JDSSV

Journal of Data Science, Statistics, and Visualisation

August 2026, Volume VI, Issue V. *doi: 10.52933/jdssv.v6i5.169*

5x3: A Practical Approach to Encouraging Thoughtful Statistical Analyses

Janet Aisbett
Meraglim Holdings

Eric J. Drinkwater
Deakin University, Geelong

Kenneth L. Quarrie
New Zealand Rugby

Stephen Woodcock
University of Technology, Sydney

Abstract

Empirical researchers who draw conclusions from statistical hypothesis tests have long been urged to be more thoughtful about their statistical designs, and to provide more modest interpretations of their findings. To support such behavioral change, we present a novel visualization tool for setting sample sizes and analyzing summary data. The core idea is to present multiple test options in the design stage of research, and to present alternative interpretations of summary data in the analysis stage. Our interactive tool is underpinned by a mathematically rigorous testing procedure featuring one-sided t-tests against the margins of material significance (i.e., effect sizes that are meaningfully different to zero). The tool is applied to studies from three different fields to illustrate how it might encourage better statistical design and more nuanced analysis and reporting.

Keywords: Significance testing, material significance, equivalence testing, NHST, multiple testing, stakeholder engagement.

1. Introduction

Concerns about how hypotheses are statistically tested and how the results are reported have been raised in both statistical and discipline-specific journals over many years (see, for example, Amrhein and Greenland 2022; Muff et al. 2022; Montero et al. 2023; Martinez Gutierrez and Cribbie 2023; White 2025; Ferr 2025). Applied researchers are accused of:

1. drawing overconfident conclusions from study findings, such as declaring a new treatment to be superior to a standard treatment on the basis of one Null Hypothesis Significance Test (NHST);
2. ignoring whether estimated effect sizes are deemed useful by practitioners, for example, declaring a new drug superior when the measured improvement is too small to be of medical value; and
3. rote application of standard tests, such as routinely testing the hypothesis that an effect size is zero at the default 0.05 level of test.

To help overcome such behaviors, we developed a visualization tool based on tests of hypotheses that effect sizes are meaningful to practitioners. It presents results of such tests at up to three test levels, which may encourage results to be reported more modestly—as a simple example, if the hypothesis that an effect size is greater than zero is rejected at level $\alpha = 0.05$ but not at the more stringent level 0.01, a researcher is not tempted to baldly claim that *the effect size is at most zero* as they might have done if only testing at level 0.05.

Our tool also supports the statistical design stage of research, during which the tests to be applied after data collection are specified, and a sample size determined that will deliver desired test powers. Simultaneously displaying sample sizes required for a range of tests and test levels, as well as a range of anticipated effect sizes, can encourage more thoughtful design (Aisbett et al. 2023).

The use of visual rather than numerical presentations by the tool is justified by findings such as Gatto et al. (2022) that visualization techniques improve overall study design and aid analysis and communication. This last point is important in applied research in which stakeholders outside the primary research team should be involved in both the statistical design and reporting stages (Greenland 2021).

1.1. Foundations

Typically, statistical tests seek to provide evidence that one treatment is better or worse than another by rejecting the hypothesis that the difference in treatment effects is zero. Hodges and Lehman (1954) recommend testing to provide evidence of a *materially significant* difference. For example, a new athletic routine might require additional training time, and so evidence of a marked improvement in performance might be needed before coaches would adopt it. Materially significant effect sizes, then, are those considered to be of practical importance by practitioners. Such effect sizes may lie either side of an interval $[L, U]$ considered equivalent to zero, or may be in a (possibly unbounded) interval.

The notion of materially significant effect sizes has seen numerous strategies that perform two or more tests of hypotheses about the same effect. These strategies have a decision theory basis in the work of Lehmann (1957). Thus, Campbell and Gustafson (2018) follow the standard null hypothesis test with a test of the hypothesis that the effect size is materially significant, a procedure they call conditional equivalence testing. Goeman et al. (2010) use the term three-sided testing to describe testing the hypothesis of no material significance while simultaneously testing to establish that the effect size is larger than U or smaller than L.

Cafri et al. (2022) augment three-sided testing with simultaneous tests to establish that the effect size is less than U or greater than L. These are the five tests implemented in our tool, as described in the next section.

In another version of five-region testing, Shih and Lee (2017) augment the NHST by testing hypotheses that the effect is above U, below L, in $[L, 0]$ or in $[0, U]$. They argue that such testing can describe the size and direction of a potential effect more efficiently and informatively than the common practice of reporting a p-value, a point estimate and a confidence interval.

A novel testing strategy in which one hypothesis about a parameter value appears to be tested at more than one alpha level is presented in Aisbett (2023). The rationale for this approach is that, to a stakeholder, a strong conclusion that a treatment is *more effective* than a control at test level $\alpha = 0.05$ together with the weaker conclusion it is *no less effective* at the more stringent level 0.005, say, may be more informative than either conclusion alone. This formalizes the common practice of highlighting smaller p-values with asterisks (Leahey 2005).

Note that, for some values, the hypothesis that the effect size is less than U, say, will be rejected by a t-test at level 0.025 while the hypothesis it is less than L won't be rejected by a t-test at alpha level 0.0001, say. Logically, we cannot conclude an effect size is non-negative without agreeing that it must be greater than -1. That is why, when performing the same test at more than one alpha level, any decision *must* be tied to the test level. Mathematically, this is achieved by replacing the original hypothesis with a family of hypotheses, one for each test level, as detailed in Aisbett (2023). The associated error rates are discussed in Aisbett (2024).

1.2. About this paper

We first explain five-region testing at up to three test levels, using displays generated by the tool that we call 5x3. Then we outline the tool's interfaces and functionalities and how it might be used in practice to help engage stakeholders in a research project. A preregistered report for a road safety study in which standard tests were planned is re-worked to illustrate how the tool could encourage broader statistical designs. Next, summary data from two studies drawn from sport science and management science are reanalyzed to show how 5x3 visualization might enrich interpretations. Finally, we address potential criticisms of our novel approach. Currently, the tool only supports t-tests.

2. Five Region Testing at One or More Alpha Levels

All figures in this section depict results from t-tests on summary data, assuming a two-group trial with 25 subjects per group.

2.1. Testing against the null

Figure 1(a) displays effect size E on the horizontal axis and standard error SE on the vertical axis. The rejection region of an α-level one-sided t-test of hypothesis $E > 0$ is bounded by the line $SE = T_n^{-1}(\alpha)E$, where T_n^{-1} denotes the inverse cumulative distribution function with n degrees of freedom. Given $\alpha = 0.025$, this is the green region in the figure, while orange represents where hypothesis $E < 0$ is rejected. The gray region is where neither hypothesis is rejected, i.e., where the standard two-sided test of the null hypothesis at level 0.05 is not rejected. The three data points plotted as numbers are all significant at a two-sided $p < 0.05$.

