Behavior Research Methods (2020) 52:489–502
https://doi.org/10.3758/s13428-019-01247-9

Nonparametric multiple comparisons

Kimihiro Noguchi[1] · Riley S. Abel[1] · Fernando Marmolejo-Ramos[2] · Frank Konietschke[3,4]

Published online: 6 May 2019

Abstract

Nonparametric multiple comparisons are a powerful statistical inference tool in psychological studies. In this paper, we review a rank-based nonparametric multiple contrast test procedure (MCTP) and propose an improvement by allowing the procedure to accommodate various effect sizes. In the review, we describe relative effects and show how utilizing the unweighted reference distribution in defining the relative effects in multiple samples may avoid the nontransitive paradoxes. Next, to improve the procedure, we allow the relative effects to be transformed by using the multivariate delta method and suggest a log odds-type transformation, which leads to effect sizes similar to Cohen's d for easier interpretation. Then, we provide theoretical justifications for an asymptotic strong control of the family-wise error rate (FWER) of the proposed method. Finally, we illustrate its use with a simulation study and an example from a neuropsychological study. The proposed method is implemented in the 'nparcomp' R package via the 'mctp' function.

Keywords Effect size · Multiple comparisons · Nonparametric statistics

Introduction

Rank-based nonparametric statistical tests are developed based on the idea of how often a randomly chosen observation from one distribution results in a smaller value than another randomly chosen observation from another distribution. To measure such effects, the original observations are converted to ranks to extract information about their empirical distribution functions of different treatments/groups/samples. Unlike the popular parametric tests which compare means, rank-based nonparametric tests require virtually no distributional assumptions on the data, making them particularly suitable for studies with non-normal distributions (e.g., reaction times data) and/or small sample sizes. However, despite their clear advantages, overall, nonparametric methods are largely underused in psychological studies (Field & Wilcox, 2017).

One possible reason for the unpopularity may come from the misconception that converting the actual observed values into ranks leads to a loss of information; however, a loss of efficiency occurs only when data are exactly or are close to being normally distributed for comparing locations. For instance, Lehmann (2009) studied the asymptotic relative efficiency (ARE) of the Mann–Whitney U test compared to the two-sample t test. Here, the ARE is the limit of the ratio of sample sizes required by the two tests being compared to achieve the same results in terms of level and power. On normal distributions, the Mann–Whitney U test is about 95% as efficient as the t test. As the underlying distribution of the data becomes less similar to a normal distribution (e.g., skewed, light-tailed, or heavy-tailed), the ARE of the Mann–Whitney U test compared to the t test may increase without bound, generally exceeding 100%. That is, in the large-sample case, the Mann–Whitney U test is typically more powerful than the t test.

Another reason why the nonparametric tests are less popular may be due to the difficulty of performing multiple

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-019-01247-9) contains supplementary material, which is available to authorized users.

✉ Kimihiro Noguchi
Kimihiro.Noguchi@wwu.edu

[1] Department of Mathematics, Western Washington University, Bellingham, WA 98225, USA

[2] School of Psychology, University of Adelaide, Adelaide, Australia

[3] Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany

[4] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany

comparisons. Traditionally, nonparametric multiple comparisons of independent samples have been performed in two steps. In the first step, the Kruskal–Wallis test is performed to evaluate the equality of distributions among different treatment groups. When a statistically significant difference is detected, the Mann–Whitney U tests are used for post hoc comparisons. However, interestingly, this two-step procedure can result in paradoxical results; i.e., it is possible to obtain results where, between three or more treatment groups, the pairwise differences are all statistically significant, yet none of them is stochastically dominant. In other words, there is no treatment group from which a random observation tends to be larger than a random observation from any of the other treatment groups. Mathematically speaking, this phenomenon is a consequence of the widely known nontransitive paradoxes. In this paper, we review the above-mentioned situation more clearly using a set of modified dice as an example with a more detailed explanation of stochastic differences. Then, we describe a method which eliminates the paradoxes by defining a reference distribution and comparing each sample to that distribution.

Lack of research in calculating an easily interpretable effect size for nonparametric multiple comparisons may be yet another reason why they are underused. In the normal-based parametric tests, Cohen's d, which divides the difference of two means by their pooled standard deviation, is often used as an effect size to understand the practical significance of the results. Supplying effect sizes in addition to (adjusted) p-values is highly recommended as, for example, Cohen (1994) famously described that using p-values with large sample sizes can show statistically significant results when no difference of practical importance is present. Similarly, even when a statistically significant result is found, p-values give little information about how different samples are. Thus, in this paper, we propose a new multiple comparison procedure that can accommodate various effect sizes to supplement p-values by generalizing the work of Konietschke et al. (2012), providing practical measures of the stochastic differences between samples. The idea resonates well with the statement released by the American Statistical Association (Wasserstein & Lazar, 2016), which strongly encourages practitioners to make decisions using various measures of significance. Furthermore, we suggest a log odds-type effect size similar to Cohen's d for nonparametric multiple comparisons, allowing the users to easily interpret the results.

Even though the importance of effect sizes has been emphasized above, p-values (or some measure of statistical significance) are also likely to remain prevalent. In psychological studies, there are many situations where one hypothesis contains several sub-hypotheses for different contrasts, requiring many tests to be performed. To ensure that research findings are replicable with a high probability, a nonparametric multiple comparison procedure for these contrasts (which shall be referred to as a nonparametric multiple contrast testing procedure (MCTP)) proposed in this paper is designed to provide a strong control the family-wise error rate (FWER) asymptotically at some prespecified $\alpha \in (0, 1)$. That is, for any configuration of true and false null hypotheses, the probability of making at least one type I error is at most α (Pesarin & Salmaso, 2010). An appropriate FWER control provides a safeguard against type I errors at the expense of failing to detect some effects that are true (Cramer et al., 2016). We give theoretical justifications of the asymptotic strong control of the FWER of the proposed nonparametric MCTP by utilizing the idea of simultaneous test procedure (STP) proposed by Gabriel (1969).

The contributions made in this paper can be summarized as follows. Firstly, we provide a concise review of key ideas and issues that occur in nonparametric multiple comparisons, including the nontransitive paradoxes and reference distribution, by expanding the brief explanations given in Konietschke et al. (2012). Then, we propose a new nonparametric MCTP that provides a strong control of the FWER and accommodates various nonparametric effect sizes and contrasts. In particular, we discuss the idea of relative effects, effect sizes for the relative effects in multiple comparisons, how to generalize the nonparametric MCTP of Konietschke et al. (2012) to accommodate various effect sizes, theoretical justifications of the strong FWER control, and small-sample approximations. Then, the newly proposed nonparametric MCTP is evaluated through a simulation study and a real-life application. Lastly, conclusions and future work are summarized, and technical details are provided in Appendix A–D. In addition, the proposed method is implemented in the R package 'nparcomp' via the function 'mctp'.

Nontransitive paradoxes

Many nonparametric tests, including the Mann–Whitney U test, measure the so-called *relative effect*, to compare different samples. As a result of that, nontransitive paradoxes can occur, making the results less interpretable. In this section, we review the relative effect and nontransitive paradoxes, and discuss a way to avoid the paradoxes.

To understand the paradox, let X_i be a random variable from the i-th sample. To measure the stochastic superiority of the i-th sample compared to the j-th sample in the two-sample case, the relative effect, which is defined as

$$p_{ij} = \int F_i dF_j = \Pr(X_i < X_j) + 0.5 \Pr(X_i = X_j),$$

is used (see Munzel and Hothorn 2001; Reiczigel et al. 2005; Wolfsegger & Jaki 2006; Ryu 2009; Umlauft et al. 2017). Specifically, if $p_{ij} > 0.5$, the j-th sample is *stochastically larger* than the i-th sample. Similarly, if $p_{ij} < 0.5$, the j-th sample is *stochastically smaller* than the i-th sample. If $p_{ij} = 0.5$, the two samples are *stochastically equal*. In other words, the relative effect p_{ij} tells us how likely it is that a random observation from the j-th sample be larger than a random observation from the i-th sample. It is also straightforward to see that $p_{ji} = 1 - p_{ij}$. Note that these relative effects have been used as ways of measuring stochastic differences (see Cliff 1993 and Vargha & Delaney 2000 for more details).

In the classical parametric setting where the means are being compared (e.g., the t test), transitivity is preserved. That is, in the case of three samples, if their means μ_i, $i = 1, 2, 3$, are such that $\mu_1 < \mu_2$ and $\mu_2 < \mu_3$, then it must be the case that $\mu_1 < \mu_3$. However, surprisingly, when the relative effects are compared, there could be a situation where $p_{21} > 0.5$ (the first sample is stochastically larger than the second sample) and $p_{13} > 0.5$ (the third sample is stochastically larger than the first sample) do not necessarily imply $p_{23} > 0.5$ (the third sample is stochastically larger than the second sample). This paradox, often referred to as nontransitive paradox, can be better understood by way of an example.

Suppose that there are three fair dice, whose faces have been modified as follows:

- Die 1 has faces 3,3,4,4,8,8;
- Die 2 has faces 2,2,6,6,7,7;
- Die 3 has faces 1,1,5,5,9,9.

