2025 ■ Vol. 21 ■ no. 2

REACT to NHST: Sensible conclusions for meaningful hypotheses

Rafael Izbicki [a], Luben M. C. Cabezas [a,b], Fernando A. B. Colugnatti [c], Rodrigo F. L. Lassance [a,b], Altay A. L. de Souza [d] & Rafael B. Stern [e]

[a]Department of Statistics, Federal University of São Carlos, São Paulo, Brazil

[b]Institute of Mathematics and Computer Sciences, University of São Paulo, São Paulo, Brazil

[c]School of Medicine, Federal University of Juiz de Fora, Minas Gerais, Brazil

[d]Psychobiology Department, Federal University of São Paulo, São Paulo, Brazil

[e]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

Abstract ■ While Null Hypothesis Significance Testing (NHST) remains a widely used statistical tool, it suffers from several shortcomings in its common usage, such as conflating statistical and practical significance, the formulation of inappropriate null hypotheses, and the inability to distinguish between accepting the null hypothesis and failing to reject it. Recent efforts have focused on developing alternatives that address these issues. Despite these efforts, conventional NHST remains dominant in scientific research due to its procedural simplicity and mistakenly presumed ease of interpretation. Our work presents an intuitive alternative to conventional NHST designed to bridge the gap between the expectations of researchers and the actual outcomes of hypothesis tests: `REACT`. `REACT` not only tackles shortcomings of conventional NHST but also offers additional advantages over existing alternatives. For instance, `REACT` applies to multiparametric hypotheses and does not require stringent significance-level corrections when conducting multiple tests. We illustrate the practical utility of `REACT` through real-world data examples.

Keywords ■ hypothesis tests, NHST, p-values, equivalence tests, three-way decision procedures.

✉ rafaelizbicki@gmail.com

10.20982/tqmp.21.2.p043

Acting Editor ■ Fernando Marmolejo Ramos (Flinders University)

Reviewers ■ Two anonymous reviewers

Introduction

Statistical hypothesis testing is a fundamental tool in scientific research, offering a structured approach to tackling research questions. In the field of clinical research, it goes beyond being a mere recommendation and becomes a nearly-mandatory requirement for publishing results and aiding important decision-making. For example, for many years, clinical trial design has remained heavily reliant on calculating sample sizes by evaluating the statistical power of hypothesis tests.

However, there has been a growing wave of criticism directed at conventional Null Hypothesis Significance Testing[1] (NHST) and p-values (Hays, 1963; Wasserman, 2013; Trafimow et al., 2018; Pike, 2019; Greenland et al., 2016; Kadane, 2016; Wasserstein et al., 2019; Diggle & Chetwynd, 2011). Much of this criticism arises from the misuse and misinterpretation of statistical tests. Even in meta-analytic studies, hypothesis tests often lead to misinterpretations that have a high impact on public policymaking. As a result, many scientific journals have taken a position against the use of conventional NHST, often discouraging its use (Campo & Lichtman, 2008; Trafimow & Marks, 2015).

Despite the ongoing criticisms and the several alternatives proposed (see Section "Connections to Existing Work" for a review), conventional NHST remains the widely ac-

[1]By "conventional NHST" we refer to the approach typically used by practitioners of NHST, which involves (*i*) the formulation of a null hypothesis usually pointing to no effects (e.g., $H_0 : \mu_A = \mu_B$ in a two-sample tests problem), and (*ii*) the establishment of a statistical test with two outcomes for H_0 that controls the type I error at a predefined level of α.

cepted standard in scientific research. This continued popularity can be explained by various factors, including its operational simplicity. Additionally, though NHST results are nuanced and complex, they often give a false impression of being easy to interpret.

It is therefore crucial to provide an alternative to conventional NHST while keeping as much of its operational simplicity as possible. This alternative should bridge the gap between researchers' expectations and the actual outcomes of tests, preventing misinterpretations. For instance, distinguishing between accepting and failing to reject the null hypothesis is essential in practice. Therefore, the main goal of this work is to introduce a framework for hypothesis testing, `REACT`, that better meets the needs of researchers. `REACT` builds upon existing solutions that aim to improve NHST, incorporating elements from three-way tests and equivalence tests like Two One-Sided Tests (TOST; Schuirmann, 1987). For a detailed comparison with these and other methods, see Section "Connections to Existing Work". We also provide an R package that implements `REACT` for common models.

Next, we revisit some of the major concerns of the standard approach to hypothesis testing. Section "`REACT`" then introduces a simplified version of `REACT` for a single hypothesis that only concerns one parameter. Section "Connections to Existing Work" shows how `REACT` relates to other methods in the literature. Section "`REACT` and its properties" introduces our full procedure, and presents its properties. Two applications are presented in Section "Applications". Section "Implications of the `REACT` Method for Machine Learning and AI" discusses `REACT` in the context of machine learning and artificial intelligence. Section "Final Remarks" concludes the paper.

Review of some NHST issues

Issue 1: Statistical significance versus practical significance. One of the primary challenges of NHST is the difficulty in distinguishing between statistical significance and practical significance (Wasserstein & Lazar, 2016). This has been noted very early in psychology (Hays, 1963). Still, low p-values are often used as a proxy for important practical significance. As an example, a study on the effectiveness of aspirin in preventing myocardial infarction (Bartolucci et al., 2011; Sullivan & Feinn, 2012) found statistically significant results (p-value $< .00001$), and therefore was stopped early due to conclusive evidence. As a result, many people were advised to take aspirin to prevent heart attacks. However, upon further investigation, the effect size was found to be practically insignificant, and the recommendation had to be revised (Sullivan & Feinn, 2012). This raises important concerns about the practical relevance of the output of such hypothesis tests and their implications for public health.

Issue 2: Implausibility of the null hypothesis. There are very few situations in which one expects precise null hypotheses to be exactly true (Edwards et al., 1963; Amrhein et al., 2017; Lecoutre & Poitevineau, 2022). For example, when comparing two medications, the primary concern is typically whether they are practically equivalent, as it is highly unlikely that any two medications will produce precisely the same effects. Therefore, establishing whether the medications have similar outcomes is often more relevant than attempting to evaluate if they have the same effect on average. Indeed, bioequivalence tests effectively modify the null hypothesis to align with practical equivalence (refer to Section "Connections to Existing Work"). This is also true in other domains. Indeed, Cohen (1992, page 1308) mentions that

> "The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null is always false, what's the big deal about rejecting it?"

A consequence of this fact is that, provided that the sample size is large enough, in most problems one will always reject the null hypothesis, making the results of NHST not meaningful (Vaughan & Corballis, 1969; Cohen, 1992, 1994; Gill, 1999; Faber & Fonseca, 2014).

Issue 3: Accept versus not-reject the null hypothesis. The absence of evidence is not evidence of absence (Altman & Bland, 1995), but NHST is not able to differentiate between failure to reject the null hypothesis and its acceptance. Indeed, Edwards et al. (1963, page 235) point out that

> "If the null hypothesis is not rejected, it remains in a kind of limbo of suspended disbelief."

This is because a null may not be rejected either because the test is not powerful enough to reject it (e.g. due to a small sample size) or because H_0 is indeed true. This has major implications for public policymaking, especially during times of crisis, such as the COVID-19 pandemic. It can be challenging to interpret research findings on the efficacy of interventions, such as the use of masks (Jefferson et al., 2023) or hydroxychloroquine (Mehra et al., 2020), when there is a lack of consensus on what constitutes evidence of absence versus no evidence of an effect (Fidler et al., 2018).

It is therefore very important for researchers to distinguish between "no evidence of effect" and "evidence of no effect" when interpreting research results (Altman & Bland, 1995; Keysers et al., 2020). Plain NHST cannot do that on its own.

In this paper, we overcome these issues by introducing the *Region of Equivalence Agnostic Confidence-based Test* (REACT).

Novelty

Numerous solutions have been proposed to tackle issues 1, 2, and 3; see Section "Connections to Existing Work" for a detailed overview of how such approaches relate to our approach. In particular, REACT uses the strengths of equivalence and three-way hypothesis testing. However, to the best of our knowledge, REACT is the first approach that simultaneously solves these and other issues. Specifically, REACT:

- Is designed to work with meaningful null hypotheses that encode practical significance
- Clearly distinguishes between "evidence of absence" and "absence of evidence"
- Does not need ad hoc procedures to perform multiple comparisons. In fact, REACT not only automatically controls the Family-Wise Error Rate (FWER) of false rejections at α (type I errors), but it also controls the FWER of false acceptances at α (type II errors)
- Can be easily applied to hypotheses that involve several parameters, such as in an ANOVA setting
- Leads to fully logically coherent solutions. For instance, in a multiple comparison problem, if REACT rejects the null $\mu_1 = \mu_2$, it will also reject the null $\mu_1 = \mu_2 = \mu_3$. This level of coherence is not typically achieved with standard procedures, resulting in epistemic confusion that complicates the reporting of test results.

REACT requires the specification of two components: (i) a confidence region for the parameters of interest, and (ii) a null hypothesis that reflects a range of parameter values considered to be practically equivalent (e.g., in an ANOVA context, $H_0 : |\mu_A - \mu_B| \leq \Delta$). Although part (ii) involves more in-depth thought compared to regular NHST, determining Δ is essentially equivalent to deciding the minimum effect size of scientific interest, as is commonly done in standard power analysis for sample size calculations. Thus, while our approach attempts to keep the operational simplicity of standard NHST, it directly meets researchers' needs.

REACT

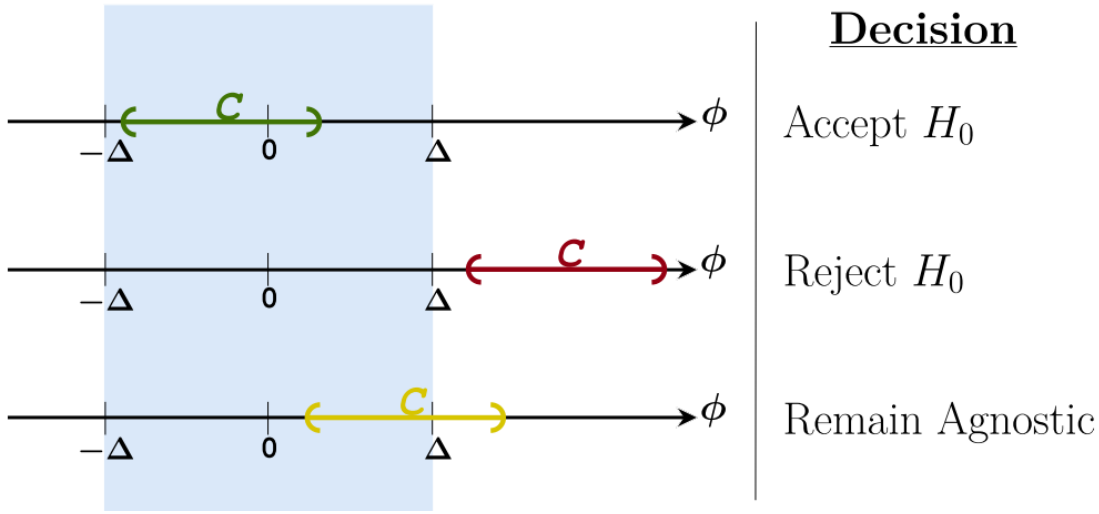
To better illustrate how our approach builds upon existing work, we first introduce it in a simplified version.

In its simplest form, REACT is composed by the following steps:

1. **[Establish the null hypothesis]** Define a null hypothesis H_0 by establishing a pragmatic hypothesis, which is a range of values considered to be practically equivalent. For example, if μ_A and μ_B are the average effects of drugs A and B on a desired outcome, the pragmatic null may be $H_0 : |\mu_A - \mu_B| \leq \Delta$, where Δ is the smallest difference of practical interest (also known as the smallest effect size of interest—SESOI (Lakens, 2017)). In this case, H_0 is usually called the equivalence range (Bauer & Kieser, 1996) or a region of practical equivalence (ROPE) (Kruschke, 2018) and is extensively used in equivalence testing. The task of setting this region may be complex, but it essentially parallels the steps taken in standard power analysis, specifically in pinpointing the significant portions of the alternative hypothesis that need high power. See Section "Connections to Existing Work" for ideas on how Δ can be chosen. Similarly, one may want to test $H_0 : |\rho| \leq \Delta$, where ρ is the correlation coefficient between two quantities of interest, or $H_0 : |\beta_i| \leq \Delta$, where β_i is the coefficient of the i-th covariate in a regression. For simplicity, we now assume that the null hypothesis has the shape
$$H_0 : |\phi| \leq \Delta$$
for some parameter ϕ, although our test is more general (see Section "REACT and its properties").
2. **[Build a confidence set]** Create C, a confidence set for the parameter of interest, ϕ. That is, C contains values of ϕ that are consistent with the dataset that was observed.
3. **[Test H_0 using C]** Test the null hypothesis using the following three-way rule:

$$\text{Decision} = \begin{cases} \textbf{Accept } H_0 \text{ if all values of } C \text{ are smaller than } \Delta \text{ in absolute value} \\ \textbf{Reject } H_0 \text{ if all values of } C \text{ are larger than } \Delta \text{ in absolute value} \\ \textbf{Remain Agnostic otherwise} \end{cases}$$

Figure 1 ■ Illustration of REACT to test hypotheses of the type $H_0 : |\phi| \leq \Delta$. C is a confidence set built using the data; the blue region represents the null hypothesis.



This procedure is illustrated in Figure 1.

Figure 2 shows an example of application of REACT to the problem of investigating whether CAMCOG scores can distinguish between three groups of patients: control (CG), mild cognitive impairment (MCI), and Alzheimer's disease (AD). In each plot, the dashed line represents the precise hypothesis of interest, which states that CAMCOG scores are equally distributed among the compared groups. The blue region represents the null hypotheses associated with each pair of groups. We display the outcome of REACT as the mean differences' confidence intervals between each pair of groups at a given sample size. In each comparison, we start by randomly sampling two observations from each group to derive the initial confidence interval. Then, in each step, we randomly add a new observation from one of the groups of interest and obtain a new confidence interval. For small sample sizes, the test remains agnostic on all three hypotheses. As the sample size increases, the pragmatic hypothesis for AD vs Control is rejected, the one for AD vs MCI is inconclusive, and the one for Control vs MCI is accepted. We obtain the same conclusions when changing the sorting order, with slight changes in MCI vs AD (Figure A1 in Appendix). More details about this example can be found in Section "Applications".

Connections to Existing Work

Next, we investigate the relationship between REACT and similar approaches.