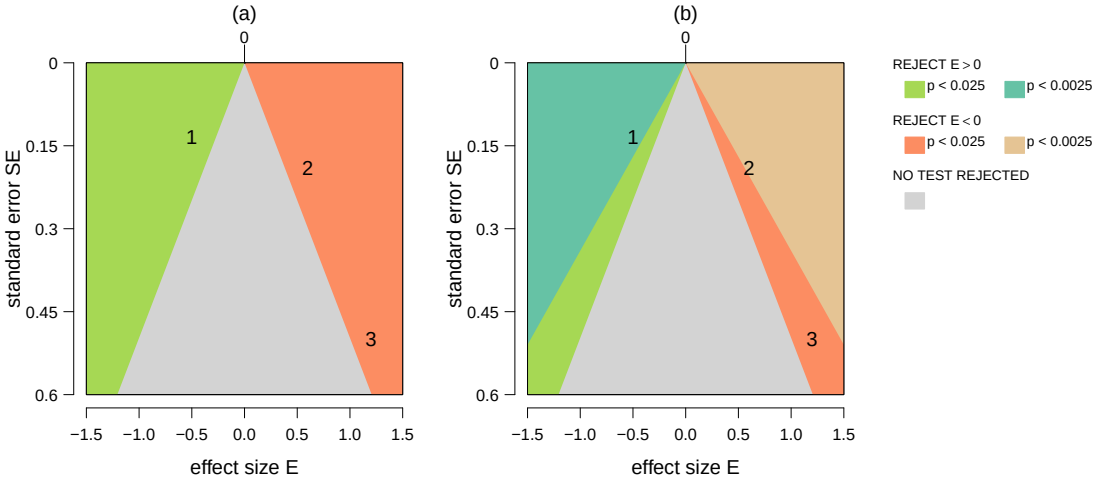

Figure 1: Rejection regions for one-sided t-tests against effect size zero. (a) Standard NHST performed as two one-sided tests against zero; (b) testing at two alpha levels. Summary data are plotted as numbers 1, 2, 3. All are interpreted as significant in standard NHST at $p < 0.05$. Findings 1 and 2 are also significant at a one-sided $p < 0.0025$, and in conventional NHST reporting might be starred as $p < 0.005$.

Figure 1(b) shows rejection regions for tests at $\alpha = 0.025$ and $\alpha = 0.0025$. Obviously, rejecting a hypothesis at level 0.0025 implies that it will be rejected at less stringent alpha levels, so the teal green region where hypothesis $E < 0$ is rejected at $\alpha = 0.0025$ must be interpreted as including the rejection region of the test at $\alpha = 0.025$.

2.2. Testing against margins of material significance

Figure 2(a) again shows the rejection regions for four tests; however, now we are testing hypotheses $E > 1$ and $E < 0$. This situation arises when effect sizes in the interval $[-1, 0]$ are deemed to be not materially significant. For example, if a current standard drug is very effective but expensive, a new cheap treatment may still be worthwhile even if it is slightly less effective.

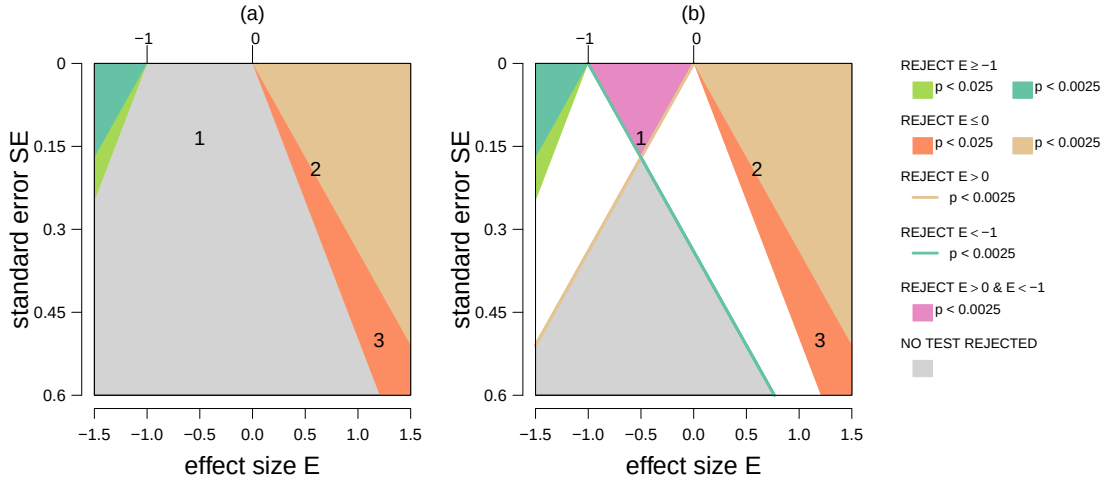

Figure 2: Rejection regions for one-sided t-tests against margins of materially significance at -1 and 0, shown as whiskers on top of charts. (a) Tests to determine materially significant values (less than -1 or greater than 0) at two alpha levels. Study 1 results are now inconclusive. (b) The region in which the effect size is found equivalent to zero (of no material significance) at $\alpha = 0.0025$ is the mauve-pink triangle, in which the Study 1 result now falls. The bounding lines of the rejection regions of hypotheses $E > 0$ and $E < -1$ at $\alpha = 0.0025$ are shown as teal and beige lines emanating from the relevant "whiskers". For example, the hypothesis $E < -1$ is rejected by findings to the right of the teal line. (This hypothesis is the only one rejected in the white region, while additional hypotheses are rejected in the mauve-pink, orange and beige regions.)

Figure 2(b) graphically depicts the richer test options available when it is recognized that effect sizes in an interval $[L, U]$ are not practically different to zero. In addition to decisions that an effect size is greater than U or less than L, we can investigate whether an effect size is *not less* than L or *not more* than U. To these four decisions can be added the decision that the effect size is practically *equivalent* to zero (i.e., within $[L, U]$, so not materially significant), made when the test that it is greater than U and the test that it is less than L are both rejected.

Deciding an effect size is greater than zero, say, implies it is greater than -1, so the latter decision can be seen as weaker. In general, tests to establish weaker findings should be more stringent (i.e., at smaller alpha levels). Thus, Figure 2(b) shows tests of hypotheses $E < -1$ and $E > 0$ only at the more stringent alpha level of 0.0025. These tests suggest that study 1 results are equivalent to zero at $\alpha = 0.0025$.

We can also now see that hypothesis $E < -1$ is rejected by the Study 3 results with $p < 0.0025$, whereas hypothesis $E \leq 0$ is only rejected at $p < 0.025$. In clinical trial terminology (see Table 1) this finding might be reported as "superior at test level 0.025 and not inferior at test level 0.0025."

3. The 5x3 Visualization Tool

This section describes use of the tool in both the statistical design and analysis stages of research.

Key components of the interface are identified in Figure 3. To obtain more precise readings, charts can be zoomed by selecting a region of interest then double clicking, and point information is displayed on hovering. Tests are nominated via a table (opened through sub-pane B of the control pane shown to the left of the Figure) whose columns are user-entered test alpha levels and whose rows describe the five test hypotheses, some or all of which may be selected.

