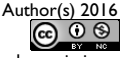
*Article*

# Performing Contrast Analysis in Factorial Designs: From NHST to Confidence Intervals and Beyond

Educational and Psychological Measurement
2017, Vol. 77(4) 690–715



DOI: 10.1177/0013164416668950
journals.sagepub.com/home/epm


**Stefan Wiens[1] and Mats E. Nilsson[1]**

**Abstract**

Because of the continuing debates about statistics, many researchers may feel confused about how to analyze and interpret data. Current guidelines in psychology advocate the use of effect sizes and confidence intervals (CIs). However, researchers may be unsure about how to extract effect sizes from factorial designs. Contrast analysis is helpful because it can be used to test specific questions of central interest in studies with factorial designs. It weighs several means and combines them into one or two sets that can be tested with *t* tests. The effect size produced by a contrast analysis is simply the difference between means. The CI of the effect size informs directly about direction, hypothesis exclusion, and the relevance of the effects of interest. However, any interpretation in terms of precision or likelihood requires the use of likelihood intervals or credible intervals (Bayesian). These various intervals and even a Bayesian *t* test can be obtained easily with free software. This tutorial reviews these methods to guide researchers in answering the following questions: When I analyze mean differences in factorial designs, where can I find the effects of central interest, and what can I learn about their effect sizes?

**Keywords**

analysis of variance, contrast analysis, confidence interval, null hypothesis significance testing, Bayesian analysis

Because of the continuing debates about how to analyze and interpret data, many researchers in psychology may feel in limbo. Are we analyzing and interpreting our

[1]Department of Psychology, Stockholm University, Stockholm, Sweden

**Corresponding Author:**
Stefan Wiens, Gösta Ekman Laboratory, Department of Psychology, Stockholm University, Frescati Hagväg 9A, Stockholm, 106 91, Sweden.
Email: sws@psychology.su.se

own data correctly, or are we conducting analyses without really understanding what they are or what their results mean (Torgerson, 1965)? Null hypothesis significance testing (NHST) has a long tradition in psychology, but there are recurrent concerns about its use (Chavalarias, Wallach, Li, & Ioannidis, 2016; Hunter, 1997; Nuzzo, 2014; Trafimow, 2014; Trafimow & Marks, 2015; Wasserstein & Lazar, 2016). From a theoretical perspective, NHST is actually an unfortunate mixture of at least two different and incompatible theories (Fisher vs. Neyman–Pearson) combined with wishful thinking (Gigerenzer, 2004; Hubbard, 2004; Perezgonzalez, 2015). Nonetheless, NHST is commonly portrayed as a single, coherent testing approach in guidelines (American Psychological Association, 2009; Wilkinson, 1999) and textbooks (for review, see Hubbard, 2004).

From a practical perspective, various aspects of NHST have been criticized because NHST is used to address questions that it cannot answer (Cumming, 2012; Fidler & Loftus, 2009; Gigerenzer, 2004; Goodman, 1999a; Ivarsson, Andersen, Stenling, Johnson, & Lindwall, 2015; Kline, 2013; Nuzzo, 2015; Wagenmakers, 2007). Among the misconceptions about what one can infer from NHST results is the notion that if the $p$ value is small (e.g., 2%), the null hypothesis ($H_0$) must be false. Although this conclusion makes intuitive sense, it is incorrect. Imagine a pregnancy test that will falsely indicate pregnancy in 1% of cases (Kalinowski & Fidler, 2010). If the test result comes out positive and suggests pregnancy ($p < 1\%$), does this mean that the individual is pregnant? Definitely not, if the tested individual is male! So, even though the data may be unlikely given that the person is not pregnant (i.e., $H_0$), they are even more unlikely given the impossibility of a man being pregnant. Formally speaking, the probability of the data given the hypothesis (i.e., P(D|H)) is not the same as the probability of the hypothesis given the data (i.e., P(H|D)). To mention another example: If somebody does not cheat ($H_0$), then the probability of winning the lottery is very low (e.g., $p < .00000001$). If you happen to win the lottery, this does not prove that you cheated! (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Another misconception is that if the $p$ value is large (e.g., $p > 0.60$), then the $H_0$ is true. This is incorrect; a large $p$ value could be obtained simply because too few subjects are tested, and as a result, the data are insensitive (Dienes, 2014). Last, researchers treat $p$ values as if they are highly reliable (Lai, Fidler, & Cumming, 2012). Unfortunately, $p$ values jump around even for perfect replications (Cumming, 2008, 2012; Cumming & Fidler, 2009; Cumming & Maillardet, 2006; Halsey, Curran-Everett, Vowler, & Drummond, 2015). These examples illustrate that even if many ingredients of NHST are inherently legitimate (Perezgonzalez, 2015; Savalei & Dunn, 2015), researchers often misinterpret the results of NHST (for further discussion, see the other articles in this special issue).

In response to this continuing debate, current guidelines in Psychology advocate the use of effect sizes and confidence intervals (CIs) (American Psychological Association, 2009; Cumming, 2014; Funder et al., 2014; Wilkinson, 1999). In the context of assessing a treatment, effect size is considered an appropriate tool for answering three questions (Faulkner, Fidler, & Cumming, 2008): Is there a treatment

effect, how large is it, and is it practically important? Whereas NHST has been used to address the first question (is there an effect, yes or no?), researchers are encouraged to go beyond this question and determine how large the actual effect is and its practical importance (Kirk, 1996; Thompson, 2002; Valentine, Aloe, & Lau, 2015).

Because researchers often use factorial designs to study mean differences between conditions or groups, it is useful to know how to extract effect sizes from factorial designs. Simple tools can be used to do this even in complex factorial designs, as long as the researchers isolate the effects of central interest. Although these tools have been described before, researchers may not be familiar with them. To fill this gap, we provide practical advice on how to analyze and interpret mean differences in factorial designs in terms of effect size.

Of course, researchers want to use the results from their sample to learn about the population value (i.e., the true value). Although NHST provides one approach to this task, other approaches (e.g., CIs, likelihood, Bayesian) may provide additional, valuable information. In briefly reviewing these alternatives and their advantages, we will assume for the sake of simplicity that the data fulfill all requirements for the use of *t* tests: Observations are randomly and independently drawn from normally distributed populations of equal variance. In sum, we hope that this tutorial will serve as a guide (Sharpe, 2013) for answering the following questions: When I analyze mean differences in factorial designs, where can I find the effects of central interest, and what can I learn about their effect sizes?

## The Common Approach to Analyzing Mean Differences

The goal of many research studies is to examine differences in means between groups or conditions. These differences between means are often analyzed and reported from the perspective of NHST (Howell, 2013). If the observed $p$ value is smaller than or equal to a critical $\alpha$ (e.g., 5%), then it is significant, and researchers should behave as if there is an effect. In contrast, if the observed $p$ value is larger than the critical $\alpha$, then it is not significant and the results are inconclusive (i.e., $H_0$ is not rejected). To conduct NHST on means, a *t* test is commonly used to compare two means, and analysis of variance (ANOVA) is used to analyze more than two means. Because researchers often want to address complex questions, they use factorial designs in which they include several independent variables to study their combined effects. In reports of these studies, the results of an ANOVA are described in terms of main and interaction effects, and the presence of effects is inferred from the $p$ values (i.e., $p < \alpha$). For example, in a $3 \times 4$ ANOVA, a significant two-way interaction may be reported as $F(df_B, df_E) = 8.08$, $p = .02$. This interaction has 6 $df$s in the numerator, that is, $df_B = (3-1) \times (4-1) = 6$; and the $df_E$ in the denominator is determined by the sample size. A standardized effect size measure such as $\eta^2$ or partial $\eta^2$ may also be reported, but it is often not discussed by the authors (Cumming, 2012; Faulkner et al., 2008; Fritz, Morris, & Richler, 2012). This may be because researchers are used to thinking in terms of mean differences rather than standardized

effects. For significant interactions, follow-up tests (e.g., simple effects or post hoc tests) are conducted in an attempt to understand the meaning of these interactions. Although this procedure for analyzing factorial designs is commonly used, it has two main drawbacks: It is inefficient in extracting the effects of central interest from the factorial design, and its focus on significance testing of the null hypothesis ($H_0$) has limited informational value.

