Journal of Statistical Theory and Practice

Statistical Reproducibility of Umbrella Alternative Tests

Norah Alalyani[2], Tahani Coolen-Maturi[1*] and Frank P. A. Coolen[1]

[1*]Department of Mathematical Sciences, Durham University, UK.
[2]Department of Mathematics and Statistics, King Faisal University, Al-Ahsa, Saudi Arabia.

*Corresponding author(s). E-mail(s): tahani.maturi@durham.ac.uk;
Contributing authors: nalalyani@kfu.edu.sa; frank.coolen@durham.ac.uk;

Abstract

The reproducibility of research findings is important across many disciplines and is a fundamental concept in scientific studies. This paper investigates the reproducibility of statistical hypothesis tests using Nonparametric Predictive Inference (NPI). NPI is a frequentist statistical framework based on minimal modelling assumptions, considering future observations to be exchangeable with the observed data. Its predictive nature makes it particularly suitable for assessing the reproducibility of a test. This paper applies NPI to study the statistical reproducibility of several umbrella alternative tests, including the Mack-Wolfe (MW), Esra and Fikri (EF), and Jonckheere-Terpstra (JT) tests. These tests evaluate the null hypothesis that location parameters are equal against the alternative hypothesis that they follow a specific order. Several examples are provided to illustrate the application of the proposed methods. The findings suggest that the reproducibility of these tests can be quite poor, especially when the test statistic is close to the critical value.

Keywords: Nonparametric Predictive Inference, Tests Reproducibility Probability, Umbrella Alternatives, Jonckheere-Terpstra test, Mack-Wolfe test, Esra and Fikri test

1 Introduction

Recently, much attention has been paid to the reproducibility of statistical hypothesis tests [1]. However, there is considerable confusion in the literature regarding the concept of reproducibility. The main question that this paper aims to address is: if a statistical test were repeated under the same circumstances, would it lead to the same conclusion regarding rejection or non-rejection of the null hypothesis? The probability that the repeated test yields the same conclusion as the original test is known as the reproducibility probability (RP).

Goodman [2] pointed out that the failure of an experiment to replicate the statistical significance achieved in previous studies often causes concern in the medical literature, primarily due to misunderstandings of the p-value. Goodman also demonstrated that the probability of replicating a statistically significant result may be lower than generally expected and that the p-value may lead to overly optimistic interpretations. In a discussion of Goodman's work, Senn [3] agreed that RP and the p-value are distinct concepts and emphasised the importance of reproducibility in statistical tests. However, he disagreed with Goodman's claim that the p-value overstates the evidence against the null hypothesis.

Shao and Chow [4] examined RP in the context of clinical trials using three approaches: a common power approach, where RP is defined by the estimated power of a future test using data from the original test; a confidence bounds approach, where RP is defined as the lower confidence bound of the estimated power of a second test; and a Bayesian approach based on the posterior predictive distribution. De Martini [5] estimated RP using the test's estimated power and the lower confidence bound of the power. De Capitani and De Martini [6, 7] further demonstrated that RP estimation can be used both to evaluate statistical test results and to define statistical tests. A comparison between RP and the p-value was discussed by De Capitani [8], while Boos and Stefanski [9] extended the work of Shao and Chow [4] by applying the estimated power approach to one-way ANOVA.

In the Nonparametric Predictive Inference (NPI) framework, reproducibility is considered a prediction problem. NPI is a frequentist statistical approach that assumes m future observations are exchangeable with given n data observations. Its predictive nature makes it particularly well-suited for studying test reproducibility. The NPI reproducibility probability enables inference by deriving lower and upper probabilities for the event that a future test, repeated under similar conditions, will reach the same conclusion as the original test—whether rejection or non-rejection of the null hypothesis. This approach is denoted by NPI-RP [10], with the corresponding lower and upper reproducibility probabilities denoted by $\underline{RP}$ and $\overline{RP}$, respectively.

In the NPI framework, we typically consider the case $m = n$, as this is a logical assumption for studying reproducibility. Research on NPI reproducibility was initiated by Coolen and BinHimd [11], who studied NPI-RP for several nonparametric tests, including the one-sample sign test, the one-sample

Wilcoxon signed-rank test, the two-sample rank sum test (Wilcoxon-Mann-Whitney test), and the two-sample Kolmogorov–Smirnov test. Coolen and Alqifari [12] extended this work to NPI-RP for a one-sample quantile test and a two-sample precedence test, both based on order statistics. Simkus et al. [13] applied NPI reproducibility analysis to the t-test in pharmaceutical research, while Marques et al. [14] studied RP in the context of likelihood ratio tests. However, as sample sizes increase, NPI-RP computations become increasingly demanding due to the growing number of orderings of m future observations among the given n observations. For some tests, deriving exact closed-form expressions for the lower and upper reproducibility probabilities is computationally challenging.

To address these computational challenges, Marques and Coolen [15] introduced the NPI Sampling of Orderings (NPI-RP-SO) method, providing an approximation for NPI lower and upper reproducibility probabilities, particularly for likelihood ratio tests. Meanwhile, Coolen and BinHimd [16] proposed the NPI Bootstrap (NPI-B) method as an alternative approximation, which is more flexible and suitable for complex test statistics and large sample sizes. However, NPI-B provides a point estimate for RP rather than estimates for NPI lower and upper reproducibility probabilities.

This paper contributes to the development of NPI for statistical reproducibility by considering tests for umbrella alternatives, specifically the Mack-Wolfe test, the Esra and Fikri test, and the Jonckheere-Terpstra test. The remainder of the paper is structured as follows: Section 2 provides an overview of classical tests for umbrella alternatives, while Section 3 presents a brief introduction to Nonparametric Predictive Inference (NPI). Section 4 introduces NPI reproducibility for the Mack-Wolfe test, including exact lower and upper reproducibility probabilities for three groups and their approximations using the NPI sampling of orderings (NPI-RP-SO) approach. To address computational challenges associated with large sample sizes, Section 5 proposes the NPI-based Bootstrap (NPI-B) method for estimating reproducibility probabilities of the three umbrella alternative tests. Several examples are provided to illustrate the proposed methods. The paper concludes with final remarks in Section 7.

2 Umbrella Alternatives Tests

The comparison of $g \geq 3$ groups in a one-way ANOVA setting may involve cases where the response increases up to a certain group and then decreases. This situation is common in many real-world problems, such as the effect of age on physical capability measures like muscle strength. Another example is the reaction to increasing drug dosage, where the response improves up to a certain dose before declining. This 'up-then-down' behaviour is known as 'umbrella ordering', a term introduced by Mack and Wolfe [17].

Let μ_i be the location parameter of the ith population. The g-sample rank tests are introduced to test the following hypothesis:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g \tag{1}$$

against the umbrella alternative:

$$H_1 : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_{p-1} \leq \mu_p \quad \text{and} \quad \mu_p \geq \mu_{p+1} \geq \ldots \geq \mu_g \tag{2}$$

for some $p \in \{1, 2, \ldots, g\}$, with at least one strict inequality. The peak of the umbrella alternative is at p, which may be either known or unknown [18].

