2025 ■ Vol. 21 ■ no. 2

When are scientific findings deemed practically or theoretically relevant? A literature review on minimum-effect testing

Paul Riesthuis [a,b] and Robert A. Cribbie [c]

[a]Faculty of Law and Criminology, KU Leuven, Leuven, Belgium

[b]Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands

[c]Faculty of Health, York University, Toronto, Canada

Abstract ■ Null-hypothesis significance testing and p-values have been criticized for their focus on whether an effect is merely different from zero. To address this limitation, Murphy and Myors (1999) introduced minimum-effect testing to psychology researchers in the Journal of Applied Psychology approximately 25 years ago. Minimum-effect tests go beyond assessing whether an effect is statistically significantly different from 0; they also examine whether it is larger in magnitude than a predefined effect size (in a positive or negative direction) that is considered practically or theoretically meaningful. The usefulness of these tests depends largely on how researchers determine and justify the smallest effect size they consider meaningful, known as the smallest effect size of interest (SESOI). As minimum-effect tests are gaining popularity, we conducted a systematic literature review to examine how researchers in psychology apply minimum-effect tests, including how they define and justify their SESOI. We found that many studies relied on unstandardized effect sizes to define their SESOI. However, appropriate justifications for SESOIs were often missing, and in some cases, no justification was provided at all. Additionally, SESOIs were sometimes determined after data collection, or it was unclear when they were established. Furthermore, many studies did not account for minimum-effect testing in their power analyses. We offer several recommendations to improve the implementation of minimum-effect tests in psychological research.

Keywords ■ minimum-effect testing; interval hypotheses; smallest effect size of interest; minimally meaningful effect size; minimally clinically important difference.

✉ paul.riesthuis@kuleuven.be

10.20982/tqmp.21.2.p082

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ Dominic Beaulieu-Prévost (Université de Montréal)

■ Adam Smiley (Belmont University)

■ and two anonymous reviewers.

Introduction

The overreliance on traditional null-hypothesis significance testing (NHST) and p-values within the field of psychology has frequently been criticized because of several limitations with the approach (Amrhein et al., 2019; Cumming, 2014). By "traditional NHST," we refer to the way it is commonly practiced in psychology, where the null hypothesis is typically specified as a nil effect (e.g., $\delta = 0$), despite the fact that NHST can, in principle, allow for different nulls (Gigerenzer, 2004). We illustrate some of these limitations through the following hypothetical study. Suppose a group of researchers conducted a study on the effects of alcohol on eyewitness memory wherein some participants received alcohol while others received a placebo. Subsequently, the participants watched a video of a crime and one week later their memory for the crime was assessed. To examine whether alcohol negatively affected eyewitness memory in terms of correct details remembered, researchers would normally conduct an independent samples t-test following the traditional NHST framework. The issues that arise with the traditional NHST stem from its focus on whether an effect exists or not (Farmus et al., 2023). Now imagine that the researchers find a $p > .05$. In this scenario, they can only conclude that there was insufficient empirical evidence to state that participants

who were given alcohol remembered fewer (or more) correct details than participants who received the placebo. In other words, the traditional NHST framework is not geared towards finding evidence for the null hypothesis; i.e., in the example above it is practically impossible to prove that an effect is precisely zero (Lakens, 2021). To provide evidence regarding a negligible association (e.g., equivalence of population means), researchers have to conduct equivalence tests wherein it is examined whether an effect is too small to matter practically or theoretically speaking (Cribbie et al., 2004).

A second issue that arises when conducting traditional NHST is when the researchers find a $p < .05$. In this scenario, researchers can only conclude that they found a non-zero difference between the alcohol and the placebo group in terms of the number of correct details remembered from the crime. This could be a practically or theoretically relevant difference, but the traditional NHST framework only addresses whether there is non-zero difference between the two groups. This becomes especially problematic with increasing sample sizes because traditional NHST can lead to statistically significant but trivial differences (Anvari & Lakens, 2021).

In response to the misuse of NHST, wherein any effect statistically different from zero was often assumed to be meaningful (Meehl, 1967, 1978), researchers developed methods to address whether an effect is practically or theoretically significant. Notable examples include the good-enough principle (Serlin & Lapsley, 1985) or the work on equivalence testing by Rogers et al. (1993). However, it was not until Murphy and Myors (1999) introduced the concept of minimum-effect testing in the Journal of Applied Psychology that such approaches were formally framed as minimum-effect tests within the field of psychology. In minimum-effect testing, we do not examine whether the difference between the alcohol and control group is greater or less than zero, but instead whether the effect is larger in magnitude than a specific effect size (e.g., raw mean difference) that is deemed practically or theoretically relevant. It follows that in both equivalence and minimum-effect tests, the effect size that separates meaningful from non-meaningful effects plays a crucial role. In the present study, we conducted a literature review to examine how minimum-effect testing has been adopted in the psychological literature with a special focus on the role of effect sizes.

