How do my distributions differ? Significance testing for the Overlapping Index using Permutation Test

Ambra Perugini [1], Giulia Calignano [1], Massimo Nucci [2], Livio Finos [3], Massimiliano Pastore [1]

December 5, 2024

Abstract

The present contribution aims to introduce the application of the permutation test to the Overlapping Index to estimate effects of interest in psychological science. Starting with most common scenarios in psychological sciences the paper highlights the importance of relying on statistical methods that are resilient to the complexities inherent in psychological data, where assumption violations are often inevitable. Subsequently, we present a Simulation study to illustrate the practical implications and reliability of the proposed test in comparison to most commonly used tests. The findings show the good control of Type I error of the ζ overlapping test and how this approach outperforms in terms of power all other tests considered in the simulation, already prom small samples. The paper provides practical guidance and demonstrates the advantages of this methodology, emphasizing its potential to enhance transparency and rigor in psychological data analysis by shifting focus from traditional significance testing to comprehensive distributional evaluations.

1 Statistical testing choices in Psychology

Methodological choices in cognitive and behavioral sciences aim to combine data richness with data collection feasibility, and at the same time they aim to land on valid interpretation based on reliable and robust statistical methods. Classic examples, like reaction times, demonstrate how specific measures have achieved such an acceptable trade-off, and for example, this is true even by comparing the framework of in lab *vs* online data collection (Semmelmann & Weigelt, 2017). Nevertheless, even in the fortunate case of reaction times which have widespread and solid epistemic rationale of use (Grosjean, Rosenbaum, & Elsinger, 2001; Proctor & Schneider, 2018; Silverman, 2010) significance testing often relies on the rigid application of a few statistical methods that have gained popularity among the scientific community and are perpetrated *perinde ac cadaver* by formal guidelines (Cumming, Fidler, Kalinowski, & Lai, 2012), even if their limits and risks have always been noted in the field of psychology and beyond (Boneau, 1960).

In fact, there is a growing caution against blindly using statistical tools and analytical methods without a deep understanding of their assumptions and implications (Scheel, Tiokhin, Isager, & Lakens, 2021). In other words, it is increasingly apparent that relying solely on significance testing as a trustworthy measure is improbable without considering the assumptions inherent to specific statistical methods, such as the t-test, across various scenarios in psychology. In fact, considering the particular circumstances of application has consistently been crucial advice when deciding on significance testing methods (Fisher, 1925).

The present contribution aims to present a novel approach for statistical testing, in particular for comparing two groups/conditions. Specifically, the present work proposes applying the Permutation test (Pesarin, 2001) alongside the Overlapping index (η, Pastore & Calcagnì, 2019) to compare empirical distributions. Importantly, by simulating different scenarios, we compare widely used statistical tests in Psychological Sciences with this novel approach to evaluate the performance in controlling type I error and power. Understanding and applying significance testing properly is crucial to deriving meaningful conclusions from psychological research. Accordingly, we offer an alternative tool to manage the assumptions underlying these analytical approaches, and increase awareness in significance testing in psychology. But above all, and even most importantly, the advantage of the Overlapping Index and its test is to take into account at the same time mean, variance and shape with a single test.

Considering that parametric tests, such as linear models, used for statistical inference require strong assumptions that are unlikely to be respected in the aforesaid field, such as normality and homoscedasticity, alternative methods become optimal. The use of non-parametric methods is particularly beneficial when these assumptions are violated (?, ?). Specifically, when using a t-test

to compare two groups or two experimental conditions using a given variable, it functions as a straightforward version of linear regression. This statistical process necessitates assumptions about the error terms, such as their independence and normal distribution, to be met. In cases such as reaction times, these assumptions might be violated if they are not properly addressed. Most importantly, two populations might present the same mean, yet their distributions largely differ in other characteristics, such as variability, leading to genuinely distinct groups (see figure 1).

The remainder of this article is structured as follows. First, we introduce the concept of the Overlapping Index, providing foundational definitions and highlighting its importance. Next, we define the Permutation approach and explore its application to the Overlapping Index, showcasing its relevance in statistical analysis. Subsequently, we present a Simulation study to illustrate the practical implications and trustworthiness of the overlapping index utilizing permutations.

The rationale behind these steps involves first introducing the concept of the Overlapping Index (η), which is crucial because it provides an intuitive measure of similarity between distributions by quantifying the overlapping area of their empirical density functions, a common question in quantitative psychology. The Permutation approach is then defined and applied to the Overlapping Index, demonstrating how nonparametric methods can offer insights without relying on typical parametric assumptions. Specifically, the Permutation test involves shuffling data points to generate a sampling distribution, allowing the calculation of a p-value and highlighting its utility in assessing the statistical significance when comparing two groups/conditions. Next, the Simulation study uses a set of different combinations of parameters to simulate various scenarios that might meet or violate the assumptions of different statistical tests, modeling a range of conditions reflective of real-world complexities in psychological research. This simulation facilitates the evaluation of the statistical power (probability of correctly rejecting a false null hypothesis) and the type I error rate (likelihood of incorrectly rejecting a true null hypothesis) of each approach. To this end, several statistical tests are compared: the t-test for independent samples, assuming equal variances; the Welch test, which does not assume equal variances; the Wilcoxon test, suitable for ordinal data or when normality is not assumed; the Permutation Test on the Overlapping Index, providing a non-parametric approach to evaluate distributional differences; the F test for examining the homogeneity of variances; and the Kolmogorov-Smirnov test for comparing two distributions regardless of their underlying forms. These results enable researchers to visualize and comprehend the reliability and utility of each approach, particularly valuable in scenarios commonly encountered in quantitative psychology, where navigating data characteristics and adhering to or deviating from test assumptions is crucial.

Finally, we discuss the results, offering insights into the strengths and limitations of the Permutation-based Overlapping index and its potential applications in psychological sciences.