Umbrella alternative tests are formulated based on the Mann-Whitney test statistic, which has been widely used in tests involving ordered alternatives [17, 19–22]. These test statistics are constructed by summing the Mann-Whitney counts to the left and right of the peak while excluding comparisons across the peak itself.

Other related approaches for constructing test statistics for umbrella alternatives were proposed by Basso and Salmaso [23], Hettmansperger and Norton [24], Chen and Wolfe [25], and Magel and Qin [26].

2.1 Mack-Wolfe (MW) Test

This section focuses on the Mack-Wolfe test for both known and unknown peak scenarios [17, 18], with the aim of investigating the statistical reproducibility of the test in each case.

To compute the Mack-Wolfe statistic A_p for a known peak p, we first determine the $p(p-1)/2$ Mann-Whitney counts U_{uv} for every pair of groups where $1 \leq u < v \leq p$, and U_{uv} is the number of observations from sample u that are smaller than the observations from sample v. Similarly, for $p \leq u < v \leq g$, we compute the $(g-p+1)(g-p)/2$ reverse Mann-Whitney counts U_{vu}. The Mack-Wolfe statistic A_p is given by:

$$A_p = \sum_{u=1}^{v-1} \sum_{v=2}^{p} U_{uv} + \sum_{u=p}^{v-1} \sum_{v=p+1}^{g} U_{vu}. \tag{3}$$

The null hypothesis in (1) is rejected at significance level α if

$$A_p \geq A_{p,\alpha}, \tag{4}$$

where $A_{p,\alpha}$ is the α-upper percentile of the null distribution of A_p. This value can be obtained using the function `cUmbrPK(`α, n, p`)` from the R package `NSM3` [27] or from published tables [17, 28].

For large sample sizes, and under the null hypothesis, the statistic A_p follows an asymptotic Normal distribution with the following mean and

variance:

$$E(A_p) = \frac{N_1^2 + N_2^2 - \sum_{i=1}^{g} n_i^2 - n_p^2}{4}, \tag{5}$$

$$\sigma^2(A_p) = \frac{1}{72}\Bigg\{2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{g} n_i^2(2n_i + 3) - n_p^2(2n_p + 3) + 12n_pN_1N_2 - 12n_p^2N\Bigg\}, \tag{6}$$

where n_i is the sample size of group i, $N_1 = \sum_{i=1}^{p} n_i$, and $N_2 = \sum_{i=p}^{g} n_i$. Since the peak group p is included in both N_1 and N_2, the total sample size is given by $N = N_1 + N_2 - n_p$.

To compute the Mack–Wolfe statistic A_p for an unknown peak p, we first estimate p using the sample data. This involves determining which group is most likely to correspond to the peak by calculating g combined sample Mann–Whitney statistics:

$$U_{\cdot q} = \sum_{i \neq q} U_{iq}, \quad q = 1, \ldots, g, \tag{7}$$

where U_{iq} represents the number of observations in sample i that are smaller than those in sample q. Under the null hypothesis, each $U_{\cdot q}$ is standardized as follows:

$$U'_{\cdot q} = \frac{U_{\cdot q} - E(U_{\cdot q})}{\sigma(U_{\cdot q})}, \quad q = 1, \ldots, g. \tag{8}$$

where $E(U_{\cdot q}) = n_q(N - n_q)/2$ and $\sigma^2(U_{\cdot q}) = n_q(N - n_q)(N + 1)/12$.

Let s denote the number of groups that are tied for the maximum value of $U'_{\cdot q}$, and let D be the subset of $\{1, 2, \ldots, g\}$ corresponding to these tied groups. The Mack–Wolfe statistic for an unknown peak is then given by

$$A'_{\hat{p}} = \frac{1}{s} \sum_{j \in D} \frac{A_j - E(A_j)}{\sigma(A_j)}, \tag{9}$$

where A_j is the peak-known statistic with peak at the jth group in Equation (3), and $E(A_j)$ and $\sigma(A_j)$ are given by Equations (5) and (6), respectively.

The null hypothesis in (1) is rejected at significance level α if:

$$A'_{\hat{p}} \geq A_{\hat{p},\alpha}. \tag{10}$$

In most cases, $s = 1$, meaning $A'_{\hat{p}}$ reduces to the standardised peak-known statistic. The critical value $A_{\hat{p},\alpha}$ is the upper percentile of the null distribution of $A'_{\hat{p}}$ (where $H_0 : \mu_1 = \mu_2 = ... = \mu_g$). It can be computed using the function `cUmbrPU(`α, n`)` from the R package `NSM3` [27] or obtained from published tables [17, 28].

2.2 Esra and Fikri (EF) Test

Esra and Fikri [21] proposed a modified Mack-Wolfe test for umbrella alternatives, addressing cases with a known or unknown peak.

For a known peak, the modified Mack-Wolfe statistic $\tilde{A}_p$ is the weighted sum of the Mann-Whitney counts to the left of the peak, $(v-u)U_{uv}$, and the weighted reverse Mann-Whitney counts to the right of the peak, $(v-u)U_{vu}$. The test statistic is given by:

$$\tilde{A}_p = \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (v-u)U_{uv} + \sum_{u=p}^{g-1} \sum_{v=u+1}^{g} (v-u)U_{vu}. \tag{11}$$

For balanced data ($n_1 = \ldots = n_g = n$), and under H_0, $\tilde{A}_p$ is asymptotically Normally distributed with mean and variance:

$$\begin{aligned} E(\tilde{A}_p) &= \frac{n^2}{2}\left[\binom{p+1}{3} + \binom{g-p+2}{3}\right], \\ \sigma^2(\tilde{A}_p) &= \frac{n^2p^2(p^2-1)(np+1)}{144} \\ &\quad + \frac{n^2(g-p+1)^2[(g-p+1)^2-1][n(g-p+1)+1]}{144} \\ &\quad + \frac{n^3p(p-1)(g-p)(g-p+1)}{24}. \end{aligned}$$

The null hypothesis H_0 is rejected at significance level α if

$$\tilde{A}_p^* = \frac{\tilde{A}_p - E(\tilde{A}_p)}{\sigma(\tilde{A}_p)} \geq Z_\alpha, \tag{12}$$

where Z_α is the upper α-quantile of the standard Normal distribution.

2.3 Jonckheere-Terpstra (JT) Test

For $p = 1$ or $p = g$, the MW test reduces to the Jonckheere-Terpstra (JT) test, which tests the ordered alternative:

$$H_1 : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_g, \tag{13}$$

with at least one strict inequality.

The Jonckheere-Terpstra (JT) test, introduced by Jonckheere [29] and Terpstra [30], requires samples to be ordered according to H_1 before data collection. The test statistic J is:

$$J = \sum_{u=1}^{v-1} \sum_{v=2}^{g} U_{uv}. \tag{14}$$

The null hypothesis in (1) is rejected in favor of the alternative in (13) at significance level α if

$$J \geq J_\alpha, \tag{15}$$

where J_α is the upper α-quantile of the null distribution of J. The values of J_α can be found in tables [31].