## Contrast Analysis in Factorial Designs

For factorial designs, it may seem difficult to extract meaningful effect sizes for main and interaction effects, which often have multiple *df*s in the ANOVA. For example, in a 3 × 4 factorial design, what would be a meaningful effect size for the two-way interaction with its 6 *df*s? The purpose of an effect size is to put an actual number on what we should be most interested in. For example, what is the mean coffee consumption (in cups per day) in Sweden? How much does mean Internet use (in hours per day) differ between old and young Swedes? How much does mean self-reported happiness change after a 3-week course in mindfulness? In all of these cases, our main interest is the means of different variables and how these means differ between conditions and groups. Critically, all of these differences are effect sizes. They are unstandardized effect sizes, because we express the effect in the original measurement units. In most cases, unstandardized effect sizes should be most informative, because researchers are familiar with their measurements and thus have a context in which to decide whether any change is relevant and noteworthy (Baguley, 2009). For example, researchers might be familiar with a particular rating scale for happiness (which ranges from −10 to +10). For them, a change of 3 points is meaningful and noteworthy whereas a change of 0.3 is not likely to be practically important. When effects are considered in terms of their original units, their meaning becomes obvious. Aside from this intuitive appeal, unstandardized effect sizes also avoid some issues with reliability, restricted range, and differences in experimental design that may distort standardized effect sizes (Baguley, 2009; Lenth, 2001).

Because factorial designs are about differences between means, effects of central interest can be captured by computing difference scores between means. This approach of comparing one set of means with another is called *contrast analysis* (Rosenthal & Rosnow, 1991; Rosenthal, Rosnow, & Rubin, 1999). The goal of any contrast analysis is to reduce the analyses of mean differences to the typical *t* tests that involve either only a single mean (one-sample *t* test) or the difference between two means (independent-samples *t* test or paired *t* test). At the maximum, only two means are compared. Therefore, any contrast analysis has only 1 *df*, the same as any *t* test. For example, if a study includes five repeated measurements (two recorded during baseline and three during the task), a contrast analysis of these data could simply compare the mean of the two baseline measurements with the mean of the three task measurements. Thus, a paired *t* test (of the mean baseline measures vs. the mean task measures) or a one-sample *t* test (of the difference scores between the

mean baseline and task measures) would directly indicate whether the mean ratings changed with the task.

In practice, contrast analysis tests for a specific pattern of means by assigning different weights (called *Lambda*) to the means (Howell, 2013; Rosenthal et al., 1999). In any contrast analysis, no more than two means are compared (as in a *t* test), but each may actually be a set of individual means that are weighted and then combined. In the example with five means, the actual weights for the five consecutive measurements would be (−0.5, −0.5, +1/3, +1/3, +1/3). In the first set, the two baseline means are weighted negatively; in the second set, the three task means are weighted positively. By applying these weights, we compute the average mean for each set and then compare the average across the two baseline measurements (negative set of means) with the average across the three task measurements (positive set of means). If the resulting value is positive, this means that the mean measurement increased from baseline to task.

Any data that can be analyzed with an ANOVA can also be analyzed by performing contrast analyses that test for specific effects. For simplicity, consider a 2 × 2 design (with 4 cells) that has two main effects. A column main effect is apparent if the mean value for the left column differs from the mean value for the right column. Of course, this difference has to be evaluated statistically, but any difference in column means would suggest a column main effect. Similarly, a row main effect is apparent if the mean value for the top row differs from that of the bottom row. The top left panel in Figure 1 shows a 2 × 2 design with some raw means. Assume that the dependent variable is a nausea rating that can range from 0 to 20. Because cell *a* in Figure 1A has a mean of 16.38, the means for the left and right columns and for the top and bottom rows differ from each other. This suggests the presence of main effects.

As shown in Figure 1B, the contrast for the column main effect weights the two left cells positively (i.e., each mean is multiplied by +0.5) and the two right cells negatively (i.e., each mean is multiplied by −0.5). These weights make intuitive sense: Take the mean of the left cells and subtract the mean of the right cells. Here, the result would be (for cells a through d) (+0.5 × 16.38) + (− 0.5 × 4.20) + (+0.5 × 4.22) + (−0.5 × 4.10) = 6.15. This is the *contrast score* for the column main effect. In general, the contrast score is obtained by multiplying each mean by its weight and then computing the sum. The contrast score is an (unstandardized) effect size measure that is in the same units as the dependent variable. Here, the column main effect reflects a nausea difference of 6.15, which means that on average, the left column has about a 6-point higher nausea rating than the right column. The row main effect can be computed similarly (see Figure 1C) and would give a nausea difference of 6.13.

Compared with main effects, interactions in factorial ANOVAs may suggest a conceptually more complicated explanation of the underlying mechanisms (Petty, Fabrigar, Wegener, & Priester, 1996). In the data, an interaction can simply be observed by noting whether the pattern (i.e., the differences between the means)

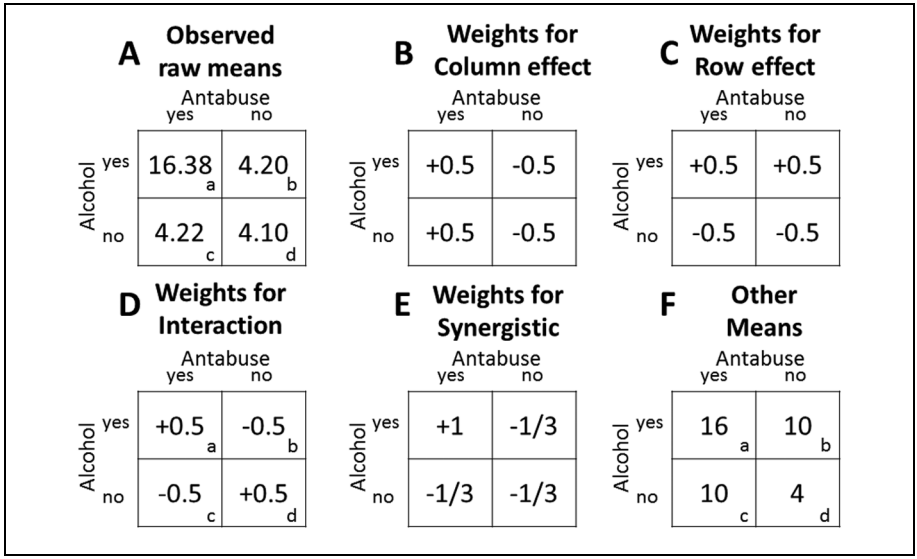



**Figure 1.** Means and contrast weights in a 2 × 2 analysis of variance.

differs from one row to the next (Howell, 2013). As shown in Figure 1A, the pattern of means within each row differs between rows. For the top row, the difference is 16.38 − 4.20 = 12.18, and for the bottom row, the difference is 4.22 − 4.10 = 0.12. This suggests an interaction (which would have to be tested statistically). Of course, this interaction is also apparent in that the pattern of means varies between columns. Figure 1D shows a set of contrast weights that tests this interaction. In a 2 × 2 ANOVA, this is the conventional, built-in interaction analysis. It tests whether the difference between cells *a* and *b* is larger than the difference between cells *c* and *d*. Alternatively, in terms of diagonals (Rosnow & Rosenthal, 1995), it tests whether the mean scores for one diagonal (*a* and *d*) are larger than the mean scores for the other diagonal (*b* and *c*).

When choosing weights, researchers should consider the following advice: Because a contrast analysis always tests for a specific direction, researchers should set the weights so that a positive contrast score fits their hypothesis. For example, if a set of baseline means is weighted negatively and another set of tasks means is weighted positively, then a positive contrast would indicate that performance increased with the task. Another consideration is that in a valid contrast, all weights must sum to zero. This ensures that the absence of an effect results in zero. Furthermore, it is often advisable to use a *standard set* (Howell, 2013). In a standard set, the sum of the positive weights is +1, and the sum of the negative weights is −1. Thus, the standard set compares directly two means with each other (although each mean is actually a combination of several means). For example, the set of weights for the interaction (Figure 1D) forms a standard set. However, it may be more

intuitive to view the interaction as the difference in the top row minus the difference in the bottom row. Thus, the weights for cells *a* to *d* might be +1, −1, −1, and +1, respectively. As a result, each mean is weighted by either +1 or −1, and two means are added together each for the positive side (sum = +2) and for the negative side (sum = −2). Although using this nonstandard set would capture the ''difference between differences'' approach (Rosnow & Rosenthal, 1995), the size of the main and interaction effects would not be directly comparable anymore because the contrast weights for the interaction sum to +2, whereas the contrast weights for each main effect sum to +1. To illustrate with another example, it would be valid to weight cells *a, b, c*, and *d* in Figure 1A with these respective weights: +5, −5, −5, and +5. Although the sum of the weights is zero, the contrast score would be 60.3. To convert this contrast score back to the units of the nausea scale, it is necessary to divide by the sum of the positive weights: 60.3/10 = 6.03. This example illustrates an important concept: For the size of two different contrasts to be comparable, the sums of their positive weights have to be identical.