Now, suppose that we are trying to find the best of these dice, or the one which rolls a higher value most often. A quick calculation shows that Die 1 rolls a higher value than Die 2 5/9 times. Similarly, Die 3 beats Die 1 5/9 times. Finally, Die 2 beats Die 3 5/9 times (see Appendix A). That is, $p_{21} = p_{13} = p_{32} = 5/9$, which implies that $p_{21} > 0.5$, $p_{13} > 0.5$ and yet $p_{23} = 4/9 < 0.5$. The rock-paper-scissors-like effect causes problems when deciding which die is the best (in the sense of finding the stochastically largest die). Unless it is apparent which die must be rolled against, there is no way of choosing the best die.

Obviously, the above situation is undesirable when performing pairwise comparisons of multiple samples using relative effects. Specifically, the above example implies that nonparametric tests, such as the Mann–Whitney U test that utilizes relative effects, should not be used for (post hoc) pairwise comparisons. To understand the problem better, by viewing the faces of the three dice as observations from three samples, their estimated relative effects ($\hat{p}_{ij}$) are given by $\hat{p}_{21} = \hat{p}_{13} = \hat{p}_{32} = 5/9$. Now, suppose that statistically significant stochastic superiority is declared when $\hat{p}_{ij} > 0.55$. Then, because $5/9 > 0.55$, each pairwise comparison tells us that the latter die is significantly larger stochastically, yet they result in contradictory statements because of the paradox.

However, we can solve the problem by defining relative effects using a reference distribution. To understand how the reference distribution works, suppose that we have a second set of the same three dice in a black box. Now, let us draw a die at random from the black box and denote its face by Y. In other words, in this case, Y can be thought of as a random variable representing the face of an 18-faced fair die containing all the faces from the three dice. We call this new die a reference die.

Let us define the relative effect of each die by $p_i = \Pr(Y < X_i) + 0.5 \Pr(Y = X_i)$, $i = 1, 2, 3$, where the comparisons are made to the common reference die. It can be shown that $p_1 = p_2 = p_3 = 1/2$, from which it can be concluded that all the three dice are stochastically equal to the reference die (see Appendix A). In this situation, the non-transitivity paradox cannot occur because all the three dice are compared to the same reference die. That implies that we can define which die is "larger" decisively by comparing the values of p_i. In addition, the distribution of the reference die is called the reference distribution, which will be defined more rigorously in the next section.

The reference distribution

We define the reference distribution by closely following the notation used in Konietschke et al. (2012). Let X_{ik} indicate the k-th random variable in the i-th independent sample, which has n_i observations, $i = 1, \ldots, a$, $k = 1, \ldots, n_i$, and let $N = \sum_{i=1}^{a} n_i$ denote the total number of observations. Moreover, let $F_i(x) = \Pr(X_{ik} < x) + 0.5 \Pr(X_{ik} = x)$, $-\infty < x < \infty$, be the normalized distribution function for the i-th sample. In general, we only require that

$$X_{ik} \sim F_i, \; k = 1, \ldots, n_i,$$

where F_i are non-degenerate distribution functions. Specifically, we do not require any relationship between the distributions; that is, some could be exponentially distributed while others may be normally or even binomially distributed, for example. Note that this allows us to consider samples which are heteroscedastic, from discrete data, and/or samples without finite means or variances (e.g., Cauchy distribution). We denote the vector of all distribution functions by $F = (F_1, \ldots, F_a)'$.

These F_i on their own cannot easily describe differences among distributions. To describe differences, let $G(x) = \frac{1}{a}\sum_{i=1}^{a} F_i(x)$ be an unweighted mean distribution. By viewing G as a distribution function, we call the composite distribution the reference distribution and use it to define treatment effects,

$$p_i = \int G dF_i = \Pr(Y < X_{ik}) + 0.5\Pr(Y = X_{ik}),\ i = 1, \ldots, a,$$

where $X_{ik} \sim F_i$ and $Y \sim G$. If $p_i < p_j$, the values from F_i tend to be smaller than those from F_j. On the other hand, if $p_i = p_j$, neither distribution tends to be smaller or larger (Noguchi et al., 2012).

As we saw in the previous section, the reference distribution has many advantages. Most importantly, because every treatment effect p_i refers to the same fixed reference distribution, there is no risk of paradoxical conclusions of the kind described in the example above. Furthermore, although the weighted mean distribution having the distribution function $\tilde{G}(x) = \frac{1}{N}\sum_{i=1}^{a} n_i F_i(x)$ has been used in the past, use of the unweighted mean distribution is recommended because it is independent of sample sizes and their allocations. Thus, the effects p_i can be used in the formulation of null hypotheses because they are model constants (Brunner et al., 2018; Gao et al., 2008; Konietschke et al., 2012).

Contrast vectors and null hypotheses

Multiple comparisons are made by specifying q contrasts of interest. A contrast is an a-dimensional vector representing the coefficients of the parameters to be used for making comparisons. In general, the contrast vector for the ℓ-th comparison can be written as $\boldsymbol{c}_\ell = (c_{\ell 1}, \ldots, c_{\ell a})'$, a non-zero vector such that $\sum_{i=1}^{a} c_{\ell i} = 0$. Without loss of generality, we add one more constraint $\sum_{i=1}^{a} |c_{\ell i}| = 2$ and describe its advantage in the next section.

To specify the parameters to be used for making comparisons, let

$$\boldsymbol{p} := (p_1, \ldots, p_a)' = \left(\int G dF_1, \ldots, \int G dF_a\right)' = \int G dF$$

be a vector of a relative effects. The vector $\boldsymbol{p}$ is then used to formulate the family of q null hypotheses:

$$\boldsymbol{\Omega} = \{H_0^\ell : \boldsymbol{c}_\ell' \boldsymbol{p} = 0,\ \ell = 1, \ldots, q\},$$

tested against their respective two-tailed alternatives.

In general, the family of hypotheses can be specified with any set of contrast vectors although, in practice, the choice of which contrasts to use is tremendously important. For example, a standard method of comparing multiple samples is that of all-pairwise comparisons, attributed to Tukey (Gabriel, 1969). This method includes all the null hypotheses of the form $p_i - p_j = 0$ for all $i < j$. In our notation, this is tested using the contrast vectors with $c_{\ell i} = 1$, $c_{\ell j} = -1$, and $c_{\ell u} = 0$ for $u \notin \{i, j\}$. For example, if we let $i = 1$, $j = 2$, and $a = 4$ for the ℓ-th comparison, we have $\boldsymbol{c}_\ell = (1, -1, 0, 0)'$. Another method, attributed to Dunnett, compares every treatment to a single, fixed treatment, usually the control group. Assuming that the fixed treatment is the first sample, this type of contrast contains all the null hypotheses of the form $p_1 - p_j = 0$ for all $j > 1$. That is, the corresponding contrast vector for the ℓ-th comparison have elements whose values are given by $c_{\ell 1} = 1$, $c_{\ell j} = -1$ for $j = \ell + 1$, and 0 otherwise.

Careful attention should be paid to which contrasts are chosen. Tukey's all-pairwise comparisons, while certainly thorough, can greatly decrease the power of a test as they may include comparisons not directly of research interest. On the other hand, Dunnett's many-to-one comparisons can result in a far more powerful test; however, they may not answer every hypothesis of interest. Also, there are many other ways of defining contrasts depending on specific research questions. We therefore favor flexible methods that allow for the use of arbitrary contrast vectors. An application section at the end of this paper includes an example of a nontrivial contrast.

It should be noted that the null hypotheses considered here are valid in the case of heteroscedasticity. This can be easily seen by exemplifying normally distributed random variables $X_{ik} \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, a$; $k = 1, \ldots, n_i$. Here, the relative effects can be computed using the parameters μ_i, σ_i^2, and the cumulative distribution function of $N(0, 1)$, $\Phi(\cdot)$, by

$$p_j = \frac{1}{a}\sum_{i=1}^{a} \int F_i dF_j = \frac{1}{a}\sum_{i=1}^{a} \Phi\left(\frac{\mu_j - \mu_i}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right),\ j = 1, \ldots, a.$$

Thus, $p_j = 0.5$ and $p_i = p_j$ hold even under heteroscedasticity, i.e., $\sigma_i^2 \neq \sigma_j^2$. Therefore, testing the null hypotheses $H_0\colon p_{ij} = 0.5$ or $H_0\colon p_1 = \cdots = p_a$ are also known as the *nonparametric Behrens–Fisher problem* (Brunner et al., 2018; Konietschke et al., 2012; Brunner et al., 2002). In general nonparametric models, $p_i = p_j$ neither implies that variances or shapes of the underlying distributions are identical. Statistical methods that do not rely on the assumption of equal variances are especially important when the distribution of a statistic under the alternative hypothesis is important, i.e., for the computation of confidence intervals. For a general discussion about heteroscedastic methods and their importance, we refer to the comprehensive textbook by Wilcox (2017).

Finally, we note that the general definition of a "treatment effect" depends on the actual research question. Again, the effects of interest considered in this paper are formulated in the sense that different variances (or even higher moments) across the groups are not considered as a treatment effect. If no treatment effect is defined in a way that treatments have no effect on the data, exchangeability of the data may be a more appropriate definition of a treatment effect (Pesarin, 2001; Calian et al., 2008; Westfall & Troendle, 2008).

Comparing relative effects

When comparing two samples, it is perhaps most intuitive to consider the difference between their relative effects. That is, the i-th sample is compared to the j-th sample by considering $p_i - p_j$. However, this simple effect size may be difficult to interpret because its magnitude is not directly comparable to the most popular effect size known as Cohen's d, which is typically given by $d_{ij} = (\bar{x}_i - \bar{x}_j)/s_p$, where s_p is the pooled standard deviation.