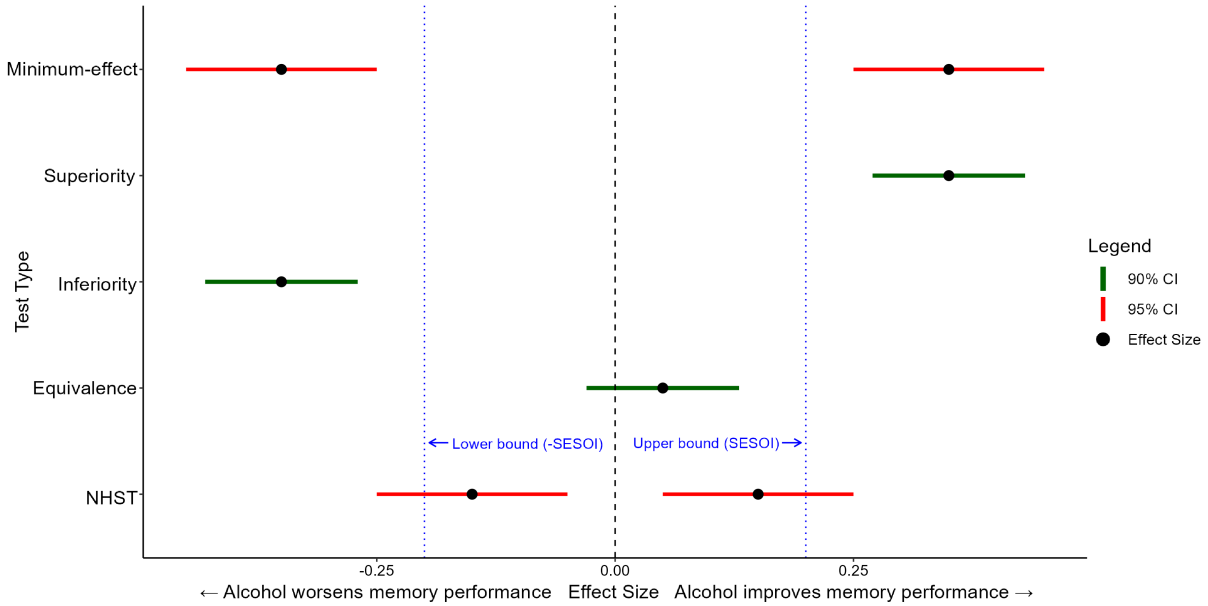
Interval Hypothesis Testing

Equivalence and minimum-effect tests are different types of interval hypothesis tests. Interval hypotheses can be regarded as extensions of the traditional NHST framework. For instance, in equivalence testing, instead of examining whether the difference between two groups is exactly zero, one tests whether the effect is smaller than a range of values deemed too small to matter (Seaman & Serlin, 1998). For instance, the researchers examining the effects of alcohol on memory can test whether the difference in memory performance between the alcohol and placebo group is smaller than one detail correctly remembered, as this could be considered too small to matter in legal cases (see Otgaar et al., 2023). In other words, the researchers define a difference of one detail as the threshold for practical relevance. For minimum-effect testing, the traditional NHST framework can also be extended wherein it is now examined whether the difference in memory performance between the alcohol and placebo groups is greater than one detail remembered instead of examining whether there is a non-zero difference. Minimum-effect testing can be examined in either direction (i.e., test of inferiority or superiority). In other words, if researchers want to determine whether participants in the alcohol group, on average, remembered one detail fewer than those in the placebo group—without considering the possibility that they might perform better—this would be classified as a test of inferiority. If the hypothesis is flipped (i.e., researchers are only interested in whether intoxicated participants have better memory performance than sober participants), it is considered a test of superiority. Taken together, the simple extension in interval hypotheses is that the null is replaced by a specific range of values (i.e., one detail remembered) that indicates when an effect is considered practically or theoretically meaningful (Murphy & Myors, 1999). Evidently, for each type of interval hypothesis test, a vital component is determining which effects are deemed practically or theoretically meaningful, also known as the smallest effect size of interest (SESOI; Lakens et al., 2018), minimally meaningful effect size (MMES; Beribisky et al., 2019), good enough value (Serlin & Lapsley, 1985), or minimally clinically important difference (MCID; McGlothlin & Lewis, 2014).[1] These concepts align conceptually with related notions such as practical significance (Kirk, 1996) and substantial effect (Beaulieu-Prévost, 2006), all of which seek to define what constitutes a meaningful effect in applied research.

The simplest way to understand interval hypothesis tests is via the confidence interval (CI) approach (Smiley et al., 2023). CIs can be used to conduct NHST, equivalence tests, and minimum-effect tests (see Figure 1). Specifically, for NHST, if the 95% CI of the effect size does not include 0, the effect can be deemed statistically significant.[2] The NHST can be conducted to examine the effect in either direction unless a specific directional hypothesis is

[1] We used SESOI throughout the article.
[2] Assuming $\alpha = .05$.

Figure 1 ■ Overview of the different interval hypotheses. The figure provides examples of when the null hypothesis for each test can be rejected. For example, in the bottom row, it is shown that the null hypothesis for a test of superiority ($H_0 : \mu_{Alc} - \mu_{NoAlc} \leq \delta$) can be rejected if the lower bound of the 90% CI falls above the upper bound of the SESOI.



given (e.g., one tailed t-test; only interested in whether alcohol impedes memory). Equivalence can be established by examining whether the entire 90% CI of the effect size falls completely within the bounds set by the SESOI.[3] In minimum-effect testing, it is examined whether the entire 95% CI of the effect size falls completely beyond the SESOI.[4] When researchers are only interested in a single direction, they can perform tests of inferiority or superiority. Let's say the research hypothesis is that alcohol worsens memory performance. The null and alternative hypotheses can be expressed as:

$$H_0 : \mu_{\text{Alc}} - \mu_{\text{NoAlc}} \geq \text{SESOI}_{\text{lower}}$$
$$H_A : \mu_{\text{Alc}} - \mu_{\text{NoAlc}} < \text{SESOI}_{\text{lower}}$$

H_0 is rejected if the upper bound of the 90% CI is less than the lower bound of the SESOI. Or, the research hypothesis could be that alcohol improves memory performance. The null and alternative hypotheses can be expressed as:

$$H_0 : \mu_{\text{Alc}} - \mu_{\text{NoAlc}} \leq \text{SESOI}_{\text{upper}}$$
$$H_A : \mu_{\text{Alc}} - \mu_{\text{NoAlc}} > \text{SESOI}_{\text{upper}}$$

H_0 is rejected if the lower bound of the 90% CI is greater than the upper bound of the SESOI.

SESOI and Interval Hypotheses

A recent review examined how equivalence tests are currently conducted in psychological research with a special focus on how the SESOI was established (Marshall et al., 2024). Interestingly, they found that most studies used unstandardized effect sizes to set their equivalence margins (i.e., SESOI; MMES; MCID) which is recommended to assess the practical or theoretical meaningfulness of an effect (Baguley, 2009; Riesthuis, 2024). However, they found suboptimal practices in how the equivalence margins were justified. Specifically, only 23 out of the 122 studies (18%) provided a justification for their equivalence margin, while the rest either provided a clear description but no justification (51/122; 42%), vague description and no justification (40/122; 33%), or no description or justification at all (8/122; 7%). If researchers haphazardly determine their SESOI without any justification, which can lead to uninformative SESOIs/tests, this could be deemed a questionable research practice (QRP; John et al., 2012). One explanation

[3] See Alter and Counsell (2023), Metzler (1974), or Seaman and Serlin (1998) for explanations of why the 90% CI is used in equivalence testing when $\alpha = .05$.