Because the contrast score is determined by the set of weights, the reader needs to know what the weights are in order to make sense of the score. For example, in a 2 × 2 ANOVA, using the contrast weights (+0.5, −0.5, −0.5, +0.5) will result in a contrast score that is half the size of that produced by using the contrast weights (+1, −1, −1, +1). Notably, this arbitrary inflation of the contrast score is not an issue with NHST. In fact, the actual contrast weights are irrelevant for NHST as long as they identify the correct pattern and sum to zero. The reason is that if the positive contrast weights sum to something other than +1, the scores are simply multiplied by this value (and the mean is shifted), but NHST gives the same result (i.e., *p* value) regardless of what contrast weights are used. To illustrate, assume that the mean improvement in happiness after a mindfulness training program equals 5 (*SD* = 2), and a one-sample *t* test shows that $p = .013$. If the individual scores are multiplied by 2, the mean improvement will be 10 (*SD* = 4), but for the one-sample *t* test, the *p* value will be identical (.013).

Analyzing data from factorial designs with contrast analyses has several advantages that the typical factorial ANOVA does not (Bird, 2002; Rosenthal & Rosnow, 1991; Rosenthal et al., 1999). First, factorial ANOVAs often have multiple *df*s in the numerator and thus are nonspecific because they capture any difference among the means (e.g., 6 *df*s for the interaction in a 3 × 4 design). Contrast analyses are more specific because they are simply conducted as *t* tests (with 1 *df*) and thus compare no more than two (sets of) means. For example, a study may include two types of mindfulness therapies and a control group (i.e., 3 levels), and happiness ratings are obtained two times before therapy and two times after therapy (i.e., 4 levels). In this 3 × 4 design, a specific contrast (with 1 *df* rather than 6 *df*s) could capture how much the therapy effects on happiness (from before to after therapy) differ between the combined therapies and the control. To extract the therapy effect on happiness for each patient, a difference score is computed by subtracting the mean happiness ratings across the two measures after therapy from the mean happiness ratings across

the two measures before therapy (i.e., weight each measure before therapy by −0.5 and each measure after therapy by +0.5). Then, the two therapy groups can be combined (each weighted by +0.5) and compared to the control (weighted by −1). Accordingly, the weights over the four consecutive measurements would be (−0.5, −0.5, +0.5, +0.5) ×+0.5 for each of the therapy group and (−0.5, −0.5, +0.5, +0.5) ×−1 for the control group. Thus, the final weights over the four consecutive measurements would be (−0.25, −0.25, +0.25, +0.25) for each therapy group and (+0.5, +0.5, −0.5, −0.5) for the control group. The resulting contrast captures the extent to which therapy improved performance relative to the control conditions. Further examples are presented below and elsewhere (Furr & Rosenthal, 2003; Jaccard & Guilamo-Ramos, 2002).

Second, the ANOVA is mainly used to conduct significance testing on main effects and interactions, and these are reported as *F* values. However, *F* values reveal nothing about the direction of the effect, whereas contrast analyses do. For example, in a *t* test of cell *a* minus *b* in Figure 1A, the positive mean difference (and *t* value) would show that *a* is larger than *b*.

Third, because the ANOVA does not provide unstandardized effect sizes (i.e., mean differences), the pattern of means needs to be inspected to understand the direction and size of the effect. Contrast analyses are more informative because the contrast score captures the actual mean difference (i.e., the unstandardized effect size) for the contrast of interest.

The flexibility of contrast analysis becomes clear if the goal is to test hypotheses that are not captured by the typical ANOVA (Abelson & Prentice, 1997; Rosenthal & Rosnow, 1991; Rosenthal et al., 1999). Reconsider the data in Figure 1A in the following context: One independent variable refers to moderate alcohol consumption (yes or no), and the other independent variable refers to Antabuse intake (yes or no). In this situation, only the combination of alcohol and Antabuse should increase nausea (cell *a*). Accordingly, the main effects seem irrelevant to the hypothesis because neither should have an effect by itself; instead, only when combined should they show a *synergistic effect* (Rosnow & Rosenthal, 1996). The typical approach to testing this hypothesis is to run an ANOVA that extracts the conventional interaction (see Figure 1D), as well as the two main effects. The advantage of this approach is that revealing such effects is what standard ANOVAs are designed to do (Petty et al., 1996). They determine whether the data can be parsimoniously explained by two main effects, or whether it is actually necessary to attribute some of the findings to an interaction.

An alternative approach is the contrast set shown in Figure 1E (Rosnow & Rosenthal, 1996). In this contrast, the value associated with the combination of alcohol and Antabuse (cell *a*) is compared with the average value of the other three cells (*b, c*, and *d*). This contrast tests the hypothesis that only the combination of Antabuse and alcohol (cell *a*) should increase nausea. This synergistic set of weights seems to be a legitimate alternative to the conventional interaction contrast (for an example, see study 2 in Bushman & Anderson, 2009). The present data support this claim,

because the contrast score for the synergistic effect exceeds that of the conventional interaction effect. For the conventional interaction, the contrast score becomes (+0.5 × 16.38) + (−0.5 × 4.20) + (−0.5 × 4.22) + (+0.5 × 4.10) = 6.03, and for the synergistic effect, the contrast score becomes (+1 × 16.38) + (−1/3 × 4.20) + (−1/3 × 4.22) + (−1/3 × 4.10) = 12.21.

However, when choosing (a priori) about the contrast weights, it is important to consider whether the contrast weights may be sensitive also to effects that would not fit the predicted pattern (Abelson, 1996; Petty et al., 1996). Indeed, in the present situation, the conventional interaction contrast (see Figure 1D) and the synergistic contrast (see Figure 1E) are differentially sensitive to cell *d* (combination of no alcohol and no Antabuse). Whereas the conventional interaction contrast weights this cell positively, the synergistic contrast weights it negatively. Accordingly, if the mean in *d* is higher (and more similar to that in *a*), the conventional interaction contrast score will increase, whereas the synergistic contrast score will decrease. In contrast, if the mean in *d* is lower (and less similar to that in *a*), the conventional interaction contrast score will decrease, whereas the synergistic contrast score will increase.

To illustrate this problem, consider the data in Figure 1F. The conventional interaction contrast score would be (+0.5 × 16) + (−0.5 × 10) + (−0.5 × 10) + (+0.5 × 4) = 0. In fact, this pattern of means can be parsimoniously explained by two main effects. In contrast, the synergistic contrast score would be (+1 × 16) + (−1/3 × 10) + (−1/3 × 10) + (−1/3 × 4) = 8. Thus, the synergistic contrast score would provide evidence for a synergistic effect, but this is debatable here because a more parsimonious explanation for the observed pattern of means may be the presence of two main effects. Because a given set of contrast weights may be sensitive to patterns of means that are not relevant to the central hypothesis, it is preferable to test a specific, narrow hypothesis rather than a global, wide hypothesis to minimize confounding of the central hypothesis (Abelson, 1996).

Even if there is only a single contrast of central interest, several authors (Abelson & Prentice, 1997; Petty et al., 1996; Rosnow & Rosenthal, 1996) have suggested performing additional contrast analyses to determine whether there are valid alternative explanations for the data (for discussion, see Richter, 2016). For example, although a synergistic effect may be predicted, a conventional interaction contrast can also be computed to determine the effect size for this contrast relative to a synergistic contrast. Another strategy is to conduct *orthogonal contrasts* in an attempt to explain all of the variance among the group means (for an example, see study 2 in Bushman & Anderson, 2009). Contrasts are orthogonal if they are statistically independent: The results of one do not reveal anything about the results of the other. In any factorial design, there are as many orthogonal contrasts as there are cells minus 1. Importantly, any complete set of orthogonal contrasts explains all of the variance among cell means. For example, in a 3 × 5 design, there are (3 × 5) − 1 = 14 orthogonal contrasts. In an ANOVA, these contrasts would be grouped together into three sets: One main effect with 2 *df* (i.e., 3 − 1), one main effect with 4 *df*s (i.e., 5 − 1), and the interaction with 8 *df*s, that is, (3 − 1) × (5 − 1). Because these tests have multiple

*df*s, they are unspecific and can be viewed as an average of several orthogonal specific contrasts (with 1 *df*). In balanced designs, contrasts are orthogonal if the cross products of the weights sum to zero (Howell, 2013). That is, for each cell, multiply the weights of the two contrasts, and then determine whether their sum is zero. As an alternative strategy, simply correlate the weights. If the correlation is zero, the contrasts are orthogonal. For example, for the conventional interaction contrast and the synergistic contrast, the cross products are $+0.5 \times +1$ (for cell *a*), $-0.5 \times -1/3$, $-0.5 \times -1/3$, and $+0.5 \times -1/3$, and their sum is 2/3. Because the sum is not zero, these contrasts are not orthogonal.