For large sample sizes, and under H_0, J is asymptotically Normally distributed with mean and variance:

$$E(J) = \frac{N^2 - \sum_{i=1}^{g} n_i^2}{4},$$
$$\sigma^2(J) = \frac{N^2(2N+3) - \sum_{i=1}^{g} n_i^2(2n_i+3)}{72},$$

where N is the total number of observations, and n_i is the sample size of group i. Thus, the null hypothesis is rejected at significance level α if

$$J^* = \frac{J - E(J)}{\sigma(J)} \geq Z_\alpha, \tag{16}$$

where Z_α is the α-upper quantile of the standard Normal distribution.

3 Nonparametric Predictive Inference (NPI)

Nonparametric Predictive Inference (NPI) is a statistical framework based on the assumption $A_{(n)}$, proposed by Hill [32, 33], which provides direct probabilities for future observations given n observations of related random quantities. Inferences based on the assumption $A_{(n)}$ are predictive and nonparametric and seem suitable when there is little or no knowledge about the random quantities of interest, other than the n observations, or when one does not want to use such information. Such inferences, based on restricted knowledge, are called 'low-structure inferences' [34].

Suppose that $X_1, X_2, \ldots, X_n, X_{n+1}$ are continuous and exchangeable random quantities. Let the ordered observations $X_1, X_2, \ldots, X_n$ be denoted by $x_1 < x_2 < \ldots < x_n$, and define $x_0 = -\infty$ and $x_{n+1} = \infty$ for convenience. We assume that ties do not occur; if they do, they can be broken by a very small amount [33]. These n observations partition the real line into $n+1$ intervals, $I_j = (x_{j-1}, x_j)$, for $j = 1, \ldots, n+1$. Given the n observations, the assumption $A_{(n)}$ for the next future observation X_{n+1} is

$$P(X_{n+1} \in I_j = (x_{j-1}, x_j)) = \frac{1}{n+1} \qquad \text{for} \quad j = 1, \ldots, n+1. \tag{17}$$

$A_{(n)}$ does not assume anything else and can be considered a post-data assumption related to exchangeability [35]. $A_{(n)}$ alone is not sufficient to derive precise probabilities for many events of interest, but it provides optimal bounds for probabilities to quantify uncertainty for all such events involving X_{n+1}. These

bounds are lower and upper probabilities in the theories of imprecise probability and interval probability [36]. Imprecise probability generalises classical probability in the sense that it describes uncertainty about events via intervals instead of single numbers. For any event A, the lower probability is denoted by $\underline{P}(A)$ and the upper probability by $\overline{P}(A)$, with $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, and the imprecision is given by $\triangle(A) = \overline{P}(A) - \underline{P}(A)$ [37]. The NPI approach has been introduced for many applications in statistics and reliability and for a range of data types, such as Bernoulli data [38] and multinomial data [39].

The NPI approach can be generalized for $m \geq 1$ future observations, X_{n+i} for $i = 1, 2, \ldots, m$, by assuming that the assumptions $A_{(n)}, \ldots, A_{(n+m-1)}$ hold for each future observation [32]. Given a dataset of n observations, the m future observations are not conditionally independent but are assumed to be exchangeable random quantities. There are $\binom{n+m}{n}$ possible orderings O_i for $i = 1, 2, \ldots, \binom{n+m}{n}$, and all possible orderings are equally likely. Let S_j denote the number of future observations that fall in the interval $I_j = (x_{j-1}, x_j)$ for $j = 1, 2, \ldots, n+1$, and define $x_0 = -\infty$ and $x_{n+1} = \infty$ for ease of notation [38, 40]. Then, inferences about these m future observations can be based on the probabilities:

$$P\Big(\bigcap_{j=1}^{n+1} \{S_j = s_j\}\Big) = \binom{n+m}{n}^{-1} \tag{18}$$

for any valid combination $(S_1, \ldots, S_{n+1})$, where S_j are non-negative integers satisfying $\sum_{j=1}^{n+1} S_j = m$ [40].

The methodology of NPI for $m \geq 1$ future observations, as given in Equation (18), will be utilised to study the reproducibility of the Mack-Wolfe test for three groups, as introduced in the next section.

4 NPI Reproducibility for Mack-Wolfe test

In this section, we consider the case of three independent groups, X, Y, and Z, with n_x observations from group X, n_y observations from group Y, and n_z observations from group Z. Let $x_1 < \ldots < x_{n_x}$ be the ordered observed values of group X, partitioning the real line into $n_x + 1$ intervals $I_j^x = (x_{j-1}, x_j)$, for $j = 1, \ldots, n_x + 1$. Similarly, let $y_1 < \ldots < y_{n_y}$ be the ordered observed values of group Y, partitioning the real line into $n_y + 1$ intervals $I_i^y = (y_{i-1}, y_i)$, for $i = 1, \ldots, n_y + 1$, and let $z_1 < \ldots < z_{n_z}$ be the ordered observed values of group Z, partitioning the real line into $n_z + 1$ intervals $I_k^z = (z_{k-1}, z_k)$, for $k = 1, \ldots, n_z + 1$. For convenience, we define $x_0 = y_0 = z_0 = -\infty$ and $x_{n_x+1} = y_{n_y+1} = z_{n_z+1} = \infty$. We assume no tied observations; if they occur, a standard tie-breaking method can be used [33].

Let the number of future observations from groups X, Y, and Z be denoted by m_x, m_y, and m_z, respectively. Here, we restrict our attention to the case where the number of future observations equals the number of data observations ($m_x = n_x$, $m_y = n_y$, and $m_z = n_z$), as this is considered a logical assumption when studying reproducibility. There are $\binom{2n_x}{n_x}$ possible orderings

of m_x future observations among the n_x data observations, with all possible orderings equally likely. Similarly, there are $\binom{2n_y}{n_y}$ possible orderings of m_y future observations among the n_y data observations, and $\binom{2n_z}{n_z}$ possible orderings of m_z future observations among the n_z data observations, all of which are equally likely.

4.1 Exact NPI Reproducibility Probabilities (NPI-RP-E)

To derive the exact NPI lower and upper reproducibility probabilities, we consider all $\binom{2n_x}{n_x}\binom{2n_y}{n_y}\binom{2n_z}{n_z}$ possible orderings, denoted by O_ℓ for $\ell = 1, 2, \ldots, \binom{2n_x}{n_x}\binom{2n_y}{n_y}\binom{2n_z}{n_z}$. For each combination of orderings O_ℓ, we consider the corresponding Mack-Wolfe test statistic, given in Equation (3), which we denote by A_{p_ℓ}. Since future observations are not known precisely, but only their counts within the intervals partitioned by the original data observations are known for a given ordering, we cannot compute an exact value of A_{p_ℓ} for a specific combination of orderings. However, we can determine the minimum and maximum possible values, denoted by $\underline{A}_{p_\ell}$ and $\overline{A}_{p_\ell}$, respectively.

Let a specific ordering of n_x future observations among the n_x data observations be denoted by $(S_1^X, \ldots, S_{n_x+1}^X)$, where S_j^X are non-negative integers satisfying $\sum_{j=1}^{n_x+1} S_j^X = n_x$, as introduced in Section 3. Similarly, let a specific ordering of n_y future observations among the n_y data observations be denoted by $(S_1^Y, \ldots, S_{n_y+1}^Y)$, with S_i^Y being non-negative integers satisfying $\sum_{i=1}^{n_y+1} S_i^Y = n_y$, and let a specific ordering of n_z future observations among the n_z data observations be denoted by $(S_1^Z, \ldots, S_{n_z+1}^Z)$, with S_k^Z being non-negative integers satisfying $\sum_{k=1}^{n_z+1} S_k^Z = n_z$.