[4] We recommend the use of a 95% CI approach for a two-tailed minimum-effect test because in extreme situations wherein the SESOI is relatively small compared to the standard deviation, the Type 1 error rate is inflated ($\approx$ 10%) using the 90% CI approach (two-one sided test with $\alpha = .05$; Schuirmann, 1987). For one-tailed minimum-effect tests (e.g., test of superiority or inferiority), the 90% CI approach can be used.

for the lack of justifications could be that researchers decided to conduct equivalence tests after they found a statistically non-significant effect ($p > .05$) using traditional NHST. In other words, they may have wanted to provide evidence for the null hypothesis after the results were known which, as discussed above, traditional NHST cannot do. Irrespective of the reasons why researchers conducted the equivalence tests this way, these practices are problematic because interval hypothesis tests are only as valuable as the chosen SESOI and its justification.

In sum, interval hypothesis tests provide numerous advantages over the traditional NHST framework, such as providing evidence for a negligible effect (equivalence testing) or showing that an effect is practically or theoretically relevant (minimum-effect testing). However, for interval hypothesis testing, the SESOI plays a crucial role. To date, there has been no review on how minimum-effect tests are conducted in the psychological literature and how the SESOIs are set and justified for these types of tests. Hence, in the present study, we conducted a systematic review of published psychological research to examine the use of minimum-effect tests, with a specific focus on the SESOI.

Systematic Review

In the present systematic review, we examined all important methodological practices within the framework of minimum-effect testing. Specifically, we examined the type of minimum-effect tests researchers conducted (e.g., superiority, inferiority, minimum-effect test), whether a minimum-effect test was directly or indirectly tested (e.g., use CI to interpret effects in terms of minimum-effect testing without specifically specifying that a minimum-effect test was conducted), the scale of the SESOI (e.g., unstandardized effect size, Cohen's d, Pearson's r, etc.), the timing of setting the SESOI (i.e., before or after the study was conducted), the justification for the SESOI (e.g., cost-benefit analysis, anchor-based method, consensus method, etc.), whether a power analysis was conducted, and, if the power analysis was specifically done for a minimum-effect test. We did not have any specific hypotheses and hence the study was exploratory in nature. The study was not preregistered.

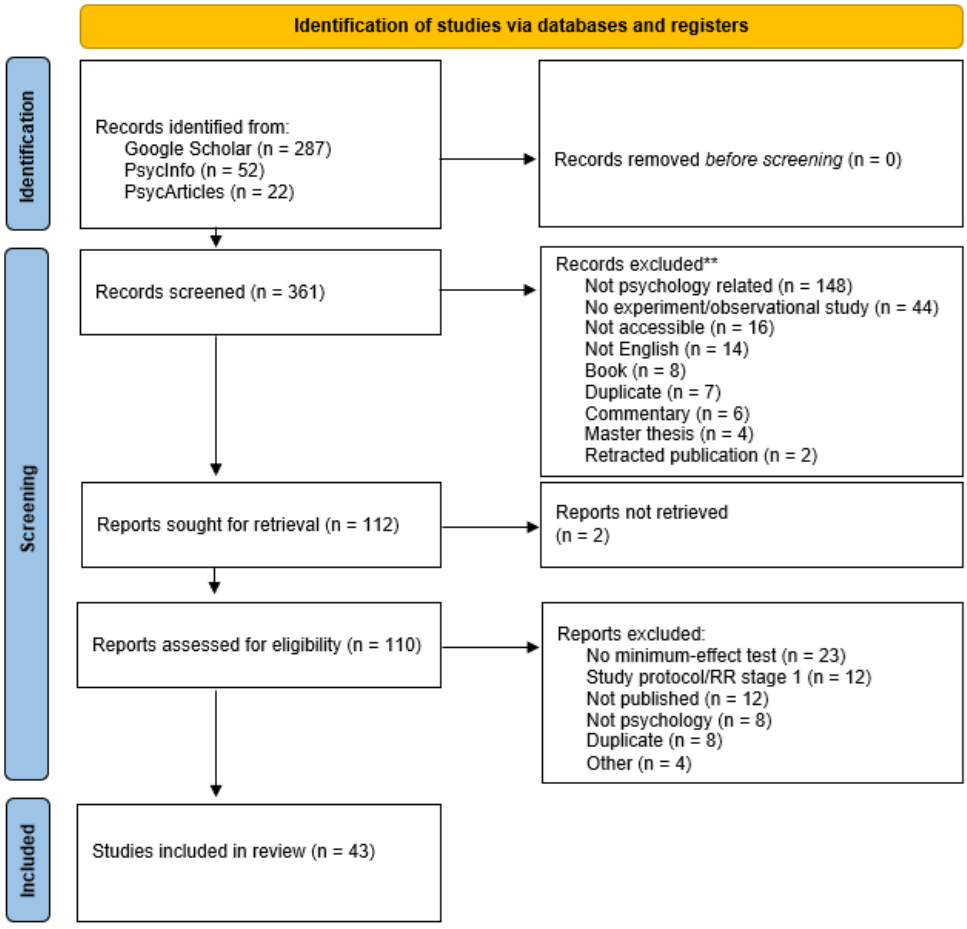
Method

Literature Search

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; see Figure 2; Page et al., 2021) workflow to transparently show the flow of information through the literature review. For our literature review, we examined the following three databases: PsycINFO, PsycArticles, and Google Scholar (see osf.io/j78dz/ for literature search results for each database). Based on the expertise of both authors, we did not expect to find many articles that conducted minimum-effect testing or variants thereof in the psychological literature. Hence, we opted for a broad search to include as many relevant articles as possible. This led to the use of the following keywords to appear anywhere in the full-text: "superiority test*" OR "inferiority test*" OR "minimum effect test*" OR "minimal effect test*" AND psychology (note that the '*' represents a wildcard which permits any text in that location; e.g., for 'test*' the algorithm will pick up 'tests', 'testing', etc.). To avoid the inclusion of the multitude of noninferiority tests which are conducted more frequently in psychological research but align conceptually with equivalence testing not minimum-effect testing, we also included the following search term: -"noninferiority". The literature search was conducted on December 19th, 2024, and all articles published before this date were included. Moreover, we only included articles that were published in a scientific journal, were in English, and conducted minimum-effect tests.