A $2 \times 2$ ANOVA allows for $2 \times 2 - 1 = 3$ orthogonal contrasts (with 1 *df* each), so there are two remaining orthogonal contrasts aside from the synergistic effect. To define a contrast that is orthogonal to the synergistic effect, compare only among means that share the same sign in their weights for the synergistic contrast (i.e., consider only the weights that were all either positive or negative). Because cells *b* through *d* share the same sign, one orthogonal contrast could be (0, +0.5, +0.5, −1); this would determine whether cell *d* is smaller than the average of cells *b* and *c* (cell *a* is not included at all because its weight is zero). The other orthogonal contrast could be (0, +1, −1, 0); this would determine whether *b* is larger than *c*. Here, the data in Figure 1F would give a positive contrast score (i.e., $10 - 4 = 6$) for one orthogonal contrast (0, +0.5, +0.5, −1) and thus suggest that a synergistic effect is not sufficient to explain the pattern of means (in preview, a Bayesian *t* test will be conducted below for this contrast analysis).

The advantages of contrast analysis are greatest in complex factorial designs (Abelson & Prentice, 1997; Rosenthal & Rosnow, 1991; Rosenthal et al., 1999). For example, in a $2 \times 6$ design that tests two age groups at six different intervals, there is a set of 11 orthogonal contrasts ($2 \times 6 - 1 = 11$). In the typical ANOVA, these are divided into three sets: a 1-*df* contrast for the main effect of age, a 5-*df* contrast for the main effect of time, and a 5-*df* contrast for the interaction. Notably, the interaction is hard to interpret because it already combines five orthogonal contrasts. But, if the hypothesis is that performance increases linearly over time, and that this linear effect is stronger for the old than the young, then a specific set of contrast weights should be used that tests specifically this hypothesis and also provides an unstandardized measure of effect size (i.e., mean difference). One possible set of weights would be (−5/9, −3/9, −1/9, +1/9, +3/9, +5/9) for the old and (+5/9, +3/9, +1/9, −1/9, −3/9, −5/9) for the young. This set captures the idea that a linear increase over time (as reflected by the equal steps of 2/9ths between time points) should be larger for older subjects than for younger ones. Also, this specific set of weights gives the unstandardized effect size (i.e., mean difference) in support of the main hypothesis of group differences in the linear trend. Note that contrast weights for other polynomial trends (e.g., quadratic) are readily available in most statistics books (Howell, 2013). However, because the weights are often not listed as standard sets (i.e., the positive weights sum to more than +1), the contrast score will be inflated.

The particular statistical methods used to conduct contrast analysis differ depending on the design (between-subjects, within-subjects, or mixed). For mixed designs, it is advisable to reduce the within-subject variables to a single variable. For example, a measure that is taken repeatedly over time can be reduced to a single variable that captures the linear trend over time. Then, between-subjects contrasts can be conducted on this single variable. A supplementary file contains R scripts of examples for different designs (that fulfill assumptions for *t* tests) together with plots of the 95% CIs (see supplementary material). Note that the 95% CIs (explained below) refer only to the individual mean (*arelational* CI) and cannot be used directly to draw conclusions about differences between means (relational CI; Rouder & Morey, 2005). Relational CIs are not considered here because several alternatives have been proposed that vary with the experimental design and the goals of the study (Baguley, 2012; Cousineau & O'Brien, 2014; Franz & Loftus, 2012; Loftus & Masson, 1994; Noguchi & Marmolejo-Ramos, 2016; O'Brien & Cousineau, 2014; Pfister & Janczyk, 2013; Tryon, 2001).

## Confidence Intervals for Individual Contrast Means

When means are evaluated with *t* tests, confidence intervals can be computed easily by hand (Pfister & Janczyk, 2013) and are standard output in statistical software. Importantly, these analyses are valid only if all assumptions for *t* tests are met, as described above (for nonparametric approaches, see Efron & Tibshirani, 1994). In research papers, CIs are reported as follows: 95% CI [LL, UL], where LL is the lower limit and UL is the upper limit (American Psychological Association, 2009). Whereas NHST is mainly used to reject the null hypothesis (Gigerenzer, 2004; Perezgonzalez, 2015), the CI (Neyman, 1935, 1937) is more informative because it rejects a range of hypotheses, namely all hypotheses that are outside of the CI (Cumming, 2014; Cumming & Finch, 2005; Dienes, 2008).

Assume that a study asks Swedes to rate their happiness (on a scale from −10 to +10), and that the mean happiness rating is exactly zero. (Indeed, the Swedish word *lagom* may capture this state of happiness: it is balanced, neither too high nor too low.) If the obtained 95% CI ranges between −2.0 and +2.0, then the hypotheses are to be rejected that the true mean is less than −2.0 or greater than +2.0, but all hypotheses inside the CI are consistent with our data.

Figure 2A uses the present example to illustrate the close relationship between CI (Neyman, 1935, 1937) and significance testing in NHST (Perezgonzalez, 2015). The left side shows the present mean (of zero) and the 95% CI, and the right side shows the sampling distribution for the same data. As shown, if the 95% CI is centered on zero, it is identical to that of the 95% nonrejection area for the sampling distribution (see area marked by dotted bracket in Figure 2A). At a two-tailed $\alpha = 5\%$, the 95% nonrejection area includes the 95% of sample means that would be expected to fall within this range (here, between −2 and +2). If the actual sample mean falls inside this nonrejection area, it is not significant (i.e., not significantly different from zero).

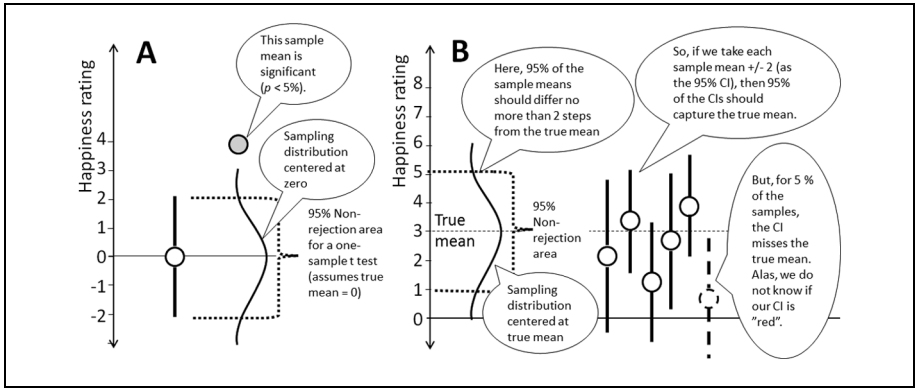



**Figure 2.** (A) Left: The actual mean happiness rating (of zero) with the 95% CI. Right: The non-rejection area (dotted) for a one-sample *t* test (at 2-tailed $\alpha$ = 5%). If a sample mean falls outside of this area, it is significantly different from zero. Note that the CI around an actual sample mean of zero (left) is identical to the nonrejection area in a one-sample *t* test (which assumes that mean = 0). (B) Across many studies, 95% of the sample means should fall within the nonrejection area (dotted) defined by the true mean and the true standard deviation. Here, several sample means (and the 95% CIs) are shown along the *x* axis. In the long run, 95% of these sample means should fall no more than $\pm$2 steps from the true mean. Because we do not know these true values, we compute CIs with a similar interval (about $\pm$2 steps). Note that the CIs differ slightly because we need to estimate the true standard deviation from each sample. Critically, the distance from the true mean to the sample means (left) is symmetric to the distance from the sample means to the true mean (right). Thus, 95% of all CIs would capture the true mean, but 5% would not (the dashed CI). Unfortunately, for individual studies, we do not know whether the CI misses the true mean (i.e., the dashed CI in the figure), that is, whether the CI is "red," as illustrated nicely by the ESCI software (Cumming, 2012).

In contrast, if the sample mean falls outside this nonrejection area, it is significant. Critically, whereas the nonrejection area is always centered on zero, the (numerically identical) CI is always centered on the sample mean. So, if the 95% CI around the sample mean does not include zero, then the (numerically identical) nonrejection area around zero does not include the sample mean either. Accordingly, if the 95% CI does not include zero, then the sample mean is significantly different from zero at a two-tailed $\alpha$ of 5%. In contrast, if the 95% CI includes zero, then the sample mean is not significantly different from zero at a two-tailed $\alpha$ of 5%. In general, a one-sample *t* test of the sample mean (at a two-tailed $\alpha$ of 5%) produces a significant result for any test value outside the 95% CI and a nonsignificant result for any test value inside the 95% CI. As a consequence, CIs can be used (by stealth) to conduct NHST (Cumming & Finch, 2005).