Power Analysis and Severity Tests. Power analysis is based on the following principle, suggested by J. Neyman (Neyman, 1957, page 16):

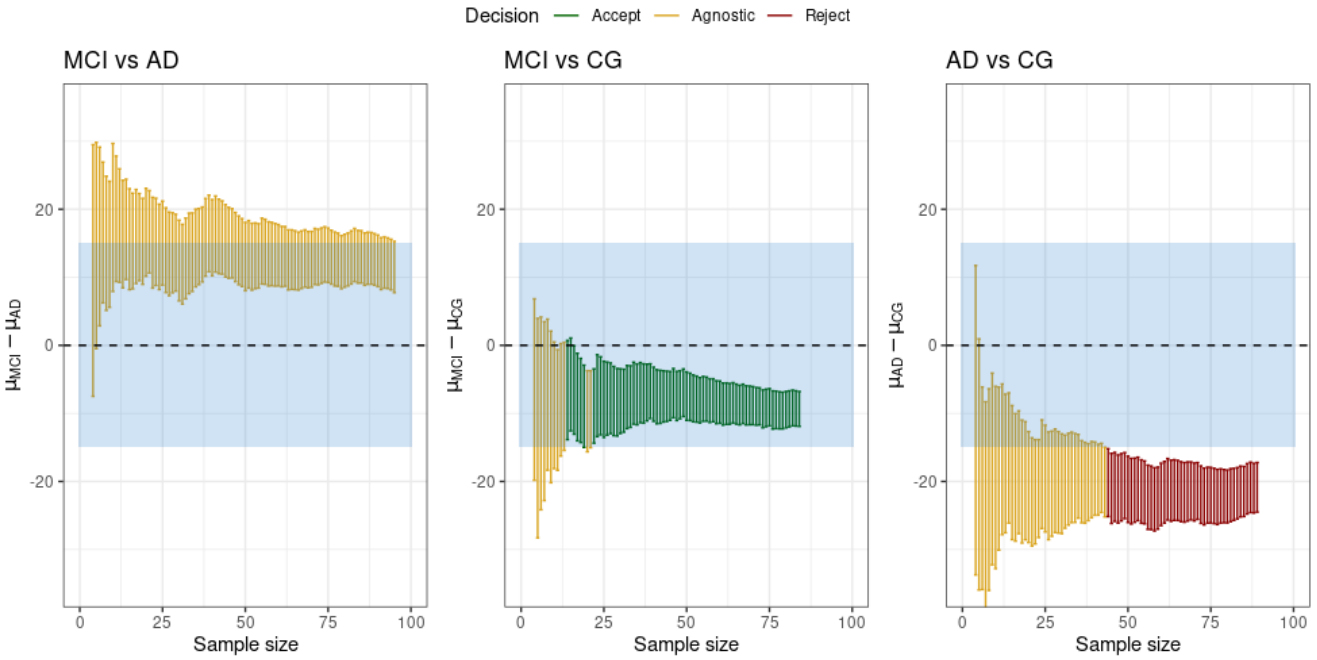
> "[If] the probability of detecting an appreciable error in the hypothesis tested was large, say .95 or greater, then and only then is the decision in favour of the hypothesis tested justifiable in the same sense as the decision against this hypothesis is justifiable when an appropriate test rejects it at a chosen level of significance."

That is, according to this approach, if the power of the study is large (at some specific point of the alternative of interest, say $\mathbb{P}_{\phi=\Delta}(\text{reject})$ in the setting presented above), then a non-rejection can be interpreted as acceptance (see also "severe testing" (Mayo & Spanos, 2006; Mayo, 2018)—which uses a similar calculation to the p-value, though assuming a point of the alternative as true instead, to verify if one is warranted to accept the null hypothesis—for a related approach). Thus, power analysis addresses issue 3 raised in the introduction by distinguishing between "evidence of absence" and "absence of evidence".

Unfortunately, power analysis is often overlooked in favor of p-values, even when reported in the analysis. To address this issue, REACT outputs a "accept", "agnostic" or "reject" decision, preventing any misinterpretation.

Scheffé's method. Scheffé's method (Scheffé, 1999) can be viewed as a special case of REACT for an analysis of variance (ANOVA) model when testing hypotheses involving only linear combinations (contrasts) of treatment effects. Scheffé's method constructs the usual confidence ellipsoid for the treatment effects. To test the null hypothesis, it

Figure 2 ■ Confidence intervals for the average difference between groups, $\mu_i - \mu_j$ as a function of the sample size. The dashed line indicates the precise hypothesis considered in each figure, $H_0 : \mu_i = \mu_j$. The blue regions delimit the null hypotheses.



checks whether the line corresponding to the null hypothesis intersects the confidence ellipsoid. If such an intersection exists, the null hypothesis is not rejected; otherwise, it is rejected.

In contrast, `REACT` allows consideration of a broader range of null hypotheses, including non-linear functions of treatment effects and pragmatic hypotheses derived from these functions. For pragmatic hypotheses, `REACT` refines the "non-rejection" outcome into either "acceptance" or "remain undecided." Additionally, `REACT` can be applied to various other models and is not restricted by the assumptions of ANOVA. See Section "Cambridge Cognition Examination" for an application of `REACT` that is similar to Scheffé's method when setting $\Delta = 0$, with the difference that `REACT` is controlling for multiple testing.

Effect size estimates. Effect sizes are an attempt to address issues 1 and 2. They measure the strength of a relationship and are often used to complement NHSTs. These include the estimated coefficients $\hat{\beta}$ of the regression function (Kutner et al., 2005), Cohen's d (Cohen, 2013) and Pearson's r (Pearson, 1920), among many others (Kirk, 2007; Ellis, 2010; Fritz et al., 2012). If the null hypothesis is rejected, one can check whether the effect size is large enough to be practically significant. Although effect sizes can provide valuable information, they are often overlooked in favor of p-values.

To address these limitations, `REACT` integrates both rejection of the null and practical significance into statistical inference. Moreover, if one designs null hypotheses using effect sizes (Kruschke & Liddell, 2018), `REACT` can become fully integrated with them. For instance, Cohen's d between two groups can be expressed as $\phi := \sigma^{-1}(\mu_A - \mu_B)$, where σ is the standard deviation of the response variable. The equivalence region can be stated in terms of ϕ as $H_0 : |\phi| \leq \Delta$. This hypothesis can then be tested using our approach via the confidence set for ϕ, $d \pm \text{SE}_d \times t_{n_A+n_B-2}$, where $d = s^{-1}(\bar{x}_1 - \bar{x}_2)$ is an estimate of ϕ, t is the Student's t-distribution, n_A and n_B are the sample sizes for each group, and SE_d is an estimate of the standard error of d (Goulet-Pelletier & Cousineau, 2018).

We note, however, that applying a test based on Cohen's d may not be ideal. As Lakens (2022, page 280) puts it,

> "Setting them [equivalence bounds] in terms of Cohen's d leads to bias in the statistical test, as the observed standard deviation has to be used to translate the specified Cohen's d into a raw effect size for the equivalence test (...) [A]s equivalence testing becomes more popu-

lar, and fields establish smallest effect sizes of interest, they should do so in raw effect size differences, not in standardized effect size differences."

Equivalence Range. The idea of working with regions of practical equivalence (instead of point null hypotheses) to address issues 1 and 2 has been presented under many different names—such as good-enough belt (Keren & Lewis, 1993), equivalence range (Bauer & Kieser, 1996), indifference zone (Hobbs & Carlin, 2007), effective null set (Gross, 2014), and even the ROPE (Kruschke, 2018) itself—but a definition with sufficient generality to cover more complex settings is more recent (pragmatic hypothesis, Esteves et al. (2019)). Let ϕ_0 be the point from which we wish to derive the region of equivalence (that is, we are interested in creating a region of equivalence around $H_0 : \phi = \phi_0$), Θ be the parameter space and $d(\cdot, \cdot)$ be a dissimilarity function. Then, the pragmatic hypothesis will be represented by

$$Pg(\phi_0, d, \Delta) := \{\phi \in \Theta : d(\phi_0, \phi) \leq \Delta\}.$$

The setting described in Section "REACT" corresponds to choosing $d(\phi_0, \phi) = |\phi_0 - \phi|$ and $\phi_0 = 0$. There are many heuristics for choosing Δ (Lakens, 2017, 2022; Wang et al., 2023; Lassance et al., 2025). These include:

- Relating Δ to another quantity in the literature about which it is easier to obtain intuition; we use this in Section "Meta-analysis".
- Identifying positive results in the literature that evaluate the same (or a similar) effect as your own study and choosing the smallest Δ such that REACT would lead to accepting the hypothesis in these previous cases. While this strategy may downplay random variability if few studies are used to derive Δ, it acts as a starting point from which researchers can propose changes later. Such an approach is particularly useful in the context of reproducibility studies since rejecting H_0 in the new study based on criteria that would have accepted the same hypothesis in the old one can be interpreted as a failure to reproduce the original finding.
- Setting Δ as the smallest change such that patients report an improvement from their original conditions. This can routinely be obtained through the use of patient reported outcome measure (PROM) scores. Even when the perception of improvement varies substantially between patients, there is a selection process available that ensures the optimality of the selected Δ (Wang et al., 2023).
- Setting $d(\phi_0, \phi)$ to be a measure of effect size (such as the standardized mean difference between two populations) and taking Δ to be the smallest effect size that is practically significant. This approach parallels traditional sample size calculations performed through power analysis.

When there are no clear strategies for determining Δ, there is available software that can help provide an automated suggestion (Makowski et al., 2019). REACT can be used within any of these approaches.

Equivalence Tests, TOST and B-values. Equivalence tests were originally developed to compare the bioequivalence of two drugs to address issues 1 and partially issue 3. Rather than testing the hypothesis of "no effect" ($\phi = 0$), the hypothesis is modified to "practically no effect" ($|\phi| < \Delta$) (Westlake, 1976; Schuirmann, 1987). Additionally, the null hypothesis is set as the hypothesis of a practical effect, namely $H_0 : |\phi| > \Delta$. Thus, under a Neyman-Pearson interpretation of the test outcomes, rejecting the null hypothesis would lead to the conclusion that there is absence of practical effect, $|\phi| \leq \Delta$. This makes equivalence tests very useful in various fields including political science (Rainey, 2014), communication research (Weber & Popova, 2012), anthropology (Smith, 2020), sensory science (Meyners, 2012), psychology (Lakens et al., 2018), and clinical trials (Walker & Nowacki, 2011; Wellek & Blettner, 2012; L. M. Friedman et al., 2015; Leung et al., 2020).

A popular method for conducting equivalence tests is the Two One-Sided Tests (TOST (Schuirmann, 1987)). The standard TOST procedure for two-sample testing involves (i) constructing a $(1 - 2\alpha)$-level confidence interval for $\mu_A - \mu_B$, and (ii) rejecting the null hypothesis that indicates the absence of a practical effect, $H_0 : |\mu_A - \mu_B| > \Delta$, only if the confidence interval falls entirely within the interval $(-\Delta, \Delta)$. While this procedure is similar to REACT in that it tests for equivalence using a confidence set, it only yields two possible outcomes: either reject the null hypothesis and conclude the absence of effect, or fail to reject the null hypothesis. In fact, TOST corresponds to applying REACT taking C to be a $(1 - 2\alpha)$-level confidence interval and merging the decisions "accept" and "agnostic". Thus, REACT allows for a more nuanced approach by also permitting acceptance of the hypothesis of a practical effect.

Unlike REACT, TOST procedures require tailoring to address each specific problem: utilizing any $(1 - 2\alpha)$-level confidence set does not necessarily control type I error rates at α (Wellek, 2010). Indeed, Berger and Hsu (1996) show several examples of TOST tests that do not control type I error rates (Berger & Hsu, 1996). Some examples of TOST methods include regression (Dixon & Pechmann, 2005; Robinson et al., 2005; Campbell, 2020; Alter & Counsell, 2021), parametric and non-parametric paired sample tests (Mara & Cribbie, 2012), and correlation coefficients (Counsell & Cribbie, 2015). REACT, however, controls type I error at α as long as $C(\mathcal{D})$ is a $(1 - \alpha)$-level confidence set (Proposition 3). Furthermore, the null and alternative

hypotheses can be interchanged without affecting the conclusions of the test: REACT is symmetric (Proposition 9). REACT can also be easily used for hypotheses that involve several parameters (see Section "Cambridge Cognition Examination" for an example), while TOST requires further work (see e.g. Yang et al., 2015, for the development of TOST to compare multiple groups).

Recently, a two-stage equivalence testing procedure that has the advantage of deriving a region of equivalence solely based on data, the empirical equivalence bound (EEB), has been proposed (Zhao et al., 2022). Let $[L_0, U_0]$ and $[L, U]$ respectively be the symmetrical $(1-\alpha)$ and $(1-2\alpha)$ confidence intervals for ϕ and set $B := \max\{|L|, |U|\}$ (the largest deviation from 0 of the interval). The EEB is given by

$$EEB_\alpha(\beta|C) = \inf_{b\in[0,\infty]}\{b : F_B(b|C, \phi = 0) \geq \beta\},$$

where C is either $0 \in [L_0, U_0]$ or $0 \notin [L_0, U_0]$, F_B is the cumulative distribution of B and β is a fixed probability. Therefore, $[-EEB, EEB]$ is the smallest symmetrical region such that one would reject $\phi = 0$ with probability β.

This serves as yet another suggestion for deriving Δ and, once its value is fixed, REACT reaches the exact same conclusions as the two-stage testing procedure in Zhao et al. (2022) when using the $100(1-2\alpha)\%$ confidence set for testing.

Non-inferiority and Superiority Tests. REACT is closer in spirit to tests that combine superiority, non-inferiority and equivalence tests, such as Julious (2004) and other variations (Tryon, 2001; Goeman et al., 2010; L. M. Friedman et al., 2015; Zhao, 2016; Lakens et al., 2018; Zhao et al., 2022). While these approaches are highly informative and gaining in popularity, they are specific to certain tests and hypotheses, and require tailoring to fit each individual problem. Therefore, there is no guarantee that they will control type I error at α (Berger & Hsu, 1996) for any given problem (indeed, they are not designed to have type I error control globally). In Section "Applications", we provide examples of scenarios where it may be difficult to adapt such an approach, whereas our approach, REACT, remains user-friendly. Moreover, as shown in Proposition 4, REACT automatically controls the Family-Wise Error Rate (Lehmann, 1957; Wang & Shen, 1999; Gupta & Huang, 1981) (FWER); there is no need to use procedures such as Bonferroni correction to account for multiple testing.

HDI+ROPE, GFBST and S-values. The HDI+ROPE (Highest Density Interval+Region of Practical Equivalence) (Kruschke, 2010, 2018; Keysers et al., 2020) represents a specific instance of REACT. Indeed, its definition (see e.g. Kruschke, 2010, page 291)) directly corresponds to REACT with the confidence set C as the $(1-\alpha)$-level Bayesian Highest Density Credible Interval.