Additionally, let $j(i) = \max\{j : x_{(j)} < y_{(i)}\}$ for $i = 1, \ldots, n_y + 1$ and $j = 0, 1, \ldots, n_x$, so that $x_{(j(i))} < y_{(i)} < x_{(j(i)+1)}$. The rank of $y_{(i)}$ in the combined ordered data from both groups X and Y is then $i + j(i)$. Likewise, let $k(i) = \max\{k : z_{(k)} < y_{(i)}\}$ for $i = 1, \ldots, n_y + 1$ and $k = 0, 1, \ldots, n_z$, so that $z_{(k(i))} < y_{(i)} < z_{(k(i)+1)}$. The rank of $y_{(i)}$ in the combined ordered data from both groups Z and Y is then $i + k(i)$.

The minimum and maximum values of A_{p_ℓ} are given by

$$\underline{A}_{p_\ell} = \sum_{i=1}^{n_y+1} S_i^Y \left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z \right], \tag{19}$$

$$\overline{A}_{p_\ell} = \sum_{i=1}^{n_y+1} S_i^Y \left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + \sum_{c=1}^{k_{(i)}-1} S_c^Z \right]. \tag{20}$$

For simplicity, the subscript ℓ is omitted on the right-hand side. A detailed justification of these results is provided in the Appendix.

The NPI lower and upper reproducibility probabilities depend on whether the original test conclusion was the rejection or non-rejection of H_0. If the original test rejects H_0, the lower reproducibility probability is obtained by

counting the number of orderings, among the total $\binom{2n_x}{n_x}\binom{2n_y}{n_y}\binom{2n_z}{n_z}$ possible orderings, for which $\underline{A}_{p_\ell} \geq A_{p,\alpha}$. The corresponding upper reproducibility probability is derived by counting the number of orderings where $\overline{A}_{p_\ell} \geq A_{p,\alpha}$. Thus, the NPI lower and upper reproducibility probabilities are

$$\underline{RP} = \frac{1}{h}\sum_{\ell=1}^{h} \mathbf{1}\{\underline{A}_{p_\ell} \geq A_{p,\alpha}\}, \tag{21}$$

$$\overline{RP} = \frac{1}{h}\sum_{\ell=1}^{h} \mathbf{1}\{\overline{A}_{p_\ell} \geq A_{p,\alpha}\}, \tag{22}$$

where $h = \binom{2n_x}{n_x}\binom{2n_y}{n_y}\binom{2n_z}{n_z}$, and $\mathbf{1}\{A\}$ is an indicator function that equals 1 if event A occurs and 0 otherwise.

If the original test does not reject H_0, the lower reproducibility probability is obtained by counting the orderings where $\overline{A}_{p_\ell} < A_{p,\alpha}$ must hold, while the upper reproducibility probability is derived from the orderings where $\underline{A}_{p_\ell} < A_{p,\alpha}$ can hold. In this case, the NPI lower and upper reproducibility probabilities are

$$\underline{RP} = \frac{1}{h}\sum_{\ell=1}^{h} \mathbf{1}\{\overline{A}_{p_\ell} < A_{p,\alpha}\}, \tag{23}$$

$$\overline{RP} = \frac{1}{h}\sum_{\ell=1}^{h} \mathbf{1}\{\underline{A}_{p_\ell} < A_{p,\alpha}\}. \tag{24}$$

This method for deriving the lower and upper reproducibility probabilities of the Mack-Wolfe test is practical for small sample sizes, and we refer to it as the Exact NPI-RP (NPI-RP-E). However, evaluating all possible orderings becomes computationally infeasible for larger sample sizes. Heuristic methods are needed; in this paper, we consider two approaches. The first is the sampling of orderings, which provides approximate estimates for the lower and upper RP. The second is the NPI bootstrap-based method, which yields a single-point estimate for RP. We first consider the sampling of orderings approach.

4.2 NPI Reproducibility Using Sampling of Orderings

As already stated in the previous section, evaluating all possible orderings can become computationally infeasible for large sample sizes. To address this, we use the NPI-RP sampling of orderings approach (NPI-RP-SO) to approximate the lower and upper reproducibility probabilities [15]. This method estimates $\underline{RP}$ and $\overline{RP}$ through simple random sampling, where each ordering has an equal probability of being selected independently of others. Increasing the number of sampled orderings improves the accuracy of these approximations.

To implement the NPI-RP-SO method for three groups X, Y, and Z, we randomly sample r orderings from the $\binom{2n_x}{n_x}$ possible orderings of m_x future

observations among the n_x data observations. Similarly, we randomly sample r orderings for groups Y and Z in the same manner [14, 15]. Each sampled ordering from one group is paired with the corresponding sampled orderings from the other groups. Using these sampled orderings, we compute the minimum and maximum values of the Mack-Wolfe test statistic, A_p, by applying Equations (19) and (20).

If the original test rejects H_0 (i.e., $A_p \geq A_{p,\alpha}$), the NPI lower and upper reproducibility probabilities are estimated as follows:

$$\widehat{\underline{RP}} = \frac{1}{r}\sum_{\ell=1}^{r} \mathbf{1}\{\underline{A}_{p_\ell} \geq A_{p,\alpha}\}, \tag{25}$$

$$\widehat{\overline{RP}} = \frac{1}{r}\sum_{\ell=1}^{r} \mathbf{1}\{\overline{A}_{p_\ell} \geq A_{p,\alpha}\}. \tag{26}$$

If the original test fails to reject H_0 ($A_p < A_{p,\alpha}$), the estimates are

$$\widehat{\underline{RP}} = \frac{1}{r}\sum_{\ell=1}^{r} \mathbf{1}\{\overline{A}_{p_\ell} < A_{p,\alpha}\}, \tag{27}$$

$$\widehat{\overline{RP}} = \frac{1}{r}\sum_{\ell=1}^{r} \mathbf{1}\{\underline{A}_{p_\ell} < A_{p,\alpha}\}. \tag{28}$$

To quantify the uncertainty in these estimates, a 95% confidence interval can be computed using the normal approximation $\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/r}$, where $\hat{\pi}$ is the estimated lower or upper reproducibility probability, r is the number of sampled orderings, and $z_{\alpha/2}$ is the $(\alpha/2)$-upper quantile of the standard normal distribution. If $\hat{\pi}$ is close to 0 or 1, the normal approximation may produce bounds outside $[0, 1]$. In such cases, the exact binomial confidence interval should be used instead; for details, see [41].

5 NPI Reproducibility Using NPI-Bootstrap

As noted earlier, computing the exact NPI lower and upper reproducibility probabilities is computationally challenging for large sample sizes due to the rapid increase in the number of orderings of future observations among the data observations. Additionally, for some statistical tests, deriving an exact closed-form expression for these probabilities is difficult. To address these challenges, the NPI Bootstrap (NPI-B) method was introduced to study the reproducibility of various statistical tests [10, 11].

