A Proposed Hybrid Effect Size Plus p-Value Criterion

A Comment on Goodman et al. (*The American Statistician*, 2019)

Peter Pütz*† Arne Kramer-Sunderbrink* Robin Tim Dreher*
Leona Hoffmann* Robin Werner*

Journal of Comments and Replications in Economics, Volume 1, 2022-4, DOI: 10.18718/81781.26

JEL: C10; C12
Keywords: NHST; Intervall null hypothesis; Minimum effect size plus p-value criterion; Thick t-test; Statistical evidence; Type I error rate; False discovery rate

Data Availability: The R code to reproduce the simulation results in this comment can be downloaded at JCRE's data archive (DOI: 10.15456/j1.2022024.101215). In addition, the analyses can be accessed and customized on github.com/drehero/goodman-replication.

Please Cite As: Pütz, Peter, Arne Kramer-Sunderbrink, Robin Tim Dreher, Leona Hoffmann, and Robin Werner (2022). A Proposed Hybrid Effect Size Plus p-Value Criterion. A Comment on Goodman et al. (2019). *Journal of Comments and Replications in Economics*, Vol 1(2022-4). DOI: 10.18718/81781.26

Abstract

In a recent simulation study, Goodman et al. (2019) compare several methods with regard to their type I and type II error rates when considering a thick null hypothesis that includes all values that are practically equivalent to the point null hypothesis. They propose a hybrid decision criterion only declaring a result "significant" if both a small p-value and a sufficiently large effect size are obtained. We successfully verify the results using our own software code in R and discuss an additional decision method that is tailored to maintain a pre-defined false positive rate. We confirm that the hybrid decision criterion has comparably low error rates in checkable settings but point out that the false discovery rate cannot be easily controlled by the researcher. Our analyses are readily accessible and customizable on github.com/drehero/goodman-replication.

*Department of Economics, Bielefeld University. None of the authors has relevant or material financial interests that relate to the research described in this paper.

†Corresponding author: Bielefeld University, Faculty of Business Administration and Economics, Universitätsstr. 25, D-33615 Bielefeld, Germany.

Received January 24, 2022; Revised July 06, 2022; Accepted July 14, 2022; Published October 20, 2022.

1 Introduction

In the context of the replication crisis, the discussion on the usefulness of null hypothesis significance testing and the correct use and interpretation of p-values has been hot in recent years. The *American Statistical Association's Statement on p-Values and Statistical Significance* (Wasserstein and Lazar, 2016) was arguably the most prominent work on pitfalls in using p-values for statistical inference. It was followed by a symposium on statistical inference organized by the *American Statistical Association* in 2017. The symposium led to a special issue of *The American Statistician* titled "Moving to a World Beyond $p < 0.05$" presenting many suggestions on how to separate signal from noise in data.

Some authors suggest that the traditional focus on point hypotheses is part of the problem (e.g., Greenland, 2019; McShane et al., 2019): When testing the point null hypothesis ($H_0^p : \mu = \mu_0$), even tiny, practically irrelevant effects can be "statistically significant" if the number of samples is big enough. Furthermore, problems in the experimental setup can easily lead to a small bias that is wrongly flagged as a statistically significant effect by the test. It was suggested to refrain from testing point hypotheses and instead to test the *thick null hypothesis* that an effect is big enough to be practically significant (e.g., Betensky, 2019; Blume et al., 2019).

Goodman et al. (2019), henceforth GSK, propose a hybrid decision criterion that is called *minimum effect size plus p-value* (MESP): It only leads to a rejection if a p-value lower than 5% is obtained in a standard t-test for the point null hypothesis and the observed effect size is large enough to be considered practically relevant, i.e. lies outside of the interval determined by the thick null hypothesis. In a simulation study they compare different decision criteria and show good overall performance of their method. More specifically, MESP generally has rather low type I error rates and has high type II error rates only in settings that the researcher can usually identify, namely low nominal power settings. However, in comparison with classical point hypothesis testing, MESP lacks a parameter to control its false positive rate with respect to the thick null hypothesis.

In this comment, we successfully verify the GSK results using independently written software code in R. We extend the analyses by a Bayesian-motivated t-test, the thick t-test, that is equipped with a parameter that controls its false positive rate with respect to the thick null hypothesis. We also compute false discovery and false omission rates for different nominal power categories and show that the MESP approach has the undesirable property that the false discovery rate does not necessarily decrease for rising nominal power. Furthermore, we provide an openly available code repository that allows the scientific community to easily verify, amend and extend the analyses conducted in GSK and in this comment.

This comment proceeds as follows. Section 2 explains the simulation setting and methods used by GSK and motivates a further decision criterion. Section 3 presents the results, while Section 4 concludes.

2 Simulation setting and decision methods

In the following, the effect size deemed practically relevant by researchers in a certain scenario is referred to as $MPSD$ (Minimum Practically Significant Distance). Accordingly, the *thick null hypothesis* is $H_0^t : |\mu - \mu_0| \leq MPSD$ or, equivalently, $H_0^t : \mu \in I_0$, where $I_0 = [\mu_0 - MPSD, \mu_0 + MPSD]$ is the *thick null interval*.