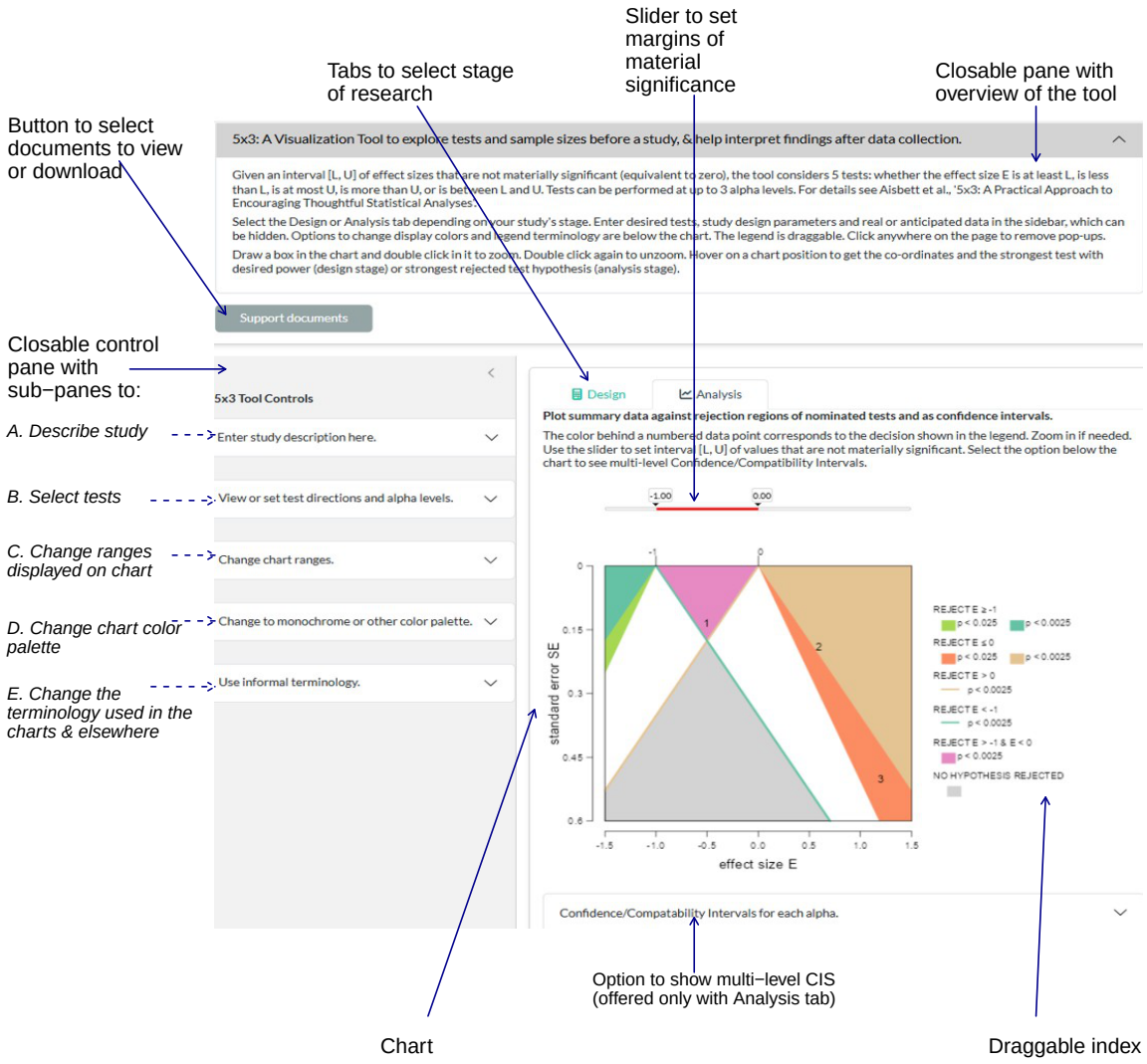

Figure 3: Components of the 5x3 interface, depicted with Analysis tab active.

3.1. Using the tool in the statistical analysis stage

The mean effect sizes and standard errors (or optionally sample variances) calculated from study data are entered through sub-pane A, and will then be plotted on charts, as in Figures 1 and 2. Compatibility (confidence) intervals for each data point may also be

displayed. These are the $100(1-2\alpha)\%$ CIs overlaid for each of up to three nominated test levels α, as illustrated in Figure 4.

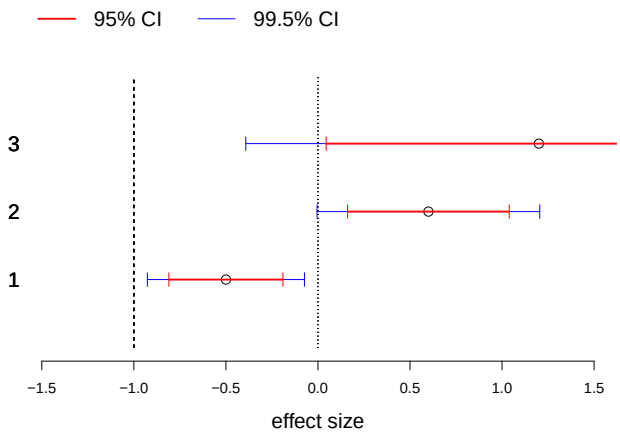

Figure 4: Compatibility/confidence intervals (CIs) corresponding to the data points plotted in Figure 2. The vertical lines are at the margins of material significance. The same conclusions can be drawn from these CIs as from the previous chart: for example, data point 3 is materially significant only at the weaker test level since the blue CI crosses the upper margin.

3.2. Using the tool in the statistical design stage

The tool can be used to prespecify tests and sample sizes prior to data collection. It has been developed to encourage researchers with constraints on feasible sample sizes to consider weaker tests (larger alpha levels, or less ambitious differentiations of effect size). When anticipated effect sizes and/or potential sample sizes are large, on the other hand, the tool could encourage researchers to test at more stringent levels than the default $\alpha = 0.05$.

The chart displayed when the Design tab is activated has anticipated sample size on the horizontal axis and total sample size on the vertical axis. The sample size scale is non-linear because sample sizes are calculated from standard errors using well known formula that also depend on the study design and the estimated variance.

The nomenclature in the Design chart legend is also changed. This is because regions depicted on the chart are where *user-nominated power* is achieved for a given test. As with standard sample size calculators, anticipated effect size(s) and sample variance must be specified, and are entered through the sub-pane A shown in Figure 3. (More than one effect size might be considered because the sample will be used in multiple studies, say, or because stakeholders have differing views about the expected effect of a treatment.) Desired power is entered through a table (accessed through an additional sub-pane of the control pane) to allow power to vary between tests and test levels, as discussed in Finner and Strassburger (2002) and Aisbett (2023, 2024).

The caption to Figure 5 describes how to read off the required sample size for each test at a given alpha level and anticipated effect size. The tool deliberately does not provide