In a similar vein, the GFBST (Stern et al., 2017) (Generalized Full Bayesian Significance Test) corresponds to choosing C as the $(1-\alpha)$-level Bayesian Highest Posterior Density Region (HPD), given by

$$C(\mathcal{D}) = \{\theta \in \Theta : f(\theta|\mathcal{D}) \geq t_{1-\alpha}\},$$

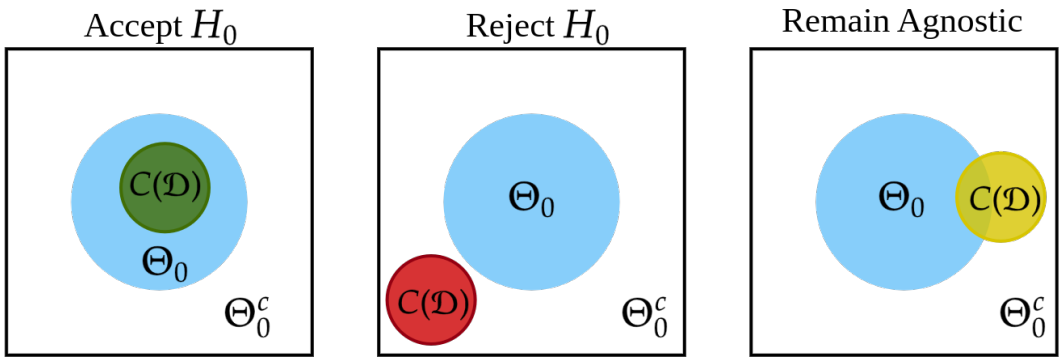
where $t_{1-\alpha}$ is chosen so that $\mathbb{P}(\theta \in C(\mathcal{D})|\mathcal{D}) = 1-\alpha$, and $f(\theta|\mathcal{D})$ is the posterior distribution of θ given data $\mathcal{D}$. New theoretical guarantees regarding the average coverage probabilities of such procedures are provided in Section "Statistical properties" and in the Appendix (Theorem 3). Furthermore, the approach presented by Patriota (2013) can also be viewed as a specific instance of REACT, where the sets C are constructed using s-values.

Three-way hypothesis tests. Three-way tests provide a more nuanced approach to hypothesis testing (Jones & Tukey, 2000; Rice & Krakauer, 2023) and were suggested *en passant* by J. Neyman, who argued that "The phrase 'do not reject H' is longish and cumbersome ... (and) should be distinguished from a 'three-decision problem' (in which the) actions are: (a) accept H, (b) reject H, and (c) remain in doubt" (Neyman, 1976, page 749). Neyman however did not develop this approach. To the best of our knowledge, Hays (1963) was the first one to present a formal three-way approach to hypothesis testing under a decision-theoretic framework.

Three-way hypothesis testing has several benefits. It can address concerns about replication and the limited publication of null results (Campbell & Gustafson, 2018) and is essential to the evolution of scientific theories (Esteves et al., 2019). Recent research has explored the many advantages of three-way tests. For example, Berg (2004) showed that three-way tests can control both types I and II error probabilities when both the null and the alternative hypotheses are precise. This is in contrast to two-way tests, where only one error can be controlled. Subsequently, Coscrato et al. (2020) generalized this approach to composite alternative hypotheses. REACT controls both type I and type II error probabilities to be no more than the same level α (Property 3). In addition, Esteves et al. (2016), Stern et al. (2017), and Esteves et al. (2023) showed that three-way tests can address logical inconsistencies that occur in standard two-way tests (Izbicki et al., 2012; Da Silva et al., 2015; Izbicki & Esteves, 2015; Fossaluza et al., 2017). This is one of the arguments made by Kruschke (2018) to justify why HDI+ROPE should be preferred over TOST+NHST.

A three-way testing approach that is very closely related to REACT is the PASS-test (Gross, 2014), which uses a region of equivalence combined with confidence intervals to reach decisions. The main differences between the methods is that REACT is more general (it can derive regions of equivalence and tests to combinations of parameters, instead of only uniparametric hypotheses) and has a series of

Figure 3 ■ Illustration of REACT to test a generic hypothesis $H_0 : \theta \in \Theta_0$. $C(\mathcal{D})$ is a confidence set built using data $\mathcal{D}$; the blue region represents the null hypothesis. Whenever $C(\mathcal{D})$ is a subset of Θ_0 (left), all plausible values for θ are in Θ_0. Hence, Θ_0 is accepted. Whenever $C(\mathcal{D})$ lies outside of Θ_0 (middle), no plausible value for θ is in Θ_0. Therefore, Θ_0 is rejected. Finally, if some points of $C(\mathcal{D})$ are inside Θ_0 and others are outside, then some plausible points for θ are compatible with Θ_0 and others are incompatible. In this situation, Θ_0 remains undecided.



useful properties particular to it (see Section "REACT and its properties"). One of them (logical coherence, Property 8) guarantees that no contradictory conclusions are reached when REACT is applied to multiple tests. The main advantage of using three-way tests such as these is that, since they are based on intervals instead of p-values, they provide greater intuitive appeal than p-value-based tests (Rainey, 2014).

Tests based on posterior probabilities and Bayes Factors. An alternative approach to testing hypotheses is the Bayesian framework. Under this perspective, one typically computes either the posterior probability of the null hypothesis, $\mathbb{P}(H_0|\mathcal{D})$, or the Bayes Factor, $\mathbb{P}(\mathcal{D}|H_0)/\mathbb{P}(\mathcal{D}|H_1)$, and rejects the null if the values are small (Berger, 1985). This approach solves many of the issues associated with NHST. For instance, Bayesian tests can accept the null hypothesis (Rouder et al., 2009; Kelter, 2020).

Furthermore, within a Bayesian framework, it is customary to formulate null hypotheses that are not precise (Edwards et al., 1963; Good, 2009; Berger, 1985), thereby circumventing the challenge of exclusively dealing with implausible hypotheses (Schervish, 1995). This, however, remains mostly restricted to uniparametric hypotheses, such as $H_0 : |\phi - \phi_0| \in [\delta_L, \delta_U]$, where $\delta_L \leq 0 \leq \delta_U$ are known beforehand (Hobbs & Carlin, 2007; Kruschke, 2018). The more popular alternative remains to assign probability masses to precise hypotheses (Jeffreys, 1961; Kass, 1993; Migon et al., 2014), done mostly due to practical reasons instead of a true representation of the researcher's beliefs.

Moreover, tests based both on posterior probabilities and on Bayes factors are not fully logically coherent in the sense described in Section "REACT and its properties" (Lavine & Schervish, 1999; Izbicki & Esteves, 2015). REACT can be used within a Bayesian context, although it is not based on computing posterior probabilities of the null hypothesis (see Section "REACT and its properties").

REACT and its properties

Next, we introduce the general version of REACT. Our goal is to test one or more hypotheses regarding parameters θ, which assume values in Θ, a subset of $\mathbb{R}^d$. In this context, a null hypothesis is of the form: $H_0 : \theta \in \Theta_0$, where Θ_0 is a subset of Θ and the alternative hypothesis, H_1, is $H_1 : \theta \in \Theta_0^c$. Whenever there is no ambiguity, Θ_0 is used instead of H_0. All definitions and proofs of the properties stated in this section are found in the Appendix.

In order to test H_0, REACT requires one to construct a region of values for θ. One possible region is obtained by using data, $\mathcal{D}$, to construct a confidence region for θ, $C(\mathcal{D})$. If a frequentist approach is used, then one requires that

$$\mathbb{P}_\theta\left(\theta \in C(\mathcal{D})\right) = 1 - \alpha, \text{ for all } \theta \in \Theta,$$

where $1-\alpha$ is the confidence level of $C(\mathcal{D})$. Notice that the randomness of this probability is on the data $\mathcal{D}$; once the data is observed there is no randomness left. For a Bayesian implementation of REACT one could use a $(1-\alpha)$-level credible region (Berger, 1985), that is, a region such that

$$\mathbb{P}\left(\theta \in C(\mathcal{D})|\mathcal{D}\right) = 1 - \alpha,$$

where the probability's randomness stems from the posterior distribution of θ given the data $\mathcal{D}$.

REACT tests H_0 using the following rule (illustrated in Figure 3):

$$\text{Decision} = \begin{cases} \textbf{Accept } H_0 \text{ if } C(\mathcal{D}) \subseteq \Theta_0, \\ \textbf{Reject } H_0 \text{ if } C(\mathcal{D}) \subseteq \Theta_0^c, \\ \textbf{Remain Agnostic} \text{ if } C(\mathcal{D}) \text{ intersects with both } \Theta_0 \text{ and } \Theta_0^c. \end{cases}$$

In what follows, `fREACT` denotes a `REACT` procedure based on a $(1-\alpha)$-level frequentist region, while `bREACT` denotes a `REACT` procedure based on a $(1-\alpha)$-level Bayesian credible region. We use `REACT` to refer to scenarios where the property in question holds for any version of $C(\mathcal{D})$. `fREACT` can also be obtained using standard p-values for point null hypotheses:

Property 1. [Computation of `fREACT` using p-values] Let p-val$_{\mathcal{D}}(\theta_0)$ be a p-value for the hypothesis $H_0 : \theta = \theta_0$. The procedure

$$\begin{cases} \textbf{Accept } H_0 \text{ if } \max_{\theta \in \Theta_0^c} \text{p-val}_{\mathcal{D}}(\theta) \leq \alpha, \\ \textbf{Reject } H_0 \text{ if } \max_{\theta \in \Theta_0} \text{p-val}_{\mathcal{D}}(\theta) \leq \alpha, \\ \textbf{Remain Agnostic} \text{ otherwise.} \end{cases}$$

is equivalent to calculating `fREACT` through the (1-α)-level confidence set: $C(\mathcal{D}) := \{\theta \in \Theta : \text{p-val}_{\mathcal{D}}(\theta) > \alpha\}$.

Oftentimes, the parameter space can be decomposed as $\Theta = \Phi \times \Psi$, where $\phi \in \Phi$ are parameters that will be tested and $\psi \in \Psi$ are nuisance parameters. Theorem 1 in the Appendix shows that, as long as a confidence (or credible) set for the parameters of interest ϕ is available, `REACT` can easily handle nuisance parameters. Moreover, Theorem 4 shows how `fREACT` can be computed using p-values in this setting, and Theorem 5 shows its Bayesian counterpart.

In many problems, one is interested in testing different hypotheses, each one dealing with a different set of parameters. For example, in a linear regression model where $\beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$, one is often interested in testing $H_0 : \beta_i = 0$ for each i at a time. `REACT` for such hypotheses can be obtained at a low computational cost as long as a confidence set for all parameters is available:

Property 2. [Easy computation of `REACT` for low-dimensional hypotheses] Assume that the parameter space can be decomposed as $\Theta = \Phi \times \Psi$ and let $H_0 : \phi \in \Phi_0$, $\Phi_0 \subset \Phi$, be the hypothesis of interest. Also, let $C(\mathcal{D})$ be a region estimator on Θ and $C_\Phi(\mathcal{D})$ be the projection of the confidence region C on the parameters ϕ, that is,

$$C_\Phi(\mathcal{D}) := \{\phi \in \Phi : \exists \psi \in \Psi \text{ such that } (\phi, \psi) \in C(\mathcal{D})\}.$$

Then, `REACT` can be computed via

$$\text{Decision} = \begin{cases} \textbf{Accept } H_0 \text{ if } C_\Phi(\mathcal{D}) \subseteq \Phi_0, \\ \textbf{Reject } H_0 \text{ if } C_\Phi(\mathcal{D}) \subseteq \Phi_0^c = \Phi - \Phi_0, \\ \textbf{Remain Agnostic} \text{ otherwise.} \end{cases}$$

In words, `REACT` only requires evaluating whether the projection of the confidence set belongs to the null hypothesis.

In the context of multiple testing, Theorem 1 in the Appendix ensures that one only needs to build a confidence region for the parameters of interest, while Property 2 guarantees that using projections of such region will not affect the conclusions or harm the properties of `REACT`. If there are nuisance parameters, using Property 2 directly on $C_\Theta(\mathcal{D})$ is not advised, as projections often lead to tests with a lower level of significance than the nominal one, potentially having lower statistical power.

The importance of Property 2 is illustrated in Section "Cambridge Cognition Examination", where we show how to implement `REACT` for multiple pairwise comparisons.

`REACT` has desirable statistical and logical properties which we explore in what follows.

Statistical properties

Whenever $C(\mathcal{D})$ is a confidence region, `REACT` has desirable frequentist statistical properties. For instance, `fREACT` controls not only the type I error rate but also the type II error rate:

Property 3. [Type I and Type II error rate control] `fREACT` has Type I and Type II errors rates of at most α.

Although the control of type I error rate is a standard criterion for hypothesis tests at point nulls, such control is seldom obtained for the type II error rates. `fREACT` obtains the latter by remaining undecided when the data is not sufficiently informative about the null hypothesis.

Next, we show that when conducting multiple hypothesis tests, `fREACT` controls a strong version of family-wise error rate (FWER), in the sense that it controls the probability of making at least one Type I or Type II error. This is more stringent than the standard notion of FWER, which typically focuses only on controlling Type I errors.

Property 4. [Family-wise error rate (FWER) control] If several hypotheses are tested, fREACT controls a generalized version of the Family-Wise Error Rate (FWER (Tukey, 1953)) over multiple hypothesis tests. That is, for every $\theta \in \Theta$,

$$\begin{cases} \text{FWER}_I := \mathbb{P}_\theta(\text{At least one correct hypothesis is rejected}) \leq \alpha, \\ \text{FWER}_{II} := \mathbb{P}_\theta(\text{At least one false hypothesis is accepted}) \leq \alpha, \\ \mathbb{P}_\theta(\text{At least one correct hypothesis is rejected or one false hypothesis is accepted}) \leq \alpha. \end{cases}$$

Remarkably, this control is achieved without necessitating the correction of significance levels for multiple testing. See Theorem 3, in the Appendix, for a Bayesian counterpart of this result using bREACT.

Next, we show that REACT is typically consistent in the sense that, as the sample size increases, REACT rejects each hypothesis that is false and accepts each one that is true. Thus, REACT does not have the issue in which the null hypothesis is rejected as the sample size increases.

Property 5. [Consistency as n increases] If the confidence set $C(\mathcal{D})$ converges to the true value of the parameter (fixed), θ, (see Definition 11), and θ does not lie on the boundary of the null hypothesis Θ_0, then, as the sample size increases, the probability that REACT accepts H_0 when it is true goes to one, and the probability that it rejects H_0 goes to zero. This property is illustrated in Figure 2.

Besides the above properties, one might also be interested in the "best" confidence set to be used in step 2 of REACT. In the following, we discuss a first development of theory for answering this question.

As a first challenge, one must answer what is a "best" test in the context of REACT. Since fREACT controls both type I and type II errors, a possible generalization of standard theory is to seek a test that minimizes the type III error, that is, the probability of remaining undecided. However, it can be hard to find such a test that minimizes this error uniformly among all parameter values and hypotheses that can be tested.

The following property provides a first answer when one restricts attention to interval confidence sets and unilateral and bilateral hypotheses. A formal mathematical development is presented in Appendix B.

Property 6. Consider that one wishes to test solely unilateral and bilateral hypotheses regarding the parameters of interest. Also, consider that there exists a standard uniformly most powerful (UMPU) test for each point hypothesis. If fREACT uses the confidence set obtained from inverting the UMPU tests, then the probability that it remains undecided is uniformly lower than that of every region test based on an interval confidence set that has the same type I and type II error rates.

Notice, however, that in more general settings there is no single best UMPU solutions.

Finally, under a Bayesian perspective, accepted hypotheses have high posterior probability, while rejected hypotheses have low posterior probability:

Property 7. [Posterior probability of the null hypothesis] If bREACT accepts H_0, then the posterior probability of H_0 is larger than $1-\alpha$. Moreover, if bREACT rejects H_0, then the posterior probability of H_0 is smaller than α.