In many situations, CIs are more informative than the results of NHST (Cumming, 2012, 2014; Cumming, Fidler, Kalinowski, & Lai, 2012; Finch & Cumming, 2009;

Nakagawa & Cuthill, 2007) and may reduce the risk of several different types of misinterpretations (Hoekstra, Johnson, & Kiers, 2012; McCormack, Vandermeer, & Allan, 2013). Furthermore, the practical importance of a finding (Kirk, 1996) can be evaluated if researchers can define a null range, that is, a range of effect sizes that has no practical importance (Dienes, 2014; Kalinowski & Fidler, 2010; Tryon, 2001). To illustrate, Figure 3 shows different CIs together with NHST results and the interpretation of CIs. For the studies *a* and *b*, NHST may simply state that a one-sample *t* test was not significant (exact $p = .99$). Having identical *p* values, the two results seem similar from an NHST perspective. However, the CIs are more informative than NHST. Because the CI in *a* is wide, it is consistent with a wide range of hypotheses (i.e., low exclusion). Further, researchers can probably select a range of hypothetical effect sizes that they would consider practically or theoretically unimportant. For example, a mean difference of less than about 1.5 points might be negligible (this null range is the gray area in the figure). Because the CI in *a* overlaps with values outside of this null range, researchers would conclude that the relevance of the result is unclear. Therefore, not much is learned from *a*, and the researchers might decide to run a new study with a larger sample size (with more power), because a large sample often leads to a small CI (Wagenmakers, Verhagen, et al., 2015). Compared with the CI in *a*, the CI in *b* is small and is thus consistent with a small range of possible hypotheses. Because the CI in *b* falls completely within the null range, the estimated effect has no practical relevance. The researchers might decide to stop here, because the effect size is of no practical importance regardless of the direction of the effect.

Similarly, for the studies *c* through *f*, NHST would suggest that all means are significant ($p < .05$). In fact, in terms of direction, the CIs for all studies suggest a positive effect (greater than zero). For studies *c* and *e*, the wide CIs mean that a wide range of hypotheses are retained (low exclusion), whereas for studies *d* and *f*, the narrow CIs mean that a narrow range of hypotheses are retained (high exclusion). In terms of relevance, *c* is unclear (because the CI overlaps with the null range), *d* has no practical relevance (because the CI falls within the null range), and *e* and *f* are important (because the CIs fall outside of the null range). Both *c* and *e* encourage running a new study with adequate power to determine the relevance of the effect (for *c*) or the size of the effect (for *e*). In contrast, both *d* and *f* encourage stopping because the effect has no practical relevance (for *d*) or has relevance in a narrow range of possible effects (for *f*); thus, there is no need for more data in either case. In sum, these examples illustrate that because CIs refer to a range of values rather than a single value (i.e., zero, the value associated with $H_0$), CIs are generally more informative than NHST results. Thus, CIs reduce dichotomous thinking and require researchers to judge the importance of an effect size in its context (Cumming, 2012, 2014; McCormack et al., 2013).

When interpreting a CI, researchers need to be aware of several common misconceptions that can lead to misinterpretation (Cumming, 2012; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). First, it is common but incorrect to conclude that we can be 95% confident

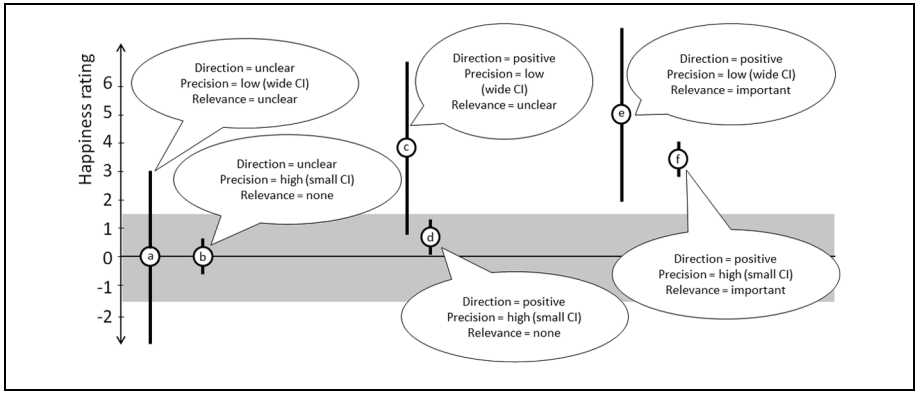



**Figure 3.** Illustration showing that confidence intervals (CIs) are more informative than null hypothesis significance testing (NHST). For each study, NHST was used to determine whether the mean differed significantly from zero. The NHST found that the results of studies a and b were not significant, whereas the results of the other studies were significant ($p < .05$). This information is less informative than what can be learned directly from the CIs (see comments). Note that effect sizes in the gray area are considered practically unimportant (null range).

(or sure) that the true value lies between the lower and upper limits of the obtained 95% CI. This interpretation is incorrect because the 95% confidence level does not refer to the CI of a single study but to the CIs across many (hypothetical) study replications. In general, the ''confidence'' in CI refers to the method of generating CIs and not to a single interval, in accordance with the repeated sampling principle. Thus, the confidence levels in CIs refer only to long-time rates over many studies, similar to the critical $\alpha$ in NHST (Perezgonzalez, 2015). For example, a 95% confidence level means that if a study were to be repeated many times, for 95% of these replications, the CIs will capture the true mean, whereas for the remaining 5% of these replications, the CIs will not. This is illustrated in Figure 2B. Assume that in the population, the mean happiness rating is 3. Further assume that if we conducted many study replications, 95% of sample means (at a given *n*) might vary between +1 and +5 (see sampling distribution). However, because we do not know the truth, we work backward: We compute the CI and center it on the study mean. Because for each study, we estimate the CI from the sample *SD*, the width of the CI will vary somewhat between studies. Critically, the distance from the true mean to the sample means is symmetric to the distance from the sample means to the true mean. So, if 95% of the sample means vary no more than about $\pm 2$ steps from the true mean = 3 (in this case, the range between +1 and +5 in happiness ratings), then 95% of the CIs with a width of about $\pm 2$ steps around the sample means will capture the true mean. However, for 5% of the studies, the CIs would not capture the true mean; these CIs would be ''red'' (Cumming, 2012, 2014). Because we conducted only a single study

and do not know whether our CI is red (i.e., misses the true mean), we cannot have any particular confidence in the CI of a single study: Either it includes the true mean or it does not. To conclude, the confidence level must not be interpreted literally as the probability that the obtained CI includes the true mean (Morey et al., 2016). Instead, it refers to the procedure: In the long run, a certain proportion of CIs (e.g., 95%) will capture the true mean. Therefore, one should place confidence in the procedure, not in the specific interval.

Second, a tempting interpretation of CIs may be that values close to the center of the CI (i.e., the mean effect) are more likely to be true than values close to the edges of the CI (Cumming, 2014; Wilkinson, 1999). For example, for the hypothetical study on happiness in Swedes, let us say that mean happiness is 0, with a 95% CI of [−2, +2]. Does this mean that the true mean is more likely to be 0 rather than to be either −2 or +2? This interpretation is incorrect according to CI theory (Neyman, 1935, 1937). Instead, researchers need to assume that all values within the CI are equally likely to be true (Dienes, 2008; Morey et al., 2016).

Third, researchers may be tempted to believe that their obtained CI includes the true value. However, similar to the Neyman–Pearson approach (Goodman, 1999a), CI theory is only a decision procedure (inductive behavior, Lew, 2012) that encourages researchers to behave as if the true value were included in the CI. Nonetheless, any post-data conclusions are not permitted because CI theory is only a pre-data theory (Feinstein, 1998; Morey et al., 2016). CI theory does not allow one to say anything about where the true value is likely to lie given a particular CI; this (posterior) probability about which hypotheses are true given the data are only permitted in Bayesian analyses (Morey et al., 2016), as explained further below.

In sum, CIs of mean differences are more informative than *p* values because CIs refer to a range of values rather than only the value associated with the null hypothesis (i.e., zero; see Figure 3). However, CIs have theoretical limitations and are often misinterpreted. A 95% confidence level does not indicate that we can be 95% certain that the CI of an individual study captures the true value. Instead, the 95% confidence level indicates that if we repeated the same study many times, 95% of the individual CIs would capture the true value. Furthermore, CIs do not allow postdata inferences about the true value given the data. For example, it is incorrect to conclude that values closer to either the upper or lower limits of the CIs are less likely to be true than values in the middle of the CIs.