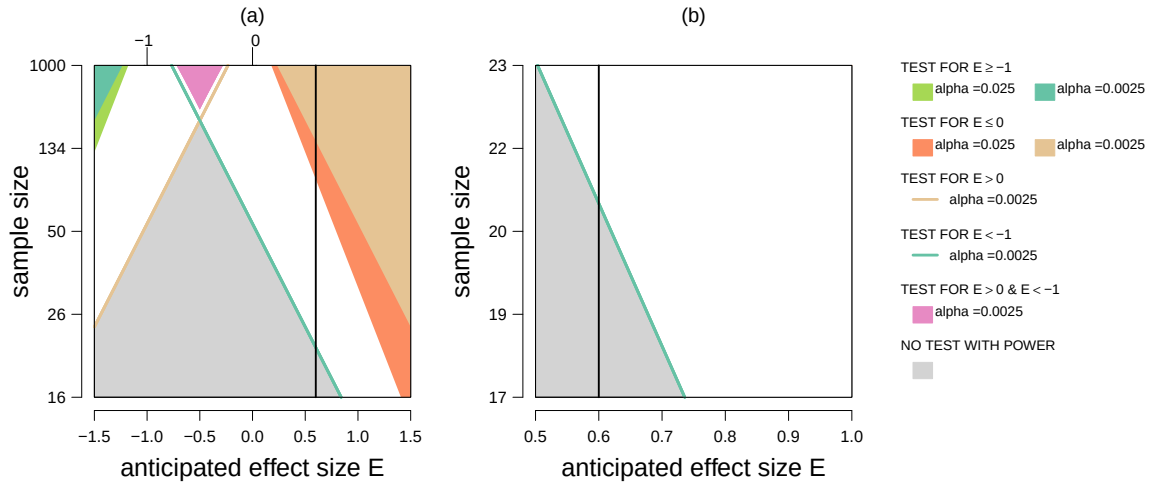

Figure 5: Design charts show sample size on a non-linear scale. Here, the anticipated effect size is 0.6 and anticipated variance is 1, with 80% power nominated for all tests. The minimum sample size to achieve this power is estimated from where the vertical line at E = 0.6 enters the region corresponding to a particular test. A sample of around 120 is needed for 80% power to reject E < 0 at $\alpha = 0.025$ (where the line enters the orange region in (a)). A sample of 22 gives 80% power to reject E < −1 at $\alpha = 0.0025$ (where the vertical line crosses the teal line, more easily read on zooming as in (b)).

numerical output of the sample size needed for a particular test with given parameters. This is to encourage users to visually explore the sensitivity of the estimated sample size to small changes in the anticipated effect size. However, reading off the sample size at which a line crosses into a region can be assisted by zooming and by hovering to get point co-ordinates.

This Figure suggests much smaller sample sizes for the test hypothesis E < −1 than for E ≤ 0, regardless of the anticipated effect size. In resource-limited trials, it may be better to select such a weaker test to reduce the chances of an inconclusive result.

3.3. More about the control pane

The tool allows for one group or two groups designs, specified through sub-pane A. This sub-pane also asks for total sample size even in the Design stage, since the tool employs t-tests rather than the normal approximations often used in sample size calculators. The default, 200, gives a normal approximation; it could be refined after an initial exploration.

Sub-panes C to E control the chart display. The chart can be toggled between monochrome and color displays or new color palettes entered by uploading a .txt or .csv file.

Correct terminology to do with statistical hypothesis testing is notoriously confusing, even to many researchers; see for example, Willigenburg and Poolman (2023). Since our tool aims to help general stakeholders engage with researchers in discussions about design parameters and study findings, we see the ability to present informal terminology as an important 5x3 feature. New terminology to describe the nominated directional tests and alpha level(s) is entered via an uploaded file. This preferred terminology will

appear in the chart legends and in the tables nominating the tests (and in the power table if applicable).

Examples of informal terminology are in Table 1, which compares formal terminology for reporting rejection of hypotheses with the terminology from clinical trials (Aberegg et al. 2018) and from a procedure known as Magnitude-Based Inference (MBI). (Although the developers of the MBI procedure described it as Bayesian (Batterham and Hopkins 2006), the method they implemented in Excel spreadsheets applied t-tests to the boundaries of an interval at one-sided alpha levels of 0.25, 0.05 and 0.005.) MBI came in two forms which used different terminology and had different workarounds to counter the very weak test level, as described in Aisbett et al. (2020).

Table 1: Some terminologies to describe test decisions.

Formal	**Clinical trial**	**MBI clinical**	**MBI**
reject $\mathrm{E} \geq L$	inferior	harmful	negative
reject $\mathrm{E} \leq U$	superior	beneficial	positive
reject $\mathrm{E} > U$	non-superior	not beneficial	not positive
reject $\mathrm{E} < L$	non-inferior	not harmful	not negative
reject both $\mathrm{E} > U$ and $\mathrm{E} < L$	equivalent	trivial	trivial
no hypothesis rejected	inconclusive	unclear	unclear

A numeric alpha level can also be described informally, for example as weak, moderate, or strong, where these terms are of course context dependent. MBI controversially described findings of one-sided tests at levels 0.25, 0.05 and 0.005 in Bayesian terms as likely, very likely and most likely (Batterham and Hopkins 2006).

4. Implementation and Use

The 5x3 tool is implemented in R v4.1.2 (R Core Team 2023) using the Shiny web application framework (Chang et al. 2023). It is available at meraglimja.shinyapps.io/5x3Tool/.

The underlying chart-drawing functionality is provided by the companion R package **Base5x3**, released under the GNU General Public License v3.0 and available at github.com/JA090/Base5x3.

The Handbook (Aisbett 2026a) describes using 5x3 testing in the context of collaborative statistical design and analysis, while the 5x3 User Manual (Aisbett 2026b) details the functionality and application of the software. These documents can be accessed through the *Support Documents* button on the app interface.

We suppose users take a Neyman–Pearson approach to statistical testing, which involves sample size calculations to attain a prespecified power or Type II error rate, at a prespecified α level or Type I error rate. This approach requires estimating the parameters of the sampling distributions, typically the mean and variance. In our situation, not only must margins of material significance be prespecified, but up to three alpha levels with matched power levels must also be prespecified (and justified). Briefly, here is how the tool is used in statistical design and analysis.

Step 1 (before data collection)

Discuss the study aims with stakeholders to help understand the applicability of the various five-region tests, using appropriate terminology to describe them (e.g., equivalence, inferiority, etc.). Collaboratively determine margins of material significance (or the interval equivalence to zero). Discuss the consequences of false positive and false negative errors. Discuss stakeholder appetite for risk with respect to errors.

Step 2 (before data collection)

Identify one or more pairs of Type I and Type II error rates to accommodate stakeholders' preferences. Identify resource constraints that might limit sample sizes. Use the tool with the Design tab active to estimate a sample size to deliver desired powers at agreed alpha levels. If resources are limited, possibly rule out some tests, or adjust power levels or test levels, in consultation with stakeholders.

Step 3 (before data collection)

Preregister the plan with a public trial registry appropriate to the discipline.

Step 4 (after data collection and pre-processing of the data)

With the tool's Analysis tab active, set up tests and test levels according to the registered plan. Plot summary statistics to determine tentative findings.

Step 5 (after data collection and analysis according to the preregistered plan)

Discuss findings with stakeholders in terms of the effect size and an appropriate statement about strength of evidence. Discuss what conclusions stakeholders would draw. Report findings in terms of both effect size and the strength of the tests.

5. Example Applications

In this section we illustrate application of the 5x3 tool to statistical design and analysis, using studies from diverse fields.

5.1. Using the 5x3 tool in statistical design

Beanland and Wynne (2019) explored the effects on car drivers of private roadside memorials such as small cairns or flowers. Subjects were paid to view video clips taken by a car-mounted camera, during which they would be asked to say what speed they would adopt as a driver, what they thought would be the sign-posted speed, and their rating of the risk of this stretch of road. Differences between responses to the clips with and without memorials was to be compared using t-tests, provided normality conditions were satisfied.

A preregistration report (Beanland et al. 2019) filed with the publisher justified anticipated standardized effect sizes of 0.53 on the basis that, for results to have policy

relevance, effects would need to be "large enough to produce a meaningful impact in the real world." To then achieve 90% power at a two-sided test level of 0.05, G*Power calculates a sample size of 40 for such a matched pairs study. (The report incorrectly says this sample size satisfies 95% power.)