The first literature search resulted in a total of 361 articles ($n_{\text{Google Scholar}} = 287$, $n_{\text{PsycInfo}} = 52$, $n_{\text{PsycArticles}} = 22$). The first round of the literature search consisted of screening the title and abstract. The most important inclusion criteria during this round were that: 1) the article was psychologically related research (e.g., the Google Scholar search included also medical articles and these were excluded), 2) was an experimental or observational study (no reviews, commentaries, books, etc.), and 3) seemed to have conducted a minimum-effect test. A liberal inclusion criteria was used in this round to include as many relevant articles as possible for the literature review. In the title and abstract screening round, 249 articles were excluded and 112 articles remained for the full-text screening round. The main reasons for excluding articles during the title and abstract screening round were as follows: not related to psychology ($n = 148$), did not report an experimental or observational study ($n = 44$), not accessible ($n = 16$), not in English ($n = 14$), books ($n = 8$), duplicates ($n = 7$), commentaries ($n = 6$), master's theses ($n = 4$), and retracted publications ($n = 2$). In the full-text screening round, articles were again assessed on whether they were: 1) psychologically related studies, 2) experimental or observational, and 3) conducted a minimum-effect test. Additionally, it was examined whether: 4) the full text was available, 5) the study was completed (e.g., no stage 1 registered reports or study protocols), 6) the study was published, and 7) the study was not a duplicate (see Figure 2). We did not include unpublished studies, as these may not have undergone peer review, and their exclusion ensures that all included studies met at least a basic standard of methodological and report-

Figure 2 ■ PRISMA Workflow of literature search



ing quality. In the end, 43 articles were included and were coded during the full-text screening round for the literature review.

Coding

The first author independently coded the 43 articles. From each article, the type of minimum-effect test researchers conducted as reported in the article was coded. Specifically, if the researchers indicated that they conducted a superiority test, then this was coded as "superiority testing". It was also coded whether the researchers conducted the minimum-effect tests directly (explicitly indicating that they used some variant of a minimum-effect test) or indirectly (comparing a confidence interval to a certain value, e.g., SESOI, without specifically indicating that they used some form of minimum-effect test). Further, for the SESOI, the scale (e.g., unstandardized, Cohen's d, Pearson's r, etc.), the exact effect size (e.g., the SESOI was a Cohen's $d = .20$), and the justifications (e.g., cost-benefit analysis, anchor-based method, consensus method, etc.) were coded. The exact paragraph or quote where the researchers indicated their SESOI was also extracted (see "`Final_Set.xlsx`" file on osf.io/j78dz/ for exact quotes). Finally, to additionally examine whether researchers planned for a minimum-effect test, it was extracted whether a power analysis was conducted and whether it was specifically for a minimum-effect test. The corresponding sample size indicated by the power analysis and the final sample size were also extracted. The final set of articles and the extracted information can be found on the Open Science Framework (OSF; osf.io/j78dz/). To conduct inter-rater reliability analyses, the second author recoded approximately 20% of the articles (8/43; see Table 1). For most coded articles there was perfect agreement between the two coders. The low reliability coefficients of "Justification SESOI", "Power analysis conducted", "Power analysis MET", and "N power" were

Table 1 ■ Overview of the inter-rater reliability estimates.

Coded variable	Lower 95% CI	Reliability coefficients	Upper 95% CI
# of experiments*	1	1	1
# of analyses*	1	1	1
MET conducted	1	1	1
MET direct or indirect	1	1	1
Type MET	1	1	1
Scale SESOI	1	1	1
Exact SESOI*	1	1	1
Timing SESOI	1	1	1
Justification SESOI	.56	.68	.79
Power analysis conducted	.11	.55	.99
Power analysis MET	.39	.61	.82
N power*	.99	.99	1.00
N Final*	1	1	1

Note. MET = minimum-effect test. *For the coded variables "# of experiments", # of analyses", "# of power analyses", "Exact SESOI", "N power" and "N Final", we conducted a two-way random model, absolute agreement, single measure intra-rater class correlation. The other coded variables were analysed using the unweighted Cohen's kappa coefficients (see osf.io/j78dz/ for the analyses). For "MET conducted" and "Exact SESOI" a formal reliability analysis could not be performed due to perfect agreement and a lack of variance and we thus indicated 1.

due to a single error that pertained to multiple analyses. When this single error was corrected, there was perfect agreement. The discrepancy in "Justification SESOI" occurred because the first coder indicated that the SESOI was based on the anchor-based method, whereas the second coder stated it was based on previous research establishing the SESOI. In fact, the SESOI was based on previous research that employed anchor-based methods (see Reliability Analyses on OSF: osf.io/j78dz/). The low reliability coefficients for "Power analysis conducted," "Power analysis MET," and "N power" were all due to a single error that affected all three variables. The first coder indicated that no power analysis had been conducted because it was not explicitly stated, whereas the second coder indicated that it had. This single discrepancy then led to divergent responses on "Power analysis MET", and "N power". Correcting the error led to perfect agreement. Last, one article was excluded because it did not conduct a superiority test but focused on the probability of superiority (i.e., common language effect size; Mastrich & Hernandez, 2021).

Data Analysis Plan

We first provide an overview of how many experiments (i.e., studies; some articles reported multiple experiments/studies) and minimum-effect tests were conducted in the included articles. Then, we provide the characteristics of the various SESOIs. However, the majority of studies reported multiple minimum-effect tests while only providing a single SESOI. Moreover, it was sometimes unclear whether the reported SESOI was relevant for all conducted analyses. If researchers indicated that they would conduct, for example, superiority tests for all analyses and only provided one SESOI, then these analyses were coded using this singular SESOI. This means that results may be overinflated as some studies conducted 30 minimum-effect tests using a singular SESOI with no justifications which may give a skewed view of the state of the use of minimum-effect tests in psychological research. Hence, for the data analysis, we present the results for every unique SESOI provided per experiment ($n = 60$). We also analyzed all conducted minimum-effect tests separately ($n = 288$), but the pattern of results was largely unchanged (see osf.io/j78dz/ for the analysis on all tests separately).