2 Overlapping Index

The overlapping index (η) is an intuitive way to define the area intesected by two or more empirical density functions (Pastore & Calcagnì, 2019). In a simple way, two distributions are similar when their distribution functions overlap, and as η diminishes, the two distributions differ. The η index varies from zero – when the distributions are completely disjoint – and one – when they are completely overlapped (Pastore, 2018). The simple interpretation of the overlapping index (η) makes its use particularly suitable for many applications (e.g. Sirbiladze, Midodashvili, & Manjafarashvili, 2024; Habibi, Achour, Bounaceur, Benaradj, & Aulagnier, 2024; Ricote et al., 2024; Rossi et al., 2024; Einbeck, Coolen-Maturi, Uwimpuhwe, & Singh, 2024; Wachendörfer & Oeberst, 2024; Rohrbach, 2024; Hawkins et al., 2024; Upadhayay, Granger, & Collins, 2024; Conversano, Frigau, & Contu, 2024; Pietrabissa et al., 2024; Nougaret et al., 2024; Greene, Guitard, Forsberg, Cowan, & Naveh-Benjamin, 2024; Schuetze & Yan, 2023; Karrobi et al., 2023; Garofalo, Giovagnoli, Orsoni, Starita, & Benassi, 2022; Jensen & Sanner, 2021).

More formally, assuming two probability density functions $f_A(x)$ and $f_B(x)$, the overlapping index $\eta : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1]$ is formally defined in the following way:

$$\eta(A, B) = \int_{\mathbb{R}^n} min[f_A(x), f_B(x)]dx \tag{1}$$

where, in the discrete case, the integer can be replaced by summation. As previously mentioned, $\eta(A, B)$ is normalized to one and when the distributions of A and B do not have points in common, meaning that $f_A(x)$ and $f_B(x)$ are disjoint, $\eta(A, B) = 0$. This index provides an intuitive way to quantify the agreement between A and B based on their density functions (Inman & Bradley Jr, 1989).

To quickly illustrate an example of the overlapping area in figure 1 are represented two different empirical densities. In panel [A], are depicted two density distributions, a Normal(10,2) and a Uniform(0,20); note that the two distributions have the same mean (10), but different variance, 4 and 33.3 respectively. In the panel [B] are represented the empirical densities of two random samples of 30 observations drawn from the two populations specified as in panel [A]; the estimated overlapping area being $\eta = 0.46$.

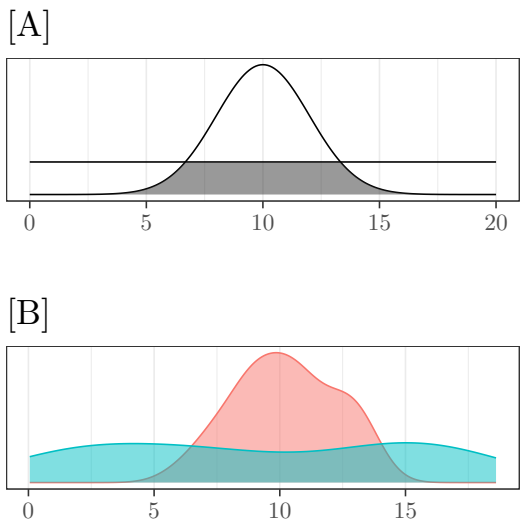

Figure 1: Comparison of a normal distribution and a uniform distribution with same mean, 10, and different variances, 4 and 33.3, respectively.

The figure 1 shows how two distributions with almost same mean could still be very different from each other with the overlapping area being $\hat{\eta} = 0.46$. In this case, the t-test focuses on mean differences, therefore correctly does not rejects the null hypothesis, even though the degree of similarity of the two densities is only 46%, in other words, the difference is about 54%. Moreover, we remind that the t-test in this case is far from ideal as the two distributions have different variances.

2.1 Permutation approach

Permutation tests, also known as randomization tests, are a class of nonparametric statistical significance tests. The concept dates back to the work of R.A. Fisher in the 1930s, in particular his book "The Design of Experiments" (Fisher, 1935). The theoretical foundations were further developed by E.J.G. Pitman in his seminal papers of 1937 (Pitman, 1937) and 1938 (Pitman, 1938). The basic principle of permutation testing is based on the idea of rearranging observed data to generate a null distribution. This approach assumes that if the null hypothesis is true, then all possible arrangements of the data are equally likely, i.e., each permuted sample has the same probability as the observed one. By resampling the data, we can obtain the distribution of the test statistic under the null hypothesis without making any assumptions about the underlying data generating process. This is particularly valuable when dealing with small sample sizes or when the assumptions of parametric tests are not met. The observed test statistic is then compared to this empirically derived null distribution to determine the probability of obtaining such a result by chance alone (Pesarin, 2001). The permutation approach allows for the adoption of any test statistic chosen by the user. For example, if we are thinking about comparing the means of the two samples, we could choose a t-test statistic; the data in the two groups are permutated and the t value is calculated each time. If the two groups come from the same population, the t-statistic computed on the observed data should be close to 0; the t-statistic computed on randomly permuted data will also give values close to zero. Therefore, the randomly generated test statistic and the observed one have the same – nonparametric – distribution. Otherwise, if the two groups come from populations with different means, the t-statistic computed on the observed data will be far from zero, while the t-statistic computed on the permuted data will be around zero.

The p-value is the probability of obtaining an equal or more extreme t-statistic compared to the observed one. observed:

$$p = \frac{(\#_{b=1}^{B}|t_b| \geq |t|) + 1}{B + 1} \tag{2}$$

where B is the number of random permutations, t is the t-statistic computed on the observed data, t_b are those computed on the permuted data.

The test will have power – i.e., the probability of getting a $p \leq \alpha$ when the two samples are really different – very close to the parametric t-test, and it will retain control over false positives even when the assumptions of normality are not met.

It is important to note that the choice of which t-statistic to use is a user choice; different test statistics (e.g., difference of mean ranks, Kolmogorov-Smirnov, etc.) will produce tests with different power. For example, if the two samples differ only in their variability and not in their mean, the permutation test based on the t-statistic will have little or no power to detect that the two samples come from different populations. In this direction, this paper proposes to use the overlap index as a test statistic that results to be powerful under a wide range of difference in distributions.

Remark 1 The choice to add a +1 in the numerator and denominator is a choice supported by many authors (Phipson & Smyth, 2010; Hemerik & Goeman, 2018) and ensures that the probability of false positives is less than or equal to α.

