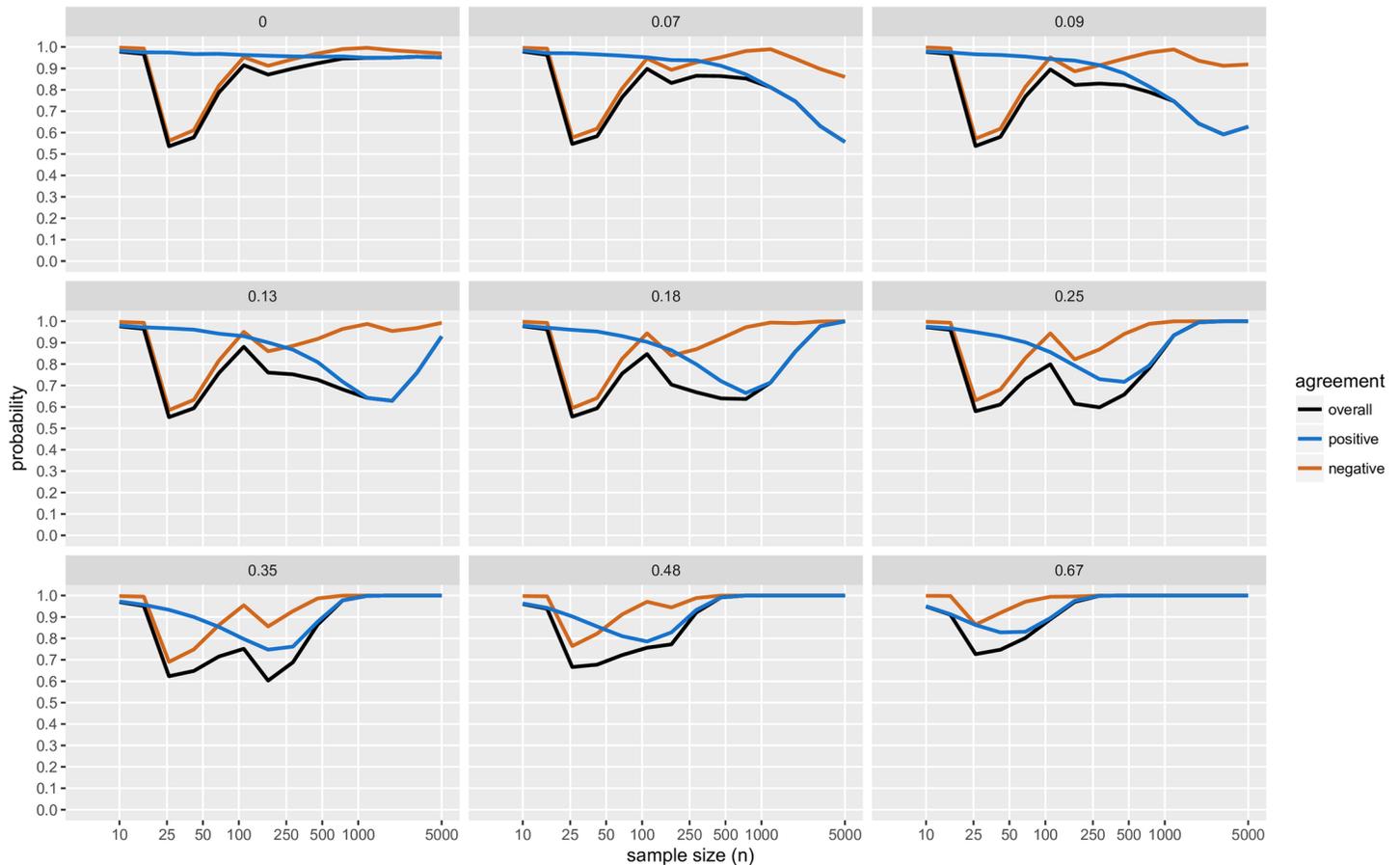
Fig 5. Simulation study results: Levels of agreement; with BF threshold of 3:1. The black line ("overall") indicates the probability that the conclusion reached by CET is in agreement with the conclusion reached by the JZS-BF Bayesian testing scheme; with JZS-BF with fixed sample size design, BF threshold of 3:1; and CET with $\alpha_1 = 0.05$, and $\alpha_2 = 0.10$. The blue line ("positive") indicates the probability that both methods agree with respect to whether or or not a result is positive. The crimson line ("negative") indicates the probability that both methods agree with respect to wether or or not a result is negative. Each panel displays the results of simulations with true mean difference, $\mu_d = 0, 0.07, 0.09, 0.13, 0.18, 0.25, 0.35, 0.48$, and 0.67.

https://doi.org/10.1371/journal.pone.0195145.g005

results, e.g. [24, 105–107]. Fig 8 (left panel) illustrates the RR procedure and Chambers et al. (2014) [31] provide an in-depth explanation answering a number of frequently asked questions. The RR policy has two central components: (1) pre-registration and (2) the "RR commitment to publish".

Pre-registration can be extremely beneficial as it reduces "researcher degrees of freedom" [51] and prevents, to a large degree, many questionable research practices including datadredging [108], the post-hoc fabrication of hypotheses, ("HARKing") [109], and *p*-hacking [110]. However, on its own, pre-registration does little to prevent publication bias. This is simply because pre-registration: (a1) cannot prevent authors from disregarding negative results [111]; (a2) does nothing to prevent reviewers and editors from rejecting studies for lack of significance; and (a3) does not guarantee that peer reviewers consider compliance with the pre-registered analysis plan [44, 112–114].

Consider the field of medicine as a case study. For over a decade, pre-registration of clinical trials has been required by major journals as a prerequisite for publication. Despite this
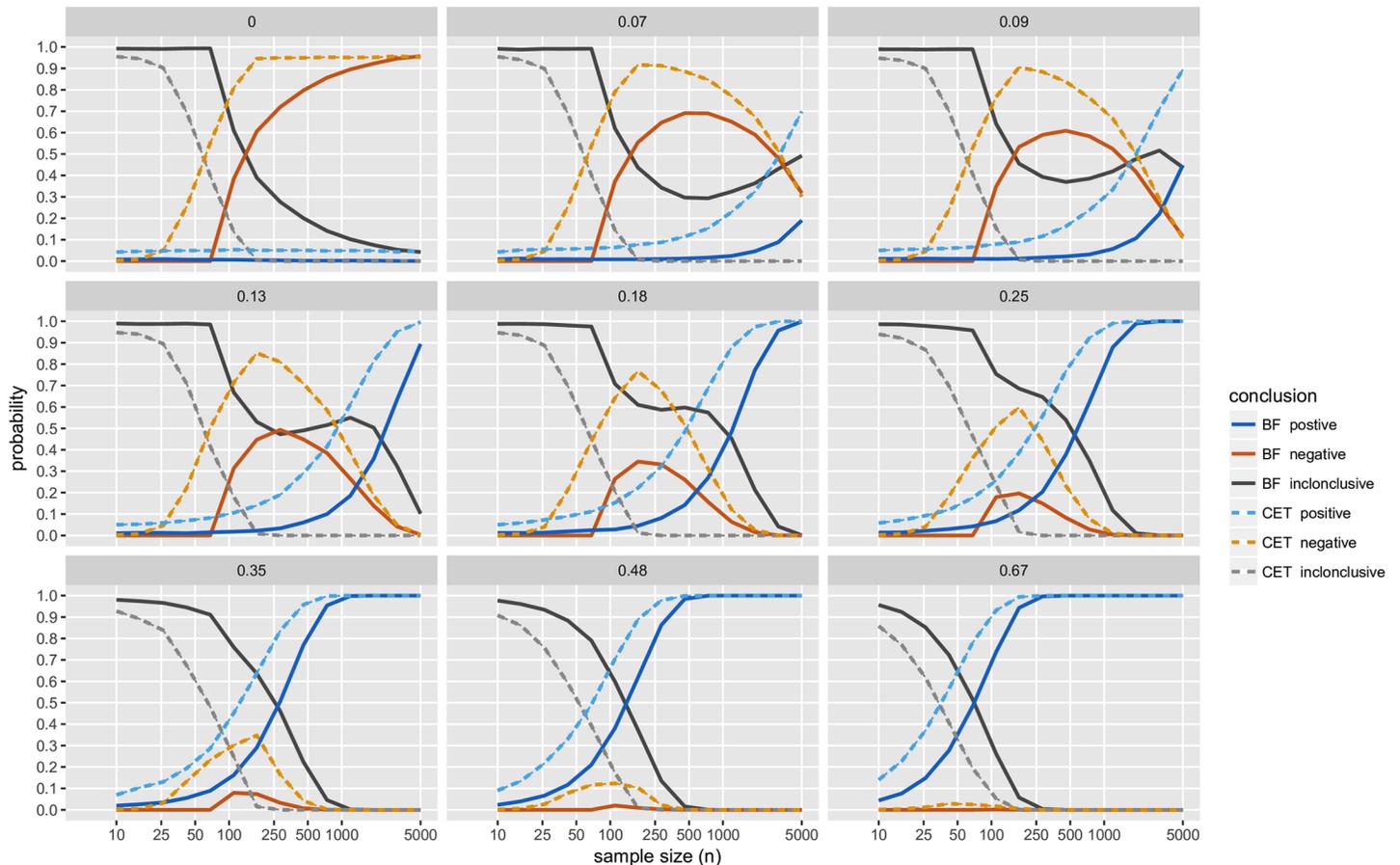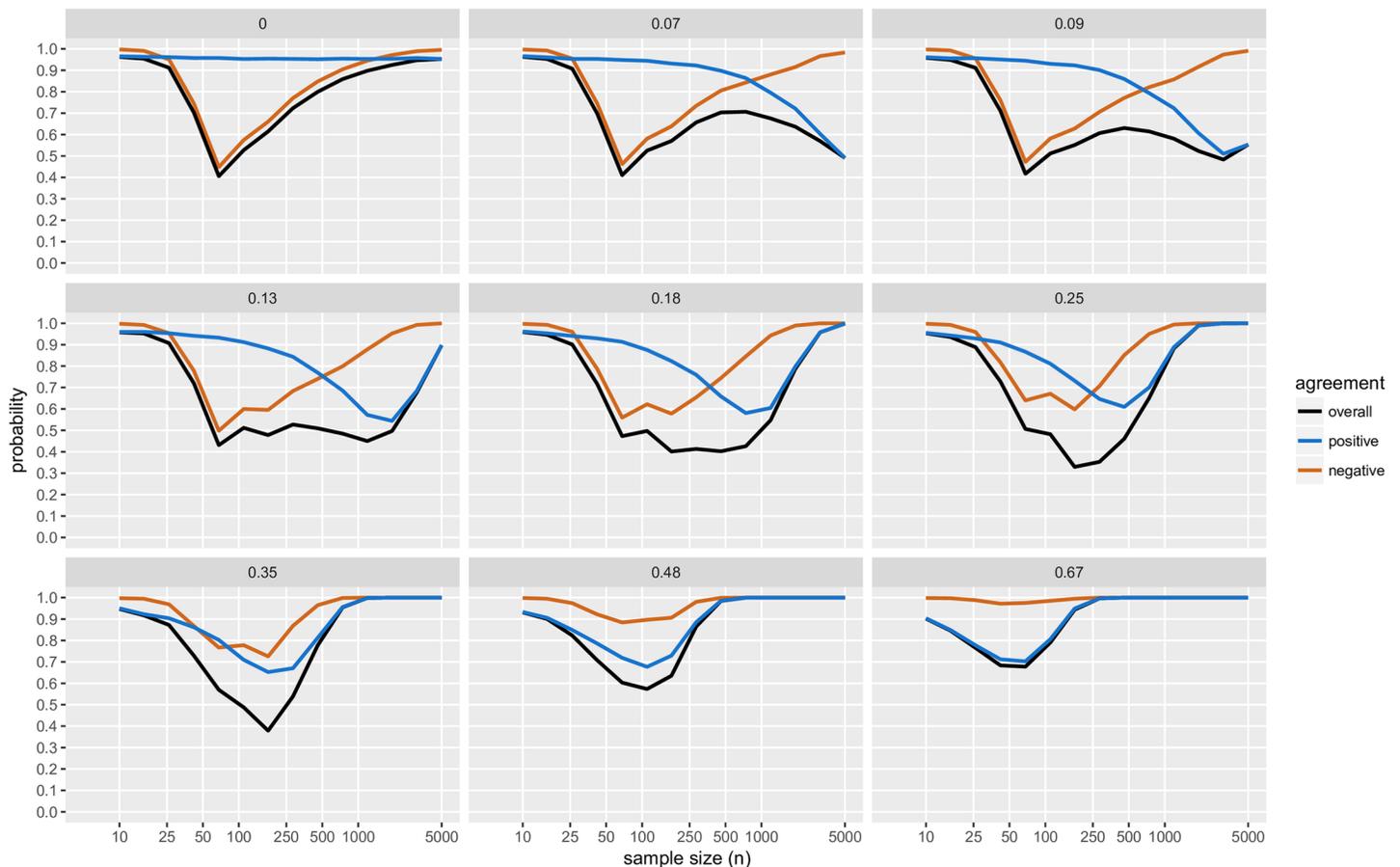
Fig 6. Simulation study results: Conclusions; with BF threshold of 6:1. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 6:1) and CET ($\alpha_1 = 0.05$, $\alpha_2 = 0.10$). Each panel displays the results of simulations with true mean difference, $\mu_d = 0$, 0.07, 0.09, 0.13, 0.18, 0.25, 0.35, 0.48, and 0.67.