Simulation We simulate 100,000 *cases*.[1] Each can be thought of as a fictional study testing whether some parameter is different from $\mu_0 = 100$. Formally, a case is characterized by a tuple $(\mu, \sigma, n, MPSD)$ describing the conditions of the study:

- μ, sampled uniformly from $\{75, \ldots, 125\}$, is the population mean of our data.
- σ, sampled uniformly from $\{4, \ldots, 60\}$, is the standard deviation of the population.
- n, sampled uniformly from $\{5, \ldots, 100\}$, is the sample size.
- $MPSD$, sampled uniformly from $\{2, \ldots, 20\}$, is the minimal effect size deemed practically relevant by researchers in our fictional scenario.

For every case, we save if $|\mu - \mu_0| > MPSD$ (i.e., if H_0^t is false). We sample the data $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma)$ and save the results of the decision methods described below based on that sample.

2.1 Original decision methods

GSK compare five different decision methods:

Conventional The first method is a t-test for the point null hypothesis H_0^p that $\mu = \mu_0$ with $\alpha = 0.05$. Note that this α is usually the desired false positive rate (the type I error): If the point null hypothesis is true it should not be falsely rejected by the test more than 5% of the time. However, since the test is used here to decide whether the thick H_0^t, not H_0^p, should be rejected, the false positive rate will exceed 5% for $MPSD > 0$.

Small-alpha The second method is a t-test as described above with $\alpha = 0.005$ instead of 0.05. Lowering α from 0.05 to 0.005 was prominently recommended to lower the rate of false positive results and irreproducible research findings (Benjamin et al., 2018). However, as discussed above, α is used to control the false positive rate with respect to H_0^p, not H_0^t.

Distance-only The third method rejects the null hypothesis if the observed effect size is at least as big as the minimum practically significant distance, i.e., $|\bar{x} - \mu_0| \geq MPSD$.

[1] In order to obtain robust results over different simulation runs also for subgroup analyses (e.g. Table 2), we increase the number of simulated cases compared to GSK who simulate 10,000 cases.

MESP GSK propose a new decision method called "Decision by Minimum Effect Size Plus p-value" that is a conjunction of the conventional and the distance-only method, i.e., the null hypothesis is rejected only if both methods would reject it, if both the p-value of the t-test is smaller than 0.05 and the observed effect size is practically significant.

Interval-based The final method considered by GSK rejects the null only if the thick null interval I_0 and the 95% confidence interval centered around the observed sample mean $\bar{x}$ do not intersect.[2]

2.2 Thick t-test

In this section we present a sixth decision method that we call the *thick t-test*. The idea is to base the decision to reject the thick null hypothesis H_0^t on the probability of observing an effect that is more extreme with respect to μ_0 than the one actually observed if H_0^t was true:

$$\begin{aligned} p &= P(|\bar{X} - \mu_0| > |\bar{x} - \mu_0| \mid H_0^t) \\ &= \sum_{\dot{\mu}=\mu_0-MPSD}^{\mu_0+MPSD} P(|\bar{X} - \mu_0| > |\bar{x} - \mu_0| \mid \mu = \dot{\mu}) P(\mu = \dot{\mu} \mid H_0^t). \end{aligned} \tag{1}$$

Note that μ is an integer in our simulation. If μ is continuous, an integral has to be used instead of the sum. Our code contains details on how to compute p in either case.

The probability in Equation (1) is known as the *prior predictive p-value* popularized by Box (1980) (see also Bayarri and Berger (2000) for a discussion on this and related concepts for composite null hypotheses). For $MPSD = 0$, i.e., $H_0^t \equiv H_0^p$, p is simply the p-value of a two-sided point hypothesis test. In that sense, p as defined above is a generalization of familiar p-values to thick null hypotheses and the same guarantee regarding the false positive rate from point hypothesis tests holds for the thick t-test: If we decide to reject H_0^t whenever $p < \alpha = 0.05$, the probability to report an effect even though the real effect is practically irrelevant is 5%. This is because, just like the familiar p-value in a two-sided point hypothesis test, p is uniformly distributed on $[0, 1]$ under H_0^t, i.e., $P(p \leq \alpha \mid H_0^t) = \alpha$. To be precise, due to the discrete distribution of μ in the simulation, p is only asymptotically uniformly distributed, and hence $P(p \leq \alpha \mid H_0^t) = \alpha$ only for $n \to \infty$.