Remark 2 As one can understand, the p-value may change depending on the permutations that are drawn. By increasing the number of permutations B, the results will change less and less. Since the number of possible permutations is finite, it is preferable, if possible, to explore the set of them (i.e., to compute the statistics on all possible permutations of the data). This set of all possible rearrangements of the data is, in fact, the orbit of the sample that allows us to compute the exact p-value - i.e., the exact probability of observing a test statistic that is as extreme or more extreme than that observed

in the data. In this case, $B = \binom{n}{n_1} = \frac{n!}{n_1!(n-n_1)!}$ and the p-value formula reduces to

$$p = \frac{(\#_{b=1}^{B} |t_b| \geq |t|)}{B}$$

since the test statistic computed on the observed data is certainly in this set.

2.2 Application of permutation test to the overlapping index

Even though the overlapping index has a simple interpretation, one could argue that it does not provide information on the significance of η, therefore, we decided to implement permutation testing to offer to the ones interested a value of significance. In particular, we implemented permutations test, to give a tool that tests differences in distributions without assumptions, offering a valid alternative in cases in which traditional assumtions are not met.

If we are reasoning from the perspective of Null Hypothesis Significance Testing (NHST), we should define the null hypothesis as follows: $H_0 : \eta = 1$, meaning that there is complete overlap between the theoretical densities in the two populations from which we sample the data. For this reason, it is more intuitive to work with the complement of η, which is $1 - \eta = \zeta$ which is the area of non-overlap, therefore, defining the null hypothesis as $H_0 : \zeta = 0$, once again meaning that there is no difference between the densities of the two populations. Obviously, this does not change the results, but only the way in which they are interpreted. When testing the difference between the two distributions, we will no longer be working with η, but with the complement ζ.

The algorithm estimates the value of ζ on the observed data ($\hat{\zeta}$). Then, through permutation, the observed values of the two groups are randomly re-assigned to the groups for B times, estimating again the new value of $\hat{\zeta}_b$. The times in which the estimate of $\hat{\zeta}_b$ on permuted data is higher or equal than the one observed on real data is estimated ($\hat{\zeta}_b \geq \hat{\zeta}$) and then the found value is divided by B, returning the p-value. A typical example of data not respecting previously said assumptions is reaction times and for this purpose we present a real case of a dataset available online (Oksuz & Rebuschat, 2024) on reaction times of word reading of high and low frequency words in English and we implement on the overlapping function the permutation test.

In the figure 2[A] are represented the densities of reaction times of word reading of high and low frequency words in English from a sample of two groups of 30 observations each; the two distributions have mean 0.78 and 1.11, variance 0.07 and 0.22, and skewness 2.32 and 2.02 respectively. The obtained value of $\hat{\eta}$ is 0.44, and consequently $\hat{\zeta}$ is 0.56. In figure 2[B] is represented the distribution of the values of $\hat{\zeta}$ obtained with 2000 permutations; we can calculate the p-*value* as follows:

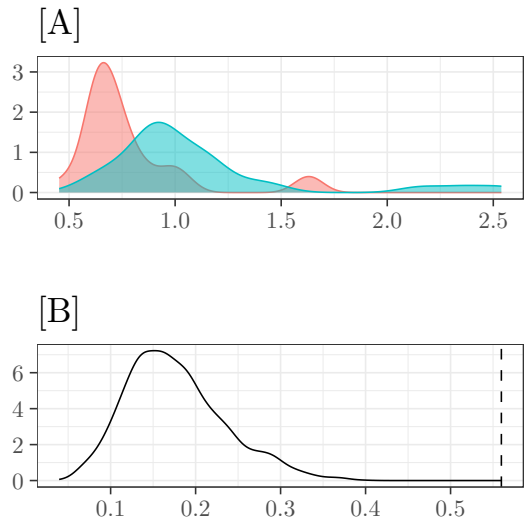

Figure 2: [A] Distribution of reaction times of word reading of high and low frequency words in English, in this case the non-overlapping area is $\hat{\zeta} = 0.56$; [B] Distribution of $\hat{\zeta}$ obtained with 2000 permutations of the data.

$$p = \frac{(\#\hat{\zeta}_b \geq \hat{\zeta}) + 1}{B + 1} = \frac{1}{2001}$$

Given that $p < .001$, we can conclude that the difference is statistically significant; in this case the t test give the same conclusion $t_{(58)} = 3.34$, $p = .001$, but we remind that the t-test evaluates only the difference between the means and requires assumtions that in this scenario are clearly not met.

2.3 Practical application

The overlapping test is easily performed using the `overlapping` R package (Pastore, Loro, Mingione, & Calcagni', 2024). In the following, we briefly present a simple example.

First, we simulate two different samples, each with $n = 50$: one from a Normal(3,2) distribution and another from a $\chi^2(3)$ distributions:

```
set.seed(1)
n <- 50
y1 <- rnorm( 50, 3, 2 )
y2 <- rchisq( 50, 3 )
```

The simulated data have means of 3.2 and 2.97, and variances of 2.76 and 4.72, respectively. The overlapping area between the two distributions is $\eta = 0.74$.

Figure 3 presents the obtained density distributions. Note that the t-test results are not significant: $t(98) = 0.6$, $p = .55$. We obtain the same result when adjusting the test for unequal variances (Welch correction): $t(91.75) = 0.6$, $p = .55$.

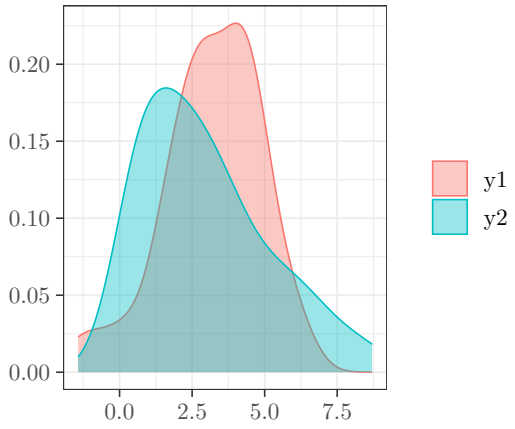

Figure 3: Simulated example.