https://doi.org/10.1371/journal.pone.0195145.g006

heralded policy change, selective outcome reporting remains ever prevalent [115–118]. (This being said, new 2017/2018 guidelines for the *clinicaltrial.gov* registry show much promise in addressing a1, a2, and a3; see [119].)

In order to prevent publication bias, RR complements pre-registration with a "commitment to publish". In practice this consists of an "in principle acceptance" policy along with the policy of publishing "withdrawn registration" (WR) studies. In order to counter authors who may simply shelve negative results following pre-registration (a1), RR journals commit to publishing the abstracts of all withdrawn studies as WR papers. By guaranteeing that, should a study follow its pre-registered protocol, it will be accepted for publication ("in principle acceptance"), RR prevents reviewers and editors from rejecting a study based on the results (a2). Finally, RR requires that a study is in strict compliance with the pre-registered protocol if it is to be published (a3). In order to keep a RR journal relevant (not simply full of inconclusive studies), RR requires, as part of registration, that a researcher commits to a sample size large enough of achieve 90% (in some cases 80%) statistical power. In a small number of RR journals, a Bayesian alternative option is offered. Instead of committing to a specific sample size, researchers commit to achieving a certain BF. (Note that currently in some RR journals, power requirements are less well defined.)

Fig 7. Simulation study results: Levels of agreement; with BF threshold of 6:1. The black line ("overall") indicates the probability that the conclusion reached by CET is in agreement with the conclusion reached by the JZS-BF Bayesian testing scheme; with JZS-BF with fixed sample size design, BF threshold of 6:1; and CET with $\alpha_1 = 0.05$, $and\alpha_2 = 0.10$. The blue line ("positive") indicates the probability that both methods agree with respect to wether or or not a result is positive. The crimson line ("negative") indicates the probability that both methods agree with respect to wether or or not a result is negative. Each panel displays the results of simulations with true mean difference, $\mu_d = 0, 0.07, 0.09, 0.13, 0.18, 0.25, 0.35, 0.48,$ and 0.67.

The policy we put forth here is not meant to be an alternative to the "pre-registration" component of RR. Its benefits are clear, and in our view is most often "worth the effort" [120]. If implemented properly, pre-registration should not "stifle exploratory work" [121]. Instead, what follows is an alternative to the second "commitment to publish" component.

Outline of a CET-based publication policy

Fig 8 (right panel) illustrates the steps of our proposed policy. What follows is a general outline.

Registration- In the first stage of a CET-based policy, before any data are collected, a researcher will register the intent to conduct a study with a journal's editor. As in the RR policy, this registration process will (1) detail the motivations, merits, and methods of the study, (2) list the defined outcomes and various hypotheses to be tested, and (3) state a target sample size, justified based on multiple grounds. However, unlike in the RR policy, failure to meet a given power threshold (e.g. failure to show calculations that give 80% power) will not generally

Registered Reports

CET publication policy
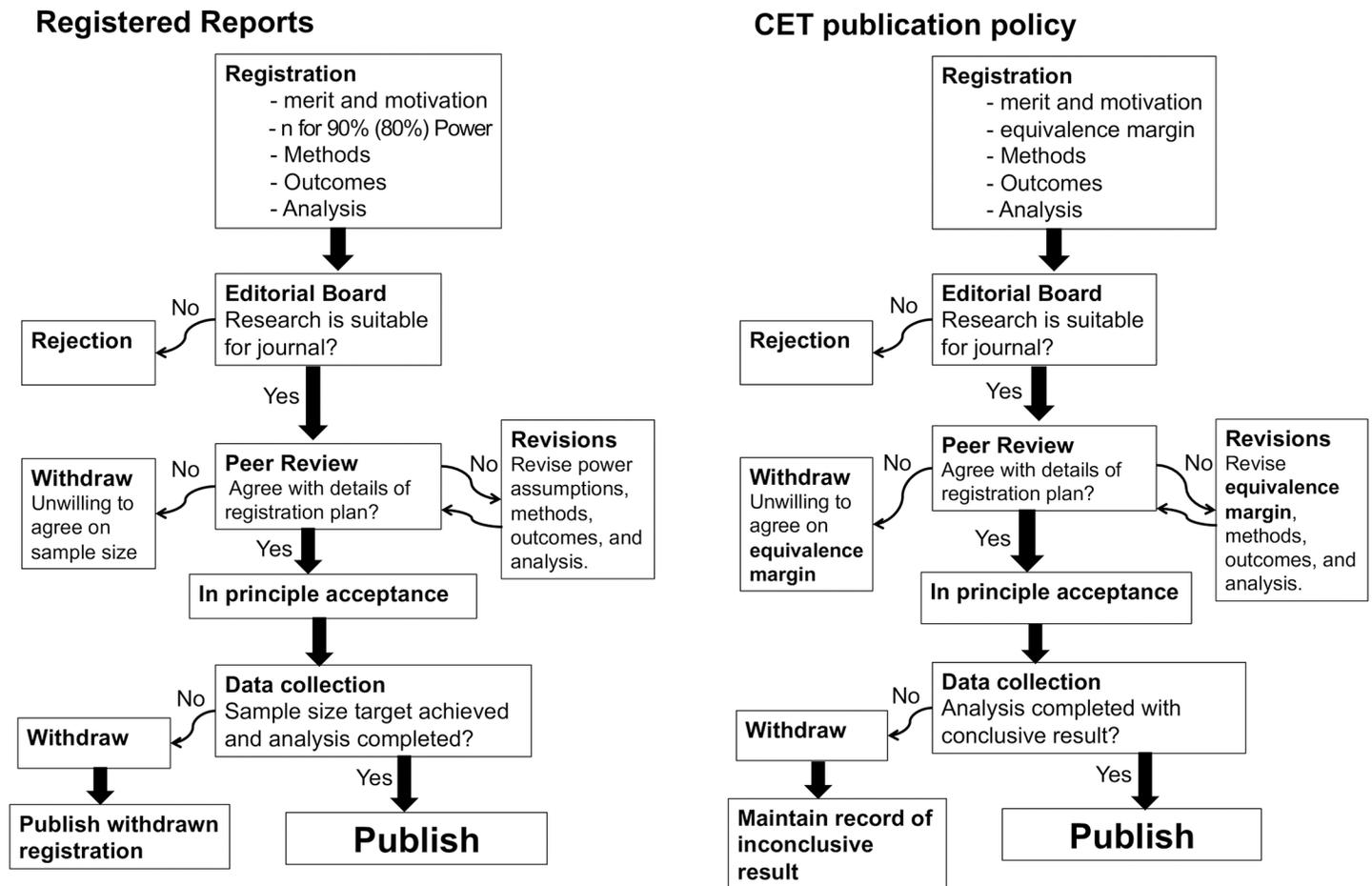
Fig 8. Left panel—The RR publication policy. The RR policy has two central components: (1) pre-registration and (2) the "RR commitment to publish". **Right panel— The CET publication policy.** The CET policy the registration will not require a sample size calculation showing a specific level of statistical power. Instead, the researcher will define an equivalence margin for each hypothesis test to be carried out. A study will only be published if a conclusive result is obtained.

https://doi.org/10.1371/journal.pone.0195145.g008

be grounds for rejection (see the Conclusion for further commentary on this point). Rather, registration under a CET-based policy will be based on defining an equivalence margin for each hypothesis test to be carried out. For example, if a researcher intends to fit a linear regression model with five explanatory variables, a margin should be defined for each of the five variables. The researcher will also need to note if there are plans for any sample size reassessments, interim and/or futility analyses.