However, we need to make an assumption about the distribution of μ under H_0^t and only under this assumption, the guarantee holds. Since we are doing a simulation we know that μ is uniformly

[2]The interval-based method is equivalent to a two-sided decision criterion presented in Betensky (2019) for rejecting a thick null hypothesis. However, Betensky (2019) allows the interval-based method also to be used to confirm thick null hypotheses. In this regard, the method applied by Betensky (2019) can be considered as a composite approach that combines the interval-based method presented in GSK and equivalence testing using two one-sided tests (Schuirmann, 1987). The latter is designed to confirm the hypothesis that an effect is not larger than some practically meaningful equivalence bounds, i.e. the boundaries of the thick null. The Bayesian counterpart to Betensky (2019) is the HDI-ROPE procedure (Kruschke, 2011) that also allows for rejection and acceptance of a thick null hypothesis. Both Betensky (2019) and Kruschke (2011) deem results inconclusive if the estimated interval (confidence or credible interval, respectively) is neither fully inside nor fully outside of the thick null hypothesis.

distributed, i.e., $P(\mu = \dot{\mu} \mid H_0^t) = 1/(2MPSD + 1)$ for $\dot{\mu} \in I_0$. In a real-world experiment on the other hand, the distribution under H_0^t used in the thick t-test is an expression of the researcher's prior knowledge or assumptions about μ. For example, if a researcher has no or decides to ignore prior knowledge, they should opt for a flat prior for μ on the thick null interval. If a researcher believes that small effects are more likely than bigger ones, they could, for example, choose a truncated normal distribution. Aside from or additional to a researcher's beliefs, prior knowledge on a parameter of interest may be gathered from expert elicitation (e.g. Albert et al., 2012) or previous studies in a specific research field (e.g. Schulz et al., 2021).

We believe that this method is useful for a number of reasons: Firstly, only two of the five original methods tested in GSK (MESP and the interval-based method) take both the $MPSD$ and the standard error of the estimated mean $\hat{\mu} = \bar{x}$ into account. Therefore, it benefits the comparison to add a decision method that uses all available information.

Secondly, it presents a framework to better understand most of the other methods: The conventional and the small-alpha method can be thought of as thick t-tests that assume $P(\mu = \mu_0 | \ H_0^t) = 1$. For a given observed effect, this distributional choice produces the smallest p-values among all possible thick t-tests with a symmetric prior. Thus, H_0^t is more often rejected by the test and may even be rejected if the observed effect is inside the thick null interval. Hence, we expect the false positive rate of these methods to be higher than α (5% and 0.5%, respectively). The interval-based method on the other hand is approximately equivalent (except for a term that is usually negligible) to the thick t-test that assumes $P(\mu = \mu_c \mid H_0^t) = 1$, where μ_c is the value in the thick null interval that is closest to the observed mean $\bar{x}$. For a given observed effect, this distributional choice produces the largest p-values and is therefore the most conservative decision with respect to rejecting H_0^t among all possible thick t-tests. Hence, we expect the false positive rate to be below 5% and the false negative rate to be high for this method. Note that the thick t-test with $P(\mu = \mu_c \mid H_0^t) = 1$ is equivalent to choosing $p = \sup_{\dot{\mu} \in I_0} P(|\bar{X} - \mu_0| > |\bar{x} - \mu_0| \mid \mu = \dot{\mu})$. Taking the supremum instead of integrating $\dot{\mu}$ out is the traditional way of dealing with composite null hypotheses (see for example Bickel and Doksum, 2015, p. 217), e.g., when conducting a one-sided t-test. Finally, the distance-only method can be thought of as a thick t-test that assumes the same distribution as the interval-based method but with $\alpha = 50\%$. Hence, we expect the false positive rate to be higher than that of the interval-based method but well below 50% since the assumed distribution is overly conservative.

Thirdly, the thick t-test could be used as a decision method in its own right. The α parameter allows controlling the risk of incorrectly rejecting a true null hypothesis directly and more accurately than the overly conservative interval-based method. Of course, this requires that the assumption about the distribution of the population mean when there is no practically relevant effect is at least approximately true. We don't believe that the need to make this assumption is necessarily a disadvantage compared with the other methods since one would have to make similar assumptions about the distribution of μ under H_0^t if one wanted to derive any guarantees regarding their false positive rates with respect to the thick null hypothesis.

Table 1: Impact of nominal power, method, and true location of the population mean on inference success.

Does the true location fall within the bounds of the thick null?	Nominal power	Number of simulated cases	Inference success rates of each method for each combination of true location and nominal power					
			Conventional	Small-alpha	MESP	Distance-only	Interval-based	Thick t-test
Yes	≥ 0.80	23,869	37.6%	53.3%	90.7%	90.7%	99.7%	95.2%
(45,193 cases)	0.30 to 0.80	12,920	76.3%	92.9%	82.6%	78.8%	99.5%	95.1%
	< 0.30	8,404	90.7%	98.4%	90.7%	55.7%	98.5%	94.7%
No	≥ 0.80	17,674	99.4%	96.7%	92.4%	92.4%	62.6%	85.8%
(54,807 cases)	0.30 to 0.80	14,507	85.5%	66.2%	83.0%	86.8%	42.9%	65.3%
	< 0.30	22,626	56.3%	35.3%	56.3%	88.0%	38.7%	50.6%

3 Results

Table 1 was created based on Table 2 of GSK. The different methods are compared in terms of their overall inference success rate, controlled for the nominal power. The nominal power for a case in the simulation relies on the power calculation for a one-sample z-test for the mean using $\alpha = 0.05$, the case's true population standard deviation σ, sample size n and minimum detectable difference $MPSD$. Table 1 also distinguishes between the true mean falling within and beyond the thick null. In addition to the methods from GSK, the thick t-test introduced in the last section is considered. The table is visualized in Figure 1.