NPI-B is based on repeated applications of Hill's assumption $A_{(n)}$, ensuring that all possible orderings of the m future values among the n original data observations are equally likely to occur [10]. Unlike Efron's bootstrap [42], which is primarily aimed at estimating population characteristics, NPI-B

is specifically designed for prediction, allowing future observations to extend beyond already observed values.

It is important to note that NPI-B provides a point estimate of the reproducibility probability rather than estimates for the NPI lower and upper probabilities. Hereafter, we refer to the reproducibility probability estimate obtained using NPI-B as NPI-RP-B.

The NPI-B method generates future observations by partitioning the real line into $n+1$ intervals based on the original n observations. A future observation is sampled by first selecting one of these intervals with equal probability $\frac{1}{n+1}$ and then drawing a value uniformly from the chosen interval. This process is repeated m times to construct an NPI-B sample of size m, with a particular focus on the case where $m = n$. The procedure is further repeated B times to generate B NPI-B samples. Special attention is required when sampling from unbounded intervals. If the selected interval is bounded, such as $I_1 = (x_0, x_1)$ or $I_{n+1} = (x_n, x_{n+1})$, the future value is drawn in the same manner as other intervals. However, if the chosen interval is unbounded, i.e., I_1 with $x_0 = -\infty$ or I_{n+1} with $x_{n+1} = \infty$, the future value is drawn with probability $\frac{1}{n+1}$ using a normal distribution tail approximation with mean $\mu = \frac{x_1+x_n}{2}$ and standard deviation $\sigma = \frac{x_n-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ^{-1} is the inverse of the normal cumulative distribution function. For datasets restricted to $(0, \infty)$, if the selected interval is (x_n, ∞), the future value is sampled from the tail of an exponential distribution with rate parameter $\lambda = \frac{\ln(n+1)}{x_{(n)}}$ [10, 11].

To approximate the reproducibility probability for umbrella alternative tests, the NPI-B method is used to generate B NPI-B samples per group. For each run i ($i = 1, 2, \ldots, T$), we compute the proportion of cases where the original dataset and the B NPI-B samples lead to the same test conclusion, that is, whether H_0 is rejected or not. Let this proportion be denoted as RP_i. The NPI-B estimate of the reproducibility probability (RP) is then given by the mean of these RP_i values. Additionally, other summary statistics, such as the minimum, median, and maximum of the RP_i values, can also be computed.

Formally, the NPI-B estimate of the reproducibility probability, denoted as NPI-RP-B, is given by

$$\widehat{RP}_{\text{boot}} = \frac{1}{T}\sum_{i=1}^{T} RP_i = \frac{1}{T}\sum_{i=1}^{T}\left[\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\{t_b^* = t^*\}\right], \tag{29}$$

where t^* is the test decision (reject or not reject the null hypothesis) based on the original data, and t_b^* is the decision based on the b-th NPI-B sample.

Algorithm 1 summarises this NPI-B approach for estimating the reproducibility probability for umbrella alternatives tests. Since the NPI-B method is highly flexible, it will be utilised for estimating reproducibility probabilities for all three umbrella alternative tests, particularly in cases where deriving exact closed-form solutions is complicated.

Algorithm 1 NPI-B algorithm for estimating reproducibility probability for umbrella alternatives tests

1: Apply the statistical test to the original g-group dataset and record the test outcome (whether H_0 is rejected or not).
2: Generate an NPI-B sample from each group based on the original g-group dataset, then apply the statistical test to these NPI-B samples.
3: Repeat Step 2 a total of B times, recording the test outcome each time.
4: Compute the proportion of cases where the original dataset and the B NPI-B samples yield the same conclusion; denote this as RP.
5: Repeat Steps 2–4 a total of T times to obtain RP_i for $i = 1, 2, \ldots, T$. The mean of these values is the NPI-B estimate of the reproducibility probability.

6 Examples

This section presents five examples that demonstrate the NPI-RP approach for assessing reproducibility across different statistical tests and scenarios. Example 1 illustrates the proposed methodology by evaluating the reproducibility probability for the Mack-Wolfe (MW) and Esra-Fikri (EF) tests using NPI-RP-E, and then comparing the results with those from the approximation methods NPI-RP-SO and NPI-RP-B to examine how well they align with the exact values. Example 2 extends the analysis to larger samples using NPI-RP-B, highlighting the influence of test statistic weighting and its relationship with p-values. Example 3 applies NPI-RP-B to investigate reproducibility for the Jonckheere-Terpstra (JT) test under ordered alternatives using large-sample simulations. Example 4 shifts focus to real-world data, applying NPI-RP-SO to analyse communication patterns in a firm using the MW and EF tests. Finally, Example 5 explores an advanced scenario in which the MW test is applied with an unknown peak, using Monte Carlo simulations to determine critical values. Together, these examples highlight the flexibility and applicability of different NPI-RP methods across a range of data settings.

Example 1 (**NPI-RP for MW and EF tests, small samples, known peak**) This example investigates the reproducibility probability for the Mack-Wolfe (MW) test and Esra-Fikri (EF) test with $g = 3$ groups (X, Y, and Z) using the NPI-RP-E approach, as introduced in Section 4.1. A comparison is then made between NPI-RP-E, NPI-RP-SO, and NPI-RP-B to evaluate whether NPI-RP-B estimates lie within the lower and upper bounds of NPI-RP-E and NPI-RP-SO.

The study considers artificial rank-based datasets with equal sample sizes: $n_x = n_y = n_z = 3$ and $n_x = n_y = n_z = 5$. The hypothesis of interest is $H_0 : \mu_x = \mu_y = \mu_z$ against the umbrella alternative $H_1 : \mu_x \leq \mu_y \geq \mu_z$, where the peak is at the second group ($p = 2$). The significance level is $\alpha = 0.05$.

For the MW test with $n_x = n_y = n_z = 3$ and $n_x = n_y = n_z = 5$, the discrete nature of the test statistic results in nominal significance levels of 0.0476 and 0.0496, respectively, as shown in Tables 1 and 2. Accordingly, the decision rule for the MW

Table 1: RP for the MW test and the EF test, with $H_1 : \mu_x \leq \mu_y \geq \mu_z$, $p = 2$, $n_x = n_y = n_z = 3$, $A_{2,0.0476} = 16$, $Z_{0.05} = 1.645$