## Beyond Confidence Intervals

In contrast to CIs, likelihood intervals identify the values that are most consistent with the results (Glover & Dixon, 2004; Johansson, 2011). Of course, we may never know the truth, but the main tenet of likelihood is that the results support those values that best predict the results (Dienes, 2008). The likelihood is the probability of obtaining the data in question given a particular (population) value. For example, in Figure 2A, the likelihood of a population value of 0 is the height of the sampling distribution at a

mean of 0. To determine the likelihood of values different from 0, we need to vary these values (from negative values to positive values), compute each sampling distribution, and extract the height of each sampling distribution for our observed mean of 0. We would then combine these values into a likelihood function that shows how the likelihood changes for various population values. Of course, in the present example, the likelihood is maximal for a population value of 0 (because this value best predicts the observed mean of 0), whereas the likelihood decreases for values that differ from 0.

Because the likelihood varies gradually, likelihood intervals are commonly selected to capture a reasonable range of likely values. For the commonly used 1/8 likelihood intervals, the value with the highest likelihood (at the center of the interval) is 8 times more likely than the value at either end of the likelihood interval. So, with likelihood intervals, the value that has maximum likelihood is probably the true value. But other values are not dismissed if they are also likely (i.e., the other values in the likelihood interval) unless their likelihood is more than 8 times less likely than the value that has maximum likelihood.

For data that meet the assumptions for *t* tests, the 1/8 likelihood intervals are identical to 96% CIs (Dienes, 2008; Morey et al., 2016). Because 95% and 96% are very close, it makes little practical difference whether the 1/8 likelihood interval or the 95% CI is reported, as requested by current guidelines (American Psychological Association, 2009; Wilkinson, 1999).

Unlike typical CIs, likelihood intervals do not need to be adjusted for multiple comparisons, stopping rules, or a priori versus post hoc analyses (Dienes, 2014). For example, if a study includes many treatments, and the data from each treatment group can be compared with those from a control group, these multiple comparisons would require adjustment of the CIs by adjusting the $\alpha$ value (e.g., from 95% to 97.5% after Bonferroni correction). In contrast, likelihood intervals do not need to be adjusted, because the result of one comparison is what it is regardless of whether you also conducted other comparisons.

Finally, the Bayesian perspective acknowledges that even though we do not know the true value, we have beliefs about plausible values (Andraszewicz et al., 2015; Goodman, 1999b; Kruschke, 2010; van de Schoot et al., 2014; Wagenmakers, Morey, & Lee, 2016; Wagenmakers et al., 2014; Zyphur & Oswald, 2015). For example, if a pregnancy test is positive, then this result is immediately dismissed as a false positive if the tested individual is a man. Similarly, if somebody conducts a study on extrasensory perception and reports finding strong effects, these results might nudge your belief toward the positive, but your final belief in extrasensory perception is also determined by your prior beliefs (Wagenmakers et al., 2011). On the one hand, if you were highly skeptical before, you will also be skeptical afterward. On the other hand, if you were a strong believer before, you will be even more convinced afterward. Therefore, the results of a study do not occur in a vacuum but are evaluated in the context of one's prior beliefs. The Bayesian framework assumes that it is possible to assign probabilities to these beliefs, and that the beliefs (as probabilities) adhere to the axioms of probability. As such, Bayesian inference provides an internally

consistent framework that captures the way that beliefs change about which values seem more likely to be true than others.

In Bayesian inference, credible intervals represent the combination of prior beliefs and likelihood into final beliefs (Dienes, 2011). For example, if the 95% credible interval is [−2, +2], this means that we are 95% certain that the true value is between −2 and +2 (i.e., P(H|D)). Likelihood (as described above) tells us all we need to know about the data, but the final beliefs (i.e., the posterior beliefs) also take into account the prior beliefs and combine them with the likelihood (i.e., the present results). Critically, NHST, CIs, and likelihood refer only to the probability of the data given a particular value (e.g., $P(D|H_0)$), and it is wishful thinking to believe that NHST provides the probability for a value given the data (Gigerenzer, 2004; Morey et al., 2016). Instead, only the Bayesian inference is a legitimate approach to capturing our final beliefs about what values are plausible (e.g., $P(H_0|D)$) after updating our prior beliefs with the data (Morey et al., 2016).

If you do not have any prior belief whatsoever, this can be described by a flat or uniform prior distribution (van de Schoot et al., 2014). As a consequence, for data that are analyzed with *t* tests, the credible interval is numerically identical to the likelihood interval (Morey et al., 2016). However, researchers typically have some beliefs about what effects are reasonable (Dienes, 2014, 2016). For example, imagine that your study uses the aforementioned happiness scale (−10 to +10) to evaluate the effects of a treatment. The maximum effect would have to be +20, because nobody's score could improve by more than 20 points. Yet, a more reasonable expectation for the true improvement effect is that it is small rather than large. This definition of the prior belief is subjective, but it is open to scrutiny by other researchers. If researchers disagree, different prior beliefs can be compared (in robustness tests) to determine whether the results are similar for different prior beliefs (van de Schoot et al., 2014). As for likelihood intervals, there is no need to adjust credible intervals for multiple comparisons, stopping rules, or a priori versus post hoc analyses (Dienes, 2008, 2014). In a situation with multiple comparisons, a Bayesian analysis would combine all of the evidence. Although some evidence may support a theory, the nonsupporting evidence is also taken into account in the weighing of all evidence (Dienes, 2016).

Before trying to estimate the size of an effect (CIs, likelihood intervals, or credible intervals), it is reasonable to determine whether there is an effect at all (Morey, Rouder, Verhagen, & Wagenmakers, 2014; Wagenmakers et al., 2016). Many researchers use NHST to try to answer this question. However, Bayesian hypothesis testing is needed if you want to compare different theories (e.g., $H_0$ vs. $H_1$) or argue that there is no effect (Dienes, 2016; Gallistel, 2009; Kruschke, 2010; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers et al., 2014; Wagenmakers et al., 2016). In Bayesian hypothesis testing, the *Bayes Factor* (BF) is the ratio of the likelihoods of two different values or hypotheses (Kass & Raftery, 1995). It states how much more likely the data are given one hypothesis rather than the other hypothesis. For example, $BF_{01} = P(D|H_0)$ divided by $P(D|H_1)$. If both $H_0$ and $H_1$ predict the data similarly, then $BF_{01} = 1$. As $BF_{01}$ increases to greater than 1, the data support

$H_0$ more strongly than $H_1$. As $BF_{01}$ decreases below 1, the data support $H_1$ more strongly than $H_0$.

The BF is helpful in interpreting results that are considered nonsignificant ($p > .05$) in NHST (Dienes, 2014, 2016). Because NHST mixes at least two approaches, one approach (Fisher) would ignore the results (fail to reject the $H_0$), whereas the other approach (Neyman–Pearson) would accept the $H_0$ if power were adequate (Hubbard, 2004; Perezgonzalez, 2015). Although CIs can be used to infer statistical equivalence (Colegrave & Ruxton, 2003; Dienes, 2014; Tryon, 2001; Walker & Nowacki, 2011), the BF can determine whether the data support the $H_0$ or are insensitive in distinguishing between the $H_0$ and an alternative hypothesis (Dienes, 2014, 2016; Wetzels et al., 2011).

To illustrate, consider the discussion of synergistic versus interaction effects (Figure 1). For the sake of simplicity, let us assume that in a repeated-measures design ($n = 40$), each individual was tested in all four combinations of alcohol and Antabuse, and the means are as shown in Figure 1. As discussed above, this pattern supports a synergistic effect. However, a critic might argue that because the synergistic effect lumps together three conditions (*b, c*, and *d*), it is possible that only one ingredient (either alcohol or Antabuse) might already increase nausea relative to the control condition (of neither alcohol nor Antabuse). In fact, mean nausea ratings tended to be higher for alcohol and Antabuse alone (4.20 and 4.22) than for neither (4.10). In a contrast analysis with the weights (0, +0.5, +0.5, −1) for cells *a* thru *d* (see Figure 1), the mean difference is 0.11; thus, it appears that alcohol and Antabuse alone increased nausea. The free software *JASP* (Love et al., 2015) was used to analyze these data (the supplement contains the raw data and the complete output). In a one-sample *t* test, $t(39) = 0.38$, $p = .70$, 95% CI [−0.47, +0.69]. The CI (or likelihood interval) was not negligibly small, and the result is consistent with two alternative possibilities: (1) that alcohol and Antabuse alone have no effect ($H_0$) and (2) that the data are too insensitive to distinguish between either no effects ($H_0$) or effects of alcohol and Antabuse alone ($H_1$).