The top half of Table 1 and the left half of Figure 1 show the success rates of the methods when the true parameter lies within the thick null interval. The interval-based method performs best in this case, as already seen in GSK. The new method of the thick t-test falsely rejects H_0^t approximately 5% of the time across all nominal power categories. The success rates of the other methods except for the interval-based method are lower in at least one case and depend on the nominal power. The MESP method exhibits overall moderate error rates.

The bottom half of Table 1 and the right half of Figure 1 show the success rates of the methods when the real parameter lies outside the null interval. The previously best method performs worst in this case. Overall, the distance-only method has high inference success rates and is topped only at high nominal power by the conventional method and the small-alpha method. However, these two methods have a much worse success rate for low nominal power. The newly considered thick t-test performs moderately. It is neither one of the best methods at high nominal power, nor one of the worst methods at low nominal power. MESP performs reasonably well for higher nominal power and exhibits low inference success only in the low nominal power setting.

Overall, the described Table 1 and Figure 1 very much resemble the corresponding illustrations in GSK. The thick t-test method performs very well if the true location falls within the bounds of the thick null. MESP has generally acceptably low error rates except for low nominal power settings

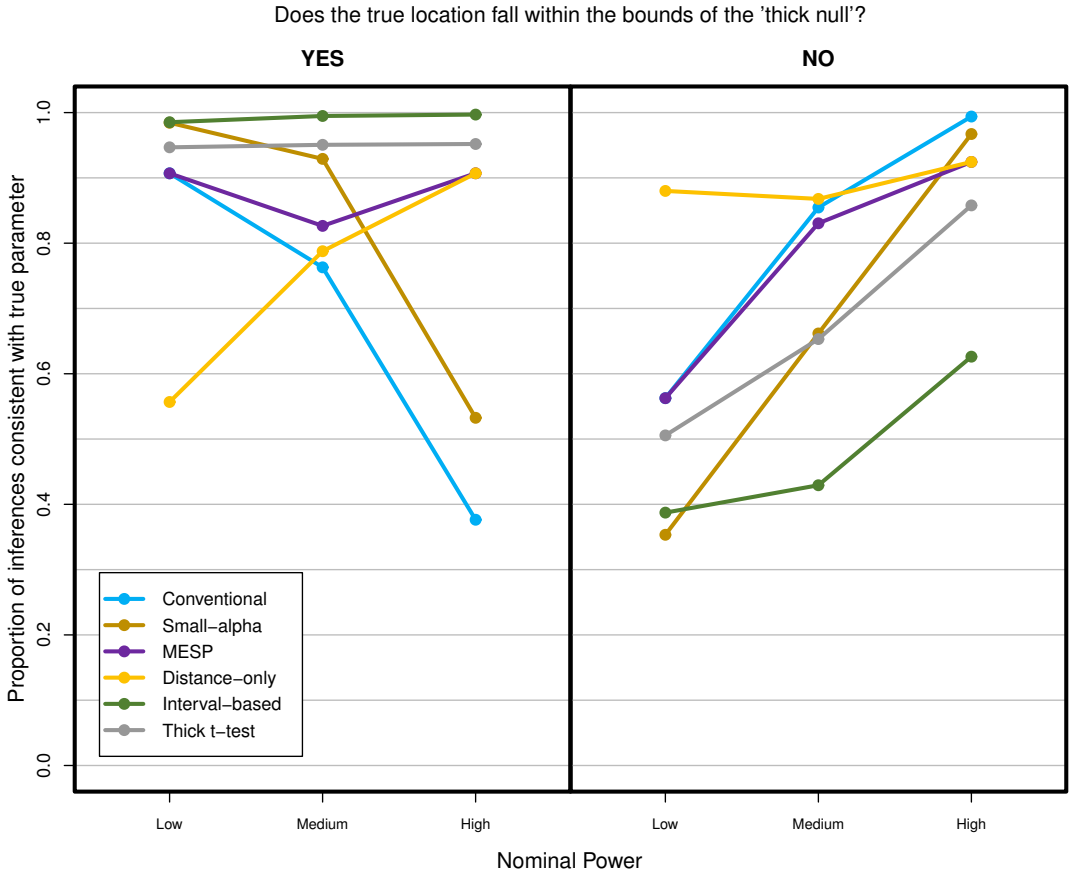

Figure 1: Graph of impact of nominal power, method, and true location of the null on inference success.

when the thick null is not true. As a composite measure of the conventional and the distance-only method it has inference success rates similar to the former when the nominal power is low and similar to the latter when the nominal power is high. The reason for this behavior is intuitive: The p-value computed for H_0^p only plays an important role for the MESP decision if effect sizes need to be really large to lead to a rejection of H_0^t. Conversely, in high nominal power settings p-values are generally low and MESP's decision is only dependent on the distance criterion. An interesting property of MESP is that due to its construction as a composite measure it has lowest inference success for medium nominal power if the thick null holds.