Editorial and Peer Review- If the merits of the study satisfy the editorial board, the registration study plan will then be sent to reviewers to assess whether the methods for analysis are adequate and whether the equivalence margins are sufficiently narrow. Once the peer reviewers are satisfied (possibly after revisions to the registration plan), the journal will then agree, in principle, to accept the study for eventual publication, on condition that either a positive or a negative result is obtained.

Data collection- Armed with this "in principle acceptance", the researcher will then collect data in an effort to meet the established sample size target. Once the data are collected and analyses complete, the study will be published if and only if either a positive or negative result

is obtained as defined by the pre-specified equivalence margins. Inconclusive results will not generally be published thus protecting a journal from becoming a "dumping ground" for failed studies. In very rare circumstances, however, it may be determined that an inconclusive study offers a substantial contribution to the field and should therefore be considered for publication. As is required practice for good reporting, any failure to meet the target sample size should be stated clearly, along with a discussion of the reasons for failure and the consequences with regards to the interpretation of results [122].

This proposed policy is most similar to the RR Bayesian option outlined in Section 3. Journals only commit ("in principle acceptance") to publish conclusive studies (small p-value/ BF above or below a certain threshold) and no a-priori sample size requirements are forced upon a researcher. As noted in Section 3, we expect that the conclusions obtained under either policy will often be the same. The CET-based policy therefore represents a frequentist alternative to the RR-BF policy.

A journal may wish to require stricter or weaker requirements for publication of results and this can be done by setting thresholds for α_1 and α_2 accordingly, (e.g. stricter thresholds: $\alpha_1 = 0.01$, $\alpha_2 = 0.05$, $\Delta = 0.5s_p$; weaker thresholds: $\alpha_1 = 0.05$, $\alpha_2 = 0.10$, $\Delta = 0.5s_p$). Also, note that a p-value may be used in one way for the interpretation of results and another way to inform publication decisions. Recently, a large number of researchers [123] have publically suggested that the threshold for defining "statistical significance" be changed from $p < 0.05$ to $p < 0.005$. However, they are careful to emphasize that while this stricter threshold should change the description and interpretation of results, it should not change "standards for policy action nor standards for publication".

"Gaming the system"- When evaluating a new publication policy one should always ask how (and how easily) a researcher could -unintentionally or not- "game the system". For the CET-policy as described above, we see one rather obvious strategy.

Consider the following, admittedly extreme, example. A researcher submits 20 different study protocols for pre-registration each with very small sample size targets (e.g. $n = 8$). Suspend any disbelief, and suppose these studies all have good merit, are all well written, and are all well designed (besides being severely underpowered), and are therefore all granted in principle acceptance. Then, in the event that the null is true (i.e. $\mu = 0$) for all 20 studies (and with $\alpha_1 = 0.05$), it is expected that at least one study out of the 20 will obtain a positive result and thus be published. This is publication bias at its worst.

In order to discourage this unfortunate practice, we suggest making a researcher's history of pre-registered studies available to the public (e.g. using ORCID [124]). However, while potentially beneficial, we do not anticipate this type of action being necessary. With the CET-policy, it is no longer in a researcher's interest to "game the system".

Consider once again, our extreme example. The strategy has an approximately 64% chance of obtaining at least one publication at a cost of 20 submissions and a total of $160 = 8 \cdot 20$ observations. If the goal is to maximize the probability of being published [125], then it is far more efficient to submit a single study with $n = 160$, in which case there is an approximately 98% chance of obtaining a publication (i.e. with $\alpha_2 = 0.10$, $Pr(inconclusive|\mu = 0, \sigma^2 = 1, \Delta = 0.5) = 0.02$; 92% chance with $\alpha_2 = 0.05$, $Pr(inconclusive|\mu = 0, \sigma^2 = 1, \Delta = 0.5) = 0.08$).

Conclusion

Publication bias has been recognized as a serious problem for several decades now [126]. Yet, in many fields, it is only getting worse [127]. An investigation by Kühberger et al. (2014) [128] concludes that the "entire field of psychology" is now tainted by "pervasive publication bias".

There remains substantial disagreement on the merits of pre-registration and result-blind peer-review (e.g. [129–131]). Yet, all can agree that innovative publication policy prescriptions can be part of the solution to the "reproducibility crisis". While some call for dropping p-values and strict thresholds of evidence altogether, we believe that it is not worthwhile to fight "the temptation to discretize continuous evidence and to declare victory" [132]. Instead, the research community should embrace this "temptation" and work with it to achieve desirable outcomes. Indeed, one way to address the "practical difficulties that reviewers face with null results" [32] is to *further* discretize continuous evidence by means of equivalence testing and we submit that CET can be an effective tool for distinguishing those "high-quality null results" [133] worthwhile of publication.

Recently, a number of influential researchers have argued that to address low reliability, scientists, reviewers and regulators should "abandon statistical significance" [134]. (Somewhat ironically, in some fields, such as reinforcement learning, the currently proposed solution is just the opposite: the adoption of "significance metrics and tighter standardization of experimental reporting" [135].) We recognize that current publication policies, in which evidence is dichotomized (without any "ontological basis") may be highly unsatisfactory. However, one benefit to adopting distinct categories based on clearly defined thresholds (as in the CET-policy) is that one can assess, in a systematic way, the state of published research (in terms of reliability, power, reproducibility, etc.). While using a "more holistic view of the evidence" [134] to inform publication decisions may (or may not?) prove effective, "meta-research" under such a paradigm is clearly less feasible. As such, the question of effectiveness may perhaps never be adequately answered.

Bayesian approaches offer many benefits. However, we see three main drawbacks. First, adopting Bayesian testing requires a substantial paradigm shift. Since the interpretation of findings deemed significant by traditional NHST may differ with Bayes, some will no doubt be reluctant to accept the shift. With CET, the traditional usage and interpretation of the p-value remains unchanged, except in circumstances when one fails to reject the null. As such, CET does not change the interpretation of findings already established as significant. Indeed, CET simply "extend[s] the arsenal of confirmatory methods rooted in the frequentist paradigm of inference" [136]. Second, as we observed in our simulation study, Bayesian testing is potentially less powerful than NHST (in the traditional frequentist sense, with a fixed sample size design) and as such could require substantially larger sample sizes. Finally, we share the concern of Morey and Rouder (2011) [137] who write that Bayesian testing "provides no means of assessing whether rejections of the nil [null] are due to trivial or unimportant effect sizes or are due to more substantial effect sizes".

There are many potential areas for further research. Determining whether Δ and/or α_1 and/or α_2 should be chosen with consideration of the sample size is important and not trivial; related work includes Pérez et al. (2014) [138] who put forward a "Bayes/non-Bayes compromise" in which the α-level of a confidence interval changes with n. Issues which have proven problematic for standard equivalence testing must also be addressed for CET. These include multiplicity control [139] and potential problems with interpretation [140]. It would also be worthwhile considering whether CET is appropriate for testing for baseline balance [141]. The impact of a CET policy on meta-analysis should also be examined [142], i.e. how should one account for the exclusion of inconclusive results in the published literature when deriving estimates in a meta-analysis? Finally, Bloomfield et al. (2018) [143] consider how different publication policies will lead to "changes [in] the nature and timing of authors' investment in their work." It remains to be seen to what extent the CET-based policy, in which the decision to publish is based both on the up-front study registration and on the results obtained after analysis, will lead to a "shift from follow-up investment to up-front investment".

The publication policy outlined here should be welcomed by journal editors, researchers and all those who wish to see more reliable science. Research journals which wish to remain relevant and gain a high impact factor should welcome the CET-policy as it offers a mechanism for excluding inconclusive results while providing a space for potentially impactful negative studies. As such, embracing equivalence testing is an effective way to make publishing null results "more attractive" [144]. Researchers should be pleased with a policy that provides a "conditional in principle acceptance" and does not insist on specific sample size requirements that may not be feasible or desirable.

It must be noted that publication policies need not be implemented as "blanket rules" that must apply to all papers submitted to a given journal. The RR policy for instance, is often only one of several options offered to those submitting research to a participating journal. Furthermore, RR is not uniformly implemented across participating journals. Each journal has adapted the general RR framework to suit its objectives and the given research field. We envision similar possibilities for a CET policy in practice. Journals may choose to offer it as one of many options for submissions and will no doubt adapt the policy specifics as needed. For example, in a field where underpowered research is of particular concern, a journal may wish to implement a modified CET publication policy that does include strict sample size/power requirements.

The requirement to specify an equivalence margin prior to collecting data will have the additional benefit of forcing researchers and reviewers to think about what would represent a meaningful effect size before embarking on a given study. While there will no doubt be pressure on researchers to "p-hack" in order to meet either the α_1 or α_2 threshold, this can be discouraged by insisting that an analysis strictly follows the pre-registered analysis plan. Adopting strict thresholds for significance can also act as a deterrent. Finally, we believe that the CET-policy will improve the reliability of published science by not only allowing for more negative research to be published, but by modifying the incentive structure driving research [145].

Using an optimality model, Higginson and Munafò (2016) [146] conclude that, given current incentives, the rational strategy of a scientist is to "focus almost all of their research effort on underpowered exploratory work [. . . and] carry out lots of underpowered small studies to maximize their number of publications, even though this means around half will be false positives." This result is in line with the views of many (e.g. [147–149]), and provides the basis for why statistical power in many fields has not improved [150] despite being highlighted as an issue over five decades ago [151]. A CET-based policy may provide the incentive scientists need to pursue higher statistical power. If CET can change the incentives driving research, the reliability of science will be further improved. More research on this question (i.e. "meta-research") is needed.

Appendix 1

CET operating characteristics and sample size calculations

It is worth noting that there is already a large literature on TOST and non-inferiority testing. See Walker and Nowacki (2011) [62] and Meyners (2012) [63] for overviews that cover the basics as well as more subtle issues. There are also many proposed alternatives to TOST for equivalence testing, what are known as "the emperor's new tests" [152]. These alternative are offered as marginally more powerful options, yet are more complex and are not widely used. Finally, note that one of the reasons equivalence testing has not been widely used, is that until recently, it has not been included in popular statistical software packages. Packages such as JAMOVI and TOSTER are now available.