Ranks			Test conclusion				NPI-RP-E		NPI-RP-B				NPI-RP-SO	
X	Y	Z	A_p	$\tilde{A}_p^*$	p-value	H_0	$\underline{RP}$	$\overline{RP}$	Min	Mean	Median	Max	$\widehat{\underline{RP}}$	$\widehat{\overline{RP}}$
1,2,3	7,8,9	4,5,6	18	2.324	0.010	R	0.125	1	0.955	0.971	0.971	0.983	0.114	1
2,3,4	7,8,9	1,5,6	18	2.324	0.010	R	0.125	1	0.760	0.789	0.790	0.822	0.123	1
1,2,3	6,8,9	4,5,7	17	2.066	0.019	R	0.106	0.930	0.645	0.676	0.676	0.720	0.107	0.934
1,2,3	5,8,9	4,6,7	16	1.807	0.035	R	0.081	0.825	0.441	0.478	0.476	0.516	0.089	0.830
1,2,7	5,8,9	3,4,6	16	1.807	0.035	R	0.086	0.832	0.369	0.398	0.397	0.429	0.089	0.824
1,2,3	6,7,9	4,5,8	16	1.807	0.035	R	0.081	0.825	0.459	0.494	0.495	0.536	0.078	0.829
1,2,3	4,8,9	5,6,7	15	1.549	0.061	NR	0.318	0.950	0.591	0.639	0.640	0.676	0.314	0.953
2,3,4	5,7,9	1,6,8	15	1.549	0.061	NR	0.273	0.939	0.603	0.645	0.646	0.684	0.275	0.930
4,6,7	3,8,9	1,2,5	14	1.291	0.098	NR	0.386	0.950	0.656	0.690	0.691	0.722	0.414	0.953
4,5,6	1,8,9	2,3,7	12	0.775	0.219	NR	0.476	0.950	0.713	0.753	0.754	0.784	0.473	0.952
1,4,8	3,5,9	2,6,7	11	0.516	0.303	NR	0.578	0.977	0.826	0.865	0.866	0.888	0.566	0.977
1,2,3	4,5,6	7,8,9	9	0.000	0.500	NR	0.790	1	0.997	0.999	1	1	0.790	1
1,2,8	4,5,6	3,7,9	9	0.000	0.500	NR	0.720	0.995	0.944	0.958	0.958	0.973	0.715	0.997
1,3,4	2,5,6	7,8,9	7	-0.516	0.697	NR	0.833	1	0.986	0.993	0.993	1	0.824	1
1,2,6	3,4,5	7,8,9	6	-0.775	0.781	NR	0.855	1	1	1	1	1	0.846	1
1,2,9	3,4,5	6,7,8	6	-0.775	0.781	NR	0.855	1	0.998	1.000	1	1	0.848	1
5,3,9	1,2,8	7,4,6	5	-1.033	0.849	NR	0.814	0.995	0.926	0.948	0.947	0.966	0.818	0.995
1,4,5	2,3,6	7,8,9	5	-1.033	0.849	NR	0.870	1	0.987	0.993	0.993	1	0.872	1
1,4,7	2,3,5	6,8,9	4	-1.291	0.902	NR	0.889	1	0.996	0.999	0.999	1	0.892	1
4,5,6	1,2,3	7,8,9	0	-2.324	0.990	NR	0.933	1	1	1	1	1	0.932	1

Table 2: RP for the MW test and the EF test, with $H_1 : \mu_x \leq \mu_y \geq \mu_z$, $p = 2$, $n_x = n_y = n_z = 5$, $\alpha = 0.05$, $A_{2,0.0496} = 39$, $Z_{0.05} = 1.645$

Ranks			Test conclusion				NPI-RP-E		NPI-RP-B				NPI-RP-SO	
X	Y	Z	A_p	$\tilde{A}_p^*$	p-value	H_0	$\underline{RP}$	$\overline{RP}$	Min	Mean	Median	Max	$\widehat{\underline{RP}}$	$\widehat{\overline{RP}}$
1,2,3,4,5	11,12,13,14,15	6,7,8,9,10	50	3.062	0.001	R	0.441	1	0.997	0.999	0.999	1	0.443	1
1,2,6,7,8	11,12,13,14,15	3,4,5,9,10	50	3.062	0.001	R	0.441	1	0.974	0.985	0.985	0.997	0.443	1
1,2,3,4,5	10,12,13,14,15	6,7,8,9,11	49	2.939	0.002	R	0.402	0.997	0.966	0.977	0.977	0.987	0.406	0.999
1,2,3,4,5	9,12,13,14,15	6,7,8,10,11	48	2.817	0.002	R	0.367	0.988	0.916	0.939	0.939	0.957	0.367	0.991
1,2,3,4,5	10,11,12,13,14	6,7,8,9,15	45	2.450	0.007	R	0.300	0.932	0.805	0.832	0.832	0.860	0.300	0.927
1,2,3,4,5	9,10,11,13,15	6,7,8,12,14	43	2.205	0.014	R	0.224	0.884	0.695	0.730	0.730	0.767	0.223	0.878
1,2,3,4,5	8,9,10,14,15	6,7,11,12,13	41	1.960	0.025	R	0.172	0.807	0.536	0.586	0.586	0.627	0.174	0.812
1,2,3,4,6	5,11,12,13,14	7,8,9,10,15	40	1.837	0.033	R	0.178	0.775	0.434	0.474	0.474	0.509	0.174	0.784
1,2,3,4,15	5,10,12,13,14	6,7,8,9,11	39	1.715	0.043	R	0.161	0.754	0.393	0.434	0.433	0.469	0.157	0.764
1,3,5,6,14	7,10,11,12,13	2,4,8,9,15	38	1.592	0.056	NR	0.284	0.858	0.571	0.602	0.600	0.648	0.278	0.853
1,3,5,7,14	6,10,11,12,13	2,4,8,9,15	37	1.470	0.071	NR	0.322	0.872	0.617	0.653	0.653	0.698	0.322	0.865
1,2,3,7,11	6,8,9,12,15	4,5,10,13,14	35	1.225	0.110	NR	0.401	0.915	0.663	0.699	0.701	0.735	0.401	0.917
1,2,3,4,5	7,8,9,10,14	6,11,12,13,15	33	0.980	0.164	NR	0.521	0.957	0.732	0.763	0.763	0.794	0.498	0.954
1,2,3,4,5	6,7,8,9,10	11,12,13,14,15	25	0.000	0.500	NR	0.821	1	0.998	1.000	1	1	0.799	1
1,2,3,6,7	4,5,8,9,10	11,12,13,14,15	21	-0.490	0.688	NR	0.866	1	0.993	0.998	0.998	1	0.850	1
1,2,10,14,15	3,4,5,9,12	6,7,8,11,13	18	-0.857	0.804	NR	0.853	0.996	0.956	0.969	0.969	0.982	0.843	0.994
1,12,13,14,15	2,3,4,5,11	6,7,8,9,10	10	-1.837	0.967	NR	0.933	1.000	0.981	0.989	0.989	0.997	0.930	1
1,6,7,11,12	2,3,4,5,9	8,10,13,14,15	8	-2.082	0.981	NR	0.950	1.000	0.995	0.999	0.999	1	0.946	1
4,7,8,9,10	1,2,3,5,6	11,12,13,14,15	2	-2.817	0.998	NR	0.969	1	1	1	1	1	0.969	1
6,7,8,9,10	1,2,3,4,5	11,12,13,14,15	0	-3.062	0.999	NR	0.972	1	1	1	1	1	0.973	1

test is to reject H_0 if the test statistic satisfies $A_p \geq A_{2,0.0476} = 16$ for $n_x = n_y = n_z = 3$, and $A_p \geq A_{2,0.0496} = 39$ for $n_x = n_y = n_z = 5$. Similarly, for the EF test, the null hypothesis is rejected if the test statistic satisfies $\tilde{A}_p^* \geq Z_{0.05} = 1.645$.