Following Table 3 of GSK, Table 2 shows inference success rates depending on deciles of the relative $MPSD$ which is defined as $MPSD/\sigma$. As $MPSD$ and σ are drawn from respective ranges of integers for the simulation of cases, ties in the relative $MPSD$ occur. When dividing the simulated cases into deciles, we add a tiny value $\epsilon \sim \mathcal{N}(0, 10^{-10})$ to the relative $MPSD$ to distribute ties on

Table 2: Impact of relative $MPSD$ and method on inference success.

Does the true location fall within the bounds of the thick null?	Decile[a] for $MPSD/\sigma$	Number of simulated cases	Inference success rates of each method for each combination of true location and relative $MPSD$					
			Conventional	Small-alpha	MESP	Distance-only	Interval-based	Thick t-test
Yes	1	1,355	93.6%	98.9%	93.6%	38.5%	97.5%	94.8%
(45,193 cases)	2	2,459	90.8%	98.3%	90.8%	54.7%	98.9%	94.6%
	3	3,305	85.4%	96.6%	85.8%	66.4%	98.9%	94.5%
	4	4,297	80.1%	94.8%	83.6%	74.1%	99.3%	95.3%
	5	5,147	75.1%	90.9%	85.2%	79.0%	99.4%	94.5%
	6	5,633	68.0%	85.8%	84.9%	80.8%	99.5%	95.3%
	7	5,605	58.3%	78.1%	86.4%	84.0%	99.5%	95.1%
	8	5,617	48.6%	65.8%	89.3%	88.1%	99.7%	95.3%
	9	5,730	34.6%	49.2%	91.6%	91.4%	99.6%	95.2%
	10	6,045	16.9%	25.7%	95.1%	95.1%	99.9%	95.3%
No	1	8,645	54.6%	34.7%	54.6%	91.8%	43.9%	51.9%
(54,807 cases)	2	7,541	63.3%	43.6%	63.3%	89.7%	43.5%	57.0%
	3	6,695	71.1%	50.4%	71.0%	87.6%	41.8%	58.8%
	4	5,703	75.9%	57.1%	74.6%	85.4%	39.2%	60.0%
	5	4,853	81.3%	64.0%	77.2%	85.1%	38.2%	61.3%
	6	4,367	86.4%	71.9%	80.2%	84.7%	40.0%	64.1%
	7	4,395	91.6%	80.8%	83.8%	86.9%	46.1%	70.7%
	8	4,383	96.0%	90.6%	88.8%	90%	55.2%	80.1%
	9	4,270	98.4%	95.5%	92.5%	93.1%	64.6%	87.7%
	10	3,955	99.9%	99.2%	97.1%	97.1%	79.5%	95.9%

[a]These ranges of values for $MPSD/\sigma$ correspond to the deciles:

1: 0.033–0.103	6: 0.345–0.419
2: 0.103–0.167	7: 0.419–0.533
3: 0.167–0.224	8: 0.533–0.739
4: 0.224–0.286	9: 0.739–1.214
5: 0.286–0.345	10: 1.214–5.000

the boundary of two deciles evenly to those deciles. The table is divided into two parts. The upper one shows success rates when the true location falls within the bounds of the thick null while the lower one shows success rates when the true location does not fall within the bounds of the thick null. Figure 2 visualizes Table 2.

Looking at the upper half of the table one can observe that the interval-based and the thick t-test method both have low error rates with the interval-based method showing almost no inference failure for high relative $MPSD$s. The MESP method has the best success rates for low and high deciles and performs moderately for medium deciles. For the distance-only method the success rate increases considerably the higher the relative $MPSD$ becomes. In contrast to that stand the conventional and the small-alpha method. They perform best for the low relative $MPSD$ values and fall off the higher it becomes.