To perform the overlapping permutation test, we can use the following code:

```
library( overlapping )
yList <- list( y1 = y1, y2 = y2 )
perm.test( yList )
```

```
$Zobs
[1] 0.261

$pval
[1] 0.00699
```

Given that $p = .007$ we can conclude that there is a statistically significant difference between the two distributions. This result suggests that the overlapping method has detected differences that the previous t-tests did not identify, highlighting the potential sensitivity of this approach.

3 Simulation study

To evaluate the performance of the permutation test applied to the overlapping index, we performed a simulation study. The aim is to generate data for a set of scenarios distinguishing mean, variance and shape of the populations and compare the ζ perm test to other commonly used tests in terms of type I error control and power.

3.1 Data generation

In the simulation, two density distributions will be compared for many different scenarios. The first distribution will always be a normal standard distribution with $\mu = 0$ and $\sigma = 1$. To simulate data for the second distribution we use the Skew-Normal distribution (Azzalini, 1985), which is defined in the following way: given $\xi \in \mathbb{R}$, $\omega \in \mathbb{R}^+$ and $\alpha \in \mathbb{R}$, then for $y \in \mathbb{R}$ we have

$$\mathcal{SN}(y|\xi,\omega,\alpha) = \frac{1}{\omega\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\xi}{\omega}\right)^2\right]\left[1+\operatorname{erf}\left(\alpha\left(\frac{y-\xi}{\omega\sqrt{2}}\right)\right)\right] \tag{3}$$

in which

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$$

is the *error function*. When $\xi = 0$, $\omega = 1$ and $\alpha = 0$ the distribution is a standard normal distribution.

ξ is the location parameter, ω is the scale parameter and α is related to the skewness of the distribution. Therefore, this distribution is suitable to generate data modelling both the distance between means (the effect size), symmetry and variance.

Mean and variance of the Skew-Normal are respectively:

$$\begin{aligned} \mu &= \xi + \omega\gamma\sqrt{2/\pi} \\ \sigma^2 &= \omega^2[1-(2\gamma^2)/\pi] \end{aligned} \tag{4}$$

in which $\gamma = \alpha/\sqrt{1+\alpha^2}$. Based on the equations (4) we can determine the values to assign to the parameters ξ e ω in function of μ and σ with the equations:

$$\begin{aligned} \xi &= \mu - \omega\gamma\sqrt{2/\pi} \\ \omega &= \sqrt{\sigma^2/[1-(2\gamma^2)/\pi]} \end{aligned} \tag{5}$$

The Skew-Normal distribution is optimal for our purpose as it allows to have control over parameters of mean, variance, skewness and kurtosis, as shown in figure 4.

3.2 Simulation design

In the simulation we confront two samples extracted from a Skew-Normal, the first one is generated from $\mathcal{SN}(0,1,0)$, which is the Standard-Normal distribution, and the second one from $\mathcal{SN}(\xi,\omega,\alpha)$. Consequently, the first sample derives always from a population with mean 0 and variance 1. To define the various scenarios, we manipulate the parameters of the second population in orther to obtain specific differences in means (δ), standard deviations (σ) and skewness (α). Four factors were sistematically varied ina complete four-factors design as follows:

- $\delta = (0, 0.2, 0.5, 0.8)$; mean of the second population, which corresponds also to the difference between the two groups, the first one has always $\mu = 0$;
- $\sigma = (1, 2, 3)$; standard deviation of the second population;
- $\alpha = (0, 2, 10)$; degree of asymmetry (skewness) of the second population;
- $n = (10, 20, 50, 100, 500)$; sample size, equal in the two samples.

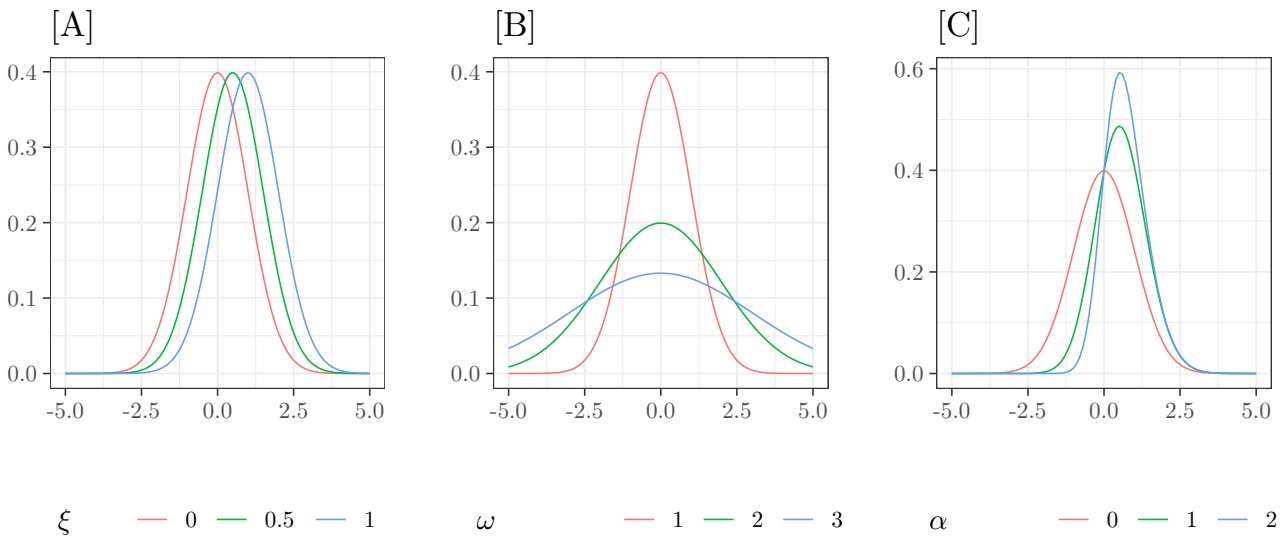

Figure 4: Examples of Skew-Normal distributions (ξ,ω,α); [A] three densities with same variance and shape but different location parameter values ($\xi = 0, 0.5, 1$), [B] three densities with same mean and shape but different scale parameter values ($\omega = 1, 2, 3$) and [C] three densities with same mean and variance but different shape parameter values ($\alpha = 0, 1, 2$).