Figure 4 shows excerpts of the *JASP* output of a default Bayesian one-sample *t* test. Because the effect of alcohol and Antabuse alone should be stronger than the effect of the control condition (of neither alcohol nor Antabuse), the prediction is directional; that is, the true effect is believed to be positive (denoted as $H_+$ in *JASP*). Further, the true effect is believed to be small rather than large (Cauchy prior width = 0.707). In Figure 4, the 95% credible interval, which is [0.006, 0.382], is standardized and needs to be multiplied by the standard deviation (1.811, see supplement) to obtain the unstandardized 95% credible interval [0.011, 0.681]. This interval is only positive because of our prior belief that alcohol and Antabuse alone can only make nausea ratings higher than those in the control condition. In the Bayesian hypothesis test, the $BF_{0+} = 4.24$ means that the data are 4 times more likely under the $H_0$ than under the $H_+$ (that presupposes a small positive effect rather than a large positive effect). The fact that $BF_{0+} > 3$ constitutes moderate support for the $H_0$. If the alternative

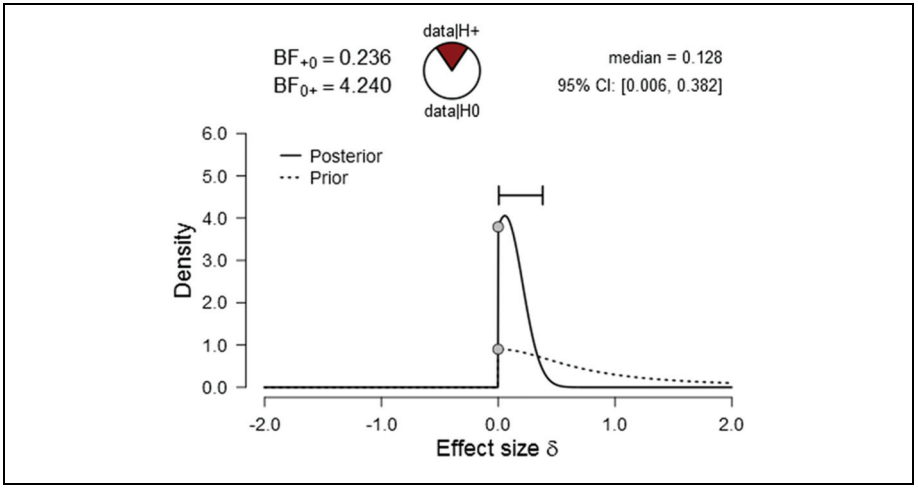



**Figure 4.** Excerpts from the output of a Bayesian one-sample *t* test in *JASP*. The Bayes Factor ($BF_{0+}$) shows that the present data are 4.2 times more likely under the $H_0$ than under the alternative hypothesis ($H_+$). In the original pie chart, the area data|$H_+$ is shown in red. In the line chart, the dotted line shows the prior distribution that captures the a priori belief that the true effect is positive and small rather than large. The solid line shows the posterior distribution that captures the beliefs after the data are incorporated. In the line chart, the $BF_{0+}$ is the likelihood for $H_0$ under the posterior (upper gray circle) divided by the likelihood for $H_0$ under the prior (lower gray circle, Wagenmakers et al., 2010). The whiskers above the distributions show the 95% credible interval for the standardized effect.

hypothesis ($BF_{+0}$) is of greater interest, it can be computed as the inverse of $BF_{0+}$ = 0.23. The $BF_{+0}$ indicates how much more likely the data are under $H_+$ than under $H_0$.

The *JASP* software nicely visualizes these results in two ways (Wagenmakers et al., 2016). In a pie chart, the red area shows the relative evidence for the data given $H_+$, and the white area shows the relative evidence for the data given $H_0$. For a $BF_{0+}$ = 1, half of the pie would be red and the other half white; for a $BF_{0+}$ = 2, two thirds would be white and one third red; and for the obtained $BF_{0+}$ = 4.24, four fifths are white, and one fifth is red. This pie chart can be understood intuitively in terms of a dartboard (E.-J. Wagenmakers, personal communication, February 5, 2016). Imagine that you, while blindfolded, throw a dart and hit the dartboard, which is rotating. Once you remove the blindfold, the less surprised you would be to hit the white area, the more you believe in the $H_0$ relative to the $H_1$. Stated differently, the more surprised you would be that you hit the red area, the less you believe that $H_1$ is a better explanation for the obtained results than is $H_0$. Accordingly, for the obtained $BF_{0+}$ = 4.24, four fifths of the dartboard is white (evidence for $H_0$), and one fifth is red (evidence for $H_1$). If you imagine that you throw a dart at this board, you would probably not be very surprised to hit the white area. This means the present data shift your

belief toward the possibility that alcohol and Antabuse alone do not make nausea ratings higher than those in the control condition.

Below the pie chart, a line plot shows the BF more formally in terms of the height of the posterior distribution at 0 relative to the height of the prior distribution at 0 (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). The height of the prior distribution at 0 shows the support for $H_0$ before the data are taken into account, and the height of the posterior distribution at 0 shows the support for the $H_0$ after the data are taken into account. Because in Figure 4 the height of the posterior distribution at 0 is larger (by a factor of 4.24) after the data are taken into account, the data provide stronger support for $H_0$ than $H_+$. Accordingly, the data shift our beliefs by a factor of 4.24 toward the $H_0$.

In Bayesian analysis, the final sample size does not have to be preplanned (as for NHST); instead, data collection continues until the researcher judges that the evidence is sufficient (Dienes, 2014; Wagenmakers et al., 2014; Wagenmakers et al., 2016). For example, for Adam Sandler movies, is there a zero correlation between movie quality and box office success (Wagenmakers et al., 2016)? It is not possible to preplan a certain number of movies for Adam Sandler; instead, the evidence is updated every time a new movie is released. *JASP* provides several additional figures that plot the development of evidence as data accumulates and the robustness of the Bayesian analyses to variations in the prior beliefs (Love et al., 2015). The supplement includes this additional output for the example in Figure 4. Furthermore, instructive examples based on actual research questions are available (Dienes, 2014, 2016; Rotteveel et al., 2015; Wagenmakers, Beek, et al., 2015).

In sum, simple and free software tools are now readily available for conducting Bayesian *t* tests and obtaining likelihood and credible intervals for contrast analyses of mean differences that isolate effects of central interest in factorial designs.

## Conclusion

In the analysis of mean differences in factorial designs, contrast analysis is helpful because it can test specific questions of central interest in factorial designs. Contrast analysis weighs several means and combines them into one or two sets that can be tested with *t* tests. The effect size is simply the difference between means. The CIs for such contrasts inform directly about direction, hypotheses exclusion, and the relevance of the effects of interest. However, any interpretation in terms of precision or likelihood requires the use of likelihood intervals or credible intervals (Bayesian). Because these intervals and even a Bayesian *t* test can be easily obtained with free software, researchers are encouraged to go beyond CIs and report the results of their contrast analyses in terms of these intervals.

### Acknowledgments

The authors thank Marco Tullio Liuzza and Anders Sand for helpful discussions, and the organizers of *Bayes at Lund 2016* for an inspirational workshop. Stefan Wiens thanks Robert M.

Kelsey for introducing him to contrast analyses during his graduate studies. We are also grateful for comments from Professors Eric-Jan Wagenmakers and Dennis Cousineau, as well as an anonymous reviewer. We also thank Stephen N. Palmer for editing.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funded in part by grant 2015-01181 from the Swedish Research Council.

## Supplementary Material

Supplementary material for this article is available online.