On the other hand, by letting $g(x) = k \log[x/(1-x)]$ for some $k > 0$, we obtain

$$g_{\log}(p_i, p_j) = g(p_i) - g(p_j) = k \log\left[\frac{p_i/(1-p_i)}{p_j/(1-p_j)}\right],$$

a constant multiple of the log odds (or log odds ratio). As for the choice of k to make the distribution of $g_{\log}$ closest to that of standard normal, Haley (1952) suggested $k = 1/1.702$, which is the most optimal choice in the minimax sense (Camilli, 1994). Thus, we adopt Haley's suggestion in this paper.

The log odds-type effect size is a favorable effect size as it resembles Cohen's d. Hasselblad and Hedges (1995) and Chinn (2000) have noted (with a slightly different choice of k) that the distribution of d_{ij} and $g_{\log}(p_i, p_j)$ are comparable under the assumption of normality with homogeneous variances. Therefore, the interpretation of $g_{\log}(p_i, p_j)$ in terms of its magnitude may be made by referring to how it would be interpreted in terms of Cohen's d. In fact, an extensive simulation study by Sánchez-Meca et al. (2003) indicates that the proposed effect size, which is in fact close to the one suggested in Cox (1970), seems to perform well under various situations.

Even though the discussion so far has been based on measuring the difference in the two-sample case, its generalization is required for the multi-sample case. For example, when comparing four samples, some may be interested in making an average comparison of the first two samples to the last two. That is, the corresponding null hypothesis assuming the additive effect is given by $H_0^\ell: (p_1 + p_2)/2 - (p_3 + p_4)/2 = 0$. To accommodate these nontrivial cases, we need to define a useful way of obtaining the effect size expressed in a form of comparing two effects, as illustrated in Tukey (1991).

To achieve its generalization, firstly, we consider separating each of the q contrast vectors $\boldsymbol{c}_\ell$, $\ell = 1, \ldots, q$, into the positive and negative parts. Specifically, let $\boldsymbol{c}_{\ell,1}$ be a vector such that its i-th element is given by $c_{\ell,1,i} = \max\{c_{\ell,i}, 0\}$. Similarly, let $\boldsymbol{c}_{\ell,2}$ be a vector such that its i-th element is given by $c_{\ell,2,i} = \max\{-c_{\ell,i}, 0\}$. That implies that $\boldsymbol{c}_\ell = \boldsymbol{c}_{\ell,1} - \boldsymbol{c}_{\ell,2}$. Also, $\sum_{i=1}^a c_{\ell i} = 0$ and $\sum_{i=1}^a |c_{\ell i}| = 2$ imply that $\sum_{i=1}^a c_{\ell,1,i} = \sum_{i=1}^a c_{\ell,2,i} = 1$. For example, using the average comparison above, for the contrast vector $\boldsymbol{c}_\ell = (1/2, 1/2, -1/2, -1/2)'$, we have $\boldsymbol{c}_{\ell,1} = (1/2, 1/2, 0, 0)'$ and $\boldsymbol{c}_{\ell,2} = (0, 0, 1/2, 1/2)'$.

Let us recall the null hypothesis $H_0^\ell: \boldsymbol{c}_\ell' \boldsymbol{p} = 0$. Using the notation above, it can be rewritten as $H_0^\ell: \boldsymbol{c}_{\ell,1}' \boldsymbol{p} - \boldsymbol{c}_{\ell,2}' \boldsymbol{p} = 0$. Moreover, because g is assumed to be strictly increasing, it is also mathematically equivalent to $H_0^\ell: g(\boldsymbol{c}_{\ell,1}' \boldsymbol{p}) - g(\boldsymbol{c}_{\ell,2}' \boldsymbol{p}) = 0$, although the latter representation is clearly preferred as it explicitly specifies the effect g to be considered. Here, we have obtained a generalization of the effect size to the multi-sample case given by $g_\ell(\boldsymbol{p}) = g(\boldsymbol{c}_{\ell,1}' \boldsymbol{p}) - g(\boldsymbol{c}_{\ell,2}' \boldsymbol{p})$. As a consequence, the family of hypotheses we are considering can be more appropriately written as

$$\boldsymbol{\Omega}^g = \{H_0^\ell: g_\ell(\boldsymbol{p}) = 0,\ \ell = 1, \ldots, q\}.$$

At the same time, it becomes clear why setting the constraint $\sum_{i=1}^a |c_{\ell i}| = 2$ is effective. Because that constraint implies that $\sum_{i=1}^a c_{\ell,1,i} = \sum_{i=1}^a c_{\ell,2,i} = 1$, both $\boldsymbol{c}_{\ell,1}' \boldsymbol{p}$ and $\boldsymbol{c}_{\ell,2}' \boldsymbol{p}$ can be interpreted as weighted averages of $p_1, \ldots, p_a$. That also ensures $\boldsymbol{c}_{\ell,1}' \boldsymbol{p} \in (0, 1)$ and $\boldsymbol{c}_{\ell,2}' \boldsymbol{p} \in (0, 1)$, implying that the generalization works for any strictly increasing and continuously differentiable g whose domain is $(0, 1)$.

As an example, let us apply the transformation $g_{\log}(x) = k \log[x/(1-x)]$ to the generalized effect size. Then, we obtain

$$g_{\log,\ell}(\boldsymbol{p}) = g_{\log}(\boldsymbol{c}_{\ell,1}' \boldsymbol{p}) - g_{\log}(\boldsymbol{c}_{\ell,2}' \boldsymbol{p}) = k \log\left[\frac{\boldsymbol{c}_{\ell,1}' \boldsymbol{p}/(1 - \boldsymbol{c}_{\ell,1}' \boldsymbol{p})}{\boldsymbol{c}_{\ell,2}' \boldsymbol{p}/(1 - \boldsymbol{c}_{\ell,2}' \boldsymbol{p})}\right].$$

In real-life situations, because p_i are unknown, they are replaced by their estimators $\hat{p}_i$ (see Konietschke et al. 2012 for details). Let the vector of relative effect estimators be denoted by $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_a)'$. Then, the generalized effect size estimator is given by $g_\ell(\hat{\boldsymbol{p}})$.

We note that the effects $p_i = \int G dF_i$ involve all of the distributions. Thus, the contrast $g(p_i) - g(p_j)$ does not only involve the distributions F_i and F_j, but also all other distributions involved in the experiment. Therefore, it should always be interpreted as a relative measure compared to the overall experiment. When the comparison of specific distributions is strictly of interest, pairwise defined effects $p_{ij} = \int F_i dF_j$ may be a better choice. These effects, however, may result in nontransitive conclusions as described above.

Test statistics

Ultimately, we are interested in finding a testing procedure that addresses each of the q individual null hypotheses $H_0^\ell\colon g_\ell(\boldsymbol{p}) = 0$, $\ell = 1, \ldots, q$, where the prespecified error rate is properly controlled. This type of testing procedure is called the multiple contrast testing procedure (MCTP). In this paper, we consider controlling the most common error rate known as the FWER. The FWER is defined as the probability of rejecting at least one true null hypothesis.

Even though the Bonferroni adjustment is the most common method for controlling the FWER, it is known to be highly conservative, leading to possibly many false non-rejections (Bender & Lange, 1999). Therefore, we firstly construct q t-type test statistics which are jointly approximately multivariate t-distributed, from which we suggest a much less conservative nonparametric MCTP that takes the correlation among the test statistics into account.

The construction of the t-type test statistics is done by an appropriate standardization of the generalized effect size estimators $g_\ell(\hat{\boldsymbol{p}})$, $\ell = 1, \ldots, q$. Let us define a vector of generalized effect size estimators by $\boldsymbol{g}(\hat{\boldsymbol{p}}) = (g_1(\hat{\boldsymbol{p}}), \ldots, g_q(\hat{\boldsymbol{p}}))'$. Then, its standardization can be derived by applying the multivariate delta method to the statistic $\sqrt{N}(\hat{\boldsymbol{p}} - \boldsymbol{p})$, which is asymptotically multivariate normal under some mild regularity conditions. In particular, it can be shown that the statistic $\sqrt{N}[\boldsymbol{g}(\hat{\boldsymbol{p}}) - \boldsymbol{g}(\boldsymbol{p})]$ is asymptotically multivariate normal with expectation $\mathbf{0}$ and some covariance matrix denoted by $\boldsymbol{V}_N^g$ (see Appendix B for details). In other words, the large-sample distribution of $\boldsymbol{g}(\hat{\boldsymbol{p}})$ is approximately multivariate normal with expectation $\boldsymbol{g}(\boldsymbol{p})$ and covariance matrix $\boldsymbol{V}_N^g/N$. Hence, by considering its marginals, the large-sample distribution of $g_\ell(\hat{\boldsymbol{p}})$ is approximately normal with expectation $g_\ell(\boldsymbol{p})$ and variance $v_{\ell\ell}^g/N$, where $v_{\ell\ell}^g = (\boldsymbol{V}_N^g)_{\ell\ell}$. By standardization, the asymptotic distribution of $\sqrt{N}[g_\ell(\hat{\boldsymbol{p}}) - g_\ell(\boldsymbol{p})]/\sqrt{v_{\ell\ell}^g}$ is standard normal.