Notice however that the reverse is not true: bREACT may not accept a hypothesis with large posterior probability. One example is when bREACT remains agnostic because at least one element of H_0 does not intersect with the credible region, due to it residing in an area with very small posterior probability.

Logical properties

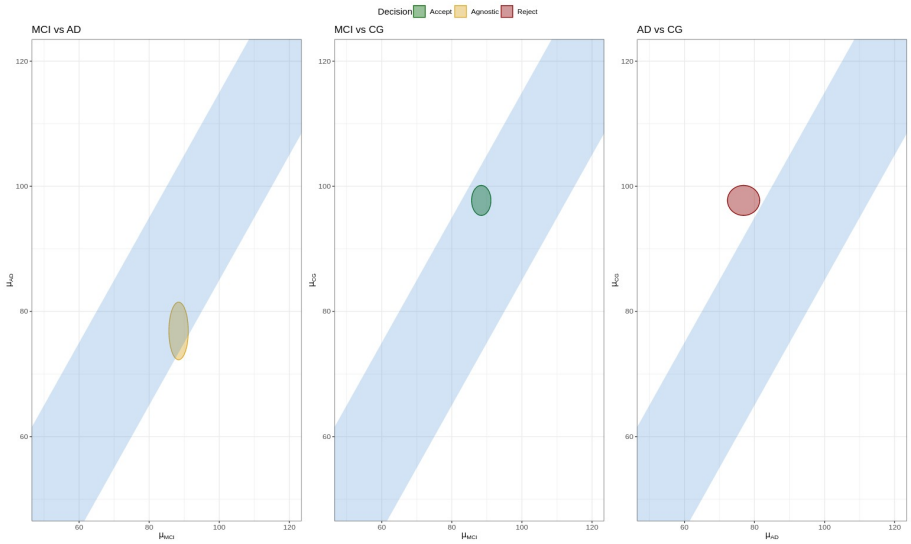
Besides having good statistical properties, REACT is also coherent from a logical perspective:

Property 8. [Logical coherence] REACT is logically coherent in the sense described in Definition 2.

That is, if one treats accepted hypotheses as true and rejected hypotheses as false, then no logical contradiction is obtained. For example:

- If $H_0 : |\mu_A - \mu_B| < \Delta$ is accepted, then so is $H_0 : |\mu_A - \mu_B| \leq \Delta$.
- If both $H_0 : \mu_A - \mu_B \leq 1$ and $H_0 : \mu_A - \mu_B \geq 0$ are accepted, then $H_0 : 0 \leq \mu_A - \mu_B \leq 1$ will be as well.
- Consider pairwise null hypotheses of the form $H_0^{i,j} : |\mu_i - \mu_j| \leq \Delta$, where μ_i's are parameters of the model, and a global null hypothesis $H_0 : \max_{i,j}\{|\mu_i - \mu_j|\} \leq \Delta$. If all $H_0^{i,j}$'s are accepted, so is H_0. Similarly, if at least one $H_0^{i,j}$ is rejected, so is H_0.

Figure 4 ■ Pairwise group comparisons given by REACT for the Cambridge Cognition Examination example. The ellipses are projections of the multivariate confidence set of Equation 2 on each pair of parameters (μ_i, μ_j), while the blue regions represent the null hypotheses $|\mu_i - \mu_j| \leq \Delta$ for the various groups. Property 2 guarantees it is enough to evaluate these projections to test these hypotheses.



Most statistical procedures do not satisfy this property, which makes their outcomes hard to interpret (Izbicki & Esteves, 2015). The reader is referred to Hansen and Rice (2023) for other types of coherence.

A consequence of Property 8 is that, unlike standard tests, where it is crucial to determine which hypothesis is labeled as the "null" and which one is labeled as the "alternative," REACT is indifferent to this choice. In other words, the choice of whether the null hypothesis is $H_0 : |\mu_A - \mu_B| \leq \Delta$ or $H_0 : |\mu_A - \mu_B| > \Delta$ does not materially affect the conclusions. If the former is rejected, it means that the latter is accepted and vice versa:

Property 9. [No need to specify which one is the null hypothesis] The null and the alternative hypotheses can be exchanged without materially affecting conclusions, and therefore there is no need to distinguish the labels "null" and "alternative" hypothesis.

Applications

Cambridge Cognition Examination

The Cambridge Cognition Examination (Roth et al., 1986) (CAMCOG) is a widely-used questionnaire for measuring the extent of dementia and assessing the level of cognitive impairment. We analyze data from Cecato et al. (2016) to investigate whether CAMCOG scores can distinguish between three groups of patients: control (CG), mild cognitive impairment (MCI), and Alzheimer's disease (AD). We assume that the score of the k-th patient in group i is given by $Y_{i,k} = \mu_i + \epsilon_{i,k}$, where μ_i is the population average for group i and $\epsilon_{i,k}$ is a Gaussian random variable with mean 0 and variance σ_i^2. Our analysis focuses on testing the hypothesis

$$H_0 : |\mu_i - \mu_j| \leq \Delta \tag{1}$$

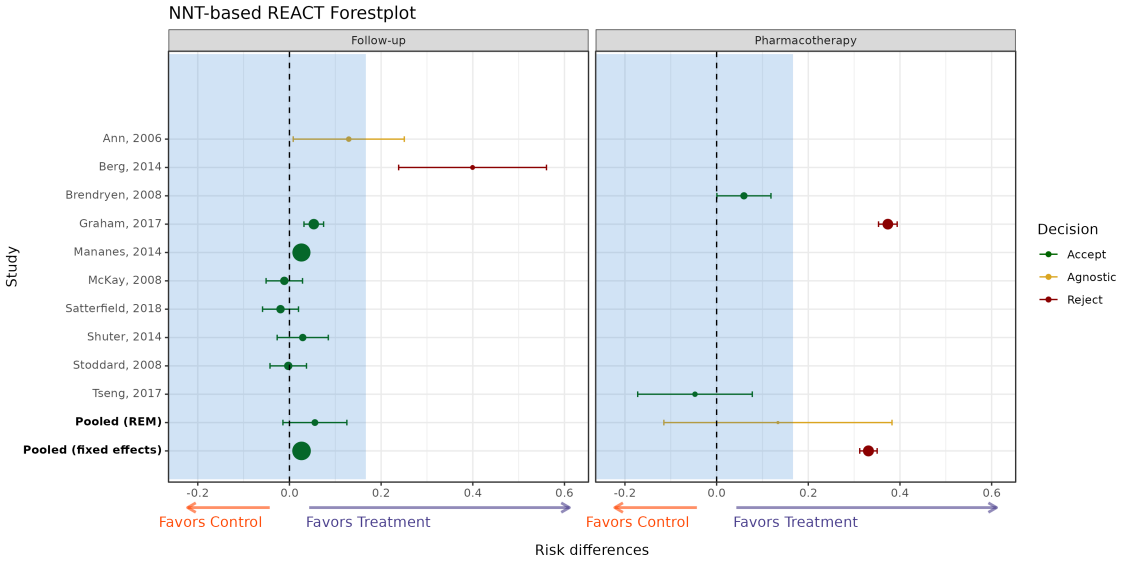
for all pairs of groups.

An initial approach to address this problem involves constructing confidence sets for individual parameters, specifically $\phi := |\mu_i - \mu_j|$. The process of building these confidence sets is illustrated in Figure 2, where we explore how the sample size influences the test results. The confidence intervals are constructed using the standard confidence set for the difference between two means, assuming a Gaussian distribution (Diez et al., 2012), where each difference may have a different variance.

The value of Δ was determined using the classification dissimilarity introduced by Esteves et al. (2019). When the pragmatic hypothesis does not hold, there exists a classifier based on the CAMCOG score which is able to discriminate the groups under comparison with an accuracy (that is, a probability of predicting the correct class) of at least 80%.

Although the previous approach is sensible, it is technically not a fully-REACT approach: REACT requires a single confidence set $C(\mathcal{D})$ to be used to test all hypotheses. A full-REACT can be obtained by using the following

Figure 5 ■ Forest plots for the meta-analyses described in Section "Meta-analysis", along with the decisions made by `REACT`. The point estimate sizes are proportional to the inverse of the variances of each study. Our approach indicates that treatment is no better than control in the follow-up studies (that is, the null hypothesis is accepted), while it remains agnostic or is rejected in the pharmacotherapy studies, depending on the choice of the pooling method. The null region of equivalence was obtained using the Number Necessary to Treat (NNT).



$100(1-\alpha)\%$ confidence region for the vector of means $\boldsymbol{\mu} = (\mu_{CG}, \mu_{MCI}, \mu_{AD})$ (Johnson & Wichern, 2002; Meeker & Escobar, 1995):

$$(\overline{\boldsymbol{x}} - \boldsymbol{\mu})' \hat{\Sigma}_{\overline{\boldsymbol{x}}}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha), \tag{2}$$

where $\overline{\boldsymbol{x}}$ is the sample mean, $\hat{\Sigma}_{\overline{\boldsymbol{x}}}$ is the estimate of the covariance matrix of $\overline{\boldsymbol{X}}$, p is the size of $\boldsymbol{\mu}$ and $\chi_\cdot^2(\cdot)$ is the quantile function of the χ^2-distribution. Moreover, $\hat{\Sigma}_{\overline{\boldsymbol{x}}}$ is a diagonal matrix in this case because the data on each subject was collected independently.

Once the confidence region has been derived, any hypothesis can be tested through the `REACT` framework by checking if the confidence region is contained in H_0. While evaluating if $C(\mathcal{D})$ is contained in H_0 is feasible, Property 2 ensures that the same conclusions can be reached by using the projection (in this case, the ellipses of Figure 4) of $C(\mathcal{D})$ on (μ_i, μ_j), requiring fewer calculations than using $C(\mathcal{D})$ and allowing for graphical visualization of the results in two dimensions. By projection, we mean the set

$$\left\{(\mu_i, \mu_j) \in \mathbb{R}^2 : \exists \mu_k \in \mathbb{R} \text{ with } (\mu_i, \mu_j, \mu_k) \in C(\mathcal{D})\right\}.$$

In Figure 4, the blue area represents the region of equivalence and each ellipse is a confidence region whose color implies the result of the test (yellow for remaining undecided, green for acceptance and red for rejection). While there is not enough evidence to conclude if MCI patients can be discriminated from those with AD (absence of evidence of a practical effect), the CAMCOG is unable to differentiate between MCI and the control group (evidence of absence of a practical effect). This is not the case when comparing AD and CG groups, supporting the idea that the CAMCOG is able to differentiate between them (evidence of presence). Finally, Property 8 and the fact that $|\mu_{AD} - \mu_{CG}| \leq \Delta$ was rejected implies that the hypothesis of no relevant difference between all groups ($H_0 : \max\{|\mu_{MCI} - \mu_{AD}|, |\mu_{MCI} - \mu_{CG}|, |\mu_{AD} - \mu_{CG}|\} \leq \Delta$) is also rejected. A Bayesian version of this analysis is presented in Figure A2 and leads to the same conclusion.

Meta-analysis

This application is an adaptation of the meta-analysis originally performed by da Silva Teixeira et al. (2022) and uses the Number Necessary to Treat (NNT) as a means to obtain the desired region of equivalence for the risk differences. The NNT is the expectation of the least people required for the treatment to present better results than the control and thus is commonly used in the literature as an indicator of clinical significance and its value can be easily provided (Citrome, 2011).

The original study evaluates patients' adherence to

tobacco cessation protocols comparing traditional approaches combined with computer-assisted health technologies to traditional approaches themselves. The treatments are compared by two main outcomes: the adherence to the follow-up period of treatments without any drug (labeled as "Follow-up") and the adherence to the pharmacotherapy, on studies that used any drug besides nicotine replacement (labeled "Pharmacotherapy"). Since the study evaluates the adherence to treatment, not the efficacy of the treatment itself, the NNT of interest should be at least smaller than the best treatments recommended by the literature (Cahill et al., 2016; Stead et al., 2012), which leads to NNT < 11.

Figure 5 presents the forest plot for both treatments, with the blue area representing the region of equivalence substituting the hypothesis that the risk difference is less or equal to 0 (treatment no better than control), the X-axis providing the confidence intervals obtained from each study and the "pooled" label representing the interval that results from aggregating all studies. The aggregation was done either through ***(i)*** a Random Effects Model (REM) with random intercepts (Bakbergenuly & Kulinskaya, 2018)—which mirrors the pooling strategy of the original study—on the risk difference of each study or ***(ii)*** a fixed effects model, which assumes that the samples from all studies come from exactly the same population.

Unlike da Silva Teixeira et al. (2022), in this case, the difference between the probabilities of success of treatment and control was used instead of the relative risk. This was due to the fact that the NNT is the inverse of such a difference, so one can be directly translated into the other. Since the interest is to evaluate clinical significance, the region of equivalence chosen was $[-1, 1/6]$, meaning that the NNT of the study has to be less than 6 for its results to be practically significant (treatment better than control). A Bayesian version of this analysis is presented in Figure A3 and leads to the same conclusion.

For the follow-up outcome studies, most of the intervals are contained in the region of equivalence, leading to the acceptance of the hypothesis that treatment is no better than control. In particular, the pooled results leads to the conclusion that the treatment is not worth pursuing.

On the other hand, the pharmacotherapy studies are scarcer and do not agree with each other. This discrepancy results in a poor estimate for the REM pooled interval, leading to the decision of remaining agnostic, i.e., more studies need to be conducted to then reach an assertive conclusion. The fixed effects pooled interval leads to rejection, as the study that rejected the hypothesis is considerably more precise than the others. Therefore, the decision to remain undecided or reject the hypothesis depends on which pooling strategy the researcher finds more appealing for this case.

Implications of the `REACT` Method for Machine Learning and AI

While this work has focused on formulations more adherent to Statistics, `REACT` can also be extended to contexts akin to Machine Learning and AI. By addressing limitations of NHST, `REACT` provides a principled way to evaluate models, handle multiple comparisons, and quantify uncertainty in automated decision-making systems. Here, we present a few of its advantages in this context.

Sample Size Sensitivity in Large-Scale Machine Learning In modern machine learning applications, datasets are often massive, which can lead traditional null hypothesis significance testing (NHST) to flag even negligible differences as statistically significant. This sensitivity to sample size is especially problematic when the goal is to assess whether a difference is practically relevant. `REACT` addresses this issue by testing **pragmatic** rather than **point-null** hypotheses. Instead of evaluating whether two models have exactly the same performance (e.g., $H_0 : \mu_1 = \mu_2$), it tests whether the difference is within a region of equivalence (e.g., $H_0 : |\mu_1 - \mu_2| \leq \Delta$), where Δ reflects a threshold of practical significance. This prevents trivial differences from leading to misleading conclusions and promotes more meaningful comparisons. The benefits of `REACT` are especially evident in supervised learning tasks, where μ_i could represent a performance metric such as AUC, F1 score, or accuracy. In A/B testing, it might represent the conversion rate of a strategy, while in reinforcement learning, it could be the expected value of a policy. By focusing on whether an observed effect is large enough to matter, rather than merely detectable, `REACT` provides a robust framework for inference that remains consistent and interpretable even as sample sizes grow (Biecek et al., 2024a; J. Friedman et al., 2001).