Figure 6 suggests how a broader presentation of design options might influence design decisions. The legend uses terminology from MBI (clinical) since the research was prompted by debate as to whether roadside memorials made driving behaviour safer or less safe.

Figure 6(a) follows the published research plan in considering t-tests against the null hypothesis. However, is this appropriate given the researchers claim a small standardized effect size would have no practical policy implication? In (b), the margins of material significance have been set to the upper bound of standardized effect size generally considered small, around ± 0.2. Power for tests of hypothesis $\mathrm{E} \leq 0.2$ was set to 90%, and power for tests of the weaker hypothesis $\mathrm{E} < -0.2$ was set to 95%.

In both Figures 6(a) and (b), it is visually obvious that the more stringent test level has a relatively modest impact compared with the choice of anticipated standardized effect size. Allowing for material significance further pushes up the sample size. For example, assuming effect size 0.53 and a one-sided alpha level of 0.025, attaining 90% power to detect a materially significant effect requires 400 subjects, up from 154 when testing against the null. The required total sample size rises to 530 at $\alpha = 0.007$.

On the other hand, smaller sample sizes of 100 and 130 give 95% power to reject hypothesis $\mathrm{E} < -0.2$ at respective levels 0.025 and 0.007. Rejection of this test suggests that the memorials don't *negatively* affect drivers' assessments—that is, it provides evidence that memorials don't cause a driver to speed or take greater risks in a materially significant way. Such a finding may be valuable, as the authors point out that roadside memorials are banned in some jurisdictions because of their perceived negative influence on driving.

With a small anticipated effect size 0.2, achieving 95% power to reject the hypothesis of harmful consequences takes a total sample size of 332 even at the more lenient alpha level.

As long as the margin(s) of material significance are prespecified, it is valid to first explore evidence that the memorials have a beneficial effect (reject $\mathrm{E} \leq 0.2$) and, if the data fail to reject that, explore evidence that they are not harmful (reject $\mathrm{E} < -0.2$). If the research budget constrains the sample size, then the first test might be made against zero, even though that would not provide any evidence that the memorials have a *meaningfully* beneficial effect on driving.

Clearly the views of stakeholders and funders are crucial to these design decisions. And, as a salutary footnote, t-tests were completely abandoned in the published statistical analysis in the completed research (Beanland and Wynne 2019).

5.2. Using the 5x3 tool in statistical analyses

We now illustrate how the 5x3 tool might be used in more thoughtful analyses of study data, drawing from two published reports.

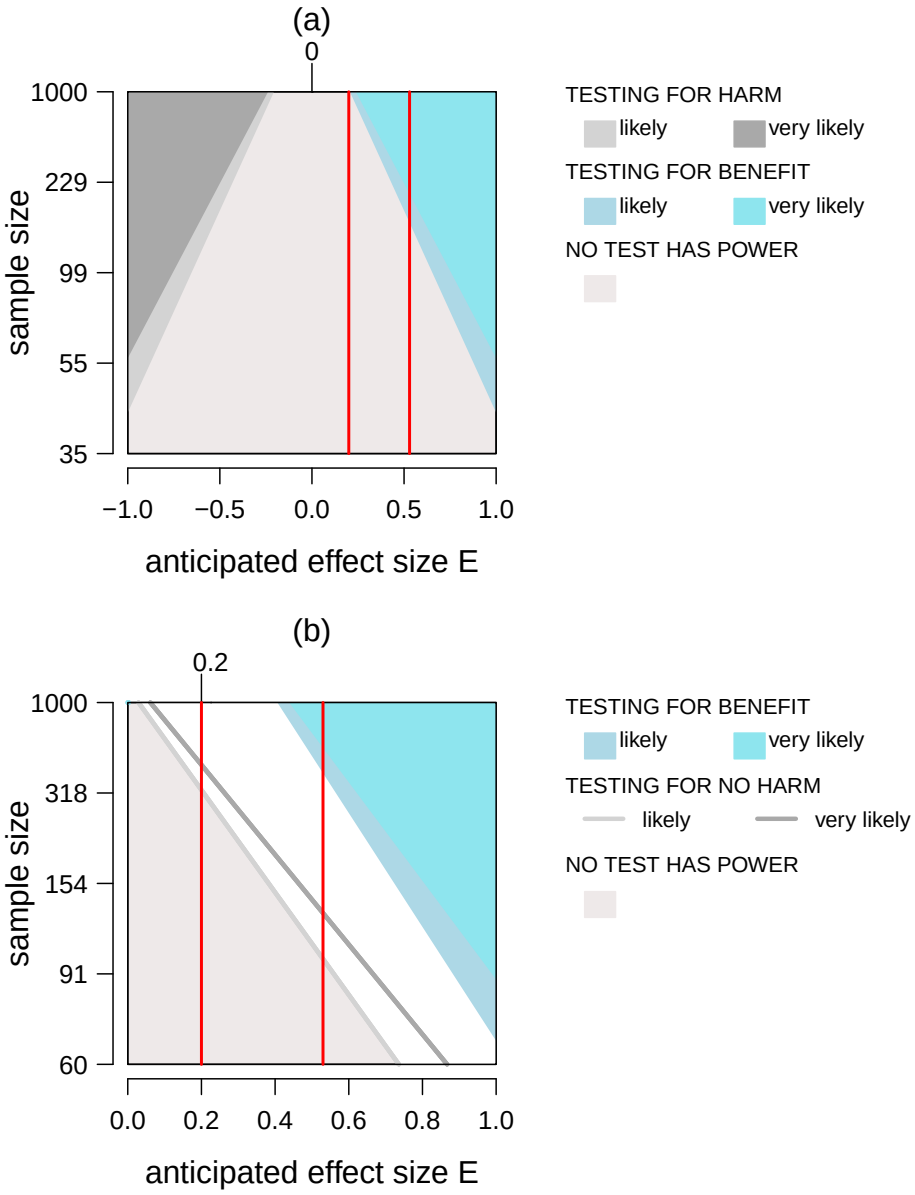

Figure 6: Design chart exploring proposal in Beanland et al. (2019). Standard t-tests were planned (one sided $\alpha = 0.025$) with an anticipated standardized effect size of 0.53 and target power 90%. We also consider a small standardized effect size of 0.2 and add tests at $\alpha = 0.007$ to allow for Bonferroni correction, given 3 measures were to be taken. Informal terminology from MBI (clinical) is used that might be helpful in presentations to stakeholders: *likely* refers to testing at $\alpha = 0.025$ and *very likely* to testing at $\alpha = 0.007$. The charts also illustrate a different color palette. Read off required sample sizes from where the vertical lines enter either blue region. (a) shows testing when material significance is ignored, while (b) considers margins of material significance of ± 0.2.

Example from sports science

In a study reported by Bartolomei et al. (2014), participants were 24 male athletes competing at a high level in sports involving extensive resistance training (for example, shot put). They were divided into two groups that performed the same suite of training exercises for the same durations; however, one group performed these exercises in a novel pattern while the second group followed a traditional pattern. Pre and post measurements of the athletes were recorded for seven strength and power tests, two of which concerned lower body strength.