The argument above shows that an appropriate t-type test statistic for H_0^ℓ is given by

$$T_\ell^g = \frac{\sqrt{N}[g_\ell(\hat{\boldsymbol{p}}) - g_\ell(\boldsymbol{p})]}{\sqrt{\hat{v}_{\ell\ell}^g}},$$

where we replaced the unknown $v_{\ell\ell}^g$ with its sample estimator $\hat{v}_{\ell\ell}^g$ in the denominator. Under H_0^ℓ, noting that $g_\ell(\boldsymbol{p}) = 0$ and $g_\ell(\hat{\boldsymbol{p}}) = g(\boldsymbol{c}_{\ell,1}'\hat{\boldsymbol{p}}) - g(\boldsymbol{c}_{\ell,2}'\hat{\boldsymbol{p}})$,

$$T_\ell^g = \frac{\sqrt{N}[g(\boldsymbol{c}_{\ell,1}'\hat{\boldsymbol{p}}) - g(\boldsymbol{c}_{\ell,2}'\hat{\boldsymbol{p}})]}{\sqrt{\hat{v}_{\ell\ell}^g}}.$$

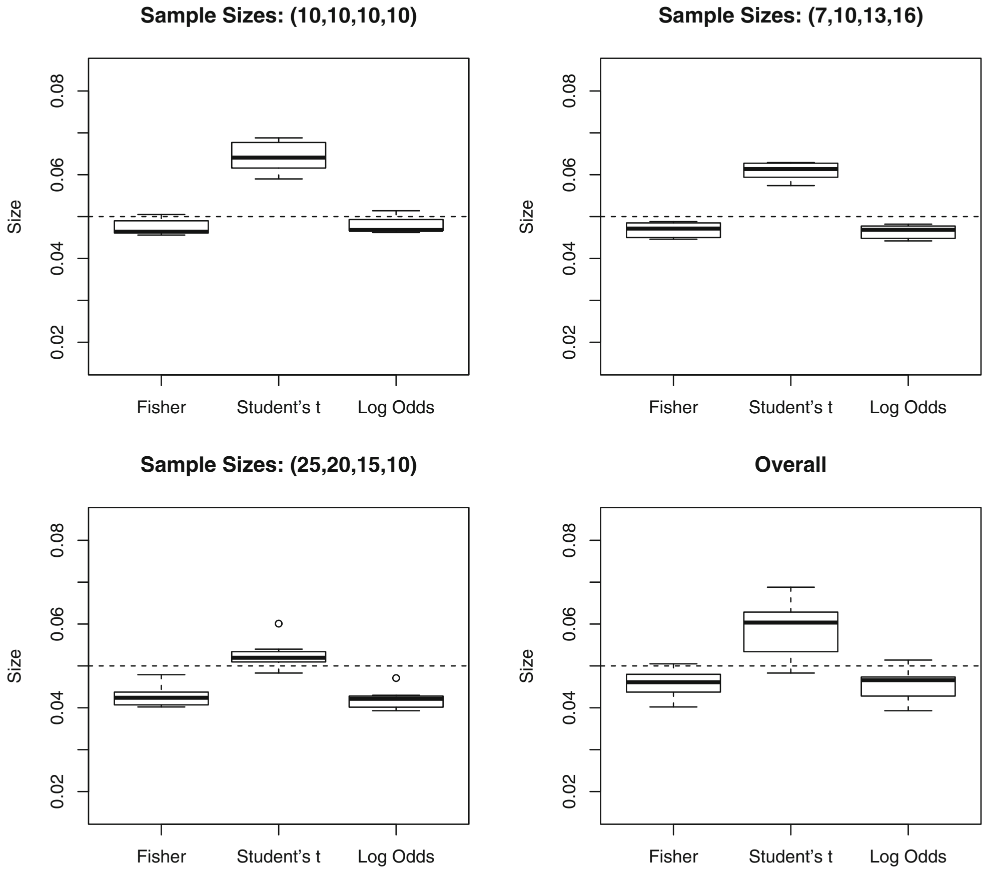
To obtain the critical values and adjusted p-values, it is necessary to understand the joint distribution of $\boldsymbol{T}^g = (T_1^g, \ldots, T_q^g)'$ under the global null hypothesis $H_0\colon \bigcap_{\ell=1}^q \{g_\ell(\boldsymbol{p}) = 0\}$. In the first step, we consider the asymptotic joint distribution of $\boldsymbol{T}^g$. By applying Slutsky's theorem, $\boldsymbol{T}^g$ asymptotically follows a multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix $\boldsymbol{R}^g$, where $(\boldsymbol{R}^g)_{\ell m} = v_{\ell m}^g/\sqrt{v_{\ell\ell}^g v_{mm}^g}$. That is, the critical values and adjusted p-values can be obtained by referring to such multivariate normal distribution. However, in practice, because $\boldsymbol{R}^g$ is unknown, it is replaced by its estimator $\hat{\boldsymbol{R}}^g$, where $(\hat{\boldsymbol{R}}^g)_{\ell m} = \hat{v}_{\ell m}^g/\sqrt{\hat{v}_{\ell\ell}^g \hat{v}_{mm}^g}$.

In reality, the asymptotic results are relevant only when large samples are available. Therefore, the results from the previous paragraph are mainly of theoretical interests. At the same time, because small sample sizes frequently occur in psychological studies (Szucs & Ioannidis, 2017), it is highly desirable to have an accurate small-sample approximation of the joint test statistics $\boldsymbol{T}^g$, which will be explored in the next section.

Small-sample approximation, adjusted *p*-values values, and simultaneous confidence intervals

An accurate small-sample approximation of the joint distribution of the test statistics $\boldsymbol{T}^g$ is essential to obtain reliable statistical results. Even though the asymptotic distribution of $\boldsymbol{T}^g$ under H_0 is multivariate normal, it is known that the multivariate normal approximation tends to produce liberal results, leading to possibly inflated false rejections. Also, psychological and behavioral data are often heteroscedastic, as emphasized in Wilcox (2017). Moreover, it is well known that the rank-transformed observations are in general heteroscedastic even if the original observations are homoscedastic (Brunner et al., 1997). Thus, we present a better approximation which is robust to heteroscedasticity using the multivariate t-distribution with appropriately modified degrees of freedom. Using the multivariate t-based approximation, we discuss how to obtain a critical value corresponding to a given FWER α, adjusted p-values, and $100(1-\alpha)\%$ simultaneous confidence intervals (SCIs).

Konietschke et al. (2012) suggested a Box-type approximation (see Box 1954; Brunner et al. 1997; Gao et al. 2008) for accurate small-sample results. Specifically, following their notation, let $\hat{\omega}_{\ell i}^2$ denote the empirical variances of the variables $A_{\ell ik} = c_{\ell i}(\widehat{G}(X_{ik}) - \frac{1}{a}\widehat{F}_i(X_{ik})) - \sum_{s\neq i} c_{\ell s}\frac{1}{a}\widehat{F}_s(X_{ik})$, where $\widehat{G}$ and $\widehat{F}$ denote the empirically estimated G and F, respectively (for more details, we refer to p. 750 of Konietschke et al., 2012). Then, an approximate small-sample distribution of $\boldsymbol{T}^g$ with $g(x) = x$ under


Sample Sizes: (10,10,10,10)

Sample Sizes: (7,10,13,16)

Sample Sizes: (25,20,15,10)

Overall

Fig. 1 Size of the test by sample size combinations. The *dashed line* indicates the significance level of 0.05

H_0 is given by the q-dimensional t-distribution with expectation $\mathbf{0}$, the correlation matrix $\hat{\boldsymbol{R}}^g$, and degrees of freedom $\nu = \max\{1, \min\{\nu_1, \ldots, \nu_q\}\}$, where

$$\nu_\ell = \frac{\left(\sum_{i=1}^a \hat{\omega}_{\ell i}^2/n_i\right)^2}{\sum_{i=1}^a \hat{\omega}_{\ell i}^4/[n_i^2(n_i-1)]}.$$

For convenience, we denote this distribution $\boldsymbol{t}(\nu, \mathbf{0}, \hat{\boldsymbol{R}}^g)$.

In our case, a slight modification is necessary to accommodate the cases where $g(x) \neq x$. To do so, following the idea of Noguchi and Marmolejo-Ramos (2016), we suggest to replace ν with $\nu^g = \max\{1, \min\{\nu_1^g, \ldots, \nu_q^g\}\}$, where

$$\nu_\ell^g = \frac{\left(\sum_{i=1}^a [\sum_{t=1}^2 \{g'(\boldsymbol{c}_{\ell,t}'\hat{\boldsymbol{p}})\}^2 I(c_{\ell,t,i} > 0)]\hat{\omega}_{\ell i}^2/n_i\right)^2}{\sum_{i=1}^a [\sum_{t=1}^2 \{g'(\boldsymbol{c}_{\ell,t}'\hat{\boldsymbol{p}})\}^4 I(c_{\ell,t,i} > 0)]\hat{\omega}_{\ell i}^4/[n_i^2(n_i-1)]}.$$

Here, $I(c_{\ell,t,i} > 0) = 1$ if $c_{\ell,t,i} > 0$ and 0 otherwise. As a remark, when $g(x) = x$, $\nu_\ell^g = \nu_\ell$ because $g'(x) = 1$.

Using ν^g, an accurate critical value corresponding to FWER $= \alpha$ and adjusted p-values can be computed. Let $t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}$ denote the two-sided equicoordinate (i.e., the quantiles in each dimension coincide) $100(1-\alpha)$-th percentile of $\boldsymbol{t}(\nu^g, \mathbf{0}, \hat{\boldsymbol{R}}^g)$, which serves as the critical value. That is, H_0^ℓ is rejected if and only if $|T_\ell^g| > t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}$. Moreover, H_0 is rejected if and only if $\max\{|T_1^g|, \ldots, |T_q^g|\} > t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}$.