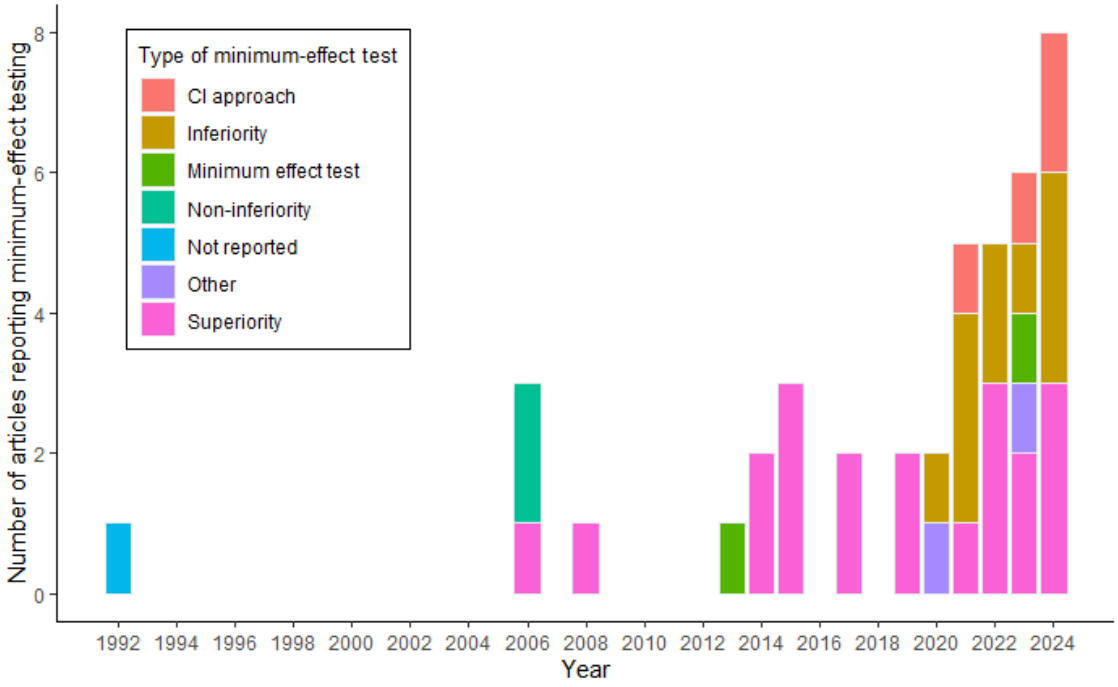
Required Software

For the data analysis, we used the following open-source research software: `R` (R Core Team, 2023), `readxl` (Wickham & Bryan, 2023), `dplyr` (Wickham et al., 2023), `tibble` (Müller & Wickham, 2023), `tidyverse` (Wickham et al., 2019), `ggplot2` (Wickham, 2016), and `gtsummary` (Sjoberg et al., 2021).

Results

In total, there were 42 articles that conducted some type of minimum-effect test. In the 42 articles, there were 48 experiments reported, 60 unique SESOIs specified, and 288 minimum-effect tests conducted. Specific findings are summarized in Table 2. The adoption of minimum-effect tests

Figure 3 ■ Visualization of the use of minimum-effect tests over time.



steadily increased since 2019 (see Figure 3).[5]

Type of Minimum-Effect Test

When examining the characteristics of the minimum-effect tests for each uniquely provided SESOI, researchers most frequently labeled their minimum-effect test as tests of superiority (27/60; 45%), whereas 27% (16/60) were labeled as tests of inferiority, 12% (7/60) were labeled as a confidence interval test, 6.7% (4/60) were labeled as a minimum-effect tests, 6.7% reported it differently (e.g., mean reduction comparison to threshold), and 3.3% (2/60) did not report which test they implemented. Additionally, 52 out of the 60 (85%) tests were directly tested as a minimum-effect test and 9 (15%) indirectly. The majority of direct minimum-effect tests were labeled as a test of superiority (27/51; 53%) or inferiority (16/51; 31%).

SESOI Characteristics

The scale of the SESOI was frequently reported as an unstandardized effect size (20/60; 33%; e.g., raw mean difference; unstandardized regression coefficient), whereas 11% (7/60) of the time it was reported as a proportion. The remaining SESOIs (26/60; 43%) were standardized effect sizes, where 18% (11/60) were reported as Cohen's d, 6.7% (4/60) as a correlation, 6.7% as Hedges' g, 3.3% (2/60) as a standardized regression coefficient, 1.7% (1/60) as an odds ratio (can also be regarded as an unstandardized effect size), 1.7% as variance explained, and for 17% (10/60) of the SESOIs there was no scale provided (see `Rmarkdown` file on OSF osf.io/j78dz/ for histogram of standardized SESOIs magnitudes). The SESOI was set in the majority of the cases before the study was conducted (44/60; 73%), whereas six times it was set after the study was conducted (10%) and in eleven cases (17%) it was unknown when the researchers set the SESOI.

The justifications for the chosen SESOIs were as follows: 25% (15/60) relied on effect sizes found in previous research, 20% (12/60) relied on previous research examining the SESOI, 10% (6/60) relied on benchmarks, 8.5% (5/60) provided subjective arguments, 3.4% (2/60) relied on multiple approaches, 1.7% (1/60) relied on consensus methods, 1.7% relied on the small telescopes approach, 1.7% relied on the effect size of the to-be replicated study, and 29% (17/60) did not provide any justifications.

Power Analysis

We also examined whether researchers conducted power analyses and whether they were specifically for a type of

[5]The study of 1992 indirectly conducted a minimum-effect test by "According to Cohen and Cohen, only correlations over .30 are interpreted".