Our proposal to systematically incorporate equivalence testing into a two-stage testing procedure has not been extensively pursued elsewhere (one exception may be [76]) and there has not been any discussion of how such a testing procedure could facilitate publication decisions for peer-review journals. One reason for this is a poor understanding of the *conditional* equivalence testing strategy whereby testing traditional non-equivalence is followed conditionally (if one fails to reject the null), by testing equivalence. In fact, whether or not such a two-stage approach is beneficial has been somewhat controversial. As the sample-size is typically determined based only on the primary test, the power of the secondary equivalence (non-inferiority) test is not controlled, thereby potentially increasing the false discovery rate [153]. Koyama and Westfall (2005) [154] investigate and conclude that, in most situations, such concern is unwarranted. A better understanding of the operating characteristics of CET is required.

What follows is a brief overview of how to calculate the probabilities of obtaining each of the three conclusions (positive, negative, and inconclusive) for CET for two-sample testing of normal data with equal variance (as described in the five steps listed in the text). More in-depth related work includes Shieh (2016) [155] who establishes exact power calculations for TOST for equivalence of normally distributed data; da Silva et al. (2009) [65] who review power and sample size calculations for equivalence testing of binary and time-to-event outcomes; and Zhu (2017) [156] who considers sample size calculations for equivalence testing of poisson and negative binomial rates.

As before, let the true population mean difference be $\mu_d = \mu_1 - \mu_2$ and the true population variance equal σ^2. Let $\sigma^* = \sigma\sqrt{(1/n_1 + 1/n_2)}$ and $s^* = s_p\sqrt{(1/n_1 + 1/n_2)}$. Fig 1 (right panel) illustrates how each of the three conclusions can be reached based on the values of s^* and $\hat{\mu}_d = \bar{x}_1 - \bar{x}_2$ obtained from the data. For this plot, $n = 90$ (with $n_1 = n_2$), $\Delta = 0.5$, $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$. The black lettered points correspond to the scenarios of Fig 1 (left panel). The interior diamond, $\diamond$, (with corners located at [0, 0], $[-\Delta/(t^*_{\alpha_2}/t^*_{\alpha_1/2} + 1), \Delta/(t^*_{\alpha_1/2} + t^*_{\alpha_2})], [0, \Delta/t^*_{\alpha_2}], [\Delta/(t^*_{\alpha_2}/t^*_{\alpha_1/2} + 1), \Delta/(t_1 + t_2)])$ covers the values for which a negative conclusion is obtained. A positive conclusion corresponds to when $|\bar{x}_1 - \bar{x}_2|$ is large and s^* is relatively small. Note how Δ and σ impact the conclusion. If the equivalence margin is sufficiently wide, one will have $Pr(inconclusive)$ approach zero as σ approaches zero. Indeed, the ratio of Δ/σ determines, to a large extent, the probability of a negative result. If the equivalence margin is sufficiently narrow and the variance relatively large (i.e. Δ/σ is very small), one will have $Pr(negative) \approx 0$.

Let us assume for simplicity that $n_1 = n_2$. Then, the sampling distributions of the sample mean and sample variance are well established: $\hat{\mu}_d \sim N(\mu_d, \sigma^{*2})$ and $\frac{(n-2)s_p^2}{\sigma^2} \sim \chi^2_{n-2}$. Therefore, given fixed values for μ and σ^2, we can calculate the probability of obtaining a positive result, $Pr(positive)$. In Fig 1 (right panel), $Pr(positive)$ equals the probability of $\hat{\mu}_d$ and s^* falling into either the left or right "positive" corners and is calculated (as in a usual power calculation for NHST):

$$Pr(positive; \mu_d, \sigma) = \left(1 - F_{n-2;\frac{\mu_d}{\sigma^*}}(t^*_{\alpha_1/2})\right) + F_{n-2;\frac{\mu_d}{\sigma^*}}(-t^*_{\alpha_1/2}) \tag{1}$$

where $F_{df,ncp}(x)$ is the cdf of the non-central t distribution with df degrees of freedom and non-centrality parameter ncp.

One can calculate the probability of obtaining a negative result, $Pr(negative)$, as the probability of $\hat{\mu}_d$ and s^* falling into the "negative" diamond, $\diamond$. Since $\hat{\mu}_d$ and s^* are independent statistics, we can write their joint density as the product of a normal probability density function, $f_N()$, and a chi-squared probability density function, $f_{\chi^2}()$. However, the resulting double integral will remain difficult to evaluate algebraically over the boundary, $\diamond$. Therefore, the

probability is best approximated numerically, for example, by Monte Carlo integration, as follows:

$$Pr(negative; \mu_d, \sigma) = \int \int_\diamond f_N(u; \mu_d, \sigma^{*2}) f_{\chi^2}(v; n-2) du dv$$

$$= \int_\diamond (\Phi(h_2(v); \mu_d, \sigma^*) - \Phi(h_1(v); \mu_d, \sigma^*)) f_{\chi^2}(v; n-2) dv$$

$$\approx \sum_{j=1}^{M} (\Phi(h_2(c_j); \mu_d, \sigma^*) - \Phi(h_1(c_j); \mu_d, \sigma^*))/M$$

where $\Phi()$ is the normal cdf and Monte Carlo draws from a chi-squared distribution provide $c_j = \sqrt{\sigma^2 q^{[j]}/(n-2)}$, with $q^{[j]} \sim \chi^2_{n-2}$ for $j = 1, \ldots, M$. The left and right-hand boundaries of the diamond-shaped "negative region", are defined by $h_1(c_j) = min(0, max(+c_j t_2 - \Delta, -c_j t_1))$ and $h_2(c_j) = max(0, min(-c_j t_2 + \Delta, +c_j t_1))$.

Defining the boundary with $h_1()$ and $h_2()$ allows for three distinct cases as seen in Fig 1 (right panel):

(1) $s^* > \Delta/t^*_{\alpha_2}$, in which case $h_1(s^*) = h_2(s^*) = 0$;

(2) $\Delta/(t^*_{\alpha_1/2} + t^*_{\alpha_2}) < s^* < \Delta/t^*_{\alpha_2}$, in which case $h_1(s^*) = -\Delta + t^*_{\alpha_2}s^*$ and $h_2(s^*) = \Delta - t^*_{\alpha_2}s^*$; and

(3) $s^* < \Delta/(t^*_{\alpha_1/2} + t^*_{\alpha_2})$, in which case $h_1(s^*) = -t^*_{\alpha_1/2}s^*$ and $h_2(s^*) = +t^*_{\alpha_1/2}s^*$.

When the equivalence boundaries are defined as a function of s_p (e.g. $\Delta = q s_p$), the calculations are somewhat different; Fig 9 illustrates. In particular, a negative conclusion requires: $q s^*/\sqrt{1/n_1 + 1/n_2} > |\hat{\mu}_d \pm s^* t^*_{\alpha_2}|$ (i.e. the $(1 - 2\alpha_2)\%$ C.I. is entirely within $[-\Delta, \Delta]$). As such, for a given sample size, it will only be possible to obtain a negative result if $q > t^*_{\alpha_2}\sqrt{1/n_1 + 1/n_2}$. Likewise, an inconclusive result will only be possible if $(q/\sqrt{1/n_1 + 1/n_2}) < (t^*_{\alpha_1/2} + t^*_{\alpha_2})$.

In order to determine an appropriate sample size for CET, one must replace μ_d with an a-priori estimate, $\tilde{\mu}_d$, the "anticipated effect size", and replace σ^2 with an a-priori estimate, $\tilde{\sigma}^2$, the "anticipated variance". Then one might be interested in calculating six values: the probabilities of obtaining each of the three possible results (positive, negative and inconclusive) under two hypothetical scenarios, (1) where $\mu_d = 0$, and (2) where μ_d equal to $\tilde{\mu}_d$, the value expected given results in the literature. One might also be interested in a hybrid approach whereby one specifies a hybrid null and alternative distribution. Since the objective of any study should be to obtain a conclusive result, sample size could also be calculated with the objective to maximize the likelihood of success, i.e. to minimize $Pr(inconclusive)$.

The quantity:

$$Pr(success) = 1 - \frac{1}{2}Pr(inconclusive; \mu_d = \tilde{\mu}_d, \tilde{\sigma}^2) - \frac{1}{2}Pr(inconclusive; \mu_d = 0, \tilde{\sigma}^2) \qquad (2)$$