For each of the $4 \times 3 \times 3 \times 5 = 180$ conditions we generated 3000 sets of data on which we performed the analysis.

In figure 5 are graphically represented the 36 scenarios of data generation, the black curves are the first population, always a $\mathcal{SN}(0, 1, 0)$, and the red curves are relative to the second population $\mathcal{SN}(\xi, \omega, \alpha)$.

For each combination $\delta \times \sigma \times \alpha \times n$, on the generated data were performed the following tests:

- t test for independent samples, assuming equal variance;
- Welch test for independent samples;
- Wilcoxon test for independent samples;
- Permutation test on the complement of the overlapping index, $\zeta = 1 - \eta$, which therefore becomes an index of difference between groups;
- F test of homogeneity of variances;
- Kolmogorov-Smirnov test for comparing two distributions.

The whole procedure generated a total of 540000 datasets as well as 3240000 of statistical tests and corresponding p-values.

3.3 Definition of Statistical tests

We introduce the chosen statistical tests summarizing the specific hypothesis and assumptions for each one.

3.3.1 t test

This is the classic case of a test for independent samples assuming equal variances and the normality of the two distributions:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ with } \sigma_1 = \sigma_2$$

3.3.2 Welch (W) test

This is the t test modified when homogeneity of variances is not respected:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ with } \sigma_1 \neq \sigma_2$$

Also this test assumes the normality.

3.3.3 Wilcoxon-Mann-Whitney (WMW) test

This is the test on ranks which evaluates the following hypothesis without assumptions on distributions:

$$H_0 : P(X_1 > X_2) = P(X_2 > X_1) = 0.5$$

in which X_1 and X_2 are the random variables representing the observations extracted from the two populations.

3.3.4 Kolmogorov-Smirnov (KS) test

This test compares the cumulative distributions

$$H_0 : F(X_1) = F(X_2)$$

without assumptions on distributions.

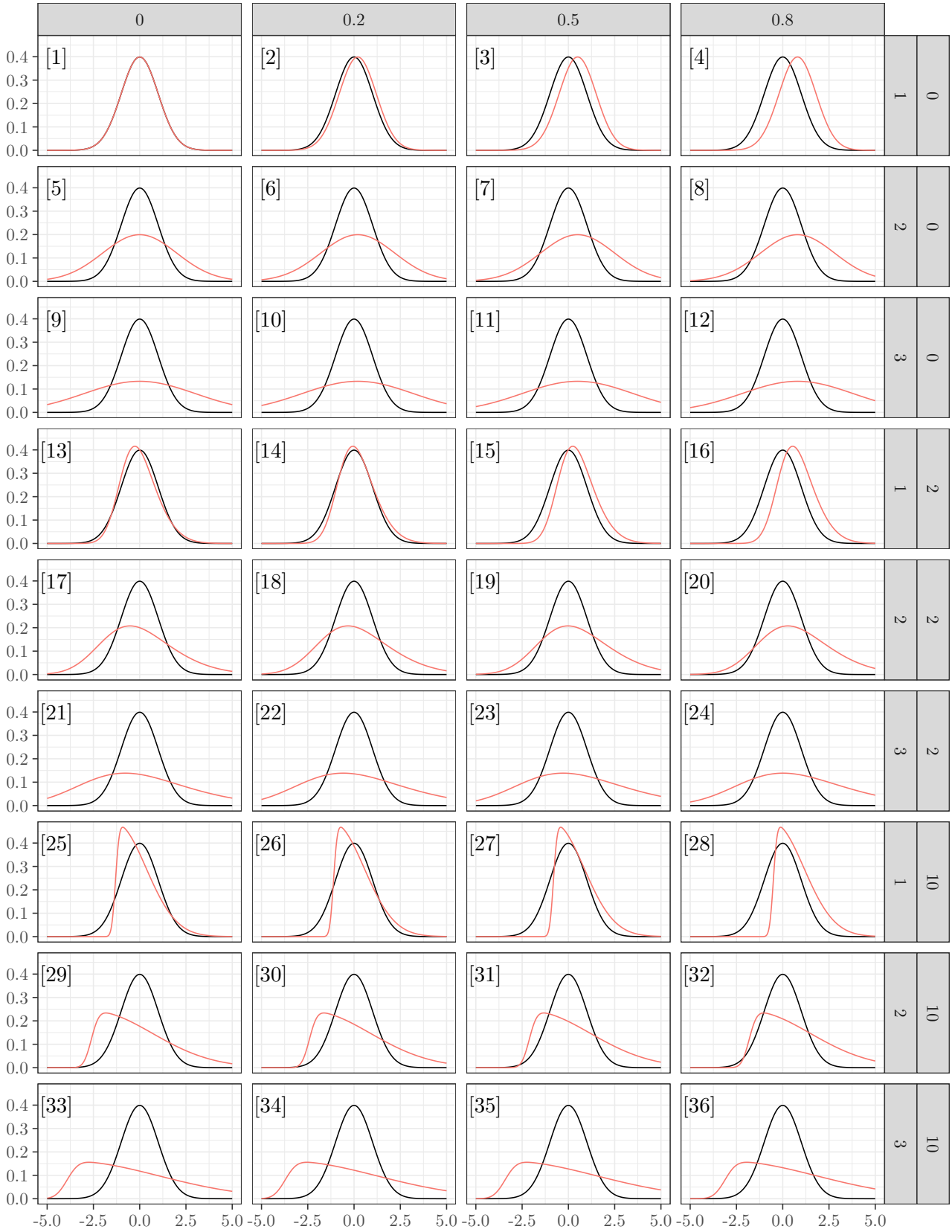

Figure 5: Generative data distributions in function of δ (column panels) and σ (row panels). The black curves are the first sample, $\mathcal{SN}(0, 1, 0)$, the red ones represent second sample.