Multiple comparison procedures having above properties are known as single-step procedures. In other words, the results for the overall comparison (H_0) and specific contrasts (H_0^ℓ) can be obtained simultaneously without any contradiction, unlike the popular procedures done in two steps. That is, rejection of at least one of H_0^ℓ, $\ell = 1, \ldots, q$, automatically implies rejection of H_0 (a property known as *coherent*), and similarly, rejection of H_0 automatically implies that at least one of H_0^ℓ, $\ell = 1, \ldots, q$, is rejected (a property known as *consonant*) (Gabriel, 1969). Coherence and consonance are not necessarily guaranteed in the popular procedures done in two steps, making the proposed single-step nonparametric MCTP much more interpretable and practical.

In addition, the adjusted p-values can be computed directly without relying on the Bonferroni adjustment. In particular, the adjusted p-value corresponding to H_0^ℓ can be calculated by finding p_ℓ for which $t_{1-p_\ell,2,\nu^g,\hat{\boldsymbol{R}}^g}$ is equal to the observed value of $|T_\ell^g|$. The overall adjusted p-value corresponding to H_0 can be calculated by $p =$

$\min\{p_1, \ldots, p_q\}$. Similar to the critical value, H_0^ℓ and H_0, respectively, are rejected if and only if $p_\ell < \alpha$ and $p < \alpha$. As a remark, computations of p_ℓ and $t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}$ can be easily done by using the R package mvtnorm (Hothorn et al., 2008).

We can also use $t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}$ to obtain approximate $100(1-\alpha)\%$ SCIs for the treatment effects (effect sizes) $\boldsymbol{g}_\ell(\boldsymbol{p})$ (see Appendix D for a derivation). Note that, whereas a traditional 100(1-α)% confidence interval for a specific $\boldsymbol{g}_\ell(\boldsymbol{p})$ includes $\boldsymbol{g}_\ell(\boldsymbol{p})$ 100(1-α)% of the time if the experiment is performed repeatedly, SCIs must contain the vector of true population parameters $\boldsymbol{g}(\boldsymbol{p})$ 100(1-α)% of the time.

In general, approximate $100(1-\alpha)\%$ SCIs for the treatment effects $\boldsymbol{g}_\ell(\boldsymbol{p})$, $\ell = 1, \ldots, q$, are given by

$$\left[\boldsymbol{g}_\ell(\hat{\boldsymbol{p}}) - t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\sqrt{\hat{v}^g_{\ell\ell}/N}, \boldsymbol{g}_\ell(\hat{\boldsymbol{p}}) + t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\sqrt{\hat{v}^g_{\ell\ell}/N}\right].$$

Simulation study

A simulation study was conducted to compare the sizes and powers of the nonparametric MCTP with the suggested log odds-type effect sizes (referred to as "Log Odds" in this section) to the ones suggested in Konietschke et al. (2012). These competing methods use $g(x) = x$ without any additional transformation (referred to as "Student's t" in this section) and with Fisher's z-transformation on $\boldsymbol{c}_\ell'\hat{\boldsymbol{p}}$ (referred to as "Fisher" in this section). All the sizes and powers are calculated using 10,000 Monte Carlo simulations.

To ensure that the simulation study covers typical cases frequently encountered in real-life situations, we have a set of different sample size combinations, distributions, and four contrasts (i.e., $a = 4$). The sample size combinations (n_1, n_2, n_3, n_4) are $(10, 10, 10, 10)$, $(7, 10, 13, 16)$, and $(25, 20, 15, 10)$, covering both equal, increasing, and decreasing sample size cases. The distributions used were the normal, (scaled and shifted) Student's t with 8 degrees of freedom, lognormal, and scaled beta with a scaling factor of 20, hence covering both symmetric and asymmetric, light- and heavy-tailed distributions. The means were chosen in such a way that $(\mu_1, \mu_2, \mu_3, \mu_4) = (10, 10, 10, x)$ where x varies from 10 to 13 with an increment of 0.5 while the variances are all set equal to 9. The contrasts were performed via Tukey's all-pairwise comparisons and Dunnett's many-to-one comparisons with the first sample being the control group. The FWER is set at $\alpha = 0.05$.

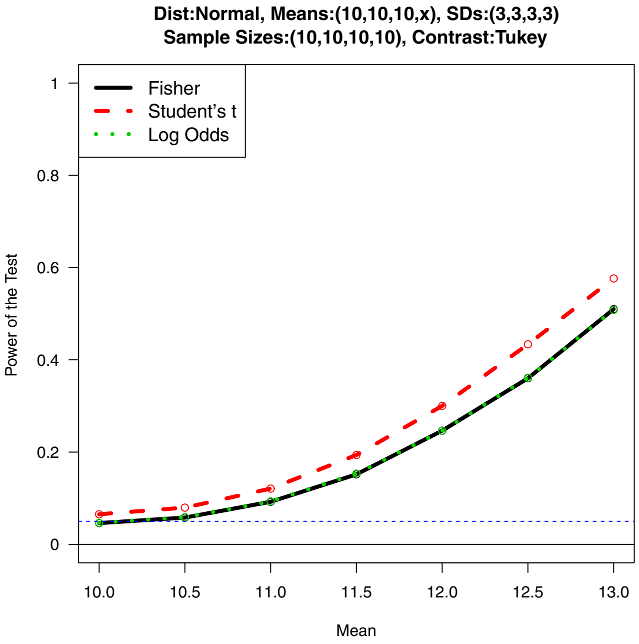


Fig. 2 Power of the test with Tukey's all-pairwise comparisons

The results are summarized in graphs for easier comparisons. Figure 1 shows, via boxplots, the sizes of the tests corresponding to the cases with $(\mu_1, \mu_2, \mu_3, \mu_4) = (10, 10, 10, 10)$. Here, size refers to the probability of falsely rejecting the global null hypothesis (H_0). Based on the simulations, the Student's t method tends to be liberal in the equal and increasing sample size combinations while the Fisher and log odds method seem slightly conservative for the decreasing sample size combinations. Overall, the Fisher and log odds methods seem more robust to various sample size combinations than the Student's t method.

For the powers of the test, each of the $3 \times 4 \times 2 = 24$ cases is compared using the power curves. Here, power refers to the probability of correctly rejecting the global

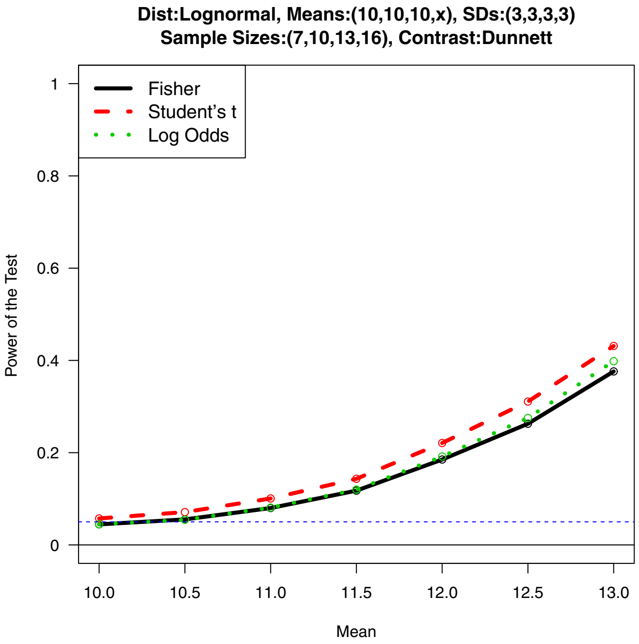


Fig. 3 Power of the test with Dunnett's many-to-one comparisons

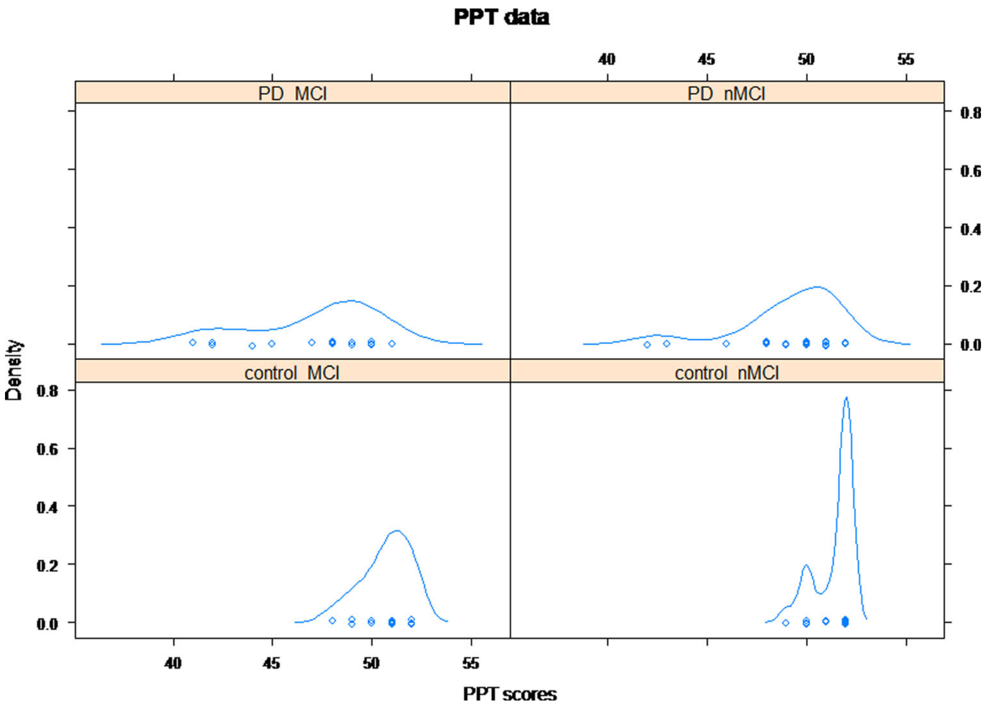


Fig. 4 Distribution of PPT scores in the four groups studied in Bocanegra et al. (2015)

null hypothesis (H_0). Figures 2 and 3 represent typical situations. That is, the Fisher and log odds methods have very similar power curves while Student's t method appears to be more powerful. However, the results need to be interpreted carefully because of the liberal nature of the Student's t method. In other words, this phenomenon can be explained by the contribution of the inflated FWER of the Student's t method. All the other results are displayed in the supplementary material.