represents the probability of a "successful" study under the assumption that the null ($\mu_d = 0$) and alternative (with specified $\mu_d = \tilde{\mu}_d$) are equally likely. (Note that both false positive-, and false negative- studies in this equation are considered "successful".) This weighted average could be considered a simple version of what is known as "assurance" [157]. Fig 10 shows how consideration of $Pr(success)$ as the criteria for determining sample size attenuates the effect of $\tilde{\mu}_d$ on the required sample size. Suppose one calculates that the required sample size is $n = 1,000$ based on the desire for 90% statistical power and the belief that $\tilde{\sigma}^2 = 1$ and $\mu_d = 0.205$, with $n_1 = n_2$ as before. This corresponds to a 61% probability of success for $\Delta = 0.1025$
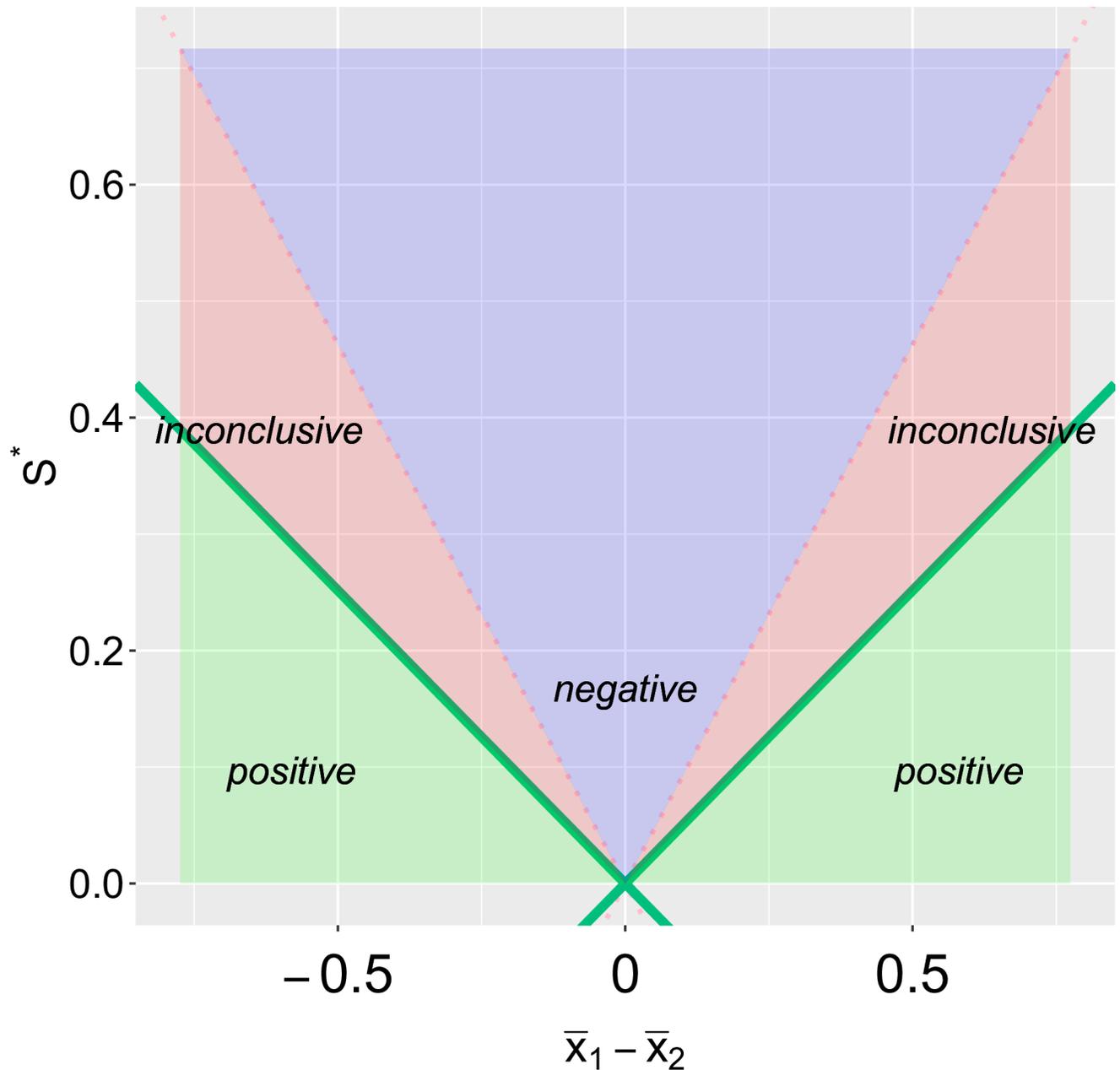
Fig 9. Rejection regions with the equivalence boundaries are defined as a function of s_p. Let $n = 90$, $\Delta = 0.5s_p$, $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$. The three conclusions, positive, negative, inconclusive correspond to the three areas shaded in green, blue and red respectively.

https://doi.org/10.1371/journal.pone.0195145.g009

$(= \frac{1}{2}\mu_d)$. If the true variance is slightly larger than anticipated, $\sigma^2 = 1.25$, and the difference in means smaller, $\mu_d = 0.123$, the actual sample size needed for 90% power is in fact $n = 3,476$, while the actual sample size needed for $Pr(success|\Delta, d, s) = 61\%$ is $n = 1,770$. On the other hand, if the true variance is slightly smaller than anticipated, $\sigma^2 = 0.75$ and the difference in means greater, $\mu_d = 0.33$, the actual sample size needed for 90% power is only $n = 288$, while the actual sample size needed for $Pr(success|\delta, d, s) = 61\%$ is $n = 638$. It follows that, if one has
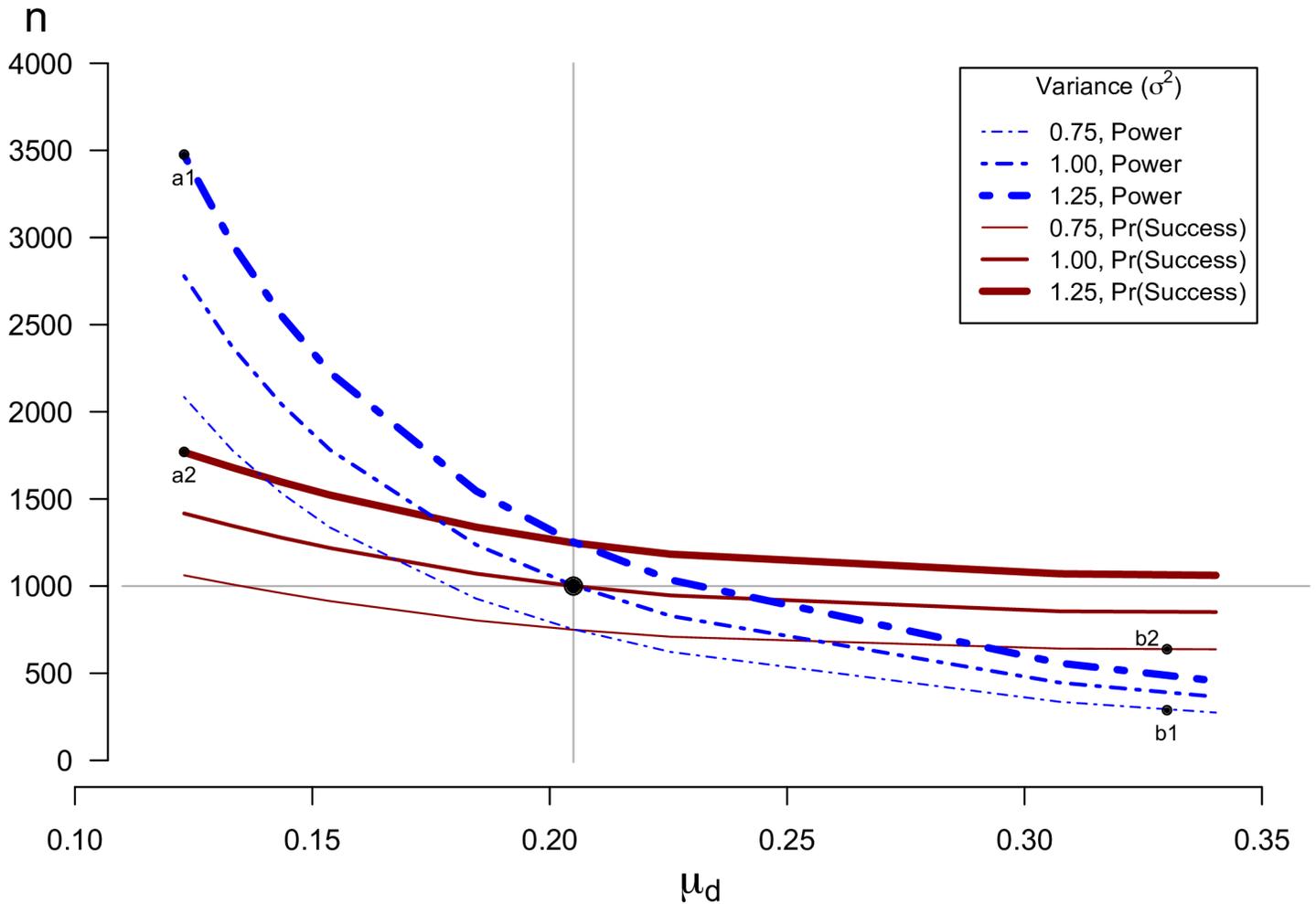
Fig 10. Required sample size for 90% Power, or Pr(Success) = 61%. Suppose the anticipated effect size is $\tilde{\mu}_d = 0.205$, the anticipated variance is $\tilde{\sigma}^2 = 1$ and the equivalence margin is pre-specified with $\Delta = 0.1025 \ (= \frac{1}{2}\tilde{\mu}_d)$. Then, based on the desire for Pr(positive) = 90% (i.e. "power" = 0.90) (dashed blue lines) or a 61% "probability of success" (solid red lines) a sample size of $n = 1,000$ (with $n_1 = n_2$) would be required. If the true variance is slightly larger than anticipated, $\sigma^2 = 1.25$, and the effect size smaller, $\mu_d = 0.123$, the actual sample size needed for 90% power is in fact 3,476, while the actual sample size needed for $Pr(success) = 61\%$ is 1,770; see points "a1" and "a2". On the other hand, if the true variance is slightly smaller than anticipated, $\sigma^2 = 0.75$ and the effect size greater, $\mu_d = 0.33$, the actual sample size needed for 90% power is only 288, while the actual sample size needed for $Pr(success|\delta, d, s) = 61\%$ is 638; see points "b1" and "b2".

https://doi.org/10.1371/journal.pone.0195145.g010

little certainty in μ_d and σ^2, calculating the required sample size with consideration of $Pr(success)$ may be less risky.

For related work on statistical power, see Shao et al. (2008) [158] who propose a hybrid Bayesian-frequentist approach to evaluate power for testing both superiority and non-inferiority. Jia and Lynn (2015) [159] discuss a related sample size planning approach that considers both statistical significance and clinical significance. Finally, Jiroutek et al. (2003) [160] advocate that, rather than calculate statistical power (i.e. the probability of rejecting $\theta = \theta_0$ should the alternative be true), one should calculate the probability that the width of a confidence interval is less than a fixed constant and the null hypothesis is rejected, given that the confidence interval contains the true parameter.

Appendix 2

JZS Bayes Factor details and additional simulation study results

Note that, for the Bayes Factor, the null hypothesis, H_0, corresponds to $\mu_d = 0$; and the alternative, H_1, corresponds to $\mu_d \neq 0$. The JZS testing scheme involves placing a normal prior on $\eta = (\mu_2 - \mu_1)/\sigma$, $\eta \sim Normal(0, \sigma_\eta^2)$, and for the hyper-parameter σ_η, placing an inverse chi-squared prior, $\sigma_\eta^2 \sim inv.\chi^2(1)$. Integrating out σ_η shows that this is equivalent to having a Cauchy prior, $\eta \sim Cauchy$. We can write the JZS Bayes Factor in terms of the standard t-statistic, $T = (\bar{x}_1 - \bar{x}_2)/s^*$, with $n^* = n_1 n_2/(n_1 + n_2)$, and $v = n_1 + n_2 - 2$, as follows:

$$B_{01} = \frac{\left(1 + \frac{T^2}{v}\right)^{-(v+1)/2}}{\int_0^\infty (1 + n^* g)^{-1/2} \left(1 + \frac{T^2}{(1+n^* g)v}\right)^{-(v+1)/2} (2\pi)^{-1/2} g^{-3/2} e^{-1/(2g)} dg} \tag{3}$$

See Rouder et al. (2009) [93] for additional details. Figs 4–7 plot results from the simulation study. Note that while we resorted to simulation, since the B_{01} is a function of the t statistic, it would be possible to use numerical methods to compute its operating characteristics.