Table 2 ■ Characteristics of the minimum-effect testing for each SESOI

Characteristics	*N* (%)
Author label for each minimum-effect test	
Superiority	27 (45%)
Inferiority	16 (27%)
CI interval approach	7 (12%)
Minimum-effect	4 (6.7%)
Other	4 (6.7%)
Not reported	2 (3.3%)
Direct or indirect minimum-effect test	
Direct	51 (85%)
Indirect	9 (15%)
Scale of the SESOI	
Unstandardized effect size	20 (33%)
Cohen's d	11 (18%)
Proportion	7 (11%)
Correlation	4 (6.7%)
Hedges' g	4 (6.7%)
Standardized regression coefficient	2 (3.3%)
Odds ratio	1 (1.7%)
Variance explained	1 (1.7%)
Not provided	10 (17%)
Timing of SESOI	
Before	44 (73%)
After	6 (10%)
Unknown	10 (17%)
Justification for the SESOI	
Previous research	15 (25%)
Previous research on SESOI	12 (20%)
Benchmark	6 (10%)
Subjective	5 (8.5%)
Multiple approaches	2 (3.4%)
Consensus panel/method	1 (1.7%)
Small telescopes approach	1 (1.7%)
Replication	1 (1.7%)
No justification	17 (29%)
Field of psychology	
Cognitive psychology	20 (33%)
Clinical psychology	16 (27%)
Health Psychology	9 (15%)
Social Psychology	8 (13%)
Other	7 (11.7%)

Note. We also examined the characteristics for each minimum-effect test conducted. However, the pattern of results only differed slightly and we refer the interested reader to OSF for analysis of all minimum-effect tests (osf.io/j78dz/).

minimum-effect test. We found that 50% of the articles conducted power analyses (30/60), 48% did not (29/60), and one study conducted a Bayes factor design analysis. Of the 30 power analyses that were conducted (including the Bayes factor design analysis), only 7 were specifically for a type of minimum-effect test (23%). When researchers conducted a power analysis, 17 times their collected sample size was greater than the sample size indicated by the power analysis (55%), 6 times it was smaller (19%), and 8 times it was exactly equal (26%).

Discussion

To improve upon the traditional NHST framework, interval hypothesis tests such as minimum-effect testing have been proposed to examine whether the observed effect size yields practical or theoretical meaningfulness. However, the added value of interval hypothesis tests relies on how a researcher determines the smallest effect size that yields practical or theoretical relevance. In the present literature review, we examined how researchers in the field of psychology conducted minimum-effect tests, including how they established and justified their SESOI. This seems to be a timely manner as our findings showed that minimum-effect tests are becoming more popular.

Type of Minimum-Effect Test

We first set out to examine which type of minimum-effect test researchers used (more specifically, what label each author used for their given minimum-effect test). We found that tests of superiority and inferiority were most

frequently reported. One explanation for this is that researchers who have determined a SESOI also have a clear idea on the direction of the effect. For instance, researchers who examine the effects of alcohol on memory are typically interested in the extent to which alcohol negatively impacts memory because the opposite direction would be counter-intuitive and against theoretical predictions. Hence, it is possible that researchers tend to think about which effect sizes are practically or theoretically meaningful when they already have a plethora of research indicating the expected direction of the effect.

Although most minimum-effect tests were conducted directly—explicitly stating that they used some form of minimum effect test—some were done indirectly, where researchers examined CIs to determine whether they surpassed a certain value (without specifically indicating that a minimum effect test was conducted). In these cases, researchers may have intended to assess the practical or theoretical relevance of their effects, but failed to explicitly report this intention. Put differently, they may have conducted a minimum-effect test without labeling it as such. To improve transparency, when a minimum-effect test is deemed appropriate researchers should explicitly state the form of minimum effect test used (e.g., superiority test).

Interesting to note is that there has been a noticeable increase in the use of minimum-effect testing in the past five years, particularly in the fields of cognitive and clinical psychology. One potential explanation for this increase can be the growing emphasis on the meaningful interpretation of effect sizes (e.g., Baguley, 2009; Funder & Ozer, 2019) and the importance of specifying the SESOI (e.g., Lakens et al., 2018). These developments may have encouraged researchers to adopt minimum-effect tests more frequently. Future research could examine whether this trend continues, particularly given the recent availability of practical guidance on implementing minimum-effect testing (Smiley et al., 2023) and conducting associated power analyses (Riesthuis, 2024; Riesthuis et al., 2025).

SESOI in Minimum-Effect Testing

In line with a recent review of equivalence testing practices (Marshall et al., 2024), we found that the many researchers relied on unstandardized effect sizes to set their SESOI. The use of unstandardized effect sizes (simple effect size; raw mean difference) is generally the recommended strategy to set the SESOI (Baguley, 2009; Riesthuis, 2024). That is, besides being more robust (i.e., does not depend on variance) and easier to calculate than standardized effect sizes, unstandardized effect sizes are also easier to interpret in terms of magnitude and practical and theoretical meaningfulness (Baguley, 2009; Greenland et al., 1986, 1991; Pek & Flora, 2018). Interestingly, previous literature reviews in the field of eyewitness memory (Riesthuis & Otgaar, 2024a) and social psychology (Farmus et al., 2023; Sun et al., 2010), wherein researchers mainly conducted NHSTs, indicated that most researchers relied on standardized effect sizes. Hence, it seems that researchers who conduct equivalence or minimum effect tests realize the necessity of interpreting unstandardized effect sizes to establish which effects are meaningful.

Although setting a SESOI using an unstandardized effect size is beneficial for interpreting its magnitude, it is crucial that proper justifications are provided. Our findings indicated that justifications were provided for 70% (43/60) of the utilized SESOIs. This practice of effect size justification seems to be slightly higher than the effect size interpretations of observed effect sizes in social psychology research (≈ 60%; Farmus et al., 2023; Sun et al., 2010) and especially higher than eyewitness memory research (≈ 10%; Riesthuis & Otgaar, 2024a, 2024b). Moreover, when examining the justifications, we found that some were in line with best practice recommendations (Anvari & Lakens, 2021; Camilleri et al., 2022; Riesthuis, 2024) such as anchor-based methods, consensus methods, or based on previous research that examined the SESOI. However, we also found that many relied on practices that are not recommended, such as effect sizes found in previous studies, benchmarks or subjective argumentation. This mix of justifications was also observed in the review of equivalence testing practices (Marshall et al., 2024) and highlights the need for researchers to meaningfully consider what might be an appropriate SESOI for any given relationship among variables. Clearly defining what constitutes a meaningful effect strengthens study design (e.g., power analysis) and helps prevent hypothesizing after results are known (Kerr, 1998).