## References

Abelson, R. P. (1996). Vulnerability of contrast tests to simpler interpretations: An addendum to Rosnow and Rosenthal. *Psychological Science*, *7*, 242-246. doi:10.1111/j.1467-9280.1996.tb00367.x

Abelson, R. P., & Prentice, D. A. (1997). Contrast tests of interaction hypotheses. *Psychological Methods*, *2*, 315-328. doi:10.1037/1082-989x.2.4.315

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*, 521-543. doi:10.1177/0149206314560412

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603-617. doi:10.1348/000712608x377117

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158-175. doi:10.3758/s13428-011-0123-7

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement*, *62*, 197-226. doi:10.1177/0013164402062002001

Bushman, B. J., & Anderson, C. A. (2009). Comfortably numb: Desensitizing effects of violent media on helping others. *Psychological Science*, *20*, 273-277. doi:10.1111/j.1467-9280.2009.02287.x

Chavalarias, D., Wallach, J., Li, A., & Ioannidis, J. A. (2016). Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA*, *315*, 1141-1148. doi:10.1001/jama.2016.1952

Colegrave, N., & Ruxton, G. D. (2003). Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology*, *14*, 446-447. doi:10.1093/beheco/14.3.446

Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: a comment on Baguley (2012). *Behavior Research Methods*, *46*, 1-3. doi:10.3758/s13428-013-0441-z

Cumming, G. (2008). Replication and p intervals p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286-300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

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

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie–Journal of Psychology*, *217*, 15-26. doi:10.1027/0044-3409 .217.1.15

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*, 138-146. doi: 10.1111/j.1742-9536.2011.00037.x

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170-180. doi:10.1037/0003-066x.60.2.170

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*, 217-227. doi:10.1037/1082-989X.11.3.217

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York, NY: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274-290. doi:10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*:781. doi:10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78-89. doi:10.1016/j.jmp.2015.10.003

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour Research and Therapy*, *46*, 270-281. doi: 10.1016/j.brat.2007.12.001

Feinstein, A. R. (1998). P-values and confidence intervals: Two sides of the same unsatisfactory coin. *Journal of Clinical Epidemiology*, *51*, 355-360. doi: 10.1016/s0895-4356(97)00295-3

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values. *Zeitschrift für Psychologie–Journal of Psychology*, *217*, 27-37. doi:10.1027/0044-3409 .217.1.27

Finch, S., & Cumming, G. (2009). Putting research in context: Understanding confidence intervals from one or more studies. *Journal of Pediatric Psychology*, *34*, 903-916. doi: 10.1093/jpepsy/jsn118

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic Bulletin & Review*, *19*, 395-404. doi:10.3758/s13423-012-0230-1

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology. General*, *141*, 2-18. doi:10.1037/a0024338

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3-12. doi:10.1177/1088868313507536

Furr, R. M., & Rosenthal, R. (2003). Evaluating theories efficiently: The nuts and bolts of contrast analysis. *Understanding Statistics*, *2*, 45-67.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439-453. doi:10.1037/a0015251

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587-606. doi: 10.1016/j.socec.2004.09.033

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*, 791-806. doi:10.3758/bf03196706

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, *130*, 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, *130*, 1005-1013.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature Methods*, *12*, 179-185. doi: 10.1038/nmeth.3288

Hoekstra, R., Johnson, A., & Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement*, *72*, 1039-1052. doi:10.1177/0013164412450297

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157-1164. doi:10.3758/s13423-013-0572-3

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth/ Cengage Learning.

Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between p's and alpha's in psychological research. *Theory & Psychology*, *14*, 295-327.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3-7. doi: 10.1111/j.1467-9280.1997.tb00534.x

Ivarsson, A., Andersen, M. B., Stenling, A., Johnson, U., & Lindwall, M. (2015). Things we still haven't learned (so far). *Journal of Sport & Exercise Psychology*, *37*, 449-461. doi: 10.1123/jsep.2015-0015

Jaccard, J., & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, *31*, 130-146. doi:10.1207/153744202753441747

Johansson, T. (2011). Hail the impossible: P-values, evidence, and likelihood. *Scandinavian Journal of Psychology*, *52*, 113-125. doi:10.1111/j.1467-9450.2010.00852.x

Kalinowski, P., & Fidler, F. (2010). Interpreting significance: The differences between statistical significance, effect size, and practical importance. *Newborn and Infant Nursing Reviews*, *10*, 50-54. doi:10.1053/j.nainr.2009.12.007

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795. doi:10.1080/01621459.1995.10476572

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

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral science* (2nd ed.). Washington, DC: American Psychological Association.

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

Lai, J., Fidler, F., & Cumming, G. (2012). Subjective p intervals: Researchers underestimate the variability of p values over replication. *Methodology–European Journal of Research Methods for the Behavioral and Social Sciences*, *8*, 51-62. doi:10.1027/1614-2241/a000037

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187-193. doi:10.1198/000313001317098149

Lew, M. J. (2012). Bad statistical practice in pharmacology (and other basic biomedical disciplines): You probably don't know p. *British Journal of Pharmacology*, *166*, 1559-1567. doi:10.1111/j.1476-5381.2012.01931.x

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence-intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476-490. doi:10.3758/BF03210951

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., … Wagenmakers, E.-J. (2015). JASP (Version 0.7.5) [Computer software].

McCormack, J., Vandermeer, B., & Allan, G. M. (2013). How confidence intervals become confusion intervals. *BMC Medical Research Methodology*, *13*, 134. doi:10.1186/1471-2288-13-134

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103-123. doi:10.3758/s13423-015-0947-8

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289-1290. doi:10.1177/0956797614525969

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*, 591-605. doi: 10.1111/j.1469-185X.2007.00027.x

Neyman, J. (1935). On the problem of confidence intervals. *Annals of Mathematical Statistics*, *6*, 111-116.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *236*, 333-380. doi:10.1098/rsta.1937.0005

Noguchi, K., & Marmolejo-Ramos, F. (2016). Assessing equality of means using the overlap of range-preserving confidence intervals. *The American Statistician*. Advance online publication. doi:10.1080/00031305.2016.1200487

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*, 150-152.

Nuzzo, R. (2015). How scientists fool themselves-and how they can stop. *Nature*, *526*, 182-185.

O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *Quantitative Methods for Psychology*, *10*, 56-67.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, *6*, 223. doi:10.3389/fpsyg.2015.00223

Petty, R. E., Fabrigar, L. R., Wegener, D. T., & Priester, J. R. (1996). Understanding data when interactions are present or hypothesized. *Psychological Science*, *7*, 247-252. doi: 10.1111/j.1467-9280.1996.tb00368.x

Pfister, R., & Janczyk, M. (2013). Confidence intervals for two sample means: Calculation, interpretation, and a few simple rules. *Advances in Cognitive Psychology*, *9*, 74-80. doi: 10.2478/v10053-008-0133-x

Richter, M. (2016). Residual tests in the analysis of planned contrasts: Problems and solutions. *Psychological Methods*, *21*, 112-120. doi:10.1037/met0000044

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York, NY: McGraw-Hill.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (1999). *Contrasts and effect sizes in behavioral research*. Cambridge, England: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (1995). Some things you learn aren't so—Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological Science*, *6*, 3-9. doi: 10.1111/j.1467-9280.1995.tb00297.x

Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interactions redux: Five easy pieces. *Psychological Science*, *7*, 253-257. doi:10.1111/j.1467-9280.1996.tb00369.x

Rotteveel, M., Gierholz, A., Koch, G., van Aalst, C., Pinto, Y., Matzke, D., … Wagenmakers, E.-J. (2015). On the automatic link between affect and tendencies to approach and avoid: Chen and Bargh (1999) revisited. *Frontiers in Psychology*, *6*, 335. doi: 10.3389/fpsyg.2015.00335

Rouder, J. N., & Morey, R. D. (2005). Relational and arelational confidence intervals: A comment on Fidler, Thomasow, Cumming, Finch, and Leeman (2004). *Psychological Science*, *16*, 77-79. doi:10.1111/j.0956-7976.2005.00783.x

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

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*, 245. doi:10.3389/fpsyg.2015.00245

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*, 572-582. doi:10.1037/a0034177

Thompson, B. (2002). ''Statistical,'' ''practical,'' and ''clinical'': How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, *80*, 64-71.

Torgerson, W. S. (1965). Multidimensional-scaling of similarity. *Psychometrika*, *30*, 379-393. doi:10.1007/bf02289530

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology*, *36*, 1-2. doi: 10.1080/01973533.2014.865505

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1-2. doi:10.1080/01973533.2015.1012991

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371-386. doi: 10.1037//1082-989x.6.4.371

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after NHST: How to describe your data without ''p-ing'' everywhere. *Basic and Applied Social Psychology*, *37*, 260-273. doi: 10.1080/01973533.2015.1060240

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*, 842-860. doi:10.1111/cdev.12169

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804. doi:10.3758/bf03194105

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., … Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, *6*, 494. doi:10.3389/fpsyg.2015.00494

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158-189. doi:10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169-176. doi: 10.1177/0963721416643289

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., … Morey, R. D. (2015). A power fallacy. *Behavior Research Methods*, *47*, 913-917. doi: 10.3758/s13428-014-0517-4

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2014). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. New York, NY: Wiley.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426-432. doi: 10.1037/a0022790

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

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*. doi:10.1080/00031305.2016.1154108

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291-298. doi: 10.1177/1745691611406923

Wilkinson, L. (1999). Statistical methods in psychology journals—Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*, 390-420. doi:10.1177/0149206313501200