Supporting information

S1 File. R-code is available to reproduce all figures within this paper on the Open Science Framework (OSF) repository under the "Conditional Equivalence Testing" project (Identifiers: DOI 10.17605/OSF.IO/UXN2Q | ARK c7605/osf.io/uxn2q).
(ZIP)

Acknowledgments

Author Contributions

Conceptualization: Harlan Campbell.

Investigation: Harlan Campbell, Paul Gustafson.

Methodology: Harlan Campbell, Paul Gustafson.

Supervision: Paul Gustafson.

Writing – original draft: Harlan Campbell, Paul Gustafson.

Writing – review & editing: Harlan Campbell, Paul Gustafson.

References

1. Ioannidis JP. Why most published research findings are false. PLoS Medicine. 2005; 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722

2. Goodman S, Greenland S. Assessing the unreliability of the medical literature: a response to 'Why most published research findings are false?'. Johns Hopkins University, Department of Biostatistics; working paper 135. 2007.

3. Leek JT, Jager LR. Is most published research really false? Annual Review of Statistics and Its Application. 2017; 4:109–122. https://doi.org/10.1146/annurev-statistics-060116-054104

4. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One. 2009; 4(5):e5738. https://doi.org/10.1371/journal.pone.0005738 PMID: 19478950

5. Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology. 2015; 37(1):1–2. https://doi.org/10.1080/01973533.2015.1012991

6. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. The American Statistician. 2016; 70(2):129–133. https://doi.org/10.1080/00031305.2016.1154108

7. Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. American journal of epidemiology. 2017; 186(6):627–635. https://doi.org/10.1093/aje/kwx261 PMID: 28938715

8. Hofmann MA. Null hypothesis significance testing in simulation. In: Proceedings of the 2016 Winter Simulation Conference. IEEE Press; 2016. p. 522–533.

9. Szucs D, Ioannidis J. When null hypothesis significance testing is unsuitable for research: a reassessment. Frontiers in human neuroscience. 2017; 11:390. https://doi.org/10.3389/fnhum.2017.00390 PMID: 28824397

10. Cumming G. The new statistics: Why and how. Psychological Science. 2014; 25(1):7–29. https://doi.org/10.1177/0956797613504966 PMID: 24220629

11. Lew MJ. To *p* or not to *p*: On the evidential nature of *p*-values and their place in scientific inference. arXiv preprint arXiv:13110081. 2013.

12. Chavalarias D, Wallach JD, Li AHT, Ioannidis JP. Evolution of reporting *p*-values in the biomedical literature, 1990-2015. JAMA. 2016; 315(11):1141–1148. https://doi.org/10.1001/jama.2016.1952 PMID: 26978209

13. Gelman A. Commentary: *p*-values and statistical practice. Epidemiology. 2013; 24(1):69–72. https://doi.org/10.1097/EDE.0b013e31827886f7 PMID: 23232612

14. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. JAMA. 1992; 267(3):374–378. https://doi.org/10.1001/jama.267.3.374 PMID: 1727960

15. Reysen S. Publication of nonsignificant results: a survey of psychologists' opinions. Psychological Reports. 2006; 98(1):169–175. https://doi.org/10.2466/pr0.98.1.169-175 PMID: 16673970

16. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. Science. 2014; 345(6203):1502–1505. https://doi.org/10.1126/science.1255484 PMID: 25170047

17. Doshi P, Dickersin K, Healy D, Vedula SS, Jefferson T. Restoring invisible and abandoned trials: a call for people to publish the findings. BMJ. 2013; 346:f2865. https://doi.org/10.1136/bmj.f2865 PMID: 23766480

18. Hartung J, Cottrell JE, Giffin JP. Absence of evidence is not evidence of absence. Anesthesiology: The Journal of the American Society of Anesthesiologists. 1983; 58(3):298–299. https://doi.org/10.1097/00000542-198303000-00033

19. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. BMJ. 1995; 311(7003):485. https://doi.org/10.1136/bmj.311.7003.485 PMID: 7647644

20. Greenwald AG. Consequences of prejudice against the null hypothesis. Psychological Bulletin. 1975; 82(1):1–20. https://doi.org/10.1037/h0076157

21. Zumbo BD, Hubley AM. A note on misconceptions concerning prospective and retrospective power. Journal of the Royal Statistical Society: Series D (The Statistician). 1998; 47(2):385–388.

22. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician. 2001; 55(1):19–24. https://doi.org/10.1198/000313001300339897

23. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. Annals of Epidemiology. 2012; 22(5):364–368. https://doi.org/10.1016/j.annepidem.2012.02.007 PMID: 22391267

24. Walster GW, Cleary TA. A proposal for a new editorial policy in the social sciences. The American Statistician. 1970; 24(2):16–19. https://doi.org/10.2307/2681924

25. Sterling TD, Rosenbaum WL, Weinkam JJ. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. The American Statistician. 1995; 49(1):108–112. https://doi.org/10.2307/2684823

26. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2008; 3(8):e3081. https://doi.org/10.1371/journal.pone.0003081 PMID: 18769481

27. Suñé P, Suñé JM, Montoro JB. Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. PLoS One. 2013; 8(1):e54583. https://doi.org/10.1371/journal.pone.0054583 PMID: 23382919

28. Greve W, Bröder A, Erdfelder E. Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. European Psychologist. 2013; 18(4):286–294. https://doi.org/10.1027/1016-9040/a000144

29. Nosek BA, Ebersole CR, DeHaven A, Mellor D. The Preregistration Revolution. Open Science Framework, preprint. 2017.

30. Chambers CD, Dienes Z, McIntosh RD, Rotshtein P, Willmes K. Registered reports: realigning incentives in scientific publishing. Cortex. 2015; 66:A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022 PMID: 25892410

31. Chambers CD, Feredoes E, Muthukumaraswamy SD, Etchells P. Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. AIMS Neuroscience. 2014; 1(1):4–17. https://doi.org/10.3934/Neuroscience.2014.1.4

32. Findley MG, Jensen NM, Malesky EJ, Pepinsky TB. Can results-free review reduce publication bias? The results and implications of a pilot study. Comparative Political Studies. 2016; 49(13):1667–1703. https://doi.org/10.1177/0010414016655539

33. Sackett DL, Cook DJ. Can we learn anything from small trials? Annals of the New York Academy of Sciences. 1993; 703(1):25–32. https://doi.org/10.1111/j.1749-6632.1993.tb26331.x PMID: 8192301

34. Bacchetti P. Peer review of statistics in medical research: the other problem. BMJ. 2002; 324 (7348):1271–1273. https://doi.org/10.1136/bmj.324.7348.1271 PMID: 12028986

35. Aycaguer LCS, Galbán PA. Explicación del tamaño muestral empleado: una exigencia irracional de las revistas biomédicas. Gaceta Sanitaria. 2013; 27(1):53–57. https://doi.org/10.1016/j.gaceta.2012.01.017

36. Matthews JN. Small clinical trials: are they all bad? Statistics in Medicine. 1995; 14(2):115–126. https://doi.org/10.1002/sim.4780140204 PMID: 7754260

37. Borm GF, den Heijer M, Zielhuis GA. Publication bias was not a good reason to discourage trials with low power. Journal of Clinical Epidemiology. 2009; 62(1):47–53. https://doi.org/10.1016/j.jclinepi.2008.02.017 PMID: 18620841

38. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. The Lancet. 2005; 365(9467):1348–1353. https://doi.org/10.1016/S0140-6736(05)61034-3

39. Bland JM. The tyranny of power: is there a better way to calculate sample size? BMJ. 2009; 339: b3985. https://doi.org/10.1136/bmj.b3985 PMID: 19808754

40. Albers C, Lakens D. When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. Journal of Experimental Social Psychology. 2018; 74:187–195. https://doi.org/10.1016/j.jesp.2017.09.004

41. Vasishth S, Gelman A. The illusion of power: How the statistical significance filter leads to overconfident expectations of replicability. arXiv preprint arXiv:170200556. 2017.

42. Djulbegovic B, Kumar A, Magazin A, Schroen AT, Soares H, Hozo I, et al. Optimism bias leads to inconclusive results—an empirical study. Journal of Clinical Epidemiology. 2011; 64(6):583–593. https://doi.org/10.1016/j.jclinepi.2010.09.007 PMID: 21163620

43. Chalmers I, Matthews R. What are the implications of optimism bias in clinical research? The Lancet. 2006; 367(9509):449–450. https://doi.org/10.1016/S0140-6736(06)68153-1

44. Chan AW, Hróbjartsson A, Jørgensen KJ, Gøtzsche PC, Altman DG. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. BMJ. 2008; 337:a2299. https://doi.org/10.1136/bmj.a2299 PMID: 19056791

45. Dienes Z, Mclatchie N. Four reasons to prefer Bayesian analyses over significance testing. Psychonomic Bulletin & Review. 2017; p. 1–12.

46. Kruschke JK, Liddell TM. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review. 2017; p. 1–29.

47. Wagenmakers EJ. A practical solution to the pervasive problems of p-values. Psychonomic Bulletin & Review. 2007; 14(5):779–804. https://doi.org/10.3758/BF03194105

48. Dienes Z. How Bayes factors change scientific practice. Journal of Mathematical Psychology. 2016; 72:78–89. https://doi.org/10.1016/j.jmp.2015.10.003

49. Etz A, Vandekerckhove J. A Bayesian perspective on the reproducibility project: Psychology. PLoS One. 2016; 11(2):e0149794. https://doi.org/10.1371/journal.pone.0149794 PMID: 26919473

50. Zhang X, Cutter G, Belin T. Bayesian sample size determination under hypothesis tests. Contemporary Clinical Trials. 2011; 32(3):393–398. https://doi.org/10.1016/j.cct.2010.12.012 PMID: 21199689

51. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011; 22 (11):1359–1366. https://doi.org/10.1177/0956797611417632 PMID: 22006061

52. Jonas KJ, Cesario J. Submission Guidelines for Authors, Comprehensive Results in Social Psychology; 2017. www.tandf.co.uk/journals/authors/rrsp-submission-guidelines.pdf

53. BMC Biology Editorial. BMC Biology—Registered Reports; March 23, 2018. https://bmcbiol. biomedcentral.com/about/registered-reports

54. Journal of Cognition Editorial. BMC Biology—Registered Reports; March 23, 2018. https://www. journalofcognition.org/about/registered-reports/

55. Lakens D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science. 2017; 8:355–362. https://doi.org/10.1177/1948550617697177 PMID: 28736600

56. Ocaña i Rebull J, Sánchez Olavarría MP, Sánchez A, Carrasco Jordan JL. On equivalence and bio-equivalence testing. Sort. 2008; 32(2):151–176.