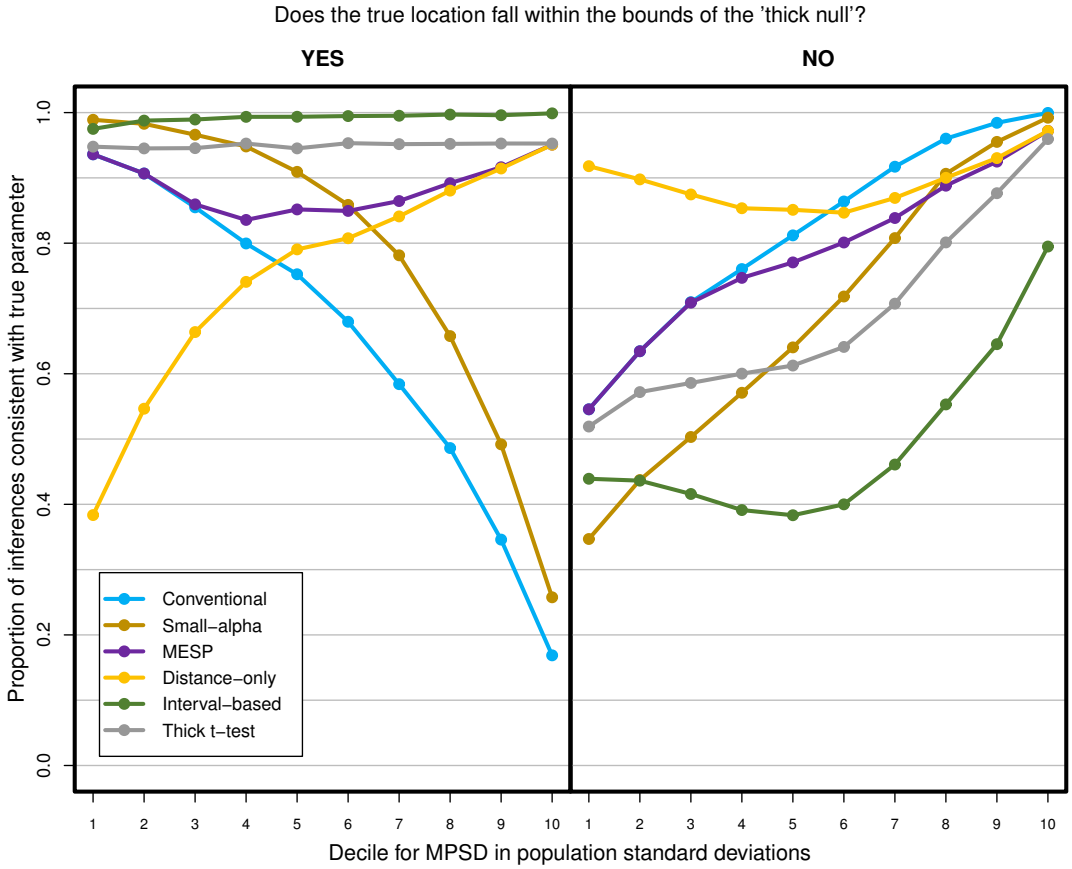

Figure 2: Impact of relative $MPSD$ and method on inference success.

If the thick null is false, the conventional and small-alpha method, MESP as well as the thick t-test perform better the higher the relative $MPSD$ becomes. The interval-based method has considerably worse success rates than the other methods in all but the lowest deciles. The distance-only method has decent success rates, especially for low and high deciles. The MESP method performs nearly as good as the conventional t-test with slightly worse performance in medium and higher deciles. It exhibits inference success rates below 70% only in settings where the relative $MPSD$ is low. In practice, such a setting could be recognized by the researcher before conducting a study if a reliable guess on the population's standard deviation is available.

Overall, these numbers confirm the results in GSK regarding the success rates for different ranges of the relative $MPSD$. As already shown in Table 1 and Figure 1, the added thick t-test falsely rejects the thick null less often compared to other methods tailored to thick null hypotheses such as the MESP and the distance-only method, but has lower inference success rates if the thick null does not hold.

Table 3: Error rates and inference success rate for each method.

	Conventional	Small-alpha	MESP	Distance-only	Interval-based	Thick t-test
False positive rate	41.4%	27.0%	11.6%	19.2%	0.6%	4.9%
False negative rate	22.1%	36.7%	25.0%	10.9%	52.5%	34.2%
False discovery rate	30.5%	26.0%	11.3%	15.1%	1.0%	5.8%
False omission rate	31.4%	37.9%	25.5%	14.1%	39.0%	30.4%
Inference success rate	69.2%	67.7%	81.1%	85.3%	71.0%	79.0%

Table 3 and Table 4 are inspired by Figure A7 of GSK. Table 3 shows error rates as well as overall inference success rates, i.e., the share of correct inference decisions over all simulations, with respect to the thick null. With an inference success rate of 69.2% for the conventional and 67.7% for the small-alpha variant, the point t-tests together with the interval-based method (71.0%) have the lowest inference success rate. With 85.3% the distance-only method has the highest overall inference success rate. Of course, in another simulation setting with the thick null being true more often than 45.2% as in our setting (Table 1), the picture might look different. Table 3 also shows that the conventional t-test has the highest false positive rate (41.4%) with respect to the thick null. Lowering the alpha of the t-test to 0.005 reduces the false positive rate to 27.0% at the expense of a higher false negative rate.

Combining the inferences of the conventional t-test and the distance-only method results in an inference success rate of 81.1% for the MESP method with a false positive rate of 11.6% and a false negative rate of 25.0%. The interval-based method has the lowest false positive rate (0.6%) but with 52.5% also the highest false negative rate of all six methods. Due to its construction the thick t-test yields a false positive rate of 4.9% and has a false negative rate that is 18.3 percentage points lower compared to the interval-based method.