Based on the observations above, we may summarize that the Fisher and log odds methods seem equally reliable and powerful while the Student's t method tends to be liberal. As the log odds method directly calculates easily interpretable effect sizes, this method may be preferred in practice.

Table 1 Hypotheses tested for data from Bocanegra et al. (2015)

Comparison	Explicit hypothesis and contrast vector
1	$H_0^1: k \log\left[\frac{\{0.5(p_1+p_2)\}/\{1-0.5(p_1+p_2)\}}{\{0.5(p_3+p_4)\}/\{1-0.5(p_3+p_4)\}}\right] = 0$ $\boldsymbol{c}_1 = (0.5, 0.5, -0.5, -0.5)'$
2	$H_0^2: k \log\left[\frac{p_1/(1-p_1)}{p_3/(1-p_3)}\right] = 0$ $\boldsymbol{c}_2 = (1, 0, -1, 0)'$
3	$H_0^3: k \log\left[\frac{p_2/(1-p_2)}{p_4/(1-p_4)}\right] = 0$ $\boldsymbol{c}_3 = (0, 1, 0, -1)'$
4	$H_0^4: k \log\left[\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right] = 0$ $\boldsymbol{c}_4 = (1, -1, 0, 0)'$
5	$H_0^5: k \log\left[\frac{p_3/(1-p_3)}{p_4/(1-p_4)}\right] = 0$ $\boldsymbol{c}_5 = (0, 0, 1, -1)'$

Note that $k = 1/1.702$

Even though the above simulations were run for homoscedastic samples, additional simulations were run for heteroscedastic samples to ensure that the above observations still hold. The results showed that, indeed, the Fisher and log odds methods seem equally reliable and powerful while the Student's t method tends to be liberal. All the details can be found in the supplementary material.

As a remark, Marozzi (2016) considered quantifying the computation error of the sizes and powers calculated via Monte Carlo simulations of permutation tests. Here, assuming that the p-values are computed exactly from the distribution under the null hypothesis, the upper bound of the root mean squared error (RMSE) of the estimated power is $0.5/\sqrt{MC}$, where MC is the number of Monte Carlo simulations. However, noting that the permutation tests provide *estimated* p-values, the actual upper bound is close to $0.6/\sqrt{MC}$, i.e., a 20% increase approximately.

In this paper, because the p-values are estimated via an approximate multivariate t-distribution, we also expect that the upper bound of the RMSE to be higher than $0.5/\sqrt{MC}$. However, because the multivariate t-distribution is considered quite accurate in approximating the distribution of $\boldsymbol{T}^g$ under H_0 (Brunner et al., 1997; Konietschke et al., 2012), we postulate that the upper bound of the RMSE to remain closer to $0.5/\sqrt{MC}$ than $0.6/\sqrt{MC}$. A more accurate assessment of the computation error will be considered in a future study.

Table 2 Results of the nonparametric MCTP analyses using data from Bocanegra et al. (2015)

Comparison	Effect Size	Effect Size SCIs	Adjusted p-Value
1	−0.851	[−1.122, −0.581]	< 0.001
2	−0.819	[−1.251, −0.387]	< 0.001
3	−0.948	[−1.377, −0.520]	< 0.001
4	0.525	[0.053, 0.997]	0.024
5	0.396	[−0.079, 0.870]	0.126

Real-life application

To illustrate the use of the modified nonparametric MCTP, we reanalyzed data from a neuropsychological study. Bocanegra et al. (2015) examined 40 patients with Parkinson's disease (PD) to determine whether cognitive deficits are language- or semantics-specific. Among them, 23 of those participants were diagnosed as not suffering from any mild cognitive impairment (PD-nMCI) and 17 were diagnosed as suffering some sort of cognitive impairment (PD-MCI). Each subgroup was matched with a control group (Control-nMCI and Control-MCI) of equal sample size, similar average age, average years of education, and proportional gender ratio (see Table 1 in Bocanegra et al. 2015). Thus, there were 40 PD patients and 40 control participants. For our purposes, we label the relative effects of PD-nMCI, PD-MCI, Control-nMCI, and Control-MCI as p_1, p_2, p_3, and p_4, respectively.

The tests the researchers used to evaluate the semantic representation of actions and objects were the Kissing and Dancing Test (KDT) and the Pyramids and Palm trees Test (PPT). We focus on the data related to the PPT, which consists of 52 cards showing triplets of images depicting a cue object-picture (the top image in each card), e.g., a pyramid, and two semantically related distractors (two side-by-side images below the cue object-picture), e.g., a palm tree and a pine tree. The participants' task is to select the picture most related to the cue object-picture (in the examples above, the correct choice is the palm tree). Normal cognitive functioning is indicated by correctly choosing 47 or more of the 52 cards (i.e., 90% of the trials), while cognitive impairment is reflected in scores lower than 47[1].

Figure 4 shows the distribution of PPT scores in each of the four groups. Note that the control groups are highly left-skewed, and that there are outliers present in the PD groups at the lower end. Thus, the nonparametric MCTP can be used to obtain reliable conclusions.

Bocanegra et al. (2015) used the two-tailed Mann–Whitney U test with a significance level of 0.05 to evaluate differences between the groups' adjusted PPT scores. They performed the following comparisons: 1. PD vs. Control, 2. PD-nMCI vs. Control-nMCI, 3. PD-MCI vs. Control-MCI, and 4. PD-nMCI vs. PD-MCI. For the first three tests, they found significant differences with Cohen's d effect sizes higher than 1. For the fourth test, they did not find significant differences.

We applied the nonparametric MCTP with the suggested log odds-type effect sizes described in this paper to the same data, and added a fifth comparison not considered in Bocanegra et al. (2015): 5. Control-nMCI vs. Control-MCI. Table 1 shows the explicit hypotheses being tested as well as the contrast vectors used to test the hypotheses. The statistical results with effect sizes, 95% SCIs, and adjusted p-values are displayed in Table 2.

For three of the comparisons considered in Bocanegra et al. (2015) (PD vs. Control, PD-nMCI vs. Control-nMCI, and PD-MCI vs. Control-MCI), our nonparametric MCTP also found significant differences at $\alpha = 0.05$, supporting their results. We also found a significant difference between PD-nMCI and PD-MCI, which their analysis did not find, suggesting a mild effect of MCI when PD patients are compared. Our fifth comparison did not yield a significant result, which strengthens the findings of Bocanegra et al. (2015), in that no difference between control groups would be expected if no neurological damage is present. In other words, if this comparison had been significant, then three of the pairwise comparisons carried out by them (those involving control groups) could have been influenced by an unknown factor underlying the control groups.

The effect sizes seen in Table 2 are slightly smaller than those found in Bocanegra et al. (2015), but this can be explained by the type of effect size used. Because Cohen's d is found using a difference of means, it can be inflated by outliers, such as those found in the PD groups. On the other hand, log odds of relative effects are less affected by these outliers. Still, the effect sizes we found are large enough to show medium-large effects for all tests which had statistically significant differences.

[1] According to a correspondence with one of the authors of the original study, the PPT was not the key test leading to the conclusions in this study. However, it is instrumental in assessing semantic representation of objects and is generally used for evaluating cases of aphasia and dementia that directly affect language.

Conclusions

In this paper, we have provided a comprehensive review of the nonparametric MCTP of Konietschke et al. (2012) and illustrated the advantages it has over traditional hypothesis testing procedures. In particular, the nonparametric MCTP uses an unweighted reference distribution to eliminate the rock-paper-scissors-like possibility of obtaining paradoxical, nontransitive results in multiple comparisons. Also, it provides a strong control of the FWER, allowing researchers to control the likelihood of type I errors appropriately. These advantages make the nonparametric MCTP a practical option for performing multiple comparisons without a need to make restrictive assumptions on the data.

Another important novel contribution discussed in this paper is a generalization of the nonparametric MCTP of Konietschke et al. (2012) to accommodate various effect size measures. In particular, the log odds-type effect size can be easily interpreted due to its similarity to Cohen's d. We have also derived a reliable small-sample approximation of the generalized nonparametric MCTP, which is effective in real-life situations when larger samples are unavailable. Using that, the calculations of adjusted p-values and SCIs of effect size measures were discussed. Furthermore, the generalized nonparametric MCTP also possesses important theoretical properties of the original nonparametric MCTP including the strong control of the FWER, and our simulation study indicates that the power and robustness of the two are comparable. Finally, our reanalysis of the neuropsychological study in Bocanegra et al. (2015) illustrates that the suggested nonparametric MCTP facilitates a rigorous understanding of multiple treatment effects. The generalized nonparametric MCTP with the log odds-type effect sizes is implemented in the `mctp` function of the `R` package `nparcomp` Version 3.0.