3.3.5 F test

This is the test of homogeneity of variances

$$H_0 : \sigma_1^2 = \sigma_2^2$$

assuming the normality.

3.3.6 ζ overlapping (ζ_{ov}) test

Since $\zeta = 1 - \eta$, in which η is the area of overlapping of the empirical distributions, the null hypothesis of the test is

$$H_0 : \zeta = 0$$

which implies that the data comes from the same population, or from populations with same shape (mean, variance and skewness) but without specific assumptions.

4 Results

First of all, we analysed correlations between the p-values of the considered tests in order to assess how much they are associated independently from the experimental condition.

Next, we considered panel [1], figure 5, the scenario in which all null hypothesis are true and assumptions are respected for all tests. Consequently, we computed type I error by counting how many times the test is significant in this scenario, and the power by counting how many times it will be significant in all other scenarios. In this way, we evaluated type I error and power based on the experimental conditions.

4.1 Correlations among tests

Figure 6 represents the correlation matrix between the p-values for the different tests in all experimental conditions. The classical tests show an order in the way they correlate. More specifically, t and W tests show a perfect correlation, WMW is highly correlated with the aforesaid tests, and the KS shows a lower but still medium-large correlation. F presents no correlation with t, W and WMW tests, and medium correlation with the ζ_{ov} and KS tests.

The ζ_{ov} test is highly correlated with the KS test, has a lower correlation with tests on means (t and W) and ranks (WMW), and a medium correlation with the F test.

4.2 Type I error and power

In figure 7, is represented type I error in panel [A] and power in panel [B] estimated considering as true null hypothesis the situation in which samples are drawn from two exactly equal populations (figure 5, panel [1]).

In relation to type I error, all tests show a good performance, whereas the KS test is too conservative for small samples.

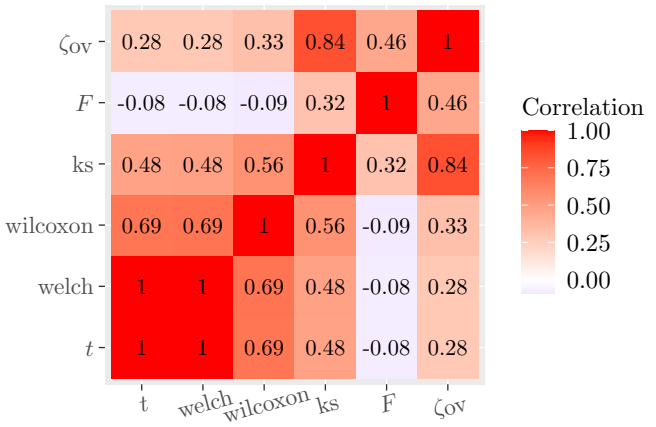

Figure 6: Correlation matrix among p-values ($N = 540000$) in chosen tests. Note: ks = Kolmogorov-Smirnov test, $\zeta_{ov} = \zeta$ overlapping test, F = variance test, wilcoxon = Wicoxon-Mann-Whitney test, welch = Welch test, t = Student's t test.

Concerning power, the ζ overlapping test outperforms all other tests, already with small sample sizes. From the graphical representation it is visible how two subgroups can be identified: one including the tests on means and ranks, not reaching adequate power even with large samples, and the second one formed by the ζ_{ov} and KS tests, reaching good power already from 100 observations, with the ζ_{ov} outperforming the KS test reaching good power already from 50 observations.

5 Discussion

The analysis of p-value correlations among the tests provides insight into their relationships. High correlation between the t-test and W test, for instance, reflects their similar objectives and shared focus on mean differences. However, lower correlations between parametric and non-parametric methods, such as the WMW and KS tests, indicate that these tests capture distinct aspects of the data, such as ranks or distribution shapes rather than means. The permutation-based tests show intermediate correlations with both parametric and non-parametric methods, which suggests that they may align with either types of tests depending on the underlying data structure, highlighting the versatility of the ζ_{ov} test.

Moreover, the present analysis evaluated the performance of various statistical significance tests across simulated scenarios comparing each test on the same null hypothesis, being no difference in the two populations from which the samples are drawn. The tests include both parametric (such as the T-test, Welch test, and F-test for variance) and non-parametric methods (such as the Wilcoxon Signed-Rank, Kolmogorov-Smirnov, and permutation-based approaches including the ζ_{ov} test). Through these scenarios, we assess each test's robust-

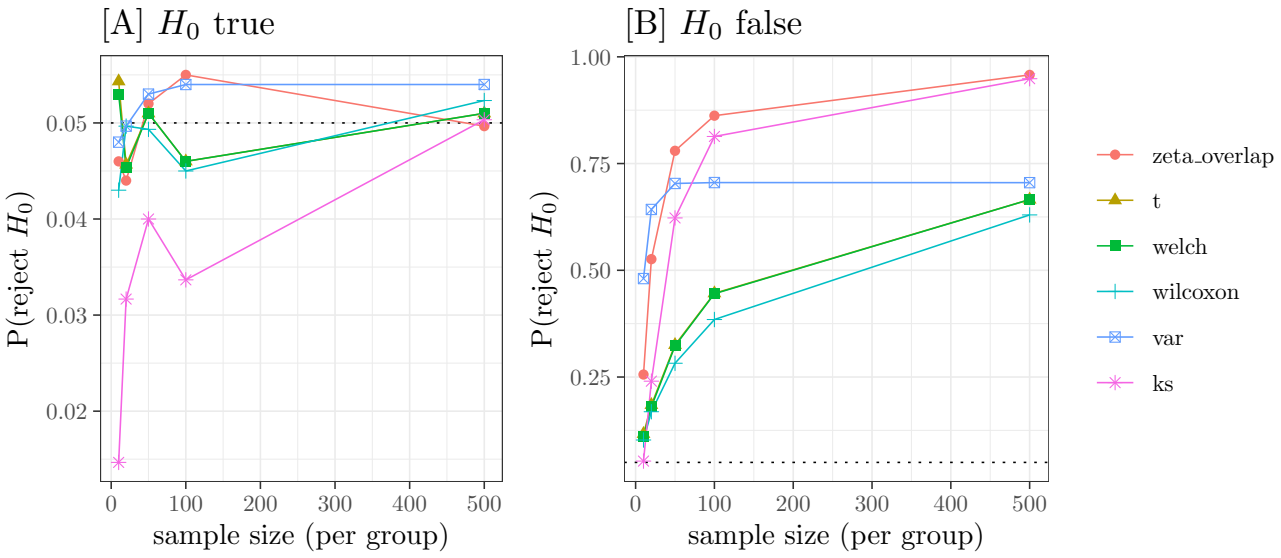

Figure 7: Control of type I error [A] and power [B] in the various tests.

ness, Type I error control, and power.