Three-Way Decision Framework in Model Selection and Ensembles Model selection in machine learning is often based on significance testing, information criteria, or validation set performance. However, these methods do not account for practical equivalence. `REACT`'s three-way decision framework (accept, reject, remain agnostic) offers a more nuanced alternative: when models are practically indistinguishable, selection can prioritize other criteria like interpretability or computational efficiency; when one model outperforms another by a meaningful margin, the better model is preferred (Biecek et al., 2024b; Domingos, 2012); and when results are inconclusive, further data collection or ensemble approaches may be warranted. In ensemble learning, `REACT` can improve model diversity by selecting models that are practically different rather than merely statistically distinct, weighting ensemble components based on their unique predictive contributions, and organizing models hierarchically by clustering those that

are practically equivalent. The logical coherence of REACT ensures that these selection processes remain internally consistent, avoiding contradictions in multi-hypothesis settings.

Error Rate Control and AutoML Reliability Automated Machine Learning (AutoML) systems make numerous decisions regarding feature selection, hyperparameter tuning, and model architecture. NHST-based methods typically focus only on controlling Type I errors, often at the expense of Type II errors. REACT improves AutoML reliability by controlling both Type I and Type II errors at level α, preventing the inclusion of irrelevant features (Type I errors) and the exclusion of relevant ones (Type II errors), avoiding incorrect conclusions about model superiority when comparing architectures, and maintaining statistical power despite multiple testing corrections. By allowing an "agnostic" decision, REACT provides AutoML systems with a principled way to defer decisions until more data is available, reducing overconfidence in uncertain cases (Zöller & Huber, 2022).

Fairness and Bias in Machine Learning Fairness evaluations in AI often rely on NHST to compare model performance across demographic groups. However, NHST-based fairness testing has notable limitations: large datasets make even small, practically irrelevant differences statistically significant; failing to reject a null hypothesis of "no difference" does not imply fairness; and multiple comparisons across demographic categories necessitate stringent corrections, reducing statistical power. REACT addresses these challenges by defining fairness through practical equivalence ($H_0 : |\mu_A - \mu_B| \leq \Delta$), where Δ represents an acceptable level of disparity; allowing three-way decisions, i.e., accepting fairness when evidence supports equivalence, rejecting fairness when disparities exceed meaningful thresholds, and remaining agnostic when evidence is inconclusive; handling multiple demographic comparisons with built-in FWER control; and ensuring logical consistency, preventing contradictory fairness conclusions (Mitchell, Shankar, et al., 2021; Mehrabi, Morstatter, et al., 2021). By shifting the focus from statistical significance to practical significance, REACT enables more interpretable and actionable fairness assessments in AI.

Technical Challenges and Future Directions. While REACT offers clear benefits, its application to machine learning settings presents notable technical challenges. Many relevant performance metrics—such as AUC and F1 score—are aggregate, making the construction of valid confidence sets difficult. Existing approaches often rely on asymptotic approximations or resampling methods, which may not yield accurate coverage, especially in small-sample scenarios. Moreover, in such cases with small samples, the resulting sets tend to be overly wide, leading frequently to agnostic outcomes. Another difficulty lies in specifying a meaningful threshold Δ for practical equivalence. For metrics like the F1 score, this choice is not straightforward and typically requires strong domain-specific knowledge. Furthermore, Δ may not scale linearly across different tasks or metrics. This motivates the development of more general dissimilarity functions that flexibly capture meaningful performance differences while being easier to calibrate. These challenges point to important directions for future research: designing computationally efficient procedures for constructing confidence sets around interpretable, task-specific metrics, and developing principled methods for choosing Δ in practice.

Final Remarks

We have shown the effectiveness of REACT in overcoming numerous difficulties with traditional NHSTs. A noteworthy aspect of REACT is its ability to differentiate between "evidence of absence" and "absence of evidence" of a practical effect, a crucial factor for conducting meta-analyses and comparing effects among different groups. Also, REACT does not lead to an automatic rejection when the sample size is large.

We have argued that REACT, building upon equivalence tests and three-way decision procedures, possesses numerous advantages over other approaches aimed at resolving NHST-related challenges. Once the null hypothesis of practical interest is specified, REACT only requires a confidence region to reach a decision, and thus researchers from diverse domains can readily apply our approach. Moreover, REACT seamlessly integrates with confidence sets, which are widely regarded as more informative than hypothesis tests and p-values (Wasserstein & Lazar, 2016). If setting an equivalence region is feasible, this integration results in a framework that surpasses traditional NHST, enhancing the interpretability of statistical inferences. Finally, REACT can be used both within the frequentist and Bayesian frameworks.

Defining the threshold associated to the pragmatic region, Δ, is a delicate and domain-specific task. Poor choices of Δ may compromise the practical interpretation of results. To address this, we recommend conducting sensitivity analyses over a range of Δ values to assess the robustness of conclusions. This anchors the choice of Δ in expert judgment and its observed consequences. In particular, it can be helpful to identify the smallest and largest Δ values for which the decision is non-agnostic. Moreover, the quality of REACT's conclusions critically depends on the properties of the confidence or credibility set. If C is constructed under violated assumptions (such as normality or independence), or if it is overly wide due to low power, the resulting decisions may be flawed or agnostic. In this

sense, REACT inherits the limitations of the underlying inferential machinery used to build C. Notice however that although REACT does not eliminate low power issues, it surfaces them transparently, avoiding misleading dichotomous conclusions.

Our examples and discussion focus on parametric models, where confidence intervals for means or differences are standard. Extending REACT to non-parametric or semi-parametric models presents additional challenges, particularly in constructing appropriate confidence/credibility regions and the associated computational costs. For the former, several strategies have been proposed to build regions with good properties (Genovese & Wasserman, 2005; Robins & van der Vaart, 2006; Davies et al., 2009; Park et al., 2023), especially in the Bayesian context, such as in Phadia (2016, Section 6.3) and Lassance et al., 2024. As for the latter, the computational cost can be circumvented when the delta method is available, but for more complex parameters using bootstrap or MCMC procedures might be required, which may be unreliable or too slow in certain contexts. In future work we will adapt REACT to these settings accounting for such setbacks.

Adopting REACT may contribute to increased confidence in scientific findings by making the decision-making process in hypothesis testing more transparent and by clearly presenting outcomes, including those cases where evidence is inconclusive. This is particularly pertinent in the realms of meta-science and replication studies, which hold significant importance in the current scientific landscape.

To support the practical implementation of REACT, we have developed an R package (available at github.com/Monoxido45/REACT) that simplifies its application for common models. This user-friendly package empowers researchers to efficiently implement our approach, fostering broader adoption and advancing scientific investigations. Embracing such tools and methodologies can contribute to more rigorous and reliable research practices, benefitting the scientific community as a whole.

Authors' note

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, FAPESP (grants 2019/11321-9, 2022/08579-7 and 2023/07068-1) and CNPq (grants 309607/2020-5, 422705/2021-7 and 305065/2023). Correspondence concerning this article should be addressed to Rafael Izbicki, Department of Statistics, Federal University of São Carlos, Rodovia Washington Luis km 235, São Carlos, São Paulo, Brazil, 13565-905. E-mail: rafaelizbicki@gmail.com.

References

Alter, U., & Counsell, A. (2021, June). *Determining negligible associations in regression*. Retrieved July 13, 2025, from https://doi.org/10.17605/OSF.IO/W96XE

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485–485. doi: 10.1136/bmj.311.7003.485.

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat ($p > 0.05$): Significance thresholds and the crisis of unreplicable research. *PeerJ*, *5*, e3544. doi: 10.7717/peerj.3544.

Bakbergenuly, I., & Kulinskaya, E. (2018). Meta-analysis of binary outcomes via generalized linear mixed models: A simulation study. *BMC Medical Research Methodology*, *18*(1). doi: 10.1186/s12874-018-0531-9.

Bartolucci, A. A., Tendera, M., & Howard, G. (2011). Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *The American Journal of Cardiology*, *107*(12), 1796–1801. doi: 10.1016/j.amjcard.2011.02.325.

Bauer, P., & Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*, *83*(4), 934–937. doi: 10.1093/biomet/83.4.934.

Berg, N. (2004). No-decision classification: An alternative to testing for statistical significance. *The Journal of Socio-Economics*, *33*(5), 631–650. doi: 10.1016/j.socec.2004.09.036.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis*. Springer-Verlag. doi: 10.1007/978-1-4757-4286-2.

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*(4), 283–319. doi: 10.1214/ss/1032280304.

Biecek, P., et al. (2024a). An experimental study on the rashomon effect of balancing methods in imbalanced classification. *arXiv preprint arXiv:2405.01557*, *99*, 1–99.

Biecek, P., et al. (2024b). Performance is not enough: The story told by a rashomon quartet. *Journal of Computational and Graphical Statistics*, *99*, 1–99.

Cahill, K., Lindson-Hawley, N., Thomas, K. H., Fanshawe, T. R., & Lancaster, T. (2016). Nicotine receptor partial agonists for smoking cessation. *Cochrane Database of Systematic Reviews*. doi: 10.1002/14651858.CD006103.pub7.

Campbell, H. (2020). Equivalence testing for standardized effect sizes in linear regression. *arXiv preprint arXiv:2004.01757*. doi: 10.48550/arXiv.2004.01757.

Campbell, H., & Gustafson, P. (2018). Conditional equivalence testing: An alternative remedy for publication

bias. *PloS one*, *13*(4), e0195145. doi: 10.1371/journal.pone.0195145.

Campo, M., & Lichtman, S. W. (2008). Interpretation of research in physical therapy: Limitations of null hypothesis significance testing. *Journal of Physical Therapy Education*, *22*(1), 43–48. doi: 10.1097/00001416-200801000-00007.

Cecato, J. F., Martinelli, J. E., Izbicki, R., Yassuda, M. S., & Aprahamian, I. (2016). A subtest analysis of the Montreal cognitive assessment (MoCA): which subtests can best discriminate between healthy controls, mild cognitive impairment and Alzheimer's disease? *International Psychogeriatrics*, *28*(5), 825–832. doi: 10.1017/S1041610215001982.

Citrome, L. (2011). The tyranny of the p-value: Effect size matters. *Klinik Psikofarmakoloji Bülteni-Bulletin of Clinical Psychopharmacology*, *21*(2), 91–92. doi: 10.5455/bcp.20110706020600.

Cohen, J. (1992). Things I have learned (so far). *98th Annual Convention of the American Psychological Association*. doi: 10.1037/0003-066X.45.12.1304.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. doi: 10.1037/0003-066x.49.12.997.

Cohen, J. (2013, May). *Statistical power analysis for the behavioral sciences* (2nd ed.) [Chapter 2, pg. 20]. Routledge. doi: 10.4324/9780203771587.

Coscrato, V., Izbicki, R., & Stern, R. B. (2020). Agnostic tests can control the type I and type II errors simultaneously. *Brazilian Journal of Probability and Statistics*, *34*(2), 230–250. doi: 10.1214/19-BJPS431.

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 292–309. doi: 10.1111/bmsp.12045.

Da Silva, G. M., Esteves, L. G., Fossaluza, V., Izbicki, R., & Wechsler, S. (2015). A Bayesian decision-theoretic approach to logically-consistent hypothesis testing. *Entropy*, *17*(10), 6534–6559. doi: 10.3390/e17106534.

da Silva Teixeira, R., Nazareth, I. F., de Paula, L. C., do Nascimento Duque, G. P., & Colugnati, F. A. B. (2022). Adherence to computational technologies for the treatment of smoking cessation: Systematic review and meta-analysis. *International Journal of Mental Health and Addiction*. doi: 10.1007/s11469-022-00839-5.

Davies, P. L., Kovac, A., & Meise, M. (2009). Nonparametric regression, confidence regions and regularization. *The Annals of Statistics*, *37*(5B), 2597–2625. doi: 10.1214/07-AOS575.

Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2012). *Openintro statistics*. OpenIntro Boston, MA, USA. doi: 10.5070/t573020084.

Diggle, P. J., & Chetwynd, A. G. (2011). *Statistics and scientific method: An introduction for students and researchers*. Oxford University Press. doi: 10.1093/acprof:oso/9780199543182.001.0001.

Dixon, P. M., & Pechmann, J. H. K. (2005). A statistical test to show negligible trend. *Ecology*, *86*(7), 1751–1756. doi: 10.1890/04-1343.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193. doi: 10.1037/h0044139.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press. doi: 10.1017/CBO9780511761676.

Esteves, L. G., Izbicki, R., Stern, J. M., & Stern, R. B. (2016). The logical consistency of simultaneous agnostic hypothesis tests. *Entropy*, *18*(7), 256. doi: 10.3390/e18070256.

Esteves, L. G., Izbicki, R., Stern, J. M., & Stern, R. B. (2023). Logical coherence in bayesian simultaneous three-way hypothesis tests. *International Journal of Approximate Reasoning*, *152*, 297–309. doi: 10.1016/j.ijar.2022.10.019.

Esteves, L. G., Izbicki, R., Stern, J. M., & Stern, R. B. (2019). Pragmatic hypotheses in the evolution of science. *Entropy*, *21*(9), 883. doi: 10.3390/e21090883.

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. doi: 10.1590/2176-9451.19.4.027-029.ebo.

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., & Kruger, A. (2018). The epistemic importance of establishing the absence of an effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244. doi: 10.1177/2515245918770407.

Fossaluza, V., Izbicki, R., da Silva, G. M., & Esteves, L. G. (2017). Coherent hypothesis testing. *The American Statistician*, *71*(3), 242–248. doi: 10.1080/00031305.2016.1237893.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (2015). *Fundamentals of clinical trials*. Springer. doi: 10.1007/978-3-319-18539-2.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2. doi: 10.1037/a0024338.

Genovese, C. R., & Wasserman, L. (2005). Confidence sets for nonparametric wavelet regression. *The Annals of Statistics*, *33*(2). doi: 10.1214/009053605000000011.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political research quarterly*, *52*(3), 647–674. doi: 10.1177/106591299905200309.

Goeman, J. J., Solari, A., & Stijnen, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in medicine*, *29*(20), 2117–2125. doi: 10.1002/sim.4002.

Good, I. J. (2009). Some logic and history of hypothesis testing. In M. Mouse (Ed.), *Good thinking: The foundations of probability and its applications* (pp. 129–148). Dover Publications.

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part i: The cohen'sd family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. doi: 10.20982/tqmp.14.4.p242.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. doi: 10.1007/s10654-016-0149-3.

Gross, J. H. (2014). Testing what matters (if you must test at all): A context-driven approach to substantive and statistical significance. *American Journal of Political Science*, *59*(3), 775–788. doi: 10.1111/ajps.12149.

Gupta, S. S., & Huang, D.-Y. (1981). *Multiple statistical decision theory: Recent developments*. Springer New York. doi: 10.1007/978-1-4612-5925-1.

Hansen, S., & Rice, K. (2023). Coherent tests for interval null hypotheses. *The American Statistician*, *77*(1), 20–28. doi: 10.1080/00031305.2022.2050299.

Hays, W. L. (1963). *Statistics for psychologists*. Holt, Rinehart &. Winston.

Hobbs, B. P., & Carlin, B. P. (2007). Practical bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(1), 54–80. doi: 10.1080/10543400701668266.