Lastly, recall that the nonparametric MCTPs discussed in this paper are single-step procedures that take the correlation among the test statistics into account. Instead of the single-step procedures, step-down procedures such as Bonferroni–Holm (Holm, 1979; Pesarin & Salmaso, 2010), can be considered using the unadjusted p-values. On the other hand, step-up procedures, e.g., Hochberg (1988), are often valid only if the joint distribution of the test statistics is of a certain multivariate order, known as multivariate of totally positive order two (MTP2). For general contrasts, the joint distribution of the test statistics does not fulfill this requirement in general. Nevertheless, the investigation of step-up procedures and their validity in the general nonparametric Behrens–Fisher situation will be part of future research.

Appendix A: Calculation of the relative effects

The three modified fair dice have the following faces:

- Die 1 has faces 3,3,4,4,8,8;
- Die 2 has faces 2,2,6,6,7,7;
- Die 3 has faces 1,1,5,5,9,9.

To calculate the probability that Die 1 rolls a higher value than Die 2, it is possible to use the conditional probability argument. Let D_i denote the random variable for the face of Die i. Then,

$$\begin{aligned} p_{21} &= \Pr(D_2 < D_1) + 0.5\Pr(D_2 = D_1) \\ &= \sum_{i\in\{2,6,7\}} \Pr(D_2 < D_1 \mid D_2 = i)\Pr(D_2 = i) \\ &= \Pr(D_1 > 2)\Pr(D_2 = 2) + \Pr(D_1 > 6)\Pr(D_2 = 6) \\ &\quad + \Pr(D_1 > 7)\Pr(D_2 = 7) \\ &= \frac{1}{3} + \frac{1}{9} + \frac{1}{9} = \frac{5}{9}. \end{aligned}$$

Similar calculations also show that $p_{13} = p_{32} = \frac{5}{9}$. To calculate the relative effect of each die, let us define Die 4 (a "super die") that has 18 faces. These 18 faces are simply the faces of the three modified fair dice. Then,

$$\begin{aligned} p_1 &= \Pr(D_4 < D_1) + 0.5\Pr(D_4 = D_1) \\ &= \sum_{i=1}^{9}[\Pr(D_4 < D_1 \mid D_4 = i)\Pr(D_4 = i) \\ &\quad +0.5\Pr(D_4 = D_1 \mid D_4 = i)\Pr(D_4 = i)] \\ &= \frac{1}{9}\sum_{i=1}^{9}\Pr(D_1 > i) + \frac{1}{18}\sum_{i=1}^{9}\Pr(D_1 = i) \\ &= \frac{1}{9}\left(2\times 1 + \frac{2}{3} + 4\times\frac{1}{3}\right) + \frac{1}{18}\left(\frac{1}{3} + \frac{1}{3} + \frac{1}{3}\right) \\ &= \frac{12}{27} + \frac{1}{18} = \frac{1}{2}. \end{aligned}$$

Similar calculations also show that $p_2 = p_3 = \frac{1}{2}$.

Appendix B: Construction of the covariance matrix

Konietschke et al. (2012) constructed a nonparametric MCTP starting from the test statistic $\sqrt{N}(\hat{\boldsymbol{p}} - \boldsymbol{p})$ whose corresponding asymptotic covariance matrix is denoted by $\boldsymbol{V}_N$. We describe how to derive the asymptotic covariance matrix of $\sqrt{N}[\boldsymbol{g}(\hat{\boldsymbol{p}}) - \boldsymbol{g}(\boldsymbol{p})]$, where $\boldsymbol{g}(\hat{\boldsymbol{p}}) = (g_1(\hat{\boldsymbol{p}}), \ldots, g_q(\hat{\boldsymbol{p}}))'$.

Let $g_{ij} = \partial g_i(\boldsymbol{p})/\partial p_j$ be the entry in the i-th row and j-th column of $\nabla \boldsymbol{g}(\boldsymbol{p})$, the matrix of gradients. Then, by applying the multivariate delta method, the asymptotic covariance matrix of $\sqrt{N}[\boldsymbol{g}(\hat{\boldsymbol{p}}) - \boldsymbol{g}(\boldsymbol{p})]$ is given by

$$V_N^g = \nabla \boldsymbol{g}(\boldsymbol{p}) \boldsymbol{V}_N \nabla \boldsymbol{g}(\boldsymbol{p})'.$$

Konietschke et al. (2012) also derived a consistent estimator for the matrix $\boldsymbol{V}_N$ and calls it $\hat{\boldsymbol{V}}_N$. We follow that convention and say a consistent estimator for V_N^g is

$$\hat{\boldsymbol{V}}_N^g = \nabla \boldsymbol{g}(\hat{\boldsymbol{p}}) \hat{\boldsymbol{V}}_N \nabla \boldsymbol{g}(\hat{\boldsymbol{p}})'.$$

A special case we are particularly interested in is when $g_i(\boldsymbol{p}) = g(\boldsymbol{c}_{i,1}'\boldsymbol{p}) - g(\boldsymbol{c}_{i,2}'\boldsymbol{p})$, where $g(x) = k \log[x/(1-x)]$. In that case, the matrix of gradients is given elementwise as follows:

$$\begin{aligned} g_{ij} &= \partial g_i(\boldsymbol{p})/\partial p_j \\ &= \partial[g(\boldsymbol{c}_{i,1}'\boldsymbol{p}) - g(\boldsymbol{c}_{i,2}'\boldsymbol{p})]/\partial p_j \\ &= \frac{k}{(\boldsymbol{c}_{i,1}'\boldsymbol{p})(1-\boldsymbol{c}_{i,1}'\boldsymbol{p})} - \frac{k}{(\boldsymbol{c}_{i,2}'\boldsymbol{p})(1-\boldsymbol{c}_{i,2}'\boldsymbol{p})}. \end{aligned}$$

Appendix C: Asymptotic strong control of the FWER

The testing family used here is carefully chosen to give an asymptotic control of the FWER. We start with a lemma by following Gabriel (1969).

Lemma 1 $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g\}$ *is a joint testing family asymptotically.*

Proof As $N \to \infty$, $\boldsymbol{T}^g$ is asymptotically multivariate normal with mean $\boldsymbol{0}$ and correlation matrix $\boldsymbol{R}^g$ as a consequence of the multivariate delta method with Slutsky's theorem. Therefore, the asymptotic joint distribution of $\boldsymbol{T}^g$ is completely specified under the null hypothesis $H_0: \cap_{\ell=1}^q \{g_\ell(\boldsymbol{p}) = 0\}$. The individual test statistics, T_ℓ^g, each converge in distribution to a standard normal random variable, so that the distribution of T_ℓ^g is independent of T_m^g when $\ell \neq m$. Thus, given a non-empty $J \subset I$, $\boldsymbol{T}^g(J) = \{T_\ell^g, \ell \in J\}$ is asymptotically completely specified under the intersection hypothesis $\tilde{H}_0^J : \cap_{\ell \in J} \{g_\ell(\boldsymbol{p}) = 0\}$. This is exactly the definition of a joint family (Gabriel, 1969) and completes the proof. □

The two-sided equicoordinate $100(1-\alpha)$-th percentile of the q-dimensional multivariate normal distribution, $\mathcal{N}_q(\boldsymbol{0}, \boldsymbol{R}^g)$, is the value $z_{1-\alpha,2,\boldsymbol{R}^g}$ such that

$$\Pr\left(\bigcap_{\ell=1}^q \{-z_{1-\alpha,2,\boldsymbol{R}^g} \le X_\ell \le z_{1-\alpha,2,\boldsymbol{R}^g}\}\right) = 1-\alpha$$

for $X = (X_1, \ldots, X_q)' \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{R}^g)$. Here, the second subscript ('2') in $z_{1-\alpha,2,\boldsymbol{R}^g}$ indicates that we are interested in the two-sided equicoordinate percentile. The computation of $z_{1-\alpha,2,\boldsymbol{R}^g}$ can be found in Bretz et al. (2001) and Genz and Bretz (2009).

In general, we do not know the asymptotic correlation matrix $\boldsymbol{R}^g$, so we replace it with its estimator $\hat{\boldsymbol{R}}^g$. Using that, we can compute $z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}$. The triple $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g, z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}\}$ now forms what is called an asymptotic STP. We can formulate the following theorem.

Theorem 1 *As* $N \to \infty$, *the STP* $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g, z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}\}$ *controls the FWER asymptotically in the strong sense. Moreover, the asymptotic control is exact.*

Proof Firstly, the STP $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g, z_{1-\alpha,2,\boldsymbol{R}^g}\}$ is coherent by the construction of $\boldsymbol{T}^g$. Moreover, by the lemma above, the STP comes from the asymptotically joint testing family $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g\}$. These two conditions suffice the requirements of Theorem 2 of Gabriel (1969) to show the asymptotic strong control of the FWER for $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g, z_{1-\alpha,2,\boldsymbol{R}^g}\}$. However, we wish to show that the conditions still hold asymptotically if we replace the critical value $z_{1-\alpha,2,\boldsymbol{R}^g}$ with $z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}$. In other words, now we consider a more realistic STP $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g, z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}\}$.