The data were analyzed with the MBI statistical procedure, which at the time was widely used in parts of sports science (Van Hooren 2018).

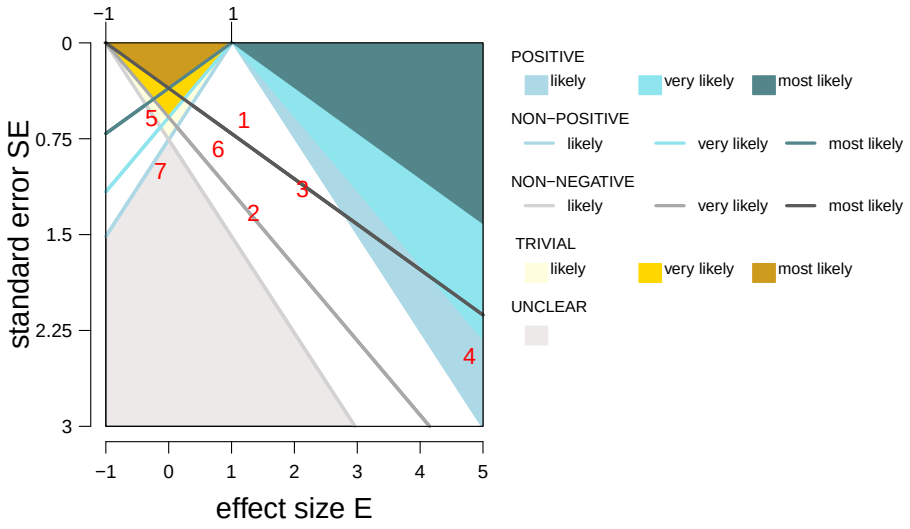

Figure 7: Reanalysis of summary data from Bartolomei et al. (2014), using one-sided alpha levels 0.10, 0.05, 0.005 and reporting terminology used in MBI. The summary data are numbered in the order the measures are listed in the cited article's Table 2.

An MBI analysis requires that researchers define a "trivial" interval, that is, values of no material significance. Bartolomei et al. (2014) set the interval margins to ±0.2 times the standard deviation of the pre-trial measurements over all participants. Because this varied with the measure, we normalized the summary statistics (standardized mean difference and standard error), obtained from the group means and standard deviations listed in Bartolomei et al. (2014), Table 2) by the relevant margins of the zero-equivalent interval before plotting them in Figure 7. The axes are therefore in units of the margins. For each of the measures, the positive direction represents a relative improvement for the novel pattern.

We raised MBI's weakest test alpha from 0.25 to 0.1, and only tested for positive/negative differences at this level. The two data points 5 and 7 with negative means were measurements of lower body strength. The remaining five points lie above the mid-gray line which bounds the rejection region of the test hypothesis $E < -1$ at alpha level 0.05. For these measures, the new exercise pattern is non-negative (non inferior in clinical trial terminology). It is likely positive (superior) at $\alpha = 0.05$ for measure 4, whereas no measure was inferior even at this weak level.

Bartolomei et al. (2014) concluded that their study results indicated that the new pattern "may be more beneficial" for upper body strength and power, whereas "no clear advantage was demonstrated" for either pattern in regard to lower-body strength. Had they used a standard null hypothesis test at a two-sided alpha of 0.05, none of the measures would have provided a significant result. (The 5x3 tool allows this to be quickly determined from the 95% CI reported in the pop-up that appears on hovering over a point.)

Despite the lack of strong findings, a coach might decide to use the new training pattern, since switching strategies carried little if any cost and even small improvements in athletic performance may be valued.

Example from management science

A study by Ladewski and Al-Bayati (2019) investigated whether the Quality Management System (QMS), described in international standard ISO 9001:2015 and developed for product quality, was also an appropriate management model for safety quality. They developed a questionnaire to elicit attitudes to the seven factors in the QMS (for example, leadership and client focus) which was completed by 45 people with roles involving product quality and 99 with roles involving safety quality.

Responses on a 6-point Likert scale were converted to integers 1 to 6 and summed across all items probing a factor. Responses from the two groups were subjected to two-sided t-tests at $\alpha = 0.007$, which the authors explained was to allow for Bonferroni correction. Although they found significant differences on two factors, the authors concluded that safety quality management could follow the management model traditionally applied to product and service quality.

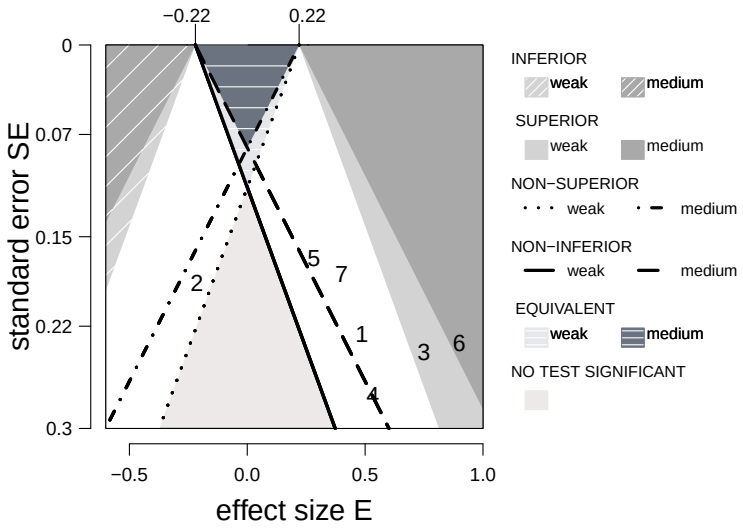

Figure 8: Reanalysis of summary data from Ladewski and Al-Bayati (2019). The plotted numbers represent differences in attitude between quality and safety management personnel to Quality Management factors, ordered as in the cited article's Table 1. Here, tests at one-sided $\alpha = 0.025$ are labeled *weak* while those at 0.0035 are *medium*. This figure also illustrates the monochrome palette that comes with the 5x3 tool.

Figure 8 uses data from the article's Table 4, and describes the 5 tests in the terminology of comparative drug trials. A positive outcome means a stronger attitude has been expressed in the product quality group. Because the factors were variously associated with 4, 6 and 7 questions, the factor scores for each group were normalized using the number of questions. This meant that each factor had a maximum possible mean score of 6 and the difference E between the groups had potential range ± 5.

Our reanalysis defined margins of material significance by considering the numerical conversion of responses from the Likert scale. (Debate about this conversion is outside our present scope, see for example, Bishop and Herron (2015). Appendix 1 justifies the margins $\pm 2 \times 0.110$ used in Figure 8). At a one-sided $\alpha = 0.0035$, no test hypothesis was rejected concerning factor 2—communication and cooperation. For all other factors, at this test level the product quality group attitude was non-inferior, while for factor 6 (client focus) it was also superior.

Under the original NHST analysis, only the factor represented by data point 5 was deemed significant, and nothing could be said about the other 5 factors. Our re-analysis presents a clearer picture in which, at the more stringent test level, the product quality group attitude was non-inferior on 6 of the 7 factors, while there was no comparable finding for the safety quality group. This suggests possible differences between the two groups, warranting further investigation into the applicability of the QMS framework to safety quality.