Throughout this paper, the original test conclusion is denoted as R (Rejection) when H_0 is rejected and NR (Non-Rejection) otherwise. Reported values in the tables are rounded to three decimal places, except for precise values of 1, which are presented without additional decimals. The NPI-RP results in Tables 1 and 2 are identical for the MW and EF tests, as both are applied to three groups with $p = 2$. Consequently, the analysis that follows applies to both tests.

To compute the exact NPI lower and upper reproducibility probabilities, we consider all possible orderings of future observations among the given data. For $n_x = n_y = n_z = 3$, there are $\binom{6}{3} = 20$ possible orderings per group, leading to a total of $20^3 = 8000$ ordering combinations. For $n_x = n_y = n_z = 5$, the number of possible orderings per group increases to $\binom{10}{5} = 252$, resulting in $252^3 = 1.600 \times 10^{17}$ ordering combinations.

The results in Tables 1 and 2 show that $\underline{RP}$ is substantially below 0.5 in several cases and tends to be lower when the test statistic is close to the threshold. Additionally, $\underline{RP}$ is lower when H_0 is rejected compared to cases of non-rejection, as the directionality of the alternative hypothesis affects reproducibility. This suggests that test results near the critical threshold, particularly those leading to rejection, do not provide strong evidence for reproducibility.

For the exact lower reproducibility probability, H_0 is rejected in the future samples only if all future Y ranks are greater than the smallest observed Y rank, meaning they do not fall in the first interval for Y; all future X ranks are smaller than the largest observed X rank, meaning they do not fall in the last interval for X; and all future Z ranks are smaller than the largest observed Z rank, meaning they do not fall in the last interval for Z.

For instance, in the first row of Table 1, where $\underline{RP} = 0.125$, this means that all future Y ranks are greater than 7, all future X ranks are less than 3, and all future Z ranks are less than 6. Since each of these individual events occurs with probability 0.5 in the NPI framework, and the three groups are independent, the overall lower reproducibility probability is calculated as $0.5 \times 0.5 \times 0.5 = 0.125$. So, there are $\binom{5}{3}$ possible orderings for each of the future X, Y, and Z ranks, leading to a total of $\binom{5}{3}\binom{5}{3}\binom{5}{3} = 1000$ ordering combinations out of the 8000 total combinations.

Conversely, if at least one future X rank exceeds the largest observed X rank, one future Y rank is smaller than the smallest observed Y rank, or one future Z rank exceeds the largest observed Z rank, then H_0 will not be rejected in all cases. This occurs because unbounded intervals allow future X ranks to exceed future Y ranks or future Z ranks to exceed Y, which affects the test outcome. Thus, for the extreme case in the last row of Table 1, where $\underline{RP} = 0.933$, all future Y ranks are greater than 3, all future X ranks are smaller than 4, and all future Z ranks are smaller than 7.

Tables 1 and 2 show that, in most cases, different rank configurations within a sample can lead to the same test statistic value, yet the NPI-RP estimates differ. This indicates that NPI-RP depends on the specific ranks rather than solely on the test statistic value. An exception occurs in Table 1 for two cases with $A_p = 6$, where the estimates coincide due to an identical number of orderings leading to the same test result. However, this is not a general property.

In Table 1, the NPI lower and upper reproducibility probabilities exhibit significant imprecision due to the small sample size per group, resulting in large differences between the corresponding lower and upper estimates. As sample sizes increase, as seen in Table 2, the imprecision decreases, reflecting the greater amount of information available.

For cases where reproducibility is close to 1, the imprecision is minimal, whereas for lower reproducibility values, the imprecision is relatively higher. Since the NPI-RP approach is data-driven, imprecision generally decreases with larger sample sizes.

For larger samples, exhaustively evaluating all possible orderings becomes computationally infeasible. For instance, with $n_x = n_y = n_z = 7$, the NPI-RP-E approach requires evaluating $\left(\binom{14}{7}\right)^3 = (3432)^3 = 4.024 \times 10^{10}$ possible orderings of future observations, making exact calculations impractical. Therefore, alternative computational methods, such as NPI-RP-SO and NPI-RP-B, are required to approximate NPI reproducibility probabilities.

For NPI-RP-SO, $r^* = 2000$ orderings were sampled. The NPI-RP-B method was applied using Algorithm 1 with $B = 1000$ and $T = 100$. Summary statistics—including the minimum, mean, median, and maximum—were computed for

$RP_i, i = 1, 2, \ldots, T$. The results were then examined to determine whether NPI-RP-B estimates fell within the bounds of NPI-RP-E and NPI-RP-SO.

From Tables 1 and 2, it is evident that 100% of NPI-RP-B estimates fall within the bounds of NPI-RP-E and NPI-RP-SO. This is expected due to the construction of NPI lower and upper probabilities, which make no assumptions about how probability masses are assigned within the intervals between consecutive observations. While this result may not hold in rare cases due to the randomness inherent in bootstrap inference, it reinforces the reliability of NPI-RP-B, demonstrating consistency with the bounds of NPI-RP-E and NPI-RP-SO.

Similar findings have been reported in previous NPI studies. In one study, it was found that 100% of NPI-RP-B estimates fell within the bounds of NPI-RP-E for the one-sample signed rank test and the Wilcoxon-Mann-Whitney test [10]. Another study found that 88% of NPI-RP-B estimates were within the bounds of NPI-RP-SO for the likelihood ratio test, demonstrating a strong consistency given the inherent variability of bootstrap-based methods [43].

Example 2 (**NPI-RP-B for MW and EF tests, known peak, large samples)** This example examines the reproducibility probability for the MW and EF tests with a known peak using simulations. The reproducibility probability is estimated using the NPI-RP-B method, as described in Algorithm 1, with $B = 1000$ and $T = 100$ iterations. The null hypothesis is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$, with the alternative hypothesis $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_{p-1} \leq \mu_p \geq \mu_{p+1} \geq \cdots \geq \mu_g$. The level of significance is $\alpha = 0.05$.

Data were simulated under both H_0 and H_1, with $g = 5$ groups, $n = 20$, and a peak at $p = 3$. Table 3 presents the reproducibility probability (RP) estimates for data generated under H_0 from a standard normal distribution. Table 4 provides the RP estimates for data generated under H_1 from normal distributions with means $\mu_x = 0.1$, $\mu_y = 0.2$, $\mu_z = 0.5$, $\mu_v = 0.2$, and $\mu_w = 0.1$, with a common standard deviation of 1.

The reproducibility probability estimates differ between the MW and EF tests due to differences in how their test statistics are computed. Specifically, in the EF test, the Mann-Whitney sums are not uniformly weighted with a value of 1, as they are in the case of $g = 3$. This results in varying reproducibility probability values between the two tests. Figure 1 explores the relationship between NPI-RP-B and the p-value for both the MW and EF tests.

From Figure 1, it is evident that NPI-RP-B estimates tend to be lower when the observed p-value is close to the significance threshold $\alpha = 0.05$, particularly in cases where the null hypothesis is rejected. This is due to the presence of directional alternatives. When the p-value is further from the threshold, the data provide stronger evidence supporting the reproducibility of the original test result. These findings align with previous studies on NPI-based reproducibility probability estimation [10, 15, 43, 44].