To establish which magnitude of effect size is practically or theoretically meaningful, researchers must contextualize their effect sizes. By contextualizing the effect size, we mean that it should be interpreted in light of the area of research, taking into account the methods used (e.g., variables, stimuli, manipulation), the population of interest, and any potential ramifications (i.e., costs) and benefits of the results/conclusions (e.g., policy decisions; Farmus et al., 2023; Riesthuis et al., 2022). This means that it is necessary to go beyond benchmarks because they lack contextualization (e.g., Cohen, 1988; Hemphill, 2003). That is because in certain research areas a small effect according to Cohen's benchmarks can be an important and meaningful effect, or vice versa (Funder & Ozer, 2019). Using effects sizes from previous research or benchmarks derived from field-specific effect size distributions are also not recommended because they do not address which effect sizes matter practically or theoretically for a particular relation-

ship of interest (i.e., they ignore important contextual factors, Panzarella et al., 2021). Given the discussion above, an important question regards what specific process researchers should use to establish an appropriate SESOI (in order to be able to conduct an informative minimum-effect test, or any other interval hypothesis test).

There are several methods that have been proposed to help researchers determine an appropriate SESOI, of which some were reported in our literature review (for an overview see Lakens et al., 2018). For instance, researchers can use anchor-based methods to determine the SESOI by examining whether an effect is meaningful based on individuals' self-reported subjective improvement or decline (Anvari & Lakens, 2021). The anchor-based method is a viable approach but requires extensive work to initially establish an appropriate SESOI (e.g., pilot study). Another method is the cost-benefit analysis in which the SESOI is determined by scrutinizing whether the benefits of a manipulation outweigh its costs within the specific research context (for an example, see Otgaar et al., 2023). Additionally, researchers can also use the consensus-method approach wherein the SESOI is determined through a consensus among experts (Riesthuis et al., 2022). The benefit of the two latter methods is that they are more feasible. That is, researchers can individually contemplate the costs and benefits of a manipulation and also conduct a consensus study among colleagues (experts) within the given field of research (as done in one of the included articles). It is very important to highlight that the estimation of the SESOI does not have to be perfect; in fact, it could be argued that the SESOI will always be an approximation and what constitutes an appropriate SESOI in one study/context might differ from that within a different study/context. Most importantly is that researchers select a SESOI and provide reasonable justifications. The selection and justification of a SESOI within a given study/context is important because this can initiate scientific discussions regarding how to finetune the SESOI estimate within particular contexts (Riesthuis, 2024).

Just as with hypotheses, establishing the SESOI is preferably done before a study is conducted to avoid any biases in determining which effect sizes are deemed meaningful (Camilleri et al., 2022; Kerr, 1998). That is, if researchers determine which effect sizes are meaningful after the results are known, they might be biased towards choosing a smaller magnitude SESOI in order to provide evidence of not only a statistical but also a practically or theoretically relevant effect. However, to reiterate, interval hypothesis tests such as minimum-effect tests depend heavily on whether the SESOI is well justified and one component of a well justified SESOI is that it is set before the study is conducted. Although our findings highlighted that most studies set the SESOI before the study was conducted, a nontrivial number of studies set it afterwards or did not clearly indicate when it was established. Hence, we recommend that researchers determine the SESOI before a study is conducted to avoid any biases.

Power Analysis

To conduct meaningful minimum-effect tests, sufficient statistical power is necessary to reliably detect the effects of interest (Bakker et al., 2012). For minimum-effect tests, larger sample sizes are required compared to NHST, as greater statistical power is needed to detect specific effect sizes rather than merely identifying nonzero differences (Riesthuis, 2024; Smiley et al., 2023). Approximately half of the articles conducted power analyses, however many were for NHST and not for minimum-effect testing. This could mean that studies were underpowered to detect an effect larger in magnitude than the SESOI. We recommend that researchers conduct power analyses for the minimum-effect test to be able to reliably detect their practically or meaningful effects (see Riesthuis, 2024; Riesthuis et al., 2025, for a tutorial on how to conduct such power analyses for various designs).

Although not explicitly coded, many articles conducted only one power analysis per experiment, despite testing multiple hypotheses and analyses—a pattern also observed in other literature reviews (Riesthuis, 2025). Moreover, we found that singular SESOIs were repeatedly used for many analyses, and it was unclear whether the given SESOI was appropriate for each minimum-effect test conducted. Ideally, researchers set SESOIs and conduct power analyses separately for each hypothesis. In other words, SESOIs should be contextualized, which means that they are relevant for the specific hypothesis of interest. Hence, we recommend that researchers transparently identify the SESOI for each hypothesis (whether it is the same for multiple hypotheses or not) and conduct a power analysis for each.

Conclusion

Approximately 25 years after the introduction of minimum-effect tests to psychology researchers, they appear to be starting to be adopted more broadly. Minimum-effect tests can overcome important limitations of the traditional NHST framework (Amrhein et al., 2019; Cumming, 2014). However, the success of minimum-effect tests depends heavily on whether a SESOI is established and appropriately justified before a study is conducted. We found that although many researchers relied on unstandardized effect sizes to set their SESOI, appropriate justifications for the SESOI were frequently lacking or not provided at all. Moreover, a nontrivial number of studies set their SESOI after the study was conducted or it was unknown when it was set. Last, researchers oftentimes did not fully plan for

minimum-effect testing in terms of their power analysis.

To summarize, we believe that minimum-effect tests are an extremely valuable methodological tool that produce results that are usually more informative than traditional NHST results. In this paper, we provided several recommendations regarding the use of minimum-effect tests in order to maximize the information gained from the tests and we hope that the increase in the use of these tests continues in the future.

Authors' note

The current manuscript has been supported by a FWO post-doctoral fellowship grant awarded to the first author (1203824N).

References

Alter, U., & Counsell, A. (2023). Determining negligible associations in regression. *The Quantitative Methods for Psychology*, *19*(1), 59–83. doi: 10.20982/tqmp.19.1.p059.

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307. doi: 10.1038/d41586-019-00857-9.

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. doi: 10.1016/j.jesp.2021.104159.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. doi: 10.1348/000712608X377117.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi: 10.1177/1745691612459060.