In figure 7, panel [A], the only test which is over-conservative in terms of Type I error is the F test, yet reaching the nominal 0.05 level with bigger samples (over 400 observations). All other tests, already from small samples, are in the range of $0.045 - 0.055$, converging closer to 0.05 as sample size grows, except for the F test remaining less conservative than the others. From this analysis we can then conclude that all tests are able to control Type I error for bigger samples, yet caution is needed for small samples when using the F test.

Concerning power, panel [B], the ζ_{ov} test outperforms all other tests. The only one competing with good power is the KS test, reaching a power of 60% with 50 observations and aligning with the ζ_{ov} at 500 observations. The F test starts at higher power for small samples but fails to reach an adequate power of 80%. W and t tests performances are identical, in line with their perfect correlation of p-values, never exceeding 60% of power and performing consistently worse than the ζ_{ov} and KS tests, especially with small samples. Lastly, the WMW test performs similarly to t and W tests with slightly less power throughout the increase of sample size. These results highlight the outstanding performance of the ζ_{ov} test already from small samples, showing an advantage in choosing this test in research settings regardless of data distribution and assumptions.

Advantages and Limitations of the ζ Permutation Test

The ζ **overlapping test**, designed to measure the degree of overlap between distributions, has specific advantages and limitations. Its main strength lies in its robustness to distributional assumptions, as it calculates p-values through permutations rather than relying on parametric assumptions like normality or equal variance. This makes it particularly useful in scenarios where other tests may fail due to assumption violations, providing a conservative Type I error rate when H_0 is true and robust power when H_0 is false.

The main limitation of the test is that it does not inform on specific parametric differences, as its design focuses on distributional overlap rather than mean differences, which means it does not directly inform on mean, variance or skewness differences specifically. But we argue that the ζ_{ov} test offers a great starting point to evaluate whether two distributions differ in the first place with high power already from small samples. Moreover, the package to compute the index also offers the possibility to plot densities and the area of overlap, therefore making it extremely intuitive to visualize how the two distributions practically differ. In cases in which the test fails to find a difference between the two distributions one can conclude that the two groups/conditions do not differ in any parameter of possible interest, but if the test finds a statistically significant difference, then the researcher can move on to test which are the specific parameters differing between the two. This approach highlights the importance of visualizing data and stresses out the invalue insight offered by descriptive statistics.

The ζ permutation test is indeed a valuable tool for non-parametric inference, particularly when distributional assumptions do not meet those required by common statistical test e.g. t-test. These are particularly relevant points given that in psychological sciences studies often involve small sample sizes, and relying on small changes in location parameters like the mean can be risky. For example, small samples are highly susceptible to the influence of extreme values, which can skew

the mean and lead to misleading conclusions about effect sizes. Even more importantly, as demonstrated in simulations, the ζ_{OV} test is less prone to being dramatically impacted by extreme values, as it directly measures the distributional overlap between groups rather than focusing solely on mean differences. This characteristic makes such test particularly valuable in small-sample contexts, especially in psychological science where robustness to outliers is critical for obtaining reliable insights into group differences.

6 Conclusion

By exploring alternative scenarios, the study offers practical indication to operate a shift in the philosophical approach to data analysis and significance testing. In fact, the Overlapping index forces the functional interpretation of the results to move beyond significance testing alone (Pastore, 2018; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; Gelman, 2018). In psychological research, considering the distribution of data rather than relying solely on significance testing offers a deeper, more nuanced understanding of results. Traditional significance testing doesn't provide information about the nature or magnitude of that effect. By visualizing and considering the entire distribution of data, researchers can observe the spread, central tendency, and shape of the data, which often reveal valuable insight about variability and individual differences within the sample. As presented in figure 3, reporting a mean difference without an understanding of the data's variability could lead to misrepresentation of the consistency or generalizability of the observed effect. Therefore, incorporating distributional analyses allows psychologists to present a fuller picture of their findings, improving both interpretability and transparency in their research conclusions.

Moreover, the present study further underscores the necessity of reasoning on the most suitable statistical tools contingent on the specific characteristics of the data and the assumptions inherent in the analytical techniques employed. Such a switch in the philosophical approach to data analysis in psychological sciences (Vasishth & Gelman, 2021) may improve the robustness and validity of psychological research findings, allowing for more aware interpretations and generalizations. We stress this by making open available data and material so that such an approach might be useful for a wide range of psychologists interested in increasing the understandability of their results.

The findings underscore the necessity of choosing statistical methods that are resilient to the complexities inherent in psychological data, where assumption violations are often inevitable. The ζ overlapping test is a robust alternative to commonly used tests in psychological science that accommodates data with unequal variances or non-normal distributions, offering reliable results even when classic parametric conditions are unmet. In this way, our test extends the flexibility of significance testing, enabling a nuanced understanding of effects in psychology.

Ultimately, statistics in psychology should reflect both theoretical knowledge and an appreciation for the distributional nuances of psychological variables. Rather than a rigid application of conventional methods, statistical analysis should be a deliberate choice that aligns with the nature of the data and the research question. The approach of the ζ overlapping test embodies this principle, capturing the depth and complexity of psychological effects in a way that is both methodologically rigorous and sensitive to the real-world structure of psychological phenomena.

Legenda

η is the area of overlap
ζ is the area of non overlap, therefore $1 - \eta$
μ is the parameter of the mean of the normal standard
σ is the standard deviation of the normal standard
δ is the difference between the two means
ξ is the location parameter of the skew-normal
ω is the scale parameter of the skew-normal
α determins the simmetry of the skew-normal

Ethical considerations

Ethical approval was not required

Conflicting interest

The authors declare no conflict of interests.