Izbicki, R., & Esteves, L. G. (2015). Logical consistency in simultaneous statistical test procedures. *Logic Journal of the IGPL*, *23*(5), 732–758. doi: 10.1093/jigpal/jzv027.

Izbicki, R., Fossaluza, V., Hounie, A. G., Nakano, E. Y., & de Braganca Pereira, C. A. (2012). Testing allele homogeneity: The problem of nested hypotheses. *BMC genetics*, *13*, 1–11. doi: 10.1186/1471-2156-13-103.

Jefferson, T., Dooley, L., Ferroni, E., Al-Ansary, L. A., van Driel, M. L., Bawazeer, G. A., Jones, M. A., Hoffmann, T. C., Clark, J., Beller, E. M., et al. (2023). Physical interventions to interrupt or reduce the spread of respiratory viruses. *Cochrane database of systematic reviews*, (1). doi: 10.1002/14651858.CD006207.pub6.

Jeffreys, H. (1961). *Theory of probability* (Third). Oxford. doi: 10.1093/oso/9780198503682.001.0001.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Prentice Hall.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological methods*, *5*(4), 411. doi: 10.1037/1082-989x.5.4.411.

Julious, S. A. (2004). Sample sizes for clinical trials with normal data. *Statistics in medicine*, *23*(12), 1921–1986. doi: 10.1002/sim.1783.

Kadane, J. B. (2016). Beyond hypothesis testing. *Entropy*, *18*(5), 199. doi: 10.3390/e18050199.

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *42*(5), 551–560. doi: 10.2307/2348679.

Kelter, R. (2020). Bayesian alternatives to null hypothesis significance testing in biomedical research: A non-technical introduction to bayesian inference with jasp. *BMC Medical Research Methodology*, *20*(1), 1–12. doi: 10.1186/s12874-020-00980-6.

Keren, G., & Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. L. Erlbaum Associates. doi: 10.4324/9781315799582.

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature neuroscience*, *23*(7), 788–799. doi: 10.1038/s41593-020-0660-4.

Kirk, R. E. (2007). Effect magnitude: A different focus. *Journal of statistical planning and inference*, *137*(5), 1634–1646. doi: 10.1016/j.jspi.2006.09.011.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676. doi: 10.1002/wcs.72.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in bayesian estimation. *Advances in methods and practices in psychological science*, *1*(2), 270–280. doi: 10.1177/2515245918771304.

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic bulletin & review*, *25*, 178–206. doi: 10.3758/s13423-016-1221-4.

Kutner, M., Nachtsheim, C. J., Neter, J., & Wasserman, L. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8*(4), 355–362. doi: 10.1177/1948550617697177.

Lakens, D. (2022). *Improving your statistical inferences*. Retrieved July 13, 2025, from https://doi.org/10.5281/ZENODO.6409077

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi: 10.1177/2515245918770963.

Lassance, R. F. L., Izbicki, R., & Stern, R. B. (2025). Adding imprecision to hypotheses: A bayesian framework for testing practical significance in nonparametric settings. *International Journal of Approximate Reasoning*, *178*, 109332. doi: 10.1016/j.ijar.2024.109332.

Lassance, R. F. L., Stern, J. M., & Stern, R. B. (2024). *Nonparametric fbst for validating linear models*. Retrieved July 13, 2025, from https://doi.org/10.48550/ARXIV.2406.15608

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, *53*(2), 119–122. doi: 10.1080/00031305.1999.10474443.

Lecoutre, B., & Poitevineau, J. (2022). The significance test controversy revisited. In B. Lecoutre (Ed.), *The significance test controversy revisited: The fiducial bayesian alternative* (pp. 41–54). Springer. doi: 10.1007/978-3-662-65705-8.

Lehmann, E. L. (1957). A theory of some multiple decision problems, i. *The Annals of Mathematical Statistics*, *28*(1), 1–25. doi: 10.1214/aoms/1177707034.

Leung, J. T., Barnes, S. L., Lo, S. T., & Leung, D. Y. (2020). Noninferiority trials in cardiology: What clinicians need to know. *Heart*, *106*(2), 99–104. doi: 10.1136/heartjnl-2019-315772.

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). Bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. doi: 10.21105/joss.01541.

Mara, C. A., & Cribbie, R. A. (2012). Paired-samples tests of equivalence. *Communications in Statistics-Simulation and Computation*, *41*(10), 1928–1943. doi: 10.1080/03610918.2011.626545.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. doi: 10.1017/9781107286184.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axl003.

Meeker, W. Q., & Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, *49*(1), 48–53. doi: 10.1080/00031305.1995.10476112.

Mehra, M. R., Desai, S. S., Ruschitzka, F., & Patel, A. N. (2020). Retracted: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: A multinational registry analysis. *The Lancet*. doi: 10.1016/s0140-6736(20)31180-6.

Mehrabi, N., Morstatter, F., et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Meyners, M. (2012). Equivalence tests–a review. *Food quality and preference*, *26*(2), 231–245. doi: 10.1016/j.foodqual.2012.05.003.

Migon, H. S., Gamerman, D., & Louzada, F. (2014). *Statistical inference: An integrated approach, second edition*. CRC Press. doi: 10.1201/b17229.

Mitchell, M., Shankar, S., et al. (2021). Ai fairness: A review of bias mitigation techniques. *Advances in Neural Information Processing Systems*, *99*, 1–99.

Neyman, J. (1957). The use of the concept of power in agricultural experimentation. *Journal of the Indian Society of Agricultural Statistics, IX*, *99*, 9–17.

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in statistics – Theory and methods*, *5*(8), 737–751. doi: 10.1080/03610927608827392.

Park, B., Balakrishnan, S., & Wasserman, L. (2023). *Robust universal inference for misspecified models*. Retrieved July 13, 2025, from https://doi.org/10.48550/ARXIV.2307.04034

Patriota, A. G. (2013). A classical measure of evidence for general null hypotheses. *Fuzzy Sets and Systems*, *233*, 74–88. doi: 10.1016/j.fss.2013.03.007.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25–45. doi: 10.1093/biomet/13.1.25.

Pereira, C. A. d. B., & Stern, J. M. (1999). Evidence and credibility: Full bayesian significance test for precise hypotheses. *Entropy*, *1*(4), 99–110. doi: 10.3390/e1040099.

Phadia, E. G. (2016). *Prior processes and their applications*. Springer International Publishing. doi: 10.1007/978-3-319-32789-1.

Pike, H. (2019). Statistical significance should be abandoned, say scientists. *BMJ*, *364*. doi: 10.1136/bmj.l1374.

Rainey, C. (2014). Arguing for a negligible effect. *American Journal of Political Science*, *58*(4), 1083–1091. doi: 10.1111/ajps.12102.

Rice, K. M., & Krakauer, C. A. (2023). Three-decision methods: A sensible formulation of significance tests—and much else. *Annual Review of Statistics and Its Application*, *10*. doi: 10.1146/annurev-statistics-033021-111159.

Robins, J., & van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, *34*(1). doi: 10.1214/009053605000000877.

Robinson, A. P., Duursma, R. A., & Marshall, J. D. (2005). A regression-based equivalence test for model validation: Shifting the burden of proof. *Tree physiology*, *25*(7), 903–913. doi: 10.1093/treephys/25.7.903.

Roth, M., Tym, E., Mountjoy, C. Q., Huppert, F. A., Hendrie, H., Verma, S., & Goddard, R. (1986). CAMDEX: A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *The British journal of psychiatry*, *149*(6), 698–709. doi: 10.1192/bjp.149.6.698.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*, 225–237. doi: 10.3758/PBR.16.2.225.

Scheffé, H. (1999). *The analysis of variance* (Vol. 72). John Wiley & Sons.

Schervish, M. J. (1995). *Theory of statistics*. Springer New York. doi: 10.1007/978-1-4612-4250-5.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15*, 657–680. doi: 10.1007/bf01068419.

Smith, R. J. (2020). P>. 05: The incorrect interpretation of "not significant" results is a significant problem. *American journal of physical anthropology*, *172*(4), 521–527. doi: 10.1002/ajpa.24092.

Stead, L. F., Perera, R., Bullen, C., Mant, D., Hartmann-Boyce, J., Cahill, K., & Lancaster, T. (2012). Nicotine replacement therapy for smoking cessation. *Cochrane Database of Systematic Reviews*. doi: 10 . 1002 / 14651858.cd000146.pub4.

Stern, J. M., Izbicki, R., Esteves, L. G., & Stern, R. B. (2017). Logically-consistent hypothesis testing and the hexagon of oppositions. *Logic Journal of the IGPL*, *25*(5), 741–757. doi: 10.1093/jigpal/jzx024.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of graduate medical education*, *4*(3), 279–282. doi: 10.4300/JGME-D-12-00156.1.

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., Bilgiç, Y. K., Bono, R., Bradley, M. T., Briggs, W. M., Cepeda-Freyre, H. A., et al. (2018). Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology*, *9*. doi: 10.3389/fpsyg.2018.00699.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2. doi: 10 . 1080 / 01973533.2015.1012991.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological methods*, *6*(4), 371. doi: 10.1037/1082-989x.6.4.371.

Tukey, J. W. (1953). *The problem of multiple comparisons*. Chapman; Hall.

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected anova designs. *Psychological bulletin*, *72*(3), 204. doi: 10.1037/h0027878.

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of general internal medicine*, *26*, 192–196. doi: 10.1007/s11606-010-1513-8.

Wang, X.-g., & Shen, H. C. (1999). Multiple hypothesis testing method for decision making. *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, *3*, 2090–2095. doi: 10 . 1109 / ROBOT.1999.770415.

Wang, Y., Devji, T., Carrasco-Labra, A., King, M. T., Terluin, B., Terwee, C. B., Walsh, M., Furukawa, T. A., & Guyatt, G. H. (2023). A step-by-step approach for selecting an optimal minimal important difference. *BMJ*, *381*. doi: 10.1136/bmj-2022-073822.

Wasserman, L. (2013). *All of statistics: A concise course in statistical inference*. Springer Science & Business Media. doi: 10.1007/978-0-387-21736-9.

Wasserstein, R. L., & Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi: 10.1080/00031305.2016.1154108.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond $p < 0.05$. *The American Statistician*, *73*(sup1), 1–19. doi: 10 . 1080 / 00031305 . 2019 . 1583913.

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication methods and measures*, *6*(3), 190–213. doi: 10.1080/19312458.2012.703834.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC press. doi: 10 . 1201 / EBK1439808184.

Wellek, S., & Blettner, M. (2012). Establishing equivalence or non-inferiority in clinical trials: Part 20 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *109*(41), 674. doi: 10 . 3238 / arztebl.2012.0674.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, 741–744. doi: 10.2307/2529259.

Yang, C., Bartolucci, A. A., & Cui, X. (2015). Multigroup equivalence analysis for high-dimensional expression data. *Cancer Informatics*, *14*, CIN–S17304. doi: 10.4137/cin.s17304.

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

Zhao, Y., Caffo, B. S., & Ewen, J. B. (2022). B-value and empirical equivalence bound: A new procedure of hypothesis testing. *Statistics in Medicine*, *41*(6), 964–980. doi: 10.1002/sim.9298.

Zöller, M.-A., & Huber, M. F. (2022). A survey on automated machine learning. *ACM Computing Surveys*, *54*(6), 1–34.

Appendix A: Additional Figures and Experiments

Figure A1 ■ Confidence intervals for the average difference between groups as a function of the sample size for three different resamplings. All conclusions as the sample size increases are the same as the original sorting from Figure 2, with slight differences in AD vs MCI for resampling 3.

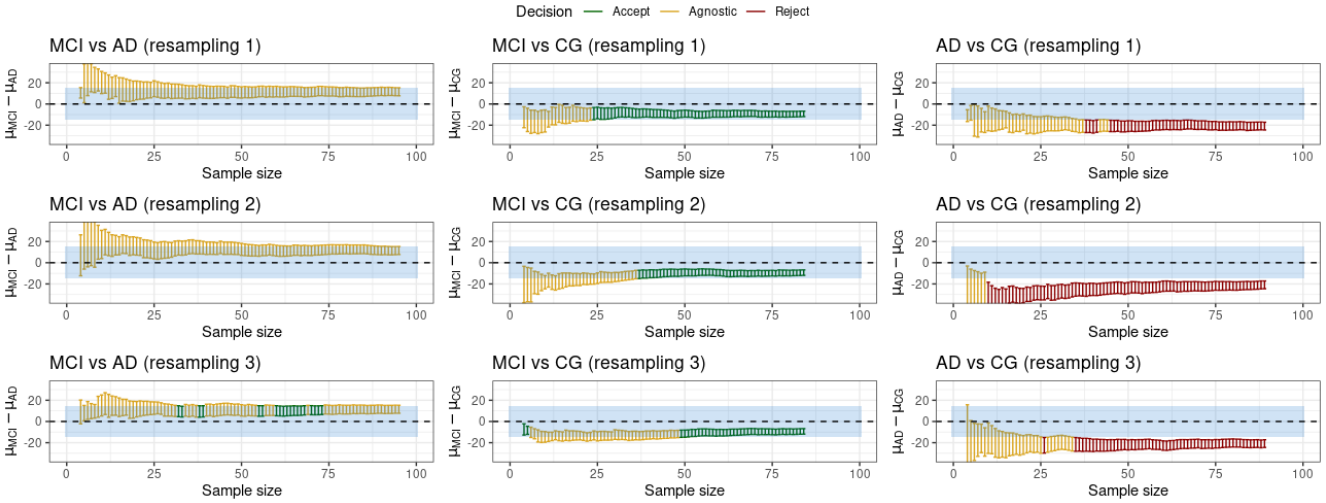


Figure A2 ■ Bayesian pairwise group comparisons of `REACT` for the CAMCOG application. The prior for each pair (μ_i, σ_i^2) is a Normal-inverse gamma with parameters $(80, 1, 3, 3)$. The credible regions are the HPD region of $(\mu_{AD}, \mu_{CG}, \mu_{MCI})$ projected on each pair (μ_i, μ_j) (dashed border for prior, solid for posterior). The blue regions represent the null hypotheses $|\mu_i - \mu_j| \leq \Delta$ for the various groups. The conclusions obtained from the posterior are the same as those in Figure 4.