Beaulieu-Prévost, D. (2006). Confidence intervals: From tests of statistical significance to confidence intervals, range hypotheses and substantial effects. *Tutorials in Quantitative Methods for Psychology*, *2*(1), 11–19. doi: 10.20982/tqmp.02.1.p011.

Beribisky, N., Davidson, H., & Cribbie, R. A. (2019). Exploring perceptions of meaningfulness in visual representations of bivariate relationships. *PeerJ*, *7*, e6853. doi: 10.7717/peerj.6853.

Camilleri, C., Beribisky, N., & Cribbie, R. (2022). *The minimally meaningful effect size: A vital component of preregistrations*. doi: 10.31234/osf.io/jbgtm.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, *60*(1), 1–10. doi: 10.1002/jclp.10217.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi: 10.1177/0956797613504966.

Farmus, L., Beribisky, N., Martinez Gutierrez, N., Alter, U., Panzarella, E., & Cribbie, R. A. (2023). Effect size reporting and interpretation in social personality research. *Current Psychology*, *42*(18), 15752–15762. doi: 10.1007/s12144-021-02621-7.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. doi: 10.1177/2515245919847202.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of socio-economics*, *33*(5), 587–606. doi: 10.1016/j.socec.2004.09.033.

Greenland, S., Maclure, M., Schlesselman, J. J., Poole, C., & Morgenstern, H. (1991). Standardized regression coefficients: A further critique and review of some alternatives. *Epidemiology*, *2*, 387–392. https://www.jstor.org/stable/20065707

Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, *123*, 203–208. doi: 10.1093/oxfordjournals.aje.a114229.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79. doi: 10.1037/0003-066X.58.1.78.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. doi: 10.1177/0956797611430953.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759. doi: 10.1177/0013164496056005002.

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648. doi: 10.1177/1745691620958012.

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.

Marshall, A. D., Occhipinti, S., & Loxton, N. J. (2024). Tipping the analytical scales, investigating the use of frequentist equivalence analyses in psychology: A scoping review. *Quality & Quantity*, *58*(3), 2929–2955. doi: 10.1007/s11135-023-01758-w.

Mastrich, Z., & Hernandez, I. (2021). Results everyone can understand: A review of common language effect size indicators to bridge the research-practice gap [https]. *Health Psychology*, *40*(10), 727–736. doi: 10 . 1037 / hea0001112.

McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: Defining what really matters to patients. *Journal of the American Medical Association*, *312*(13), 1342–1343. doi: 10.1001/jama.2014.13128.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115. doi: 10.1086/288135.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. doi: 10.1037/0022-006X.46.4.806.

Metzler, C. M. (1974). Bioavailability - a problem in equivalence. *Biometrics*, *30*(2), 309–317. doi: 10 . 2307 / 2529651.

Müller, K., & Wickham, H. (2023). *tibble: Simple data frames* [R package version 3.2.1]. https://CRAN.R-project.org/package=tibble

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*(2), 234. doi: 10.1037/0021-9010.84.2.234.

Otgaar, H., Riesthuis, P., Neal, T. M. S., Chin, J., Boskovic, I., & Rassin, E. (2023). If generalization is the grail, practical relevance is the nirvana: Considerations from the contribution of psychological science of memory to law. *Journal of Applied Research in Memory and Cognition*, *12*, 176–179. doi: 10.1037/mac0000116.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Moher, D., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. doi: 10 . 1136 / bmj.n71.

Panzarella, E., Beribisky, N., & Cribbie, R. A. (2021). Denouncing the use of field-specific effect size distributions to inform magnitude. *PeerJ*, *9*, e11383. doi: 10 . 7717/peerj.11383.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*, 208. doi: 10.1037/met0000126.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Riesthuis, P. (2024). Simulation-based power analyses for the smallest effect size of interest: A confidence-interval approach for minimum-effect and equivalence testing. *Advances in Methods and Practices in Psychological Science*, *7*(2), 1–14. doi: 10 . 1177 / 25152459241240722.

Riesthuis, P. (2025). Assessing power analysis practices in eyewitness memory research: A review and simulation study. *Journal of Applied Research in Memory and Cognition*. doi: 10.1037/mac0000237.

Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology*, *36*(1), 203–215. doi: 10.1002/acp.3911.

Riesthuis, P., & Otgaar, H. (2024a). On the use of receiver operating characteristic area under the curve in eyewitness memory research. *Legal and Criminological Psychology*, *00*, 1–19. doi: 10.1111/lcrp.12300.

Riesthuis, P., & Otgaar, H. (2024b). An overview of the replicability, generalizability and practical relevance of eyewitness testimony research in the journal of criminal psychology. *Journal of Criminal Psychology*, *15*(2), 176–194. doi: 10.1108/JCP-04-2024-0031.

Riesthuis, P., Otgaar, H., & Bücken, C. (2025). Ready to ROC? a tutorial on simulation-based power analyses for null hypothesis significance, minimum-effect, and equivalence testing for roc curve analyses. *Behavior Research Methods*, *57*, 120. doi: 10 . 3758 / s13428 - 025 - 02646-x.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. doi: 10.1037//0033-2909.113.3.553.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*(4), 403–411. doi: 10 . 1037 / 1082-989X.3.4.403.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83. doi: 10.1037//0003-066x.40.1.73.

Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A., & Larmarange, J. (2021). Reproducible summary tables with the gtsummary package. *The R Journal*, *13*, 570–80. doi: 10.32614/RJ-2021-053.

Smiley, A. H., Glazier, J. J., & Shoda, Y. (2023). Null regions: A unified conceptual framework for statistical inference. *Royal Society Open Science*, *10*(11), 221328. doi: 10.1098/rsos.221328.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*(4), 989–1004. doi: 10.1037/a0019507.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Yutani, H., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686.

Wickham, H., & Bryan, J. (2023). *readxl: Read excel files* [R package version 1.4.3]. https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* [R package version 1.1.3]. https://CRAN.R-project.org/package=dplyr

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on osf.io/j78dz/

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/j78dz/

Citation

Riesthuis, P., & Cribbie, R. A. (2025). When are scientific findings deemed practically or theoretically relevant? a literature review on minimum-effect testing. *The Quantitative Methods for Psychology*, *21*(2), 82–94. doi: 10.20982/tqmp.21.2.p082.

Received: 31/03/2025 ~ Accepted: 19/06/2025