57. Goeman JJ, Solari A, Stijnen T. Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. Statistics in Medicine. 2010; 29(20):2117–2125. https://doi.org/10.1002/ sim.4002 PMID: 20658478

58. Zhao G. Considering both statistical and clinical significance. International Journal of Statistics and Probability. 2016; 5(5):16. https://doi.org/10.5539/ijsp.v5n5p16

59. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. Annals of Internal Medicine. 2006; 145(1):62–69. https://doi.org/10.7326/0003-4819-145-1-200607040-00011 PMID: 16818930

60. Greene CJ, Morland LA, Durkalski VL, Frueh BC. Noninferiority and equivalence designs: issues and implications for mental health research. Journal of Traumatic Stress. 2008; 21(5):433–439. https://doi. org/10.1002/jts.20367 PMID: 18956449

61. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: A tutorial. pre-print Retrieved from the Open Science Framework. 2017.

62. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. Journal of General Internal Medicine. 2011; 26(2):192–196. https://doi.org/10.1007/s11606-010-1513-8 PMID: 20857339

63. Meyners M. Equivalence tests–A review. Food Quality and Preference. 2012; 26(2):231–245. https:// doi.org/10.1016/j.foodqual.2012.05.003

64. Chen JJ, Tsong Y, Kang SH. Tests for equivalence or noninferiority between two proportions. Drug Information Journal. 2000; 34(2):569–578. https://doi.org/10.1177/009286150003400225

65. da Silva GT, Logan BR, Klein JP. Methods for equivalence and noninferiority testing. Biology of Blood and Marrow Transplantation. 2009; 15(1):120–127. https://doi.org/10.1016/j.bbmt.2008.10.004

66. Wiens BL, Iglewicz B. Design and analysis of three treatment equivalence trials. Controlled Clinical Trials. 2000; 21(2):127–137. https://doi.org/10.1016/S0197-2456(99)00052-5 PMID: 10715510

67. Wellek S. Testing statistical hypotheses of equivalence and noninferiority. CRC Press; 2010.

68. Dixon PM, Pechmann JH. A statistical test to show negligible trend. Ecology. 2005; 86(7):1751–1756. https://doi.org/10.1890/04-1343

69. Dannenberg O, Dette H, Munk A. An extension of Welch's approximate t-solution to comparative bio-equivalence trials. Biometrika. 1994; 81(1):91–101. https://doi.org/10.2307/2337053

70. Hauschke D, Steinijans V, Diletti E. A distribution-free procedure for the statistical analysis of bioequivalence studies. International Journal of Clinical Pharmacology, Therapy, and Toxicology. 1990; 28 (2):72–78. PMID: 2307548

71. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. CMAJ: Canadian Medical Association Journal. 1995; 152 (2):169–173. PMID: 7820798

72. Gardner MJ, Altman DG. Confidence intervals rather than *p*-values: estimation rather than hypothesis testing. BMJ (Clin Res Ed). 1986; 292(6522):746–750. https://doi.org/10.1136/bmj.292.6522.746

73. Cumming G. Replication and *p*-intervals: *p*-values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science. 2008; 3(4):286–300. https://doi.org/10. 1111/j.1745-6924.2008.00079.x PMID: 26158948

74. Reichardt CS, Gollob HF. When confidence intervals should be used instead of statistical tests, and vice versa. Routledge; 2016.

75. Parkhurst DF. Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation Equivalence tests improve the logic of significance testing when demonstrating similarity is

important, and reverse tests can help show that failure to reject a null hypothesis does not support that hypothesis. Bioscience. 2001; 51(12):1051–1057.

76. Hauck WW, Anderson S. A proposal for interpreting and reporting negative studies. Statistics in Medicine. 1986; 5(3):203–209. https://doi.org/10.1002/sim.4780050302 PMID: 3526499

77. Pocock SJ, Stone GW. The primary outcome fails -what next? New England Journal of Medicine. 2016; 375(9):861–870. https://doi.org/10.1056/NEJMra1510064 PMID: 27579636

78. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p < 0.05): Significance thresholds and the crisis of unreplicable research. PeerJ Preprints; 2017.

79. Hung H, Wang SJ, O'Neill R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. Biometrical Journal. 2005; 47(1):28–36. https://doi.org/10.1002/bimj.200410084 PMID: 16395994

80. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. Controlled Clinical Trials. 2002; 23(1):2–14. https://doi.org/10.1016/S0197-2456(01)00196-9 PMID: 11852160

81. Keefe RS, Kraemer HC, Epstein RS, Frank E, Haynes G, Laughren TP, et al. Defining a clinically meaningful effect for the design and interpretation of randomized controlled trials. Innovations in Clinical Neuroscience. 2013; 10(5-6 Suppl A):4S. PMID: 23882433

82. Bedard PL, Krzyzanowska MK, Pintilie M, Tannock IF. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. Journal of Clinical Oncology. 2007; 25(23):3482–3487. https://doi.org/10.1200/JCO.2007.11.3670 PMID: 17687153

83. Meyners M. Least equivalent allowable differences in equivalence testing. Food Quality and Preference. 2007; 18(3):541–547. https://doi.org/10.1016/j.foodqual.2006.07.005

84. Ng TH. Noninferiority hypotheses and choice of noninferiority margin. Statistics in Medicine. 2008; 27 (26):5392–5406. https://doi.org/10.1002/sim.3367 PMID: 18680173

85. Tsou HH, Hsiao CF, Chow SC, Yue L, Xu Y, Lee S. Mixed noninferiority margin and statistical tests in active controlled trials. Journal of Biopharmaceutical Statistics. 2007; 17(2):339–357. https://doi.org/10.1080/10543400601183861 PMID: 17365228

86. Barker L, Rolka H, Rolka D, Brown C. Equivalence testing for binomial random variables: which test to use? The American Statistician. 2001; 55(4):279–287. https://doi.org/10.1198/000313001753272213

87. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, Group C, et al. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. 2006; 295 (10):1152–1160. https://doi.org/10.1001/jama.295.10.1152 PMID: 16522836

88. Konijn EA, van de Schoot R, Winter SD, Ferguson CJ. Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. Communication Methods and Measures. 2015; 9(4):280–302. https://doi.org/10.1080/19312458.2015.1096332

89. Mulder J, Wagenmakers EJ. Editors' introduction to the special issue 'Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments'. Journal of Mathematical Psychology. 2016; 72:1–5. https://doi.org/10.1016/j.jmp.2016.01.002

90. Gönen M. The Bayesian t-test and beyond. Statistical Methods in Molecular Biology. 2010; 620:179–199. https://doi.org/10.1007/978-1-60761-580-4_4

91. Dienes Z. Using Bayes to get the most out of non-significant results. Frontiers in Psychology. 2014; 5:781. https://doi.org/10.3389/fpsyg.2014.00781 PMID: 25120503

92. Hoekstra R, Monden R, van Ravenzwaaij D, Wagenmakers EJ. Bayesian reanalysis of null results reported in the New England Journal of Medicine: Strong yet variable evidence for the absence of treatment effects. Manuscript submitted for publication. 2017.

93. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review. 2009; 16(2):225–237. https://doi.org/10.3758/PBR.16.2.225

94. Berger JO, Delampady M. Testing Precise Hypotheses. Statistical Science. 1987; 2(3):317–335. https://doi.org/10.1214/ss/1177013238

95. Greenland S, Poole C. Living with p-values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiology. 2013; 24(1):62–68. https://doi.org/10.1097/EDE.0b013e3182785741 PMID: 23232611

96. Marsman M, Wagenmakers EJ. Three insights from a Bayesian interpretation of the one-sided p-value. Educational and Psychological Measurement. 2017; 77(3):529–539. https://doi.org/10.1177/0013164416669201

97. Wagenmakers EJ, Wetzels R, Borsboom D, Van Der Maas HL. Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). Journal of personality and social psychology. 2011; 100(3):426–432. https://doi.org/10.1037/a0022790 PMID: 21280965

98. Rouder JN. Optional stopping: No problem for Bayesians. Psychonomic Bulletin & Review. 2014; 21 (2):301. https://doi.org/10.3758/s13423-014-0595-4

99. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. CRC Press; 1999.

100. Lakens D. Performing high-powered studies efficiently with sequential analyses. European Journal of Social Psychology. 2014; 44(7):701–710. https://doi.org/10.1002/ejsp.2023

101. Schönbrodt FD, Wagenmakers EJ. Bayes factor design analysis: Planning for compelling evidence. Psychonomic Bulletin & Review. 2016; p. 1–15.

102. Ioannidis JP. Journals should publish all null results and should sparingly publish 'positive' results. Cancer Epidemiology Biomarkers & Prevention. 2006; 15(1):186–186. https://doi.org/10.1158/1055-9965.EPI-05-0921

103. Shields PG. Publication Bias Is a Scientific Problem with Adverse Ethical Outcomes: The Case for a Section for Null Results. Cancer Epidemiology and Prevention Biomarkers. 2000; 9(8):771–772.

104. Dirnagl U, et al. Fighting publication bias: introducing the Negative Results section. Journal of Cerebral Blood Flow and Metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism. 2010; 30(7):1263–1264. https://doi.org/10.1038/jcbfm.2010.51

105. Lawlor DA. Quality in epidemiological research: should we be submitting papers before we have the results and submitting more hypothesis-generating research? International Journal of Epidemiology. 2007; 36(5):940–943. https://doi.org/10.1093/ije/dym168 PMID: 17875575

106. Mell LK, Zietman AL. Introducing prospective manuscript review to address publication bias. International Journal of Radiation Oncology -Biology Physics. 2014; 90(4):729–732. https://doi.org/10.1016/j.ijrobp.2014.07.052

107. Smulders YM. A two-step manuscript submission process can reduce publication bias. Journal of Clinical Epidemiology. 2013; 66(9):946–947. https://doi.org/10.1016/j.jclinepi.2013.03.023 PMID: 23845183

108. Berry A. Subgroup Analyses. Biometrics. 1990; 46(4):1227–1230. PMID: 2085637

109. Kerr NL. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review. 1998; 2(3):196–217. https://doi.org/10.1207/s15327957pspr0203_4 PMID: 15647155

110. Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University. 2013.