6. Discussion

For too many researchers, statistical design and analysis is a part of the research process that is conducted by rote (this type of data = this type of analysis = this type of results). Researchers with limited formal statistical education may focus on computer outputs without appreciating the theoretical basis of the results. Applying statistical tests of the null hypothesis and $p < 0.05$ becomes routine. These habits are harder to break when reviewers and editors look for standard tests (Wasserstein et al. 2019). New approaches are needed to change these behaviours.

The 5x3 visualization tool has the potential to encourage a more thoughtful approach because it presents the research team with a set of hypotheses that concern the margins of material significance and cover the entire parameter space, in contrast to simply testing for zero effect. Our tool also presents researchers with a range of alpha levels, which can (and should) be adjusted, but which encourages exploration beyond the default setting of 0.05 that is so widely observed.

Effective use of the tool requires researchers and stakeholders to collaboratively answer questions such as:

— how large an effect do we care about?

— how comfortable are we about making an incorrect decision and what are the consequences if we do?

— what sample sizes are feasible given available resources and therefore what tests can deliver satisfactory power?

Such questions can only be answered in context and involve expert knowledge of the application domain. Even then, experts may differ, and so the tool helps accommodate their differences. The 5x3 companion Handbook (Aisbett 2026a) provides guidance on how to incorporate stakeholder differences at the statistical design stage.

Rather than steering researchers into more modest claims about their findings, could our tool lead to even wilder claims? Simultaneous testing of opposing hypotheses at more than one alpha level is in some sense a scatter gun approach, liable to return a "significant" finding that researchers could seize upon as proof of an effect regardless of the test level.

A key facet of our approach, stemming directly from the underlying theory, is that a finding about an effect size is irrevocably linked to the alpha level of the test associated with the finding. An intervention cannot be reported as "more effective" than current practice, say, because it must be qualified by a descriptor of the test level.

The reporting language still matters. Results of statistical analyses in applied research are eventually assessed for their suitability by stakeholders and so "results need to be articulated to the key stakeholders at an appropriate level of detail/language so that all understand" (Bishop et al. 2006, p. 163). The role of stakeholders as well as researchers therefore needs to be considered when communicating study findings, just as in the design stage.

For this reason, concerns about imprecision of verbal labels need to be evaluated against the goal of effective communication with stakeholders who are not statistically trained and who may be turned off by formal language. The most useful terminology for describing any statistical finding will depend on context (Amrhein and Greenland 2022). The 5x3 tool allows users to nominate context-appropriate labels at both design and reporting stages. This is an important feature. When using tool output in scholarly publications, formally correct terminology is expected. When using the tool with stakeholders, informal terminology may be more productive, because, as Buchheit (2017, p. 40) argues, unless researchers have effective communication with stakeholders "their fancy reports with high quality stats will end up in the bin."

With our tool, researchers can obtain "conclusive" findings from studies that would be seriously underpowered with conventional analyses. Could it therefore lead to studies with smaller sample sizes? We argue there is little incentive for a researcher to collect fewer data than they would had they planned a conventional NHST analysis, just because our tool offers the opportunity for a weak conclusion such as "Our data are compatible with finding the new intervention to be no worse than existing practice at $p < 0.10$. Nothing can be said about comparative performance at $p < 0.05$."

On the other hand, researchers often face real-world constraints on sample sizes. While weaker test levels must then be planned, these may still fit the research purpose. Moreover, findings will add to a body of knowledge, especially when incorporated into subsequent meta-analyses. Ethical considerations are the stock answer to why small samples are bad, yet ethics questions may equally arise when research projects are canned simply because a sample size calculator set to $p < 0.05$ and 80% power returns a number that is not feasible given project resources.

Our suggested approach to the design of five region testing at multiple alpha levels using the tool needs to be trialed in diverse scenarios and modified as required to ensure

researchers and stakeholders remain engaged. Possible pitfalls must be identified to help educate researchers and avoid misuse that would impede acceptance of the approach by reviewers and journal editors. In addition, the tool needs extension to a broader range of statistical tests, including the ANOVA family.

Research reporting will continue to be immodest until researchers (and indeed, statisticians) appreciate that qualifications need to be attached to the conclusions drawn from any statistical analysis. This is why we believe that strategies that perform multiple tests on the same effect, enhanced with visualization, have the potential to break ingrained habits and improve statistical decision-making.

Supporting Material

1. Web application publicly available at meraglimja/shinyapps.io/5x3Tool/.

2. Project repository available at github.com/JA090/Base5x3/ containing: User Manual for the 5x3 tool; Handbook on the use of 5x3 in collaborative statistical design and reporting; and code for both interactive and batch versions.

Acknowledgments

We thank Will Hopkins and Alan Batterham, the developers of MBI, for inspiring a debate about conventional statistical analyses that led to the development of this tool.

Generative AI Statement

Grok 4.1 (xAI, November 2025) and Claude Sonnet 4.5 (Anthropic, September 2025) were used to create minor snippets of code.

References

Aberegg, S. K., Hersh, A. M., and Samore, M. H. (2018). Empirical consequences of current recommendations for the design and interpretation of non-inferiority trials. *Journal of General Internal Medicine*, 33:88–96, DOI: 10.1007/s11606-017-4161-4.

Aisbett, J. (2023). Interpreting tests of a hypothesis at multiple alpha levels within a Neyman–Pearson framework. *Statistics and Probability Letters*, 201, DOI: 10.1016/j.spl.2023.109899.

Aisbett, J. (2024). Can expected error costs justify testing a hypothesis at multiple alpha levels rather than searching for an elusive optimal alpha? *PLOS ONE*, 19(9), DOI: 10.1371/journal.pone.0304675.

Aisbett, J. (2026a). Handbook on the use of 5x3 in collaborative statistical design and reporting. https://github.com/JA090/Base5x3.

Aisbett, J. (2026b). User manual for the 5x3 tool. https://github.com/JA090/Base5x3.

Aisbett, J., Drinkwater, E. J., Quarrie, K. L., et al. (2023). Applying generalized funnel plots to help design statistical analyses. *Statistical Papers*, 64:355–364, DOI: 10.1007/s00362-022-01322-y.

Aisbett, J., Lakens, D., and Sainani, K. L. (2020). Magnitude based inference in relation to one-sided hypotheses testing procedures. *SportRxiv*, DOI: 10.31236/osf.io/pn9s3.

Amrhein, V. and Greenland, S. (2022). Rewriting results in the language of compatibility. *Trends in Ecology and Evolution*, 37(7):567–568, DOI: 10.1016/j.tree.2022.02.001.

Bartolomei, S., Hoffman, J. R., Merni, F., and Stout, J. R. (2014). A comparison of traditional and block periodized strength training programs in trained athletes. *Journal of Strength and Conditioning Research*, 28(4):990–997, DOI: 10.1519/JSC.0000000000000366.

Batterham, A. M. and Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *International Journal of Sports Physiology and Performance*, 1:50–57. https://pubmed.ncbi.nlm.nih.gov/19114737/.

Beanland, V. and Wynne, R. A. (2019). Effects of roadside memorials on drivers' risk perception and eye movements. *Cognitive Research: Principles and Implications*, 4(1):32, DOI: 10.1186/s41235-019-0184-1.

Beanland, V., Wynne, R. A., and M. Salmon, P. (2019). Effects of roadside memorials on drivers' risk perception and eye movements (registered report, stage 1). figshare. Journal contribution, DOI: 10.6084/m9.figshare.6181937.v1.