Since $\hat{\boldsymbol{R}}^g$ is a consistent estimator of $\boldsymbol{R}^g$, we have that $(\hat{\boldsymbol{R}}^g - \boldsymbol{R}^g)_{\ell m} \xrightarrow{p} 0$ for any (ℓ, m). Now, let us consider the continuous map $\boldsymbol{f}(\boldsymbol{R}^g) = z_{1-\alpha,2,\boldsymbol{R}^g}$. By continuity of $\boldsymbol{f}$, we must also have that $\boldsymbol{f}(\hat{\boldsymbol{R}}^g) - \boldsymbol{f}(\boldsymbol{R}^g) \xrightarrow{p} 0$ as $N \to \infty$. Thus, $z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}$ is a consistent estimator for $z_{1-\alpha,2,\boldsymbol{R}^g}$. Therefore, as $N \to \infty$, the STP $\{\boldsymbol{\Omega}^g, \boldsymbol{T}^g, z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}\}$ asymptotically controls the FWER in the strong sense by Theorem 2 of Gabriel (1969). That is, given any non-empty $J \subset I$,

$$\lim_{N\to\infty} \Pr\left(\bigcup_{\ell \in J} \left\{|T_\ell^g| > z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}\right\} \,\middle|\, \tilde{H}_0^J\right) \le \alpha.$$

Also, because

$$\lim_{N\to\infty} \Pr\left(\bigcup_{\ell=1}^q \left\{|T_\ell^g| > z_{1-\alpha,2,\hat{\boldsymbol{R}}^g}\right\} \,\middle|\, H_0\right) = \alpha,$$

the asymptotic FWER control is exact. □

Appendix D: Computing SCIs for the treatment effects

As before, we write $\boldsymbol{g}_\ell(\boldsymbol{p}) = g(\boldsymbol{c}_{\ell,1}'\boldsymbol{p}) - g(\boldsymbol{c}_{\ell,2}'\boldsymbol{p})$ to denote the treatment effects, and $\boldsymbol{g}_\ell(\hat{\boldsymbol{p}}) = g(\boldsymbol{c}_{\ell,1}'\hat{\boldsymbol{p}}) - g(\boldsymbol{c}_{\ell,2}'\hat{\boldsymbol{p}})$ to denote the sample treatment effects. In this computation, we rewrite the statistics T_ℓ^g, $\ell = 1, \ldots, q$, using their definition and then solve for the treatment effect. Note that the probability is in fact an approximation because we are using the multivariate t-based approximation.

$$\begin{aligned}1-\alpha &\approx \Pr\left(\bigcap_{\ell=1}^{q}\left\{|T_\ell^g| \le t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\right\}\right)\\ &= \Pr\left(\bigcap_{\ell=1}^{q}\left\{\frac{\sqrt{N}|\boldsymbol{g}_\ell(\hat{\boldsymbol{p}})-\boldsymbol{g}_\ell(\boldsymbol{p})|}{\sqrt{\hat{\upsilon}_{\ell\ell}^g}} \le t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\right\}\right)\\ &= \Pr\left(\bigcap_{\ell=1}^{q}\left\{|\boldsymbol{g}_\ell(\hat{\boldsymbol{p}})-\boldsymbol{g}_\ell(\boldsymbol{p})| \le t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\sqrt{\hat{\upsilon}_{\ell\ell}^g/N}\right\}\right)\\ &= \Pr\left(\bigcap_{\ell=1}^{q}\left\{\boldsymbol{g}_\ell(\boldsymbol{p}) \in \left[\boldsymbol{g}_\ell(\hat{\boldsymbol{p}}) \pm t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\sqrt{\hat{\upsilon}_{\ell\ell}^g/N}\right]\right\}\right).\end{aligned}$$

Therefore, approximate $100(1-\alpha)\%$ SCIs for $\boldsymbol{g}_\ell(\boldsymbol{p})$, $\ell = 1, \ldots, q$, are given by

$$\left[\boldsymbol{g}_\ell(\hat{\boldsymbol{p}}) - t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\sqrt{\hat{\upsilon}_{\ell\ell}^g/N}, \boldsymbol{g}_\ell(\hat{\boldsymbol{p}}) + t_{1-\alpha,2,\nu^g,\hat{\boldsymbol{R}}^g}\sqrt{\hat{\upsilon}_{\ell\ell}^g/N}\right].$$

References

Bender, R., & Lange, S. (1999). Multiple test procedures other than Bonferroni's deserve wider use. *BMJ: British Medical Journal*, *318*(7183), 600–601. https://doi.org/10.1136/bmj.318.7183.600a.

Bocanegra, Y., Garcia, A., Pineda, D., Buritica, O., Villegas, A., Lopera, F., & Gomez, D. (2015). Syntax, action verbs, action semantics, and object semantics in Parkinson's disease: Dissociability, progression, and executive influences. *Cortex*, *69*, 237–254. https://doi.org/10.1016/j.cortex.2015.05.022.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*(2), 290–302. https://doi.org/10.1214/aoms/1177728786.

Bretz, F., Genz, A., & A Hothorn, L. (2001). On the numerical availability of multiple comparison procedures. *Biometrical Journal*, *43*(5), 645–656. https://doi.org/10.1002/1521-4036(200109)43:5<645::AID-BIMJ645>3.0.CO;2-F.

Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, *92*(440), 1494–1502. https://doi.org/10.2307/2965420.

Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. New York: Wiley.

Brunner, E., Konietschke, F., Pauly, M., & Puri, M. (2018). Rank-based procedures in factorial designs: hypotheses about nonparametric treatment effects. *Journal of the Royal Statistical Society, Series B*. https://doi.org/10.1111/rssb.12222.

Calian, V., Li, D., & Hsu, J. C. (2008). Partitioning to uncover conditions for permutation tests to control multiple testing error rates. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(5), 756–766. https://doi.org/10.1002/bimj.200710471.

Camilli, G. (1994). Teacher's corner: Origin of the scaling constant d= 1.7 in item response theory. *Journal of Educational Statistics*, *19*(3), 293–295. https://doi.org/10.2307/1165298.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:223.3.CO;2-D.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494. https://doi.org/10.1037/0033-2909.114.3.494.

Cohen, J. (1994). The earth is round (p <.05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997.

Cox, D. R. (1970). *Analysis of Binary Data*. Boston: Chapman & Hall/CRC.

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2)), 640–647. https://doi.org/10.3758/s13423-015-0913-5.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: a primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013.

Gabriel, K. R. (1969). Simultaneous test procedures-Some theory of multiple comparisons. *The Annals of Mathematical Statistics*, 224–250. https://doi.org/10.1214/aoms/1177697819

Gao, X., Alvo, M., Chen, J., & Li, G. (2008). Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. *Journal of Statistical Planning and Inference*, *138*(8), 2574–2591. https://doi.org/10.1016/j.jspi.2007.10.015.

Genz, A., & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-01689-9.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error*. Applied Mathematics and Statistics Laboratories, Stanford University.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*(1), 167–178. https://doi.org/10.1037/0033-2909.117.1.167.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802. https://doi.org/10.1093/biomet/75.4.800.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425.

Konietschke, F., Hothorn, L. A., & Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, *6*, 738–759. https://doi.org/10.1214/12-EJS691.

Lehmann, E. L. (2009). Parametric versus nonparametrics: Two alternative methodologies. *Journal of Nonparametric Statistics*, *21*(4), 397–405. https://doi.org/10.1080/10485250902842727.

Marozzi, M. (2016). Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Statistical Methods in Medical Research*, *25*(6), 2593–2610. https://doi.org/10.1177/0962280214529104.

Munzel, U., & Hothorn, L. A. (2001). A unified approach to simultaneous rank test procedures in the unbalanced one-way layout. *Biometrical Journal*, *43*(5), 553–569. https://doi.org/10.1002/1521-4036(200109)43:5<553::AID-BIMJ553>3.0.CO;2-N.

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software*, 50(i12). https://doi.org/10.18637/jss.v050.i12.

Noguchi, K., & Marmolejo-Ramos, F. (2016). Assessing equality of means using the overlap of range-preserving confidence intervals. *The American Statistician*, *70*(4), 325–334. https://doi.org/10.1080/00031305.2016.1200487.

Pesarin, F. (2001). *Multivariate Permutation Tests: With Applications in Biostatistics*. Wiley Chichester.

Pesarin, F., & Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley.

Reiczigel, J., Zakariás, I., & Rózsa, L. (2005). A bootstrap test of stochastic equality of two populations. *The American Statistician*, *59*(2), 156–161. https://doi.org/10.1198/000313005X23526.

Ryu, E. (2009). Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. *Statistics in Medicine*, *28*(25), 3179–3188. https://doi.org/10.1002/sim.3700.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–467. https://doi.org/10.1037/1082-989X.8.4.448.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*(1), 100–116. https://doi.org/10.1214/ss/1177011945.

Umlauft, M., Konietschke, F., & Pauly, M. (2017). Rank-based permutation approaches for non-parametric factorial designs. British Journal of Mathematical and Statistical Psychology. https://doi.org/10.1111/bmsp.12089

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. https://doi.org/10.3102/10769986025002101.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108.

Westfall, P. H., & Troendle, J. F. (2008). Multiple testing with minimal assumptions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(5), 745–755. https://doi.org/10.1002/bimj.200710456.

Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing*, (4th ed.). New York: Academic Press.

Wolfsegger, M. J., & Jaki, T. (2006). Simultaneous confidence intervals by iteratively adjusted alpha for relative effects in the one-way layout. *Statistics and Computing*, *16*(1), 15–23. https://doi.org/10.1007/s11222-006-5197-1.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.