Funding statement

No Funding supported this project.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological bulletin*, *57*(1), 49.

Conversano, C., Frigau, L., & Contu, G. (2024). Overlapping coefficient in network-based semi-supervised clustering. *Computational Statistics*, 1–24.

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*(3), 138–146.

Einbeck, J., Coolen-Maturi, T., Uwimpuhwe, G., & Singh, A. (2024). A comparison of threshold-free measures for assessing the effectiveness of educational interventions. *The Journal of Experimental Education*, 1–18.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical proceedings of the cambridge philosophical society* (Vol. 22, pp. 700–725).

Fisher, R. A. (1935). *The design of experiments*. Macmillan.

Garofalo, S., Giovagnoli, S., Orsoni, M., Starita, F., & Benassi, M. (2022). Interaction effect: Are you doing the right thing? *PLoS One*, *17*(7), e0271668.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16–23.

Greene, N. R., Guitard, D., Forsberg, A., Cowan, N., & Naveh-Benjamin, M. (2024). Working memory limitations constrain visual episodic long-term memory at both specific and gist levels of representation. *Memory & Cognition*, 1–25.

Grosjean, M., Rosenbaum, D. A., & Elsinger, C. (2001). Timing and reaction time. *Journal of Experimental Psychology: General*, *130*(2), 256.

Habibi, I., Achour, H., Bounaceur, F., Benaradj, A., & Aulagnier, S. (2024). Predicting the future distribution of the barbary ground squirrel (atlantoxerus getulus) under climate change using niche overlap analysis and species distribution modeling. *Environmental Monitoring and Assessment*, *196*(11), 1–18.

Hawkins, S. J., Gärtner, Y., Offner, T., Weiss, L., Maiello, G., Hassenklöver, T., & Manzini, I. (2024). The olfactory network of larval Xenopus laevis regenerates accurately after olfactory nerve transection. *European Journal of Neuroscience*.

Hemerik, J., & Goeman, J. (2018). Exact testing with random permutations. *Test*, *27*(4), 811–825.

Inman, H. F., & Bradley Jr, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-theory and Methods*, *18*(10), 3851–3874.

Jensen, T. M., & Sanner, C. (2021). A scoping review of research on well-being across diverse family structures: Rethinking approaches for understanding contemporary families. *Journal of Family Theory & Review*, *13*(4), 463–495.

Karrobi, K., Tank, A., Fuzail, M. A., Kalidoss, M., Tilbury, K., Zaman, M., ... Roblyer, D. (2023). Fluorescence Lifetime Imaging Microscopy (FLIM) reveals spatial-metabolic changes in 3D breast cancer spheroids. *Scientific Reports*, *13*(1), 3624.

Nougaret, S., Ferrucci, L., Ceccarelli, F., Sacchetti, S., Benozzo, D., Fascianelli, V., ... Genovesio, A. (2024). Neurons in the monkey frontopolar cortex encode learning stage and goal during a fast learning task. *Plos Biology*, *22*(2), e3002500.

Oksuz, D. C., & Rebuschat, P. (2024, Mar). *Collocational processing in typologically different languages, english and turkish.* OSF. Retrieved from osf.io/muwjz

Pastore, M. (2018). Overlapping: a R package for estimating overlapping in empirical distributions. *Journal of Open Source Software*, *3*(32), 1023.

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: a distribution-free overlapping index. *Frontiers in psychology*, *10*, 1089.

Pastore, M., Loro, P. A. D., Mingione, M., & Calcagni', A. (2024). overlapping: Estimation of overlapping in empirical distributions [Computer software manual]. (R package version 2.2)

Pesarin, F. (2001). *Multivariate permutation tests: with applications in biostatistics*. Wiley.

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, *9*(1).

Pietrabissa, G., Semonella, M., Marchesi, G., Mannarini, S., Castelnuovo, G., Andersson, G., & Rossi, A. A. (2024). Validation of the Italian version of the web screening questionnaire for common mental disorders. *Journal of Clinical Medicine*, *13*(4), 1170.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, *4*(2), 225–232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: Iii. the analysis of variance test. *Biometrika*, *29*(3/4), 322–335.

Proctor, R. W., & Schneider, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, *71*(6), 1281–1299.

Ricote, N., Weinberger, C., Ramírez-Otarola, N., Bustamante, S., Málaga, M. L., Barceló, G., ... Maldonado, K. (2024). The interplay of resource availability and parent foraging strategies on juvenile sparrow individual specialization. *Journal of Avian Biology*, e03391.

Rohrbach, T. (2024). Are women politicians kind and competent? disentangling stereotype incongruity in candidate evaluations. *Political Behavior*, 1–24.

Rossi, A. A., Panzeri, A., Fernandez, I., Invernizzi, R., Taccini, F., & Mannarini, S. (2024). The impact of trauma core dimensions on anxiety and depression: a latent regression model through the Post-

Traumatic Symptom Questionnaire (PTSQ). *Scientific Reports*, *14*(1), 23036.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755.

Schuetze, B. A., & Yan, V. X. (2023). Psychology faculty overestimate the magnitude of cohen'sd effect sizes by half a standard deviation. *Collabra: Psychology*, *9*(1), 74020.

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*, 1241–1260.

Silverman, I. W. (2010). Simple reaction time: It is not what it used to be. *The American journal of psychology*, *123*(1), 39–50.

Sirbiladze, G., Midodashvili, B., & Manjafarashvili, T. (2024). Divergence and similarity characteristics for two fuzzy measures based on associated probabilities. *Axioms*, *13*(11), 776.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Upadhayay, H. R., Granger, S. J., & Collins, A. L. (2024). Comparison of sediment biomarker signatures generated using time-integrated and discrete suspended sediment samples. *Environmental Science and Pollution Research*, *31*(15), 22431–22440.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342.

Wachendörfer, M. M., & Oeberst, A. (2024). Differences between true and false memories using the criteria-based content analysis. *Applied Cognitive Psychology*, *38*(5), e4246.