The false discovery rate is important from a practical perspective, as it is defined as the share of false positive findings among all positive findings (rejections of the null hypothesis). In the context of a thick null hypothesis, this is the probability that there is actually no practically relevant effect in the population if the decision criterion indicates that there is one. In our simulation, the false discovery rate is lowest for the interval-based method, while the false omission rate as its counterpart is lowest for the distance-only method. While the false discovery and false omission rate computed here are dependent on the share of true thick null hypotheses and can be derived from Table 1, one would have to decide on a prior probability of the thick null being true for a single study in practice.

Table 4 compares the six methods regarding the false discovery rate and the false omission rate in more detail. In this table, a distinction is made between high, medium and low nominal power. As the false discovery and omission rates are crucially dependent on the share of thick null hypotheses being true, they are normalized by using true/false positive/negative rates to compute

them instead of absolute numbers, i.e.,

$$\text{normalized false discovery rate} = \frac{\text{false positive rate}}{\text{false positive rate} + \text{true positive rate}} \text{ and}$$

$$\text{normalized false omission rate} = \frac{\text{false negative rate}}{\text{false negative rate} + \text{true negative rate}},$$

where the true/false positive/negative rates can be derived from Table 1. To this end, comparisons between different nominal power categories are allowed and it can be investigated whether methods improve with increasing nominal power. All rates should only be interpreted relative to each other since, in general, the false discovery rate decreases and the false omission rate increases with more true locations falling beyond the thick null. Thus, a different simulation setting, e.g., with more true locations falling beyond the thick null, would lead to different results.

The interval-based method has false discovery rates which are well below 5% for all nominal power categories. The thick t-test method also exhibits overall low false discovery rates. The first three methods shown in Table 4 do not have the generally desirable property of a decreasing false discovery rate for rising nominal power. While the MESP has overall medium false discovery rates, the false discovery rate rises when moving from low to medium nominal power. The reason is the aforementioned construction of the MESP as a composite measure of the conventional and the distance-only method which have contradicting false discovery rates when the nominal power increases.

Regarding the false omission rate, a consistent structure can be recognized across all methods: The false omission rate increases with decreasing nominal power. For high nominal power, the value is lowest for the conventional method (1.6%) and highest for the interval-based method (27.3%). For low nominal power, the distance-only method shows the lowest value of 17.7%.

When considering the false discovery rate and the false omission rate, there is no clear best method even when disregarding the low nominal power category. For medium and high nominal power, many methods have false discovery rates that arguably exceed desirable values. The interval-based method works best in this regard at the expense of a considerable false omission rate even in the highest nominal power category. To this end, the individual context of a study plays a crucial role in the choice of the appropriate decision criterion.

Table 4: (Normalized) false discovery rate and (normalized) false omission rate depending on nominal power.

	Nominal power	Number of simulated cases	Conventional	Small-alpha	MESP	Distance-only	Interval-based	Thick t-test
(Normalized) false discovery rate	≥ 0.80	41,543	38.6%	32.6%	9.1%	9.1%	0.5%	5.3%
	0.30 to 0.80	27,427	21.7%	9.7%	17.3%	19.7%	1.2%	7%
	< 0.30	31,030	14.2%	4.2%	14.2%	33.5%	3.7%	9.5%
(Normalized) false omission rate	≥ 0.80	41,543	1.6%	5.8%	7.7%	7.7%	27.3%	13%
	0.30 to 0.80	27,427	16%	26.7%	17%	14.4%	36.5%	26.7%
	< 0.30	31,030	32.5%	39.6%	32.5%	17.7%	38.3%	34.3%

4 Conclusion and Discussion

In this comment, we confirmed the results of GSK using our own software code written in R that enables the easy implementation of further decision criteria. We added such a decision criterion that allows to control for a pre-defined type I error rate with respect to a thick null hypothesis.

We confirmed that the MESP method as proposed by GSK has comparably low type I and type II error rates with respect to a thick null hypothesis in settings the researcher can check for. Only in low nominal power settings the MESP method has difficulties to detect a practically relevant effect and, accordingly, has a quite high false omission rate. The false discovery rate is moderate in comparison with other decision criteria but has the undesirable property that it does not decrease monotonically with rising nominal power (assuming a fixed prior probability for the thick null being true). If the researcher's aim is to keep the false discovery rate low, the MESP should only be used in high nominal power settings.

The MESP method could also be applied for *point* null hypothesis testing: As it augments the conventional t-test by taking the minimum practically significant effect size into account, MESP is a simple and effective approach for adding an additional layer of protection against false positives without strongly increasing the false negative rate. However, if researchers are serious about testing whether the *thick* null hypothesis is true, MESP lacks a parameter to adapt the method to contexts with different costs of false negatives and positives with respect to the thick null hypothesis. In contrast, the interval-based method and the thick t-test (assuming the chosen distribution of $\mu \mid H_0^t$ is approximately true) provide such a parameter with α being the upper bound of the false positive rate or approximately equal to it, respectively.