111. Song F, Loke Y, Hooper L. Why are medical and health-related studies not being published? A systematic review of reasons given by investigators. PLoS One. 2014; 9(10):e110418. https://doi.org/10.1371/journal.pone.0110418 PMID: 25335091

112. van Lent M, IntHout J, Out HJ. Differences between information in registries and articles did not influence publication acceptance. Journal of Clinical Epidemiology. 2015; 68(9):1059–1067. https://doi.org/10.1016/j.jclinepi.2014.11.019 PMID: 25542517

113. Mathieu S, Chan AW, Ravaud P. Use of trial register information during the peer review process. PLoS One. 2013; 8(4):e59910. https://doi.org/10.1371/journal.pone.0059910 PMID: 23593154

114. Academia StackExchange. Why isn't pre-registration required for all experiments?; March 23, 2018. https://academia.stackexchange.com/questions/74711/why-isnt-pre-registration-required-for-all-experiments

115. Ramsey S, Scoggins J. Commentary: practicing on the tip of an information iceberg? Evidence of underpublication of registered clinical trials in oncology. The Oncologist. 2008; 13(9):925–929. https://doi.org/10.1634/theoncologist.2008-0133 PMID: 18794216

116. Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA. 2009; 302(9):977–984. https://doi.org/10.1001/jama.2009.1242 PMID: 19724045

117. Ross JS, Mulvey GK, Hines EM, Nissen SE, Krumholz HM. Trial publication after registration in ClinicalTrials.gov: a cross-sectional analysis. PLoS Medicine. 2009; 6(9):e1000144. https://doi.org/10.1371/journal.pmed.1000144 PMID: 19901971

118. Huić M, Marušić M, Marušić A. Completeness and changes in registered data and reporting bias of randomized controlled trials in ICMJE journals after trial registration policy. PLoS One. 2011; 6(9):e25258. https://doi.org/10.1371/journal.pone.0025258 PMID: 21957485

119. Zarin DA, Tse T, Williams RJ, Carr S. Trial reporting in ClinicalTrials.gov -the final rule. New England Journal of Medicine. 2016; 375(20):1998–2004. https://doi.org/10.1056/NEJMsr1611785 PMID: 27635471

120.   Wager E, Williams P. "Hardly worth the effort" -Medical journals' policies and their editors' and publishers' views on trial registration and publication bias: quantitative and qualitative study. BMJ. 2013; 347: f5248. https://doi.org/10.1136/bmj.f5248 PMID: 24014339

121.   Gelman A. Preregistration of studies and mock reports. Political Analysis. 2013; 21(1):40–41. https://doi.org/10.1093/pan/mps032

122.   Toerien M, Brookes ST, Metcalfe C, De Salis I, Tomlin Z, Peters TJ, et al. A review of reporting of participant recruitment and retention in RCTs in six major journals. Trials. 2009; 10(1):52. https://doi.org/10.1186/1745-6215-10-52 PMID: 19591685

123.   Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2017.

124.   Haak LL, Fenner M, Paglione L, Pentz E, Ratner H. ORCID: a system to uniquely identify researchers. Learned Publishing. 2012; 25(4):259–264. https://doi.org/10.1087/20120404

125.   Charlton BG, Andras P. How should we rate research?: Counting number of publications may be best research performance measure. BMJ. 2006; 332(7551):1214–1215. https://doi.org/10.1136/bmj.332.7551.1214-c PMID: 16710008

126.   Rosenthal R. The file drawer problem and tolerance for null results. Psychological Bulletin. 1979; 86 (3):638–641. https://doi.org/10.1037/0033-2909.86.3.638

127.   Pautasso M. Worsening file-drawer problem in the abstracts of natural, medical and social science databases. Scientometrics. 2010; 85(1):193–202. https://doi.org/10.1007/s11192-010-0233-5

128.   Kühberger A, Fritz A, Scherndl T. Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. PLoS One. 2014; 9(9):e105825. https://doi.org/10.1371/journal.pone.0105825 PMID: 25192357

129.   Coffman LC, Niederle M. Pre-analysis plans have limited upside, especially where replications are feasible. The Journal of Economic Perspectives. 2015; 29(3):81–97. https://doi.org/10.1257/jep.29.3.81

130.   de Winter J, Happee R. Why selective publication of statistically significant results can be effective. PLoS One. 2013; 8(6):e66463. https://doi.org/10.1371/journal.pone.0066463 PMID: 23840479

131.   van Assen MA, van Aert RC, Nuijten MB, Wicherts JM. Why publishing everything is more effective than selective publishing of statistically significant results. PLoS One. 2014; 9(1):e84896. https://doi.org/10.1371/journal.pone.0084896 PMID: 24465448

132.   Gelman A, Carlin J. Some natural solutions to the *p*-value communication problem and why they won't work. 2017.

133.   Shields PG, Sellers TA, Rebbeck TR. Null results in brief: meeting a need in changing times. Cancer Epidemiology and Prevention Biomarkers. 2009; 18(9):2347–2347. https://doi.org/10.1158/1055-9965.EPI-09-0684

134.   McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. arXiv preprint arXiv:170907588. 2017.

135.   Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D. Deep Reinforcement Learning that Matters. arXiv preprint arXiv:170906560. 2017.

136.   Wellek S. A critical evaluation of the current 'p-value controversy'. Biometrical Journal. 2017; 59 (5):854–872. https://doi.org/10.1002/bimj.201700001 PMID: 28504870

137.   Morey RD, Rouder JN. Bayes factor approaches for testing interval null hypotheses. Psychological Methods. 2011; 16(4):406–419. https://doi.org/10.1037/a0024377 PMID: 21787084

138.   Pérez ME, Pericchi LR. Changing statistical significance with the amount of information: The adaptive α significance level. Statistics & Probability Letters. 2014; 85:20–24. https://doi.org/10.1016/j.spl.2013.10.018

139.   Lauzon C, Caffo B. Easy multiplicity control in equivalence testing using two one-sided tests. The American Statistician. 2009; 63(2):147–154. https://doi.org/10.1198/tast.2009.0029 PMID: 20046823

140.   Aberegg SK, Hersh AM, Samore MH. Empirical consequences of current recommendations for the design and interpretation of noninferiority trials. Journal of general internal medicine. 2018; 33(88):1–9.

141.   Senn S. Testing for baseline balance in clinical trials. Statistics in Medicine. 1994; 13(17):1715–1726. https://doi.org/10.1002/sim.4780131703 PMID: 7997705

142.   Hedges LV. Modeling publication selection effects in meta-analysis. Statistical Science. 1992; 7 (2):246–255. https://doi.org/10.1214/ss/1177011364

143.   Bloomfield RJ, Rennekamp KM, Steenhoven B. No system is perfect: understanding how registration-based editorial processes affect reproducibility and investment in research quality–Free Responses to Survey of Conference Participants. 2018.

144. O'Hara B. Negative results are published. Nature. 2011; 471(7339):448–449. https://doi.org/10.1038/471448e PMID: 21430758

145. Nosek BA, Spies JR, Motyl M. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science. 2012; 7(6):615–631. https://doi.org/10.1177/1745691612459058 PMID: 26168121

146. Higginson AD, Munafò MR. Current incentives for scientists lead to underpowered studies with erroneous conclusions. PLoS Biology. 2016; 14(11):e2000995. https://doi.org/10.1371/journal.pbio.2000995 PMID: 27832072

147. Bakker M, van Dijk A, Wicherts JM. The rules of the game called psychological science. Perspectives on Psychological Science. 2012; 7(6):543–554. https://doi.org/10.1177/1745691612459060 PMID: 26168111

148. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013; 14 (5):365–376. https://doi.org/10.1038/nrn3475 PMID: 23571845

149. Gervais WM, Jewell JA, Najle MB, Ng BK. A powerful nudge? Presenting calculable consequences of underpowered research shifts incentives toward adequately powered designs. Social Psychological and Personality Science. 2015; 6(7):847–854. https://doi.org/10.1177/1948550615584199

150. Smaldino PE, McElreath R. The natural selection of bad science. Royal Society Open Science. 2016; 3(9):160384. https://doi.org/10.1098/rsos.160384 PMID: 27703703

151. Cohen J. The statistical power of abnormal-social psychological research: a review. The Journal of Abnormal and Social Psychology. 1962; 65(3):145–153. https://doi.org/10.1037/h0045186 PMID: 13880271

152. Perlman MD, Wu L, et al. The emperor's new tests. Statistical Science. 1999; 14(4):355–369. https://doi.org/10.1214/ss/1009212516

153. Ng TH. Issues of simultaneous tests for noninferiority and superiority. Journal of Biopharmaceutical Statistics. 2003; 13(4):629–639. https://doi.org/10.1081/BIP-120024199 PMID: 14584713

154. Koyama T, Westfall PH. Decision-theoretic views on simultaneous testing of superiority and noninferiority. Journal of Biopharmaceutical Statistics. 2005; 15(6):943–955. https://doi.org/10.1080/10543400500265694 PMID: 16279353

155. Shieh G. Exact power and sample size calculations for the two one-sided tests of equivalence. PLoS One. 2016; 11(9):e0162093. https://doi.org/10.1371/journal.pone.0162093 PMID: 27598468

156. Zhu H. Sample size calculation for comparing two Poisson or negative binomial rates in noninferiority or equivalence trials. Statistics in Biopharmaceutical Research. 2017; 9(1):107–115. https://doi.org/10.1080/19466315.2016.1225594

157. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. Pharmaceutical Statistics. 2005; 4(3):187–201. https://doi.org/10.1002/pst.175

158. Shao Y, Mukhi V, Goldberg JD. A hybrid Bayesian-frequentist approach to evaluate clinical trial designs for tests of superiority and non-inferiority. Statistics in Medicine. 2008; 27(4):504–519. https://doi.org/10.1002/sim.3028 PMID: 17854052

159. Jia B, Lynn HS. A sample size planning approach that considers both statistical significance and clinical significance. Trials. 2015; 16(1):213. https://doi.org/10.1186/s13063-015-0727-9 PMID: 25962998

160. Jiroutek MR, Muller KE, Kupper LL, Stewart PW. A new method for choosing sample size for confidence interval–based inferences. Biometrics. 2003; 59(3):580–590. https://doi.org/10.1111/1541-0420.00068 PMID: 14601759