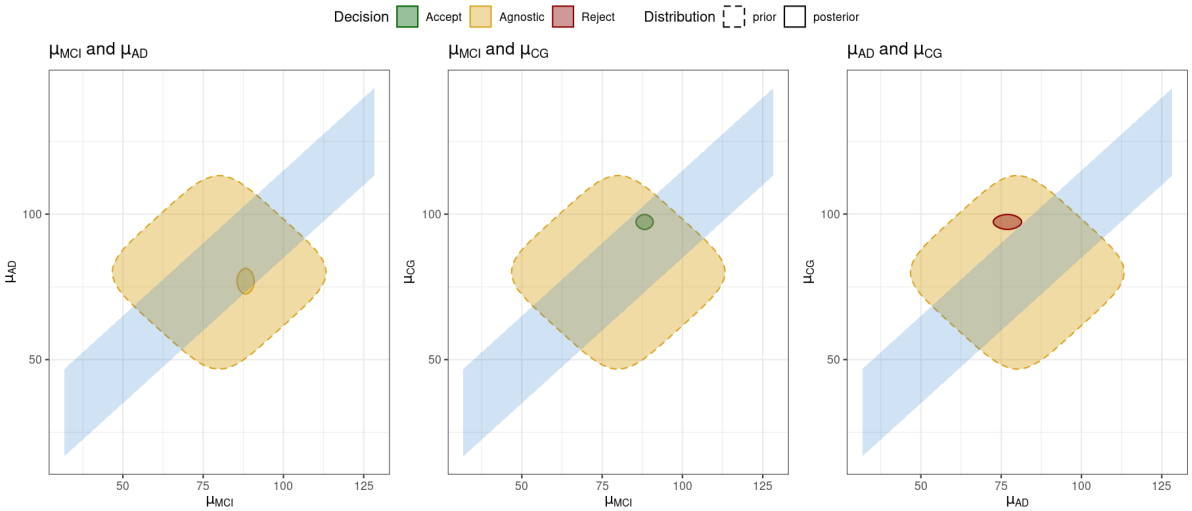
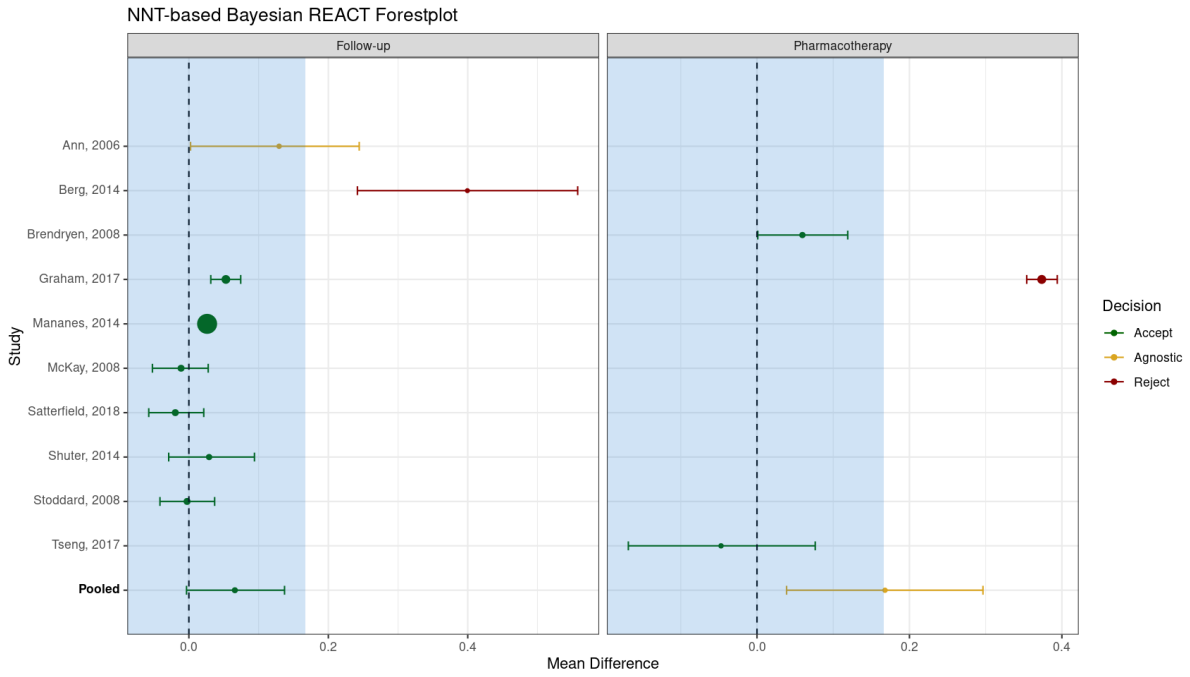


Figure A3 ■ Bayesian REACT using the Jeffreys prior $Beta(1/2, 1/2)$ for the proportion of successes for the meta-analyses described in Section "Meta-analysis". The intervals represent the 95% HPD region, while the point estimates are the posterior means and their respective sizes are the inverse of the posterior variances. The pooled model assumes a hierarchical structure, where the i-th study with the j-th intervention (control or treatment) has probability of success $\theta_{i,j}|\theta_j \sim$ Logit-Normal(logit$(\theta_j), 0.1)$, while $\theta_j \sim Beta(1/2, 1/2)$. The conclusions are similar to those obtained in Figure 5.



Appendix B: Additional Theorems and Definitions

Theorem 1. [Computation of REACT in problems with nuisance parameters] Assume that the parameter space can be decomposed as $\Phi \times \Psi$, where $\phi \in \Phi$ denote parameters of interest and $\psi \in \Psi$ are nuisance parameters. Let $C_\phi(\mathcal{D}) \subseteq \Phi$ be a region estimator of ϕ only. Consider the following procedure to test $H_0 : \phi \in \Phi_0$, where $\Phi_0 \subset \Phi$:

$$\text{Decision} = \begin{cases} \textbf{Accept } H_0 \text{ if } C_\phi(\mathcal{D}) \subset \Phi_0 \\ \textbf{Reject } H_0 \text{ if } C_\phi(\mathcal{D}) \subset \Phi_0^c \\ \textbf{Remain Agnostic} \text{ otherwise.} \end{cases}$$

This is a proper REACT procedure.

Definition 1. A hypothesis test for a hypothesis $H \subset \Theta$ is a function $\mathcal{R}_H : \mathcal{X} \longrightarrow \{0, 1/2, 1\}$, where $\mathcal{X}$ is the sample space, such that 0 represents the acceptance of H, 1 its rejection, and $1/2$ is the agnostic decision.

Definition 2. Let $\sigma(\Theta)$ be a subset of Θ with several hypotheses to be tested. For each $H \in \sigma(\Theta)$, let $\mathcal{R}_H$ denote a test for H. The collection of hypothesis tests $(\mathcal{R}_H)_{H\in\sigma(\Theta)}$ is defined to be logically coherent if

1. (Propriety) $\mathcal{R}_\Theta \equiv 0$,
2. (Monotonicity) If $H \subseteq H_*$, $\mathcal{R}_{H_*} \leq \mathcal{R}_H$,
3. (Invertibility) For every H, $\mathcal{R}_H \equiv 1 - \mathcal{R}_{H^c}$,
4. (Intersection consonance) If $\mathcal{H}$ is a collection of hypothesis such that $\mathcal{R}_H(D) = 0$, for every $H \in \mathcal{H}$, then $\mathcal{R}_{\cap_{H\in\mathcal{H}}H}(D) = 0$.

Definition 3. A region test $\mathcal{R}$, has level (α, β) if $\sup_{H_0} \sup_{\theta\in H_0} \mathbb{P}_\theta(\mathcal{R}_{H_0} = 1) \leq \alpha$ and $\sup_{H_0} \sup_{\theta\notin H_0} \mathbb{P}_\theta(\mathcal{R}_{H_0} = 0) \leq \beta$, that is the test controls the type I error by α and the type II error by β.

Definition 4. Let $\mathcal{R}$ be a region test. The decisiveness function of $\mathcal{R}$, $\beta_{\mathcal{R}}$, determines how frequently the test does not remain undecided, for each value of θ and each hypothesis H_0, that is,

$$\beta_{\mathcal{R}}(\theta, H_0) = \mathbb{P}_\theta\left(\mathcal{R}_{H_0} \neq \frac{1}{2}\right).$$

Definition 5. A (α, α)-level region test, $\mathcal{R}$, is unbiased if $\inf_{\theta, H_0} \beta_{\mathcal{R}}(\theta, H_0) \geq \alpha$.

Definition 6. An interval region estimator, $\mathcal{C}$, is a region estimator such that there exist real-valued functions, $a(\mathcal{D})$ and $b(\mathcal{D})$ and for every $\mathcal{D}$ $\mathcal{C}(\mathcal{D}) = (a(\mathcal{D}), b(\mathcal{D}))$.

Definition 7. A region test, $\mathcal{R}_{\mathcal{C}}$ is an interval region test if $\mathcal{C}$ is an interval region estimator.

Definition 8. Let $\mathcal{R}$ be an unbiased interval region test and $\mathcal{H}$ be a collection of hypotheses. $\mathcal{R}$ is uniformly most decisive on $\mathcal{H}$ among unbiased (α, α)-level region tests based on intervals if, for every unbiased, (α, α)-level, interval region test, $\mathcal{R}^*$,

$$\beta_{\mathcal{R}}(\theta, H_0) \geq \beta_{\mathcal{R}^*}(\theta, H_0), \text{ for every } \theta \text{ and } H_0 \in \mathcal{H}.$$

Lemma 1. If $\mathcal{R}^*$ is an unbiased, (α, α)-level test based on the region $\mathcal{C}^*$, then $\phi^* = \mathbb{I}(\theta_0 \notin \mathcal{C}^*)$ is an unbiased α-level binary test.

Theorem 2. Let $\mathcal{H} = \{\{\theta_0\} : \theta_0 \in \Theta\} \cup \{(-\infty, \theta_0) : \theta_0 \in \Theta\} \cup \{(\theta_0, \infty) : \theta_0 \in \Theta\}$, that is, the collection of all unilateral and bilateral hypotheses. For each θ, assume ϕ_{θ_0} is a UMPU α-level test for testing $H_0 : \theta = \theta_0$. If $\mathcal{C}(\mathcal{D}) = \{\theta_0 : \phi_{\theta_0}(\mathcal{D}) = 0\}$ is an interval region estimator, then $\mathcal{R}_{\mathcal{C}}$ is uniformly most decisive on $\mathcal{H}$ among unbiased (α, α)-level interval region tests.

Example 1. Let Φ be the cumulative density function of a standard normal, $X_1, \ldots, X_n$ be i.i.d. and $X_1 \sim N(\theta, \sigma_0^2)$, where θ is unknown and σ_0^2 is known. For each $H_0 : \theta = \theta_0$, $\phi_{\theta_0} = \mathbb{I}\left(\frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma_0} \geq -\Phi(0.5\alpha)\right)$ is the UMPU α-level test for H_0. Note that

$$\begin{aligned}\mathcal{C} &:= \{\theta_0 : \phi_{\theta_0} = 0\} \\ &= \left(\bar{X} + \frac{\Phi(0.5\alpha)\sigma_0}{\sqrt{n}}, \bar{X} - \frac{\Phi(0.5\alpha)\sigma_0}{\sqrt{n}}\right)\end{aligned}$$

Hence, it follows from Theorem 2 that $\mathcal{R}_{\mathcal{C}}$ is uniformly most decisive on $\mathcal{H}$ among unbiased (α, α)-level interval region tests.

Appendix C: Proofs

Lemma 2. Let $C(\mathcal{D}) := \{\theta \in \Theta : \text{p-val}_{\mathcal{D}}(\theta) > \alpha\}$. For every $H \subseteq \Theta$, $C(\mathcal{D}) \cap H = \emptyset$ if and only if $\max_{\theta \in H} \text{p-val}_{\mathcal{D}}(\theta) \leq \alpha$.

Proof. If $\max_{\theta \in H} \text{p-val}_{\mathcal{D}}(\theta) \leq \alpha$, then for every $\theta \in H$, $\text{p-val}_{\mathcal{D}}(\theta) \leq \alpha$ and, by construction, $\theta \notin C(\mathcal{D})$. That is, $C(\mathcal{D}) \cap H = \emptyset$. If $\max_{\theta \in H} \text{p-val}_{\mathcal{D}}(\theta) > \alpha$, then there exists $\theta \in H$, such that $\text{p-val}_{\mathcal{D}}(\theta) > \alpha$ and, by construction, $\theta \in C(\mathcal{D})$. Therefore, $C(\mathcal{D}) \cap H \neq \emptyset$. □

Proof of Property 1. First, observe that

$$\mathbb{P}_\theta(\theta \notin C(\mathcal{D})) = \mathbb{P}_\theta(\text{p-val}_{\mathcal{D}}(\theta) \leq \alpha) \leq \alpha.$$

Hence, $C(\mathcal{D})$ is a $1 - \alpha$ confidence interval. The rest of the proof follows directly from Lemma 2. □

Definition 9. Let $C(\mathcal{D})$ be a region estimator for θ, that is, a function $C : \mathcal{X} \to \sigma(\Theta)$, where $\mathcal{X}$ is the sample space. The region test for H_0 based on C, $\mathcal{R}_{C,H_0}$, is

$$\mathcal{R}_{C,H_0} = \begin{cases} 0 & \text{, if } C \subseteq H_0 \\ 1 & \text{, if } C \subseteq H_0^c \\ \frac{1}{2} & \text{, otherwise.} \end{cases}$$

Proof of Property 2. The proof is equivalent to showing that $\mathcal{R}_{C,H_0} = \mathcal{R}_{C_\Phi,\Phi_0}$. Also, when looking at the parameter space as a whole, we note that H_0 possesses the following property:

$$\left\{ (\phi,\psi) \in H_0 \implies \bigcup_{\psi\in\Psi} (\phi,\psi) \subset H_0 \right\}, \quad \forall \phi \in \Phi. \tag{3}$$

Now, let us consider the possible settings of $\mathcal{R}_{C,H_0}$. First, when the original test accepts H_0,

$$\mathcal{R}_{C,H_0} = 0 \overset{\text{def. 9}}{\iff} C(\mathcal{D}) \subseteq H_0 \overset{(3)}{\iff} C_\Phi(\mathcal{D}) \times \Psi \subseteq H_0 = \Phi_0 \times \Psi \overset{\text{def. 9}}{\iff} \mathcal{R}_{C_\Phi,\Phi_0} = 0. \tag{4}$$

As for when the test rejects H_0,

$$\mathcal{R}_{C,H_0} = 1 \overset{\text{prop. 9}}{\iff} \mathcal{R}_{C,H_0^c} = 0 \overset{(4)}{\iff} \mathcal{R}_{C_\Phi,\Phi_0^c} = 0 \overset{\text{prop. 9}}{\iff} \mathcal{R}_{C_\Phi,\Phi_0} = 1. \tag{5}$$

Lastly, since (4) and (5) are necessary and sufficient conditions for respectively accepting and rejecting H_0, it follows that $\mathcal{R}_{C,H_0} = 1/2 \iff \mathcal{R}_{C_\Phi,\Phi_0} = 1/2$, thus concluding the proof.