The performance of the decision criteria generally depends on whether the thick null hypothesis is true or not. In practice, one can make an informed guess on this probability, i.e., specifying a prior probability, or select the decision criterion according to the estimated nominal power and the type of error one wants to avoid.

If the prior probability on the thick null being true is set high, it is advisable to select a decision criterion with low false positive rate, namely the interval-based method or the thick t-test. Vice versa, if one rather believes the true location lies outside the thick null interval, the distance-only

method is generally the method of choice as it minimizes the false negative rate. However, the prior is ultimately unknown and usually one rather knows the sample size and the MPSD and can make a good estimate of the population's standard deviation. Thus, the nominal power can be approximated and used for selecting a decision criterion: In high nominal power settings, the MESP and the distance-only are good choices irrespective of whether the thick null hypothesis is true or not. If the nominal power is low, there is no criterion that can be generally recommended: If the aim is to minimize the false positive rate, all approaches except for the distance-only method work well. The latter, however, is the only criterion that has a decent false negative rate in such a case.

In general, it is recommendable for hypothesis testing to specify and publish the (thick) null hypothesis, planned analyses methods, and the decision criterion before conducting the study to reduce options for p-hacking and increase transparency in science. Irrespective of the decision criterion used, every empirical study is subject to limitations and assumptions and thus uncertainty beyond sampling noise. While an extensive debate on statistical inference including dos and don'ts as well as merits and pitfalls can be found in Wasserstein et al. (2019) and the corresponding issue of the *The American Statistician*, we conclude by citing a generic piece of advice to the research community using statistics: "Accept uncertainty. Be thoughtful, open, and modest" (Wasserstein et al., 2019).

References

Albert, I., S. Donnet, C. Guihenneuc-Jouyaux, S. Low-Choy, K. Mengersen, and J. Rousseau (2012). "Combining expert opinions in prior elicitation." *Bayesian Analysis*, 7(3), 503–32. DOI: 10.1214/12-BA717.

Bayarri, M. J. and J. O. Berger (2000). "*P* values for composite null models." *Journal of the American Statistical Association*, 95(452), 1127–42. DOI: 10.2307/2669749.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2018). "Redefine statistical significance." *Nature Human Behaviour*, 2(1), 6–10. DOI: s41562-017-0189-z.

Betensky, R. A. (2019). "The p-value requires context, not a threshold." *The American Statistician*, 73(1), 115–17. DOI: 10.1080/00031305.2018.1529624.

Bickel, P. J. and K. A. Doksum (2015). *Mathematical statistics: Basic ideas and selected topics, Volume I.* 2nd ed. Chapman and Hall/CRC.

Blume, J. D., R. A. Greevy, V. F. Welty, J. R. Smith, and W. D. Dupont (2019). "An introduction to second-generation p-values." *The American Statistician*, 73(1), 157–67. DOI: 10.1080/00031305.2018.1537893.

Box, G. E. P. (1980). "Sampling and bayes' inference in scientific modelling and robustness." *Journal of the Royal Statistical Society. Series A (General)*, 143(4), 383–430. DOI: 10.2307/2982063.

Goodman, W. M., S. E. Spruill, and E. Komaroff (2019). "A proposed hybrid effect size plus p-value criterion: Empirical evidence supporting its use." *The American Statistician*, 73(1), 168–85. DOI: 10.1080/00031305.2018.1564697.

Greenland, S. (2019). "Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values." *The American Statistician*, 73(1), 106–14. DOI: 10.1080/00031305.2018.1529625.

Kruschke, J. K. (2011). "Bayesian assessment of null values via parameter estimation and model comparison." *Perspectives on Psychological Science*, 6(3), 299–312. DOI: 10.1177/1745691611406925.

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett (2019). "Abandon statistical significance." *The American Statistician*, 73(1), 235–45. DOI: 10.1080/00031305.2018.1527253.

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 Pharmacodynamics*, 15(6), 657–80. DOI: 10.1007/BF01068419.

Schulz, S., M. Zondervan-Zwijnenburg, S. A. Nelemans, D. Veen, A. J. Oldehinkel, S. Branje, and W. Meeus (2021). "Systematically defined informative priors in bayesian estimation: An empirical application on the transmission of internalizing symptoms through mother-adolescent interaction behavior." *Frontiers in Psychology*, 12, 620802. DOI: 10.3389/fpsyg.2021.620802.

Wasserstein, R. L. and N. A. Lazar (2016). "The ASA's statement on p-values: Context, process, and purpose." *The American Statistician*, 70(2), 129–33.
DOI: 10.1080/00031305.2016.1154108.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). "Moving to a world beyond "p<0.05"." *The American Statistician*, 73(sup1), 1–19. DOI: 10.1080/00031305.2019.1583913.