Example 3 (**NPI-RP-B for JT test, large samples)** This example examines the reproducibility probability for the Jonckheere-Terpstra (JT) test for five groups, X, Y, Z, V, and W, using simulations with large sample sizes. The reproducibility probability is estimated using the NPI-RP-B method, as described in Algorithm 1. The hypothesis under consideration is $H_0 : \mu_x = \mu_y = \mu_z = \mu_v = \mu_w$ versus the ordered alternative $H_1 : \mu_x \leq \mu_y \leq \mu_z \leq \mu_v \leq \mu_w$, with equal sample sizes

Table 3: RP for the MW test and the EF test under H_0, $n = 20$, $\alpha = 0.05$, $A_{3,0.05} = 1410$, $Z_{0.05} = 1.645$

A_p	p-value	H_0	Min	Mean	Median	Max	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
1406	0.053	NR	0.515	0.546	0.546	0.600	1.572	0.058	NR	0.524	0.558	0.560	0.606
1371	0.090	NR	0.576	0.611	0.610	0.647	1.352	0.088	NR	0.572	0.606	0.606	0.646
1328	0.157	NR	0.690	0.719	0.720	0.755	1.111	0.133	NR	0.650	0.679	0.679	0.712
1244	0.365	NR	0.799	0.826	0.827	0.853	0.362	0.359	NR	0.797	0.827	0.827	0.851
1206	0.481	NR	0.842	0.867	0.867	0.894	0.058	0.477	NR	0.840	0.868	0.868	0.889
1189	0.534	NR	0.851	0.880	0.879	0.899	-0.131	0.552	NR	0.851	0.884	0.885	0.905
1146	0.664	NR	0.907	0.927	0.926	0.947	-0.398	0.655	NR	0.908	0.926	0.927	0.942
1123	0.727	NR	0.930	0.948	0.948	0.961	-0.598	0.725	NR	0.929	0.946	0.946	0.961
1086	0.815	NR	0.949	0.962	0.963	0.974	-0.881	0.811	NR	0.945	0.961	0.962	0.973
994	0.947	NR	0.974	0.984	0.984	0.991	-1.625	0.948	NR	0.975	0.985	0.985	0.993

Table 4: RP for the MW test and the EF test under H_1, $n = 20$, $\alpha = 0.05$, $A_{3,0.05} = 1410$, $Z_{0.05} = 1.645$

A_p	p-value	H_0	Min	Mean	Median	Max	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
1578	0.001	R	0.746	0.779	0.779	0.811	2.841	0.002	R	0.734	0.763	0.763	0.790
1520	0.006	R	0.696	0.731	0.733	0.762	2.511	0.006	R	0.690	0.729	0.730	0.761
1467	0.018	R	0.564	0.610	0.611	0.647	2.201	0.014	R	0.593	0.641	0.641	0.678
1431	0.035	R	0.475	0.515	0.514	0.545	1.829	0.034	R	0.493	0.525	0.525	0.556
1415	0.046	R	0.445	0.489	0.490	0.524	1.730	0.042	R	0.456	0.504	0.506	0.538
1364	0.099	NR	0.591	0.623	0.624	0.658	1.279	0.100	NR	0.587	0.623	0.624	0.656
1328	0.157	NR	0.683	0.709	0.710	0.733	0.975	0.165	NR	0.680	0.711	0.711	0.738
1292	0.235	NR	0.722	0.756	0.757	0.783	0.750	0.227	NR	0.716	0.746	0.747	0.773
1167	0.602	NR	0.884	0.907	0.907	0.925	-0.294	0.615	NR	0.896	0.915	0.917	0.934

$n_x = n_y = n_z = n_v = n_w = 20$ and a significance level of $\alpha = 0.05$. The null hypothesis is rejected at the nominal level $\alpha = 0.0499$ if $J \geq 2271$.

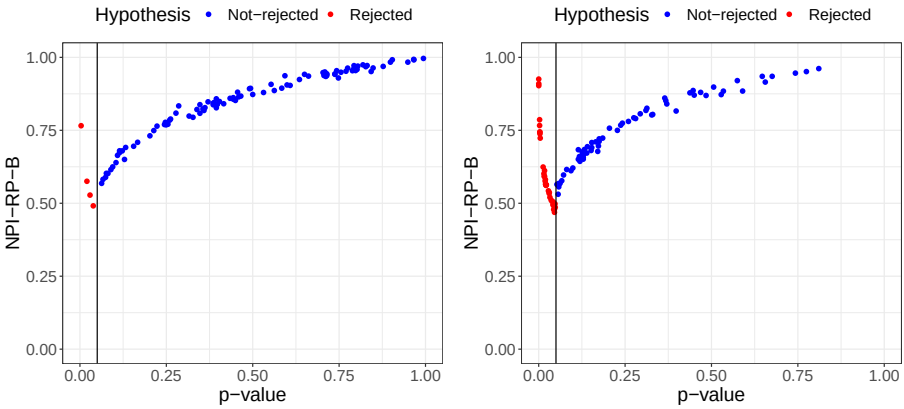
Data were simulated under both H_0 and H_1. In Table 5, data under H_0 were generated from the standard normal distribution. In Table 6, data under H_1 were generated from normal distributions with increasing means: $\mu_x = 0$, $\mu_y = 0.1$, $\mu_z = 0.2$, $\mu_v = 0.3$, and $\mu_w = 0.4$, with a common standard deviation of 1.

The simulation study follows these steps: Algorithm 1 is applied with $B = 1000$ and $T = 100$. For each iteration, a sample is generated for each group, the JT test is performed, and the test outcomes are recorded. The reproducibility probability estimates for 20 simulated datasets per scenario are reported in Tables 5 and 6.

The relationship between NPI-RP-B and the p-value of the JT test is examined. The p-value is used for visualization rather than the critical value, as each simulation scenario has a different threshold based on sample size and the number of groups. Despite this distinction, both approaches lead to the same conclusion about rejecting or not rejecting H_0. The vertical line in Figure 2 represents the significance level $\alpha = 0.05$.

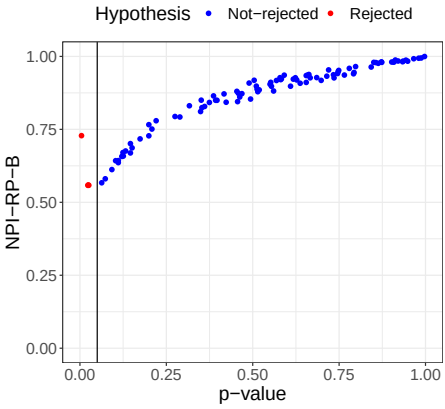
Figure 2a shows that when samples are generated under H_0, i.e., from $N(0, 1)$, the null hypothesis is rarely rejected. Conversely, Figure 2b illustrates that under H_1, where groups are sampled with increasing means, the null hypothesis is rejected more frequently, as expected. This aligns with the concept that higher power increases the likelihood of correctly rejecting H_0 when H_1 is true.

As expected, reproducibility probabilities are low when the observed p-value is close to the threshold 0.05, with RP estimates tending to be lower in rejection cases than in non-rejection cases—often substantially below 0.5. This is due to the directional nature of the ordered alternatives. The RP estimates increