City Research Online

City St George's, University of London

Citation: Riesthuis, P., Howe, M. L. & Otgaar, H. (2025). Meaningful approaches to assessing the size of effects in memory research: applications and recommendations for study design, interpretation, and analysis. Memory, 33(5), pp. 485-494. doi: 10.1080/09658211.2025.2492496

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/35071/

Link to published version: https://doi.org/10.1080/09658211.2025.2492496

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Meaningful approaches to assessing the size of effects in memory research: Applications and recommendations for study design, interpretation, and analysis

Paul Riesthuis[1,2], Mark L. Howe[3], & Henry Otgaar[1,2]

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

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

[3] Department of Psychology and Neuroscience, City St. George's, University of London, UK

Correspondence should be addressed to Paul Riesthuis: paul.riesthuis@kuleuven.be. The current manuscript has been supported by a FWO post-doctoral fellowship granted to the first author (1203824N) and a FWO research project grant awarded to the third author (G0D3621N).

Meaningful approaches to assessing the size of effects in memory research: Applications and recommendations for study design, interpretation, and analysis

In an earlier editorial in this journal (Howe & Otgaar, 2020), we highlighted the importance of publishing results that, although not statistically significant by traditional null-hypothesis significance testing (NHST) techniques, were practically or theoretically important. There are a number of well-known problems with the use of traditional NHST procedures (e.g., see Amrhein et al., 2019), and in this editorial we touch on some of these issues with respect to memory research. More importantly, we extend our discussion of alternatives to NHST techniques when determining whether findings are practically or theoretically important, namely, through an examination of effect sizes.

Effect sizes, which quantify the magnitude of a phenomenon (Cohen, 1988), are important in psychology and thus also memory research. Unlike statistical significance ($p < .05$), which can lead to arbitrary dichotomous decisions, effect sizes provide a more informative measure of practical and theoretical relevance (e.g., Funder & Ozer, 2019). Specifically, the traditional NHST framework has been criticized because of its focus on finding non-zero differences but not addressing whether the difference is meaningful on a practical or theoretical level (e.g., Amrhein et al., 2019; Cumming, 2014). This issue is particularly relevant in research areas that can influence policy decisions and can have real-world implications such as the field of eyewitness memory (e.g., Wells et al., 2020). Additionally, a fundamental issue with the NHST framework is that the null hypothesis (i.e., effect size is exactly 0) is practically always false (Cohen, 1990), and a $p > .05$ (i.e., statistical nonsignificance) does not mean absence of an effect, despite this common misinterpretation (Murphy et al., 2025).

To illustrate these issues of the NHST framework, imagine a group of researchers examining whether a new drug will improve memory. To do so, participants are either given

the new drug (experimental group) or a placebo (control group) and then study a list of words. After a certain delay, memory performance of both groups is assessed. Following the NHST framework, when researchers conduct their statistical analyses, two conclusions can emerge: 1) $p < .05$ suggesting that the observed difference of this magnitude or greater between the two sample means is unlikely if there truly was no difference in the population, or 2) $p > .05$ indicating that there is insufficient empirical evidence to conclude that the observed difference between the two sample means would be unlikely under the null hypothesis (Lakens, 2021). Effect sizes can be useful to examine whether effects are truly meaningful and not simply different from zero. Moreover, effect sizes are useful to find support for the null-hypothesis (i.e., effect is too small to matter). In this editorial, we highlight the importance of interpreting effect sizes meaningfully in memory research, preferably before a study is conducted. Moreover, we show how effect sizes can be used in power and data analyses for studies in the field of memory.

Effect sizes

There are different ways to quantify the magnitude of a phenomenon (i.e., effect size). For instance, in the hypothetical experiment examining the effects of a new drug on memory, the effect size could be expressed in the raw mean difference in words retrieved. Specifically, let's say that participants in the experimental group remembered, on average, 10 words while participants in the control group remembered, on average, 9 words. In this case, the effect size would be the remembrance of 1 word. Put differently, participants in the experimental group remembered, on average, 1 more word than participants in the control group. This would be an example of an unstandardized effect size (i.e., simple or raw effect size) in simple between or within subject designs which is expressed in the units of the original measurement (i.e., words remembered).

However, sometimes the outcome variable in memory research is binary (e.g., false memory formed or not) or count (e.g., number of retrieval attempts). In these cases, nonlinear models such as logistic (for binary outcomes) and Poisson regression (for count outcomes) are more appropriate. For logistic and Poisson regression, the average marginal effect (AME) serves as an interpretable unstandardized effect size. Specifically, AMEs provide the average change in the outcome (binary, count) for a one-unit change in the predictor (e.g., control to the experimental group; Norton et al., 2019). For logistic regression, the AME represents the average change in probability to observe a success between the control and experimental group while in Poisson regression, the AME represents the average change in the expected counts between the experimental and control group. Taken together, an unstandardized effect size is simply the effect size expressed in the original units of measurement used in the study.

Alternatively, effect sizes can be expressed in standardized units (i.e., standardized effect size), typically in terms of standard deviations. For example, assuming that the standard deviations in the hypothetical experiment for both the experimental and control group are two, a possible standardized effect size could be Cohen's *d*

$$= \frac{M_1 - M_2}{SD_{pooled}} = \frac{10 - 9}{2} = .5$$

wherein M_1 is the mean of the experimental group, M_2 is the mean of the control group, and the SD_{pooled} is calculated as follows:

$$= \sqrt{\frac{SD_1^2 + SD_2^2}{2}} = \sqrt{\frac{2^2 + 2^2}{2}} = 2$$

Cohen's *d* is an example of a standardized effect size for comparing mean differences. Many different types of standardized effect sizes exist for various types of analyses (e.g., [in-]dependent sample t-test, ANOVA, regression, etc.; see Jané et al., 2024 for a complete overview of standardized effect sizes). Common effect sizes for comparing mean differences include Cohen's *d*, Cohen's d_z (for within-subject [paired] designs), and Hedges' *g*. When

measuring associations between variables, Pearson's r, Spearman's ρ, and Kendall's τ are some examples of standardized effect sizes. Finally, for ANOVA or regression analyses, standardized effect sizes such as R^2, ω^2, η^2, and η_p^2 are frequently used to quantify variance explained.

Interestingly, memory researchers frequently rely on standardized effect sizes when planning studies and analyzing data (Morris & Fritz, 2013; Riesthuis, 2025; Riesthuis & Otgaar, 2024a; 2024b). However, their unit-free nature (e.g., standard deviations) makes them difficult to interpret (Brooks et al., 2014). This challenge is reflected in the memory literature, where standardized effect sizes are frequently reported but rarely interpreted in terms of practical (or even theoretical) meaning (Riesthuis, 2025; Riesthuis & Otgaar, 2024a; 2024b). When researchers did interpret their effect sizes, they mostly relied on Cohen's benchmarks for small, medium, and large effects (Cohen, 1988; p. 24). However, the use of Cohen's benchmarks has been criticized (Correl et al., 2020; Funder & Ozer, 2019; Pek & Flora, 2018). Even Cohen himself argued that researchers should not use these benchmarks because they lack contextualization (Cohen, 1988; p.25, pp 532-535). In other words, whether an effect size is small or large depends on the context of research, the methods used (e.g., variables, stimuli, manipulation), the population of interest, and any potential associated costs and benefits of the outcome (Funder & Ozer, 2019; Riesthuis & Cribbie, 2025).

Effect sizes in the memory literature have sometimes been interpreted based on prior research or field-specific effect size benchmarks (i.e., benchmarks based on the distribution of published literature; Panzarella et al., 2021) to consider the context of the study (Morris & Fritz, 2013; Riesthuis, 2025; Riesthuis & Otgaar, 2024a; 2024b). However, the fact that a previous study detected a certain effect size does not necessarily imply that it is practically or theoretically meaningful. Similarly, establishing benchmarks based on published effect sizes does not guarantee that an effect is genuinely meaningful (Panzarella et al., 2021).

Furthermore, publication bias (i.e., the tendency to publish statistically significant findings over non-significant ones; Fanelli, 2012) can lead to either over- or underestimation of effect size benchmarks. For instance, if there is a bias toward publishing significant findings, with larger effect sizes reported due to small sample sizes, the effect size is overestimated. Hence, researchers ought to contextualize their effect sizes to understand its practical or theoretical relevance.

To better assess whether an effect size is practically or theoretically meaningful (see section "Smallest effect size of interest" below on approaches on how to determine which effects are meaningful), it is recommended to focus on unstandardized effect sizes (e.g., Baguley, 2009; Funder & Ozer, 2019; Riesthuis, 2024). Because unstandardized effect sizes are expressed in the original measurement units, they provide a clear sense of the effect's magnitude and significance. For instance, stating that the experimental group remembered, on average, one more word than the control group is more intuitive than saying that they remembered 0.5 standard deviations more. To maximize the interpretability of effect sizes, researchers should ideally also report descriptive statistics (e.g., means, standard deviations). This context helps clarify what an effect actually means. For example, a raw mean difference of 1 on a 1–6 stress scale (*1 = not stressed* and *6 = very stressed*) may be more meaningful when the difference in group means goes from 3 to 4 than from 1 to 2, as the former reflects a change in moderate stress, whereas the latter may still represent low stress overall.

Additionally, unstandardized effect sizes are easier to calculate and more robust, as they do not depend on variance (Baguley, 2009). For example, two experiments might yield the same unstandardized effect size (e.g., raw mean differences of 1), but the standardized effect sizes could differ significantly due to differences in variance. This illustrates that standardized effect sizes might be over or underestimated depending on the variance (e.g., poor measurement increases variance) which may not reflect the true magnitude or practical

significance of the effect. In some circumstances, such as when comparing conceptually similar effects (e.g., comparing the misinformation effect and spontaneous false memories) or when the metric of the measurement has no meaning (e.g., Likert scale), it can be beneficial to also examine the standardized effect size (but see Greenland et al., 1991; 1986). Hence, we recommend that researchers report both the unstandardized and standardized effect sizes to help them interpret whether an effect is meaningful or not.

Additionally, researchers can use common language effect sizes (CLES; Mastrich & Hernandez, 2021) to help interpret effect sizes. CLES are intuitive effect sizes that transform the standardized effect size into a comprehensible probability or number. Specifically, research has demonstrated that the use of CLES and graphical representations facilitated researchers and laypeople to meaningfully interpret effect sizes compared to standardized effect sizes (Brooks et al., 2014; Krasikova et al., 2018). An example of a CLES for the hypothetical experiment above is the probability of superiority (equivalent to the area under the curve[1]). The probability of superiority refers to the probability that a randomly chosen participant from the experimental group has a higher value (e.g., more words remembered) than a randomly chosen participant from the control group. For instance, for the hypothetical experiment of the new drug on memory, the probability of superiority can be calculated as follows:

$$= \phi\left(\frac{Cohen's\ d}{\sqrt{2}}\right) = = \phi\left(\frac{.5}{\sqrt{2}}\right) = .64$$

in which ϕ is the cumulative distribution function. In this case, the probability that a randomly chosen participant in the experimental group remembered more words compared to a randomly chosen participant in the control group is .64.

[1] The area under the curve can also be interpreted as the average sensitivity across all possible values of 1 − specificity. Another interpretation is the proportion correct in a two-alternative forced choice (2AFC) task

For correlational analyses, the Dunlap CLES (Dunlap, 1994) can be calculated which indicates the probability that in a pair of randomly selected people, the participant with the higher score on one variable will also have a higher score on the other variable. The Dunlap CLES is calculated as follows:

$$\text{Dunlap CLES} = \frac{\arcsin(r)}{\pi} + .5$$

in which arcsin refers to the arcsine transformation and *r* is the correlation coefficient. If in a study a correlation coefficient of .6 is observed then this would translate to a Dunlap CLES of .78. In other words, there is a 78% chance that in a pair of randomly selected people, the participant with the higher score on one variable will also have a higher score on the other variable.

Another type of CLES is the Number Needed to Treat (NNT) which is especially relevant when researchers use logistic regressions. In logistic regression, the effect size odds ratio is commonly used to compare the odds of an event occurring in the experimental group versus the control group (Szumilas, 2010). However, odds ratios are frequently misinterpreted, which is why researchers are encouraged to use the NNT instead (e.g., Norton et al., 2018; another option could be to use probit regression; e.g., DeCarlo, 1998). The NNT indicates how many participants need to be treated (e.g., giving participants a memory-enhancing drug) before one additional success outcome (e.g., a beneficial memory effect) is observed compared to the control group. The NNT for a simple two group experimental study is calculated as follows:

$$\text{NNT} = \frac{1}{P_1 - P_2}$$

in which P_1 is the probability of observing a success in the experimental group and P_2 is the probability of observing a success in the control group ($P_1 - P_2$ = absolute risk reduction). If the probability of observing a memory enhancing effect in the experimental group is .7 and the probability of observing a memory enhancing effect in the control group is .5, then the

NNT is 5. A NNT of 5 means that 5 participants in the experimental group need to receive the drug to observe one additional memory enhancing effect compared to the control group (for an overview of all the possible CLES; see Mastrich & Hernandez, 2021).

Finally, the magnitude of an effect can be expressed through the persons as effect size approach (Grice et al., 2020). That is, instead of focusing on group differences, persons as effect sizes refers to how many participants behave in accordance with the anticipated effect. For instance, Ying and colleagues (2023) applied this approach to assess how many laypersons perceived suspect identifications from biased lineups as more accurate than those from unbiased lineups. Specifically, they found that, in Experiment 1, 74% of participants rated identifications from biased lineups as more accurate. Hence, the person as effect size approach indicates how many participants actually show the proposed effect (for a complete overview see Grice et al., 2020). Taken together, (un-)standardized effect sizes, CLES, and persons as effect sizes are different ways to interpret the magnitude of an effect.

To improve effect size interpretation, we recommend researchers to take a multiverse approach and use different types of effect sizes to interpret when an effect size is deemed practically or theoretically meaningful. To facilitate the interpretation of effect sizes, we created a Shiny app (https://paulriesthuis.shinyapps.io/interpreteffect/) which can also be accessed locally via the R package "*interpreteffect*" (https://github.com/PaulRiesthuis/interpreteffect). In the Shiny app, we provide an application wherein researchers can indicate the parameters of their study (e.g., means, standard deviations, sample sizes, etc.) in frequently used experimental designs (i.e., [in-]dependent sample t-tests, correlation, logistic regression, Poisson regression; see Figure 1). Based on the provided parameters, the (un-)standardized effect sizes and CLES are calculated along with graphs indicating how the distribution (e.g., density plot, dotplot, scatterplot, histogram) may look. For within-subject designs, we also provide the persons as effect size for which

researchers can indicate which range of values are deemed equivalent across timepoints (default is set to 0). The application can be used to determine which effect sizes researchers deem practically or theoretically meaningful.

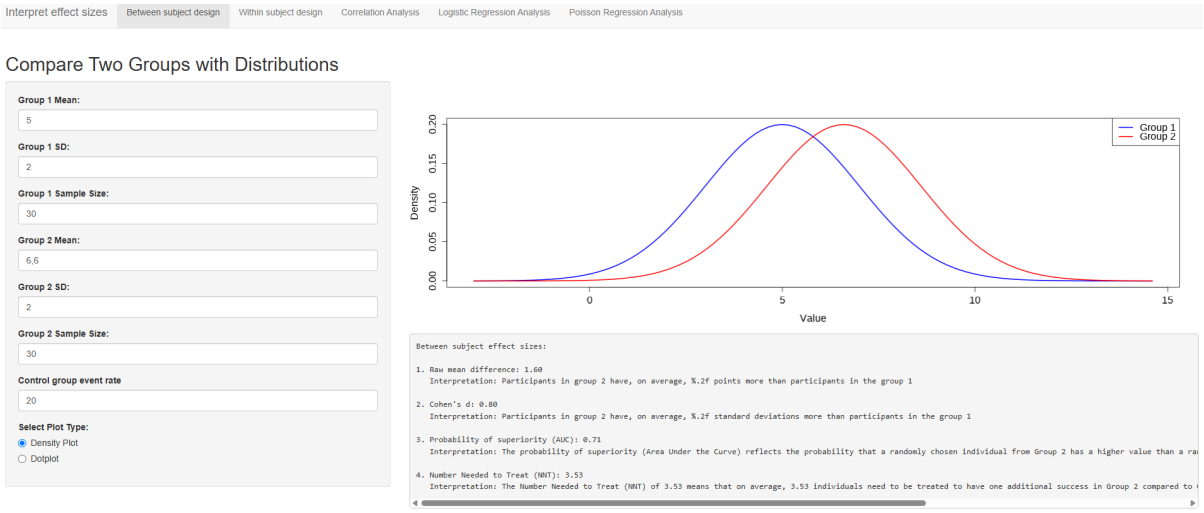

Figure 1. A screenshot of the Shiny app.

Smallest effect size of interest

As highlighted above, interpreting effect sizes meaningfully is important to understand the practical and theoretical relevance of the findings. In other words, interpreting observed effect sizes is crucial for understanding the meaning of a study and whether it was a successful or not (Camilleri et al., 2021). However, ideally, researchers determine a priori what the smallest effect size is that yields practical or theoretical significance, also known as the smallest effect size of interest (SESOI; Lakens et al., 2018), minimally meaningful effect size (MMES; Beribisky et al., 2019), or minimally clinically important difference (MCID; McGlothlin & Lewis, 2014)[2]. Estimating a SESOI a priori prevents researchers from making post hoc justifications about why their observed effect size is (or is not) practically or theoretically relevant (Riesthuis, 2024). Just as when interpreting effect sizes, setting the SESOI should take the context of the study into consideration. We will briefly highlight

[2] We used the term SESOI throughout the manuscript.

several approaches to determine the SESOI (for a complete overview see Anvari & Lakens, 2021).

One method to estimate the SESOI is through the just-noticeable difference approach (Lakens et al., 2018). In the just-noticeable difference approach, the SESOI is estimated based on what the smallest effect size is that can be detected by an observer (e.g., participant). One example is the study by Burriss and colleagues (2015) who examined whether men are more attracted to women during the fertile phase of their ovulatory cycle due to an increase in redness of their facial skin. Interestingly, women's facial skin turned slightly more red, but men did not perceive this change and thus the effect was too small to be of practical significance. A memory-related example could be investigating whether subtle differences in background noise affects recall. Suppose researchers test whether reducing background noise from 40 dB to 38 dB improves memory performance. If participants do not perceive this small reduction in noise, any observed improvement in memory may be too small to be practically meaningful.

Another method to estimate the SESOI is the anchor-based method (Anvari & Lakens, 2021), which relies on participants' self-reports to assess what constitutes a meaningful change. In this method, a subjective assessment—such as a feeling of physical improvement or decline—serves as an anchor to establish a meaningful difference on a related measurement. For instance, Lemay and colleagues (2019) estimated the SESOI for the Hospital Anxiety and Depression Scale in a cardiovascular rehabilitation study by asking participants to assess their overall health compared to a year ago. Then, they calculated the mean difference on the Hospital Anxiety and Depression Scale for participants who reported feeling "somewhat better" and participants who reported feeling "about the same" which was used as the SESOI. A similar approach can be applied to memory research. For example, if researchers want to examine whether a new memory training program improves word recall,

they can estimate the SESOI by calculating the mean difference in the number of words remembered between participants who reported that their memory was “about the same” and those who reported that it had “improved” in response to the question, “Compared to before the training, how would you rate your memory?”.

Alternatively, researchers can also conduct consensus-based studies to estimate the SESOI (Mouelhi et al., 2020; van der Heijde et al., 2001; Wells et al., 2001). In consensus method studies, stakeholders (e.g., researchers, patients, legal professionals) with relevant expertise for the study are asked for their opinions on what the SESOI should be. Hence, this approach relies on collecting input from a group of experts or relevant parties to estimate what effect size would be considered practically or theoretically meaningful within the context of the study. For instance, Riesthuis and colleagues (2022) provided memory scientists with several hypothetical experimental designs for false memory studies and asked them what they would consider the SESOI to be. There are several different types of consensus studies such as the Delphi method or the nominal group technique to create a consensus (e.g., face-to-face discussions, anonymous response, chance to revise answers, etc.; for an overview see Black et al., 1999). While conducting large-scale consensus studies may not be feasible for researchers, smaller-scale consensus studies within research groups or departments might be possible which can provide a reasonable estimate of the SESOI.

Finally, one approach that researchers can implement is the cost-benefit analysis (Lakens et al., 2018; Riesthuis, 2024). In a cost-benefit analysis, researchers can highlight what the associated costs (e.g., expensive training, time extensive) and benefits (e.g., improved memory; easier to implement) are of their study outcome. One example of a cost-benefit analysis for the SESOI in eyewitness memory research was provided by Otgaar and colleagues (2023). They estimated that a possible SESOI could already be a mean difference of 1 detail remembered because the misremembering or forgetting of 1 detail can be and has

been problematic in (previous) legal cases. Another example could be if researchers want to examine whether a simple instruction (e.g., only report what you remember) improves memory accuracy. In this scenario, any increase in memory accuracy might be deemed practically relevant because there are no costs and only benefits. However, if instead of a simple instruction, a costly and time extensive training is used, then a larger SESOI might be warranted. In the cost-benefit analysis, it becomes clear that estimates of the SESOI depend largely on the context of the study which should be taken into account in whichever approach is taken. We encourage researchers to conduct these cost-benefit analyses when estimating the SESOI, also when they used any of the previously mentioned methods (e.g., just-noticeable difference, anchor-based, consensus). That is because a cost-benefit analysis can provide the required contextualization to transparently estimate a reasonable SESOI. Finally, determining an estimate of a SESOI does not have to be perfect, as long as researchers provide appropriate justifications (Riesthuis, 2024; Riesthuis & Cribbie, 2025). Indeed, one could argue that the SESOI will always be an approximation, and what is considered an appropriate SESOI in one study or context may differ from that in another. When researchers determine a priori what their SESOI is and provide justifications, scientific discussions can be held to finetune this estimate. Furthermore, establishing the SESOI a priori will allow researchers to conduct more informative statistical analyses and more accurately plan their studies, something we discuss next.

SESOI and statistical analyses

When a SESOI is estimated and justified, (memory) researchers can analyze data beyond the traditional NHST framework and conduct more informative statistical tests. Specifically, as alluded to before, in the NHST framework it is not possible to find evidence for the null hypothesis and the focus is on finding non-zero differences. Using the SESOI, it is possible to conduct equivalence tests where it can be determined whether an effect is too

small to care about (Cribbie et al, 2004; Westlake, 1972) or minimum-effect tests where it can be determined whether an effect is not just statistically significant but whether it is also of practical or theoretical relevance (Murphy & Myors, 1999; Riesthuis & Cribbie, 2025). Moreover, the SESOI can be used to determine whether a replication attempt was successful or not (Riesthuis et al., 2024). However, as will become clear, the insights gained from these analyses depend on the SESOI and whether it is properly estimated and justified. We will highlight each statistical analysis and show the role of the SESOI.

Equivalence testing. In some studies, researchers might want to examine whether an effect is too small to matter (Cribbie et al., 2004). For instance, imagine researchers that want to examine whether a new investigative interviewing technique which is easier to implement than the standard investigative interviewing technique, performs equivalently well in terms of memory performance. Or when researchers want to examine whether the high-confidence suspect-identification accuracy is not affected by variations in memory strength (e.g., Smith et al., 2021). Despite the many equivalent types of hypotheses, many researchers mistakenly interpret a $p > .05$ as evidence of no effect (Murphy et al., 2024). However, under the traditional NHST framework, it is not possible to prove this null hypothesis (Cohen, 1990). To determine whether an effect size is too small to matter, researchers need to conduct equivalence tests.

There are several equivalence tests such as the two-one sided test procedure (Schuirmann, 1997) or the Anderson and Hauck (1983) procedure, both of which are readily available in the *negligible* (Cribbie et al., 2024) or *TOSTER* (Caldwell, 2022) packages in *R* (R Core Team, 2023), or in JASP (2024). A straightforward way to understand equivalence testing is through the confidence interval approach (CI; Riesthuis, 2024; Riesthuis et al., 2025; Smiley et al., 2023). That is, CIs can be used to conduct NHST and equivalence tests. Specifically, in the traditional NHST framework, when the 95% CI of the effect size does not

include 0, it can be considered statistically significant ($p < .05$; given an $\alpha = .05$). For equivalence tests, the 90% CI is used because it corresponds to two one-sided tests at $\alpha = .05$ (Alter & Counsell 2023; Seaman & Serlin, 1998) and determines whether the entire 90% CI falls completely within the equivalence bounds set by the SESOI. For example, imagine that the researchers examining whether a new interviewing technique performs equivalently well compared to the standard interviewing technique, set their SESOI to 1 mean difference in terms of details remembered. In this case, the lower equivalence bound is -1 (SESOI in the negative direction) and the upper equivalence bound is set to 1 (SESOI in positive direction; see Figure 2). When conducting equivalence tests, there are four possible scenarios: 1) the 95% CI does include 0 (statistically nonsignificant) and the 90% CI falls completely within the equivalence bounds (statistically equivalent), 2) the 95% does not include 0 (statistically significant) and the 90%CI falls completely within the equivalence bounds (statistically equivalent), 3) the 95% CI does not include 0 (statistically significant) but the 90%CI includes the SESOI in either direction (statistically nonequivalent), and 4) the 95% CI does include 0 (statistically nonsignificant) and the 90% CI includes the SESOI in either direction (statistically nonequivalent; see Figure 2).

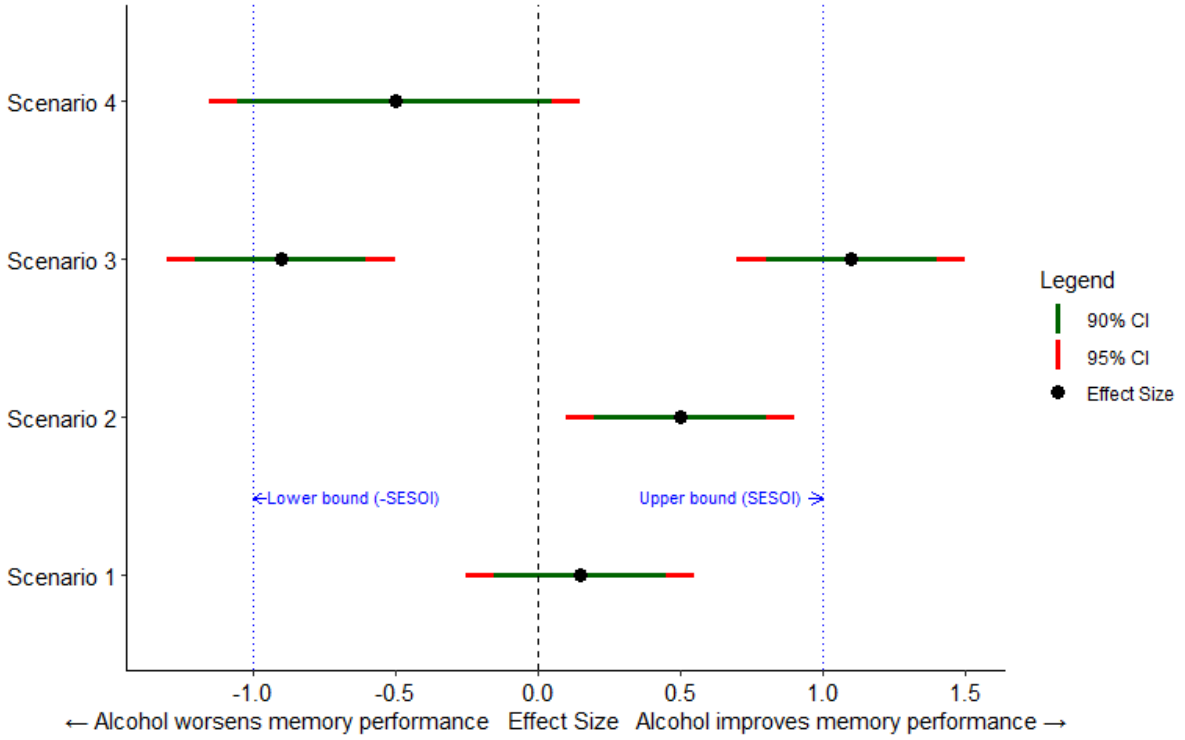

Figure 2. Possible outcome scenarios in equivalence testing.

In Scenario 1, there is clear evidence that the effect is equivalent because the effect is statistically nonsignificant and statistically equivalent. In Scenario 2, the effect would be deemed too small to care about (i.e., equivalent) even though it is statistically significant. In Scenario 3, the effect (in either direction) is statistically significant and statistically nonequivalent. This does not mean that the effect is also practically relevant (see the next section on minimum-effect testing). Last, in Scenario 4, the results are simply inconclusive. Specifically, the effect is not statistically significant but also not equivalent and more data is necessary to make any conclusions. Hence, the CI approach is an easy method to determine whether an effect is statistically significant and equivalent.

Minimum-effect testing. In minimum-effect testing, the idea is to not just examine whether an effect is significantly different from 0 but whether it is practically or theoretically meaningful (Murphy & Myors, 1999; Riesthuis & Cribbie, 2025). In other words, a minimum-effect test is an extension of the NHST wherein the nil is replaced by the SESOI. Although minimum-effect tests are available in the *TOSTER* package (Caldwell, 2022;

Lakens, 2017), the idea behind minimum-effect testing is easily explained using the CI approach. Specifically, in minimum-effect testing it is examined whether the 95% CI is greater in magnitude (in either direction) than the SESOI. To illustrate the idea behind minimum-effect tests, imagine that researchers want to examine whether a new, but expensive, drug can improve memory. Because the drug is expensive, the researchers want to make certain that the effect it has on memory is not just statistically significant but also practically meaningful. Again, they use a SESOI of 1 detail remembered and conduct minimum-effect tests. For minimum-effect tests, there are three possible scenarios: 1) the 95%CI does not include 0 (statistically significant) but is smaller in magnitude than the SESOI, 2) the 95%CI does not include 0 (statistically significant) but also includes the SESOI in either direction (inconclusive), or 3) the 95%CI does not include 0 (statistically significant) and is also greater in magnitude than the SESOI (practically significant; see Figure 3).

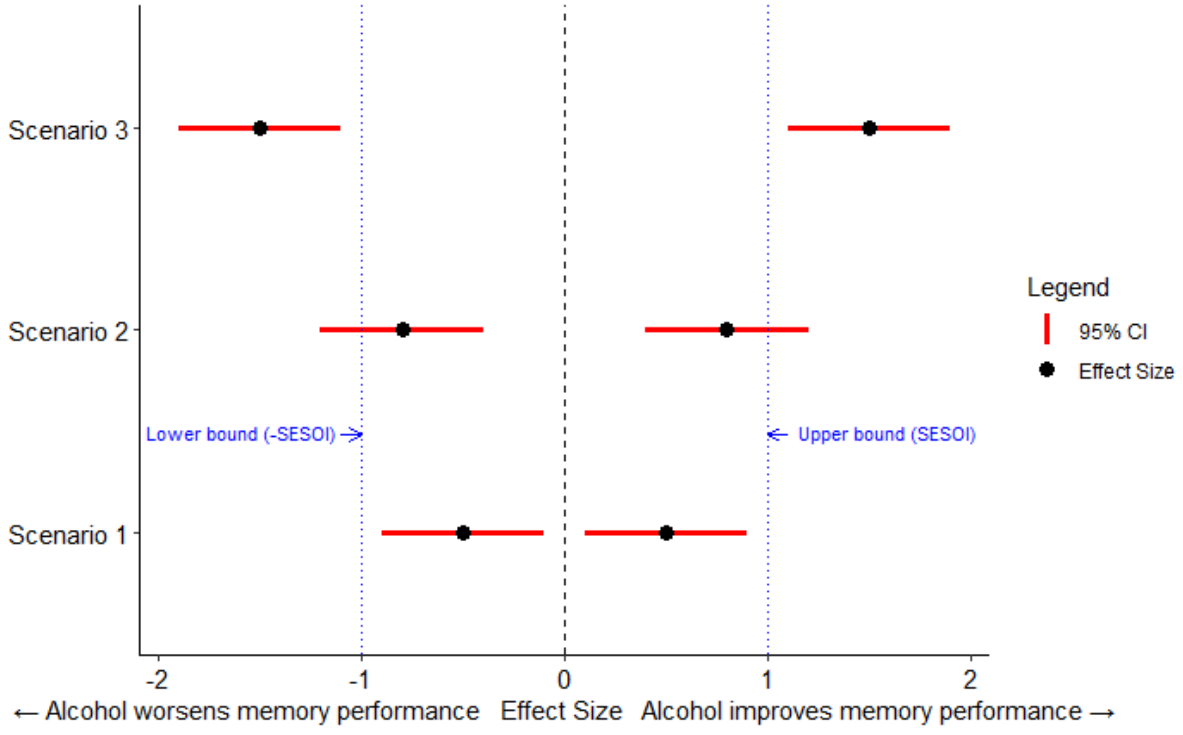

Figure 3. Possible outcome scenarios in minimum-effect testing.

In Scenario 1, this means that the effect is statistically significant but not practically or theoretically relevant (i.e., equivalent). In Scenario 2, the result is inconclusive because even

though the effect is statistically significant (95% CI does not include 0), the 95% CI does include the SESOI which means that it is uncertain whether the effect is practically or theoretically relevant. In Scenario 3, there is clear evidence that the effect is not only statistically relevant but also practically or theoretically meaningful.

If the SESOI is not properly estimated and justified, the information gained from the equivalence and minimum-effect tests are questionable. However, current practices in equivalence and minimum-effect testing have shown that although many researchers followed best practices in using unstandardized effect sizes to indicate their SESOI (Baguley, 2009; Riesthuis, 2024), justifications were often lacking or suboptimal (Marshall et al., 2024; Riesthuis & Cribbie, 2025). Hence, we urge researchers to start thinking about contextualized SESOIs, estimate them via the earlier proposed methods, and conduct cost-benefit analyses for their studies to maximize information gained by equivalence and minimum-effect testing.

Power analyses

To further increase the information gained from equivalence and minimum-effect testing, power analyses should accurately reflect the hypothesis of interest and be conducted using the SESOI. However, previous research has demonstrated that power analyses in memory research are currently conducted sub optimally (Riesthuis & Otgaar, 2024a; 2024b; Riesthuis, 2025). Specifically, most power analyses do not use meaningful or justified effect sizes and most power analyses are not reproducible. Moreover, researchers typically only conduct one power analysis even when they have multiple hypotheses. Ideally, researchers set a SESOI for each hypothesis of interest and conduct power analyses specifically for these different hypotheses to accurately determine the necessary sample size for each hypothesis. Further, it is recommended that researchers also conduct sensitivity analyses to examine how statistical power is affected when key parameters (e.g., standard deviation, correlation between measures, etc.) are different. Additionally, if researchers have hypotheses in line with

equivalence or minimum-effect testing, this should be reflected in the power analysis. This is especially relevant because larger sample sizes are required to conduct equivalence or minimum-effect tests compared with NHST. We refer the reader to several tutorial papers on how to conduct these power analyses for a wide range of study designs (i.e., [in-]dependent sample t-tests, correlations, multilevel models, Receiver Operating Characteristic curves; Caldwell, 2022; Lakens, 2017; Riesthuis, 2024; Riesthuis et al., 2025).

Conclusion & Recommendations

In this editorial, we introduced effect sizes and highlighted the importance of properly interpreting them in the field of memory. Specifically, we encourage memory researchers to start thinking about what the smallest effect size is that yields practical or theoretical relevance for their studies. To do so, we have provided several methods on how such a SESOI can be estimated. To help interpret effect sizes and estimating a SESOI, we created a Shiny app (https://paulriesthuis.shinyapps.io/interpreteffect/) and *R* package "*interpreteffect"* (https://github.com/PaulRiesthuis/interpreteffect) which provides the (un-)standardized effect sizes, CLES, and persons as effect sizes for commonly used experimental designs. We also showed how the SESOI can be used to overcome limitations of the traditional NHST framework through equivalence and minimum-effect testing and provided information on how to improve power analyses. Finally, we offer general recommendations for researchers to consider when using effect sizes in their studies. Although our perspective is that these recommendations can make memory research more informative, deviations from them may be entirely appropriate. First, researchers should report their unstandardized and standardized effect sizes along with descriptive statistics. Second, effect sizes should be interpreted meaningfully by contextualizing them. That is, we argue that instead of relying on Cohen's or field-specific effect size distributions to infer practical or theoretical meaning, we recommend a multiverse approach (Steegen et al., 2016) wherein the (un-)standardized, CLES, and

persons as effect sizes are assessed (e.g., https://paulriesthuis.shinyapps.io/interpreteffect/). However, a specific focus on unstandardized effect sizes is encouraged. Third, another recommendation is to determine the SESOI a priori and provide justifications for the SESOI (e.g., anchor-based, consensus, etc.; see above). We especially draw attention to conducting cost-benefit analyses to contextualize effect sizes as they can provide an informative answer on why a chosen effect size might be practically or theoretically relevant. Fourth, we encourage researchers to conduct power analyses for each hypothesis of interest using contextualized and justified SESOIs. Moreover, the power analysis should reflect the eventually conducted test (i.e., equivalence, minimum-effect, NHST) and analysis (e.g., independent sample t-test, ANOVA, multilevel model). Relatedly, it is recommended to conduct sensitivity analyses to examine how statistical power is affected when key parameters change. Finally, we encourage researchers to share, for instance, their R code or all parameters of G*Power (Faul et al., 2009) and its output to ascertain that the power analyses are reproducible. Taken together, the appropriate use of effect sizes will improve the statistical and practical inferences of memory research.

References

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

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*(7748), 305–307.https://doi.org/10.1038/d41586-019-00857-9

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics-Theory and Methods*, *12*(23), 2663-2692. https://doi.org/10.1080/03610928308828634

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

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

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

Black, N., Murphy, M., Lamping, D., McKee, M., Sanderson, C., Askham, J., & Marteau, T. (1999). Consensus development methods: a review of best practice in creating clinical guidelines. *Journal of Health Services Research & Policy*, *4*(4), 236-248. https://doi.org/10.1177/135581969900400410

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology, 99*(2), 332–340. https://doi.org/10.1037/a0034745

Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J., Stevens, M., Quigley, R., ... & Rowland, H. M. (2015). Changes in women's facial skin color over the ovulatory cycle are not detectable by the human visual system. *PLoS One*, *10*(7), e0130093. https://doi.org/10.1371/journal.pone.0130093

Caldwell, A.R. (2022). Exploring equivalence testing with the updated TOSTER R package. *PsyArXiv*. https://doi.org/10.31234/osf.io/ty8de

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small','medium', and 'large' for power analysis. *Trends in Cognitive Sciences*, *24*, 200-207. https://doi.org/10.1016/j.tics.2019.12.009

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. https://doi.org/10.1002/jclp.10217

Cribbie, R., Alter, U., Beribisky, N., Chalmers, P., Counsell, A., Farmus, L., & Martinez Gutierrez, N. (2024). *negligible: A Collection of Functions for Negligible Effect/Equivalence Testing*. R package version 0.1.9. https://CRAN.R-project.org/package=negligible

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

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin, 116*(3), 509–511. https://doi.org/10.1037/0033-2909.116.3.509

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904. https://doi.org/10.1007/s11192-011-0494-7

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

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

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*(5), 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. https://doi.org/10.1093/oxfordjournals.aje.a114229

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443-455. https://doi.org/10.1177/2515245920922982

Howe, M. L., & Otgaar, H. (2020). Theoretically important failures to reject the null hypothesis: Introduction to a special section of *Memory*. *Memory, 28*, 837-838.

JASP Team (2024). JASP (Version 0.19.3)[Computer software].

Jané, M. B., Xiao, Q., Yeung, S. K., Ben-Shachar, M. S., Caldwell, A. R., Cousineau, D., … Feldman, G. (2024b). Guide to Effect Sizes and Confidence Intervals. https://doi.org/10.17605/OSF.IO/D8C4G

Krasikova, D. V., Le, H., & Bachura, E. (2018). Toward customer-centric organizational science: A common language effect size indicator for multiple linear regressions and regressions with higher-order terms. *Journal of Applied Psychology, 103*(6), 659–675. https://doi.org/10.1037/apl0000296

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*, 33267. https://doi.org/10.1525/collabra.33267

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on psychological science*, *16*(3), 639-648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920951503

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. https://doi.org/10.1177/2515245918770963

Lemay, K. R., Tulloch, H. E., Pipe, A. L., & Reed, J. L. (2019). Establishing the minimal clinically important difference for the hospital anxiety and depression scale in patients with cardiovascular disease. *Journal of Cardiopulmonary Rehabilitation and Prevention*, *39*(6), E6-E11. https://doi.org/10.1097/HCR.0000000000000379

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. https://doi.org/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. *Health Psychology*, *40*(10), 727-736. https://doi.org/10.1037/hea0001112

McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: defining what really matters to patients. *Jama*, *312*, 1342-1343. https://doi.org/10.1001/jama.2014.13128

Morris, P. E., & Fritz, C. O. (2013). Effect sizes in memory research. *Memory*, *21*(7), 832-842. https://doi.org/10.1080/09658211.2013.763984

Mouelhi, Y., Jouve, E., Castelli, C., & Gentile, S. (2020). How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. *Health and Quality of Life Outcomes*, *18*, 1-17. https://doi.org/10.1186/s12955-020-01344-w

Murphy, S. L., Merz, R., Reimann, L. E., & Fernández, A. (2025). Nonsignificance misinterpreted as an effect's absence in psychology: prevalence and temporal analyses. *Royal Society Open Science*, *12*(3), 242167. https://doi.org/10.1098/rsos.242167

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*, 234-248. https://doi.org/10.1037/0021-9010.84.2.234

Norton, E. C., Dowd, B. E., & Maciejewski, M. L. (2019). Marginal effects—quantifying the effect of changes in risk factors in logistic regression models. *Jama*, *321*(13), 1304-1305. https://doi.org/10.1001/jama.2019.1954

Norton, E. C., & Dowd, B. E. (2018). Log odds and the interpretation of logit models. *Health services research*, *53*(2), 859-878. https://doi.org/10.1111/1475-6773.12712

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. https://doi.org/10.1037/mac0000116

Panzarella, E., Beribisky, N., & Cribbie, R. A. (2021). Denouncing the use of field-specific effect size distributions to inform magnitude. *PeerJ*, *9*, e11383. https://doi.org/10.7717/peerj.11383

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208. https://doi.org/10.1037/met0000126

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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). https://doi.org/10.1177/25152459241240722

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

Riesthuis, P., & Cribbie, R. (2025). When are scientific findings deemed practically or theoretically relevant? A literature review on minimum-effect testing. *Under review at The Quantitative Methods for Psychology*

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*, 203-215. https://doi.org/10.1002/acp.3911

Riesthuis, P., Mesquida, C., & Cribbie, R. (2024c). Statistical (non) Significance≠(un) Successful Replication: The Importance of the Smallest Effect Size of Interest. *PsyArxiv*. https://doi.org/10.31234/osf.io/s3zfy

Riesthuis, P., & Otgaar, H. (2024a). An overview of the replicability, generalizability and practical relevance of eyewitness testimony research in the Journal of Criminal Psychology. *Journal of Criminal Psychology*. Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JCP-04-2024-0031

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

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*(6), 657–680. https://doi.org/10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*(4), 403–411. https://doi.org/10.1037/1082-989X.3.4.403

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. https://doi.org/10.1098/rsos.221328

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712. https://doi.org/10.1177/1745691616658637

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, *19*(3), 227–229. https://pubmed.ncbi.nlm.nih.gov/20842279/

Van der Heijde, D., Lassere, M., Edmonds, J., Kirwan, J., Strand, V., & Boers, M. (2001). Minimal clinically important difference in plain films in RA: group discussions, conclusions, and recommendations. OMERACT Imaging Task Force. *The Journal of Rheumatology, 28*, 914-917.

Wells, G., Beaton, D., Shea, B., Boers, M., Simon, L., Strand, V., ... & Tugwell, P. (2001). Minimal clinically important differences: review of methods. *The Journal of Rheumatology*, *28*, 406-412.

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*(1), 3. https://doi.org/10.1037/lhb0000359

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences*, *61*, 1340-1341. https://doi.org/10.1002/jps.2600610845