Bishop, D., Burnett, A., Farrow, D., Gabbett, T., and Newton, R. (2006). Sports science roundtable: Does sports science research influence practice? *International Journal of Sports Physiology and Performance*, 1(2):161–168, DOI: 10.1123/ijspp.1.2.161.

Bishop, P. A. and Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International Journal of Exercise Science*, 8(3):297–302, DOI: 10.70252/LANZ1453.

Buchheit, M. (2017). Want to see my report, Coach? A battle worth fighting: A comment on the vindication of magnitude-based inference. *Aspetar Journal*, 6:34–43, www.aspetar.com/journal/viewarticle.aspx?id=350#.X2R_2RAzbix.

Cafri, G., Wood, J., and Gagne, J. (2022). Partition testing for real-world evidence studies. *Pharmacoepidemiology and Drug Safety*, 31(12):1280–1286, DOI: 10.1002/pds.5540.

Campbell, H. and Gustafson, P. (2018). Conditional equivalence testing: An alternative remedy for publication bias. *PLOS ONE*, DOI: 10.1371/journal.pone.0195145.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2023). *shiny: Web Application Framework for R*, https://CRAN.R-project.org/package=shiny. R package version 1.8.0.

Ferr, H. (2025). Misinterpretations of the p-value in psychological research: Implications for mental health and psychological science. *PLOS Mental Health*, 2(2):e0000242, DOI: 10.1371/journal.pmen.0000242.

Finner, H. and Strassburger, K. (2002). The partitioning principle. A powerful tool in multiple decision theory. *The Annals of Statistics*, 30(4):1194–1213. https://projecteuclid.org/journals/annals-of-statistics/volume-30/issue-4/The-partitioning-principle--a-powerful-tool-in-multiple-decision/10.1214/aos/1031689023.full.

Gatto, N. M., Wang, S. V., Murk, W., et al. (2022). Visualizations throughout pharmacoepidemiology study planning, implementation, and reporting. *Pharmacoepidemiology and Drug Safety*, 31(11):1140–1152, DOI: 10.1002/pds.5529.

Goeman, J. J., Solari, A., and Stijnen, T. (2010). Three-sided hypothesis testing. *Statistics in Medicine*, 29:2117–2125. https://pubmed.ncbi.nlm.nih.gov/20658478/.

Greenland, S. (2021). Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. *Paediatric and Perinatal Epidemiology*, 3(5):8–23, DOI: doi:10.1111/ppe.12711.

Hodges, J. L. and Lehman, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society Series B (Methodological)*, 16(2):261–268, DOI: 10.1111/j.2517-6161.1954.tb00169.x.

Ladewski, B. J. and Al-Bayati, A. (2019). Quality and safety management practices: The theory of quality management approach. *Journal of Safety Research*, 69:193–200, DOI: 10.1016/j.jsr.2019.03.004.

Leahey, E. (2005). Alphas and asterisks: The development of statistical significance testing standards in sociology. *Social Forces*, 84(1):1–24, DOI: 10.1353/sof.2005.0108.

Lehmann, E. (1957). A theory of some multiple decision problems. Parts I and II. *The Annals of Mathematical Statistics*, 28:1–25; 547–572. DOI: 10.1214/aoms/1177707034.

Martinez Gutierrez, N. and Cribbie, R. (2023). Effect sizes for equivalence testing: Incorporating the equivalence interval. *Methods in Psychology*, 9, DOI: 10.1016/j.metip.2023.100127.

Montero, O., Hedeland, M., Balgoma, D., et al. (2023). Trials and tribulations of statistical significance in biochemistry and omics. *Trends in Biochemical Sciences*, 48(6):503–512, DOI: 10.1016/j.tibs.2023.01.009.

Muff, S., Nilsen, E., O'Hara, R., and Nater, C. (2022). Rewriting results sections in the language of evidence. *Trends in Ecology and Evolution*, 37:203–210, DOI: 10.1016/j.tree.2021.10.009.

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

Shih, H. Y. and Lee, W. C. (2017). A five-region hypothesis test for exposure-disease associations. *Scientific Reports*, 7:5131, DOI: 10.1038/s41598-017-05301-4.

Van Hooren, B. (2018). Magnitude-based inference: What is it? How does it work and is it appropriate? *Sports Performance and Science Reports*, 1:1–2. https://www.researchgate.net/publication/325217813_Magnitude-based_inference_What_is_it_How_does_it_work_and_is_it_appropriate.

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

White, C. (2025). The perilous P value. *Journal of the American Veterinary Medical Association*, 263(2):267–270, DOI: 10.2460/javma.24.10.0645.

Willigenburg, N. W. and Poolman, R. W. (2023). The difference between statistical significance and clinical relevance: The case of minimal important change, non-inferiority trials, and smallest worthwhile effect. *Injury*, 54:110764, DOI: 10.1016/j.injury.2023.04.051.

Appendix 1

A Likert response converted to integer x can be interpreted as a true response somewhere in the interval $[x-0.5, x+0.5)$. Implicitly, all values in this interval are equivalent, as they all are represented by x. If none are preferred, then the response is a random variable following $x - 1/2 + U(0,1)$, for U the uniform distribution. The sum of n uniformly distributed random variables has the Irwin–Hall distribution $IH(n)$ of order n. This has mean $n/2$ and standard deviation $\sigma(n) = \sqrt(n/12)$, and the probability that values lie within two standard deviations of the mean exceeds 95%.

Now consider the normalized sum s of a subject's Likert responses to a factor with n questions. The sum represents a true sum that follows the distribution $s - 1/2 + IH(n)/n$, and it is straightforward to calculate that the probability that the true sum lies in the interval $[s - 2\sigma(n), s + 2\sigma(n))$ is greater than 95%. Values in this interval are treated as equivalent to s. The values represent the subject's attitude to the role of a factor such as leadership. It follows that differences in attitude of $\pm 2\sigma(n)$ are not considered materially significant (otherwise a different scale or more questions would have been incorporated).

For factors with 4 questions, $\sigma(n) = 0.144$ and for 7 factors it is 0.110. We used the smaller value in re-analyzing data from Ladewski and Al-Bayati (2019). With margins set to $\pm 2 \times 0.144$, the communication and co-operation factor falls in the non-superior region; all other decisions remain the same. However, as the communication factor had 7 questions, there is no justification for using wider margins.

Affiliation:

Janet Aisbett
Meraglim Holdings Corporation
777 South Flagler Drive, West Palm Beach,
Florida, United States 33401
E-mail: janet.aisbett@meraglim.com

Eric J. Drinkwater
School of Exercise and Nutrition Sciences, Deakin University,
221 Burwood Highway, Burwood,
Victoria, Australia 3125
E-mail: eric.drinkwater@deakin.edu.au

Ken Quarrie
Chief Scientist, New Zealand Rugby
PO Box 2172 Wellington
New Zealand 6140
E-mail: Ken.Quarrie@nzrugby.co.nz

Stephen Woodcock
School of Mathematical and Physical Sciences, University of Technology Sydney
PO Box 123, Ultimo
New South Wales, Australia 2007
E-mail: Stephen.Woodcock@uts.edu.au

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/

ISSN 2773-0689
August 2026, Volume VI, Issue V *Submitted:* 2026-04-10
doi:10.52933/jdssv.v6i5.169 *Accepted:* 2026-07-21