□

Proof of Property 3. (Adapted from Coscrato et al. (2020)). Since $C(\mathcal{D})$ has confidence $1-\alpha$, $\mathbb{P}_\theta(\theta \notin C(\mathcal{D})) \leq \alpha$, for every $\theta \in \Theta$. Therefore,

$$\begin{aligned}
\sup_{\theta_0\in H_0} \mathbb{P}_{\theta_0}(\mathcal{R}_{C,H_0} = 1) &= \sup_{\theta_0\in H_0} \mathbb{P}_{\theta_0}(C(\mathcal{D}) \subseteq H_0^c) \leq \sup_{\theta_0\in H_0} \mathbb{P}_{\theta_0}(\theta_0 \notin C(\mathcal{D})) \leq \alpha \\
\sup_{\theta_1\in H_0^c} \mathbb{P}_{\theta_1}(\mathcal{R}_{C,H_0} = 0) &= \sup_{\theta_1\in H_0^c} \mathbb{P}_{\theta_1}(C(\mathcal{D}) \subseteq H_0) \leq \sup_{\theta_1\in H_0^c} \mathbb{P}_{\theta_1}(\theta_1 \notin C(\mathcal{D}) \leq \alpha
\end{aligned}$$

□

Definition 10. The family-wise type I error, $FWER_I$, is the probability that some truly null hypothesis is incorrectly rejected. Similarly, the family-wise type II error, $FWER_{II}$, is the probability that some truly non-null hypothesis is incorrectly accepted. That is,

$$\begin{aligned}
FWER_I(\theta) &:= \mathbb{P}_\theta(\cup_{H:\theta\in H} \mathcal{R}_{C,H} = 1) \\
FWER_{II}(\theta) &:= \mathbb{P}_\theta(\cup_{H:\theta\notin H} \mathcal{R}_{C,H} = 0).
\end{aligned}$$

Proof of Property 4. For every $H \in \sigma(\Theta)$,

$$\begin{aligned}
FWER_I(\theta) &= \mathbb{P}_\theta(\cup_{H:\theta\in H} \mathcal{R}_{C,H} = 1) \\
&= \mathbb{P}_\theta(\cup_{H:\theta\in H} C \subseteq H^c) \\
&= \mathbb{P}_\theta(\cup_{H:\theta\in H} C \cap H = \emptyset) \\
&= \mathbb{P}_\theta(\theta \notin C) \leq \alpha.
\end{aligned}$$

$$\begin{aligned}
FWER_{II}(\theta) &= \mathbb{P}_\theta(\cup_{H:\theta\notin H} \mathcal{R}_{C,H} = 0) \\
&= \mathbb{P}_\theta(\cup_{H:\theta\notin H} C \subseteq H) \\
&= \mathbb{P}_\theta(\theta \notin C) \leq \alpha.
\end{aligned}$$

□

Definition 11. A confidence set, $C(\mathcal{D})$, converges to the true θ if, for every $\theta_0 \in \Theta$ and ϵ-ball around θ_0, $B(\theta_0,\epsilon)$,

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}(C(\mathcal{D}) \subseteq B(\theta_0,\epsilon)) = 1.$$

Definition 12. For $A \subseteq \Theta$, let $\mathring{A}$ denote the interior of A. A hypothesis test for H_0, $\mathcal{R}_{H_0}$, is consistent if, for every $\theta \in \mathring{H}_0$, $\lim_{n\to\infty} \mathbb{P}_\theta(\mathcal{R}_{H_0} = 0) = 1$ and, for every $\theta \in \mathring{H}_0^c$, $\lim_{n\to\infty} \mathbb{P}_\theta(\mathcal{R}_{H_0} = 1) = 1$.

Proof of Property 5. Let $\theta \in \mathring{H}_0$. There exists $\epsilon > 0$ such that $B(\theta,\epsilon) \subseteq H_0$. Hence,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}_\theta(\mathcal{R}_{H_0,C} = 0) &= \lim_{n\to\infty} \mathbb{P}_\theta(C(\mathcal{D}) \subseteq H_0) & \\
&\leq \lim_{n\to\infty} \mathbb{P}_\theta(C(\mathcal{D}) \subseteq B(\theta,\epsilon)) & B(\theta,\epsilon) \subseteq H_0 \\
&= 1 & \text{Definition 12}
\end{aligned}$$

Similarly, let $\theta \in \mathring{H}_0^c$. There exists $\epsilon > 0$ such that $B(\theta, \epsilon) \subseteq H_0^c$. Hence,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}_\theta(\mathcal{R}_{H_0,C} = 1) &= \lim_{n\to\infty} \mathbb{P}_\theta(C(\mathcal{D}) \subseteq H_0^c) \\
&\leq \lim_{n\to\infty} \mathbb{P}_\theta(C(\mathcal{D}) \subseteq B(\theta, \epsilon)) && B(\theta, \epsilon) \subseteq H_0^c \\
&= 1 && \text{Definition 12}
\end{aligned}$$

□

Proof of Property 8. (Adapted from Esteves et al. (2016)). Let $(\mathcal{R}_H)_{H\in\sigma(\Theta)}$ be a collection of tests based on confidence set C.

1. Since $C(\mathcal{D}) \subseteq \Theta$, $\mathcal{R}_\Theta \equiv 0$,
2. Let $H \subseteq H_*$. If $\mathcal{R}_H(D) = 0$, then $C(D) \subset H$. Hence, $C(D) \subset H_*$, that is, $\mathcal{R}_{H_*}(D) = 0$. Also, if $\mathcal{R}_{H_*}(D) = 1$, then $C(D) \subseteq H_*^c$. Hence, $C(D) \subseteq H^c$ and $\mathcal{R}_H(D) = 1$. Conclude that $\mathcal{R}_{H_*} \leq \mathcal{R}_H$.
3. It is sufficient to prove that, for every $H \in \sigma(\Theta)$, $\mathcal{R}_H(D) = 0$ if and only if $\mathcal{R}_{H^c}(D) = 1$. The proof follows from the fact that $\mathcal{R}_H(D) = 0$ when $C(D) \subseteq H$ and $\mathcal{R}_{H^c}(D) = 1$ when $C(D) \subseteq (H^c)^c$, that is, $C(D) \subseteq H$.
4. If $\mathcal{R}_H(D) = 0$, for every $H \in \mathcal{H}$, then $C(D) \subseteq H$, for every $H \in \mathcal{H}$. Hence, $C(D) \subseteq \cap_{H\in\mathcal{H}} H$. That is, $\mathcal{R}_{\cap_{H\in\mathcal{H}} H}(D) = 0$.

□

Proof of Property 9. Property 9 is a consequence of invertibility in Definition 2 Hence, this property is a corollary of Property 8. □

Proof of Theorem 1. Let $C(\mathcal{D}) := C_\phi(\mathcal{D}) \times \Psi$. Notice that

$$C_\phi(\mathcal{D}) \subset \Phi_0 \iff C(\mathcal{D}) \subset \Phi_0 \times \Psi,$$

and therefore the procedure stated on the theorem is equivalent to the following `REACT` procedure:

$$\text{Decision} = \begin{cases} \textbf{Accept } H_0 \text{ if } C(\mathcal{D}) \subset \Phi_0 \times \Psi \\ \textbf{Reject } H_0 \text{ if } C(\mathcal{D}) \subset \Phi_0^c \times \Psi \\ \textbf{Remain Agnostic} \text{ otherwise.} \end{cases}$$

The conclusion follows. □

Proof of Lemma 1. Since $\mathcal{R}^*$ is a region-based test, $\{\mathcal{R}^* \text{ rejects } \{\theta_0\}\}$ is the same as $\{\phi^* = 1\}$. Hence, since $\mathcal{R}^*$ is a (α, α)-level test

$$\mathbb{P}_{\theta_0}(\phi^* = 1) = \mathbb{P}_{\theta_0}(\mathcal{R}^* \text{ rejects } \{\theta_0\}) \leq \alpha.$$

That is, ϕ^* is a α-level binary test. Also, since $\mathcal{R}^*$ is unbiased,

$$\mathbb{P}_\theta(\phi^* = 0) = \mathbb{P}_{\theta_0}(\mathcal{R}^* \text{ remains undecided about } \{\theta_0\}) \leq 1 - \alpha$$

Hence, ϕ^* is unbiased. □

Proof of Theorem 2. Let $\mathcal{R}^*$ be an unbiased, (α, α)-level based on the interval region $\mathcal{C}^*$. For every $H_0 \in H$, define $\phi^* = \mathbb{I}(\theta_0 \notin \mathcal{C}^*)$.

$$\begin{aligned}
\beta_{\mathcal{R}}(\theta, H_0) &= \mathbb{P}_\theta(\theta_0 \notin \mathcal{C}) \\
&= \mathbb{P}_\theta(\phi_{\theta_0} = 1) \\
&\geq \mathbb{P}_\theta(\phi^* = 1) && \text{Lemma 1 and } \phi_{\theta_0} \text{ is UMPU} \\
&= \mathbb{P}_\theta(\theta_0 \notin \mathcal{C}^*) \\
&= \beta_{\mathcal{R}^*}(\theta, H_0).
\end{aligned}$$

□

Appendix D: Additional Results

Definition 13. The Bayesian family-wise false conclusion error of $(\mathcal{R}_H)_{H\in\sigma(\Theta)}$, γ, is:

$$\gamma = \mathbb{P}(\exists H \in \sigma(\Theta) : (\theta \in H \text{ and } \mathcal{R}_H = 1) \text{ or } (\theta \notin H \text{ and } \mathcal{R}_H = 0))$$

Theorem 3. If R is a credibility region for θ with credibility $1-\alpha$ and $(\mathcal{R}_H)_{H\in\sigma(\Theta)}$ is based on R, then $\gamma \leq \alpha$.

Proof of Theorem 3.

$$\begin{aligned}\gamma &= \mathbb{P}(\exists H \in \sigma(\Theta) : (\theta \in H \text{ and } \mathcal{R}_H = 1) \text{ or } (\theta \notin H \text{ and } \mathcal{R}_H = 0))\\ &= \mathbb{P}(\exists H \in \sigma(\Theta) : (\theta \in H \text{ and } R \subseteq H^c) \text{ or } (\theta \notin H \text{ and } R \subseteq H))\\ &\leq \mathbb{P}(\theta \notin R) = \alpha\end{aligned}$$

□

Theorem 4. [Computation of `fREACT` using p-values in problems with nuisance parameters] Assume that the parameter space can be decomposed as $\Phi \times \Psi$, where $\phi \in \Phi$ denote parameters of interest and $\psi \in \Psi$ are nuisance parameters. Let p-val$_{\mathcal{D}}(\phi_0)$ be a p-value for the hypothesis $H_0 : \phi = \phi_0$. Consider the following procedure to test $H_0 : \phi \in \Phi_0$, where $\Phi_0 \subset \Phi$:

$$\text{Decision} = \begin{cases}\textbf{Accept } H_0 \text{ if } \sup_{\phi\in\Phi_0^c} \text{p-val}_{\mathcal{D}}(\phi) \leq \alpha\\ \textbf{Reject } H_0 \text{ if } \sup_{\phi\in\Phi_0} \text{p-val}_{\mathcal{D}}(\phi) \leq \alpha\\ \textbf{Remain Agnostic} \text{ otherwise}\end{cases}$$

This procedure is equivalent to the following `fREACT` procedure:

$$\text{Decision} = \begin{cases}\textbf{Accept } H_0 \text{ if } C(\mathcal{D}) \subset \Phi_0 \times \Psi\\ \textbf{Reject } H_0 \text{ if } C(\mathcal{D}) \subset \Phi_0^c \times \Psi\\ \textbf{Remain Agnostic} \text{ otherwise.}\end{cases}$$

where $C(\mathcal{D}) := C_\phi(\mathcal{D}) \times \Psi$ and $C_\phi(\mathcal{D}) := \{\phi \in \Phi : \text{p-val}_{\mathcal{D}}(\phi) > \alpha\}$. Moreover, it can be more easily written as

$$\text{Decision} = \begin{cases}\textbf{Accept } H_0 \text{ if } C_\phi(\mathcal{D}) \subset \Phi_0\\ \textbf{Reject } H_0 \text{ if } C_\phi(\mathcal{D}) \subset \Phi_0^c\\ \textbf{Remain Agnostic} \text{ otherwise.}\end{cases}$$

Also, $C_\phi(\mathcal{D})$ is a (1-α)-level confidence set for ϕ, and $C(\mathcal{D})$ is a (1-α)-level confidence set for (ϕ, ψ).

Proof. Notice that

$$\begin{aligned}\sup_{\phi\in\Phi_0^c} \text{p-val}_{\mathcal{D}}(\phi) \leq \alpha &\iff \text{For every } \phi \in \Phi_0^c, \text{p-val}_{\mathcal{D}}(\phi) \leq \alpha\\ &\iff \text{For every } \phi \in \Phi_0^c \text{ and } \psi \in \Psi, (\phi,\psi) \notin C(\mathcal{D})\\ &\iff C(\mathcal{D}) \subset \Phi_0 \times \Psi\end{aligned}$$

Thus, the procedure accepts H_0 if, and only if, $C(\mathcal{D}) \subset \Phi_0 \times \Psi$. Similarly, the procedure rejects H_0 if, and only if, $C(\mathcal{D}) \subset \Phi_0^c \times \Psi$. It follows that this procedure is a `REACT`-type procedure. Now, if the p-values are valid, for every $(\phi, \psi) \in \Phi \times \Psi$,

$$\mathbb{P}_{(\phi,\psi)}\left((\phi,\psi) \in C(\mathcal{D})\right) = \mathbb{P}_{(\phi,\psi)}\left(\text{p-val}_{\mathcal{D}}(\phi) > \alpha\right) = 1-\alpha,$$

which concludes the proof. □

Definition 14 (Pereira and Stern (1999))**.** The e-value for a hypothesis $H_0 : \theta = \theta_0$ is the posterior probability

$$\text{e-val}_{\mathcal{D}}(\theta_0) = 1 - \mathbb{P}\left(\theta \in T_{\mathcal{D}} | \mathcal{D}\right),$$

where

$$T_{\mathcal{D}} = \{\theta : f(\theta|\mathcal{D}) \geq f(\theta_0|\mathcal{D})\}$$

Theorem 5. [Computation of `bREACT` using e-values] Let e-val$_{\mathcal{D}}(\theta_0)$ be an e-value (Pereira & Stern, 1999) for the hypothesis $H_0 : \theta \in \Theta_0$. Then the following is a `bREACT` procedure:

$$\text{Decision} = \begin{cases} \textbf{Accept } H_0 \text{ if } \max_{\theta \in \Theta_0^c} \text{e-val}_{\mathcal{D}}(\theta) \leq \alpha \\ \textbf{Reject } H_0 \text{ if } \max_{\theta \in \Theta_0} \text{e-val}_{\mathcal{D}}(\theta) \leq \alpha \\ \textbf{Remain Agnostic} \text{ otherwise} \end{cases}$$

The (1-α)-level Bayes set that corresponds to this procedure is the $(1-\alpha)$-level Highest Posterior Density (HPD) region for θ:

$$C(\mathcal{D}) := \{\theta \in \Theta : \pi(\theta|\mathcal{D}) > C\},$$

where C is such that

$$\mathbb{P}(\theta \in C(\mathcal{D})|\mathcal{D}) = 1 - \alpha.$$

This procedure is equivalent to the GFBST (Stern et al., 2017) and, if C is an interval, to ROPE (Kruschke, 2018).

Proof. Can be found in Esteves et al. (2016, Example 8). □

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on github.com/Monoxido45/REACT

Citation

Izbicki, R., Cabezas, L. M. C., Colugnatti, F. A. B., Lassance, R. F. L., de Souza, A. A. L., & Stern, R. B. (2025). REACT to NHST: Sensible conclusions for meaningful hypotheses. *The Quantitative Methods for Psychology*, *21*(2), 43–68. doi: 10.20982/tqmp.21.2.p043.

Received: 11/03/2025 ∼ Accepted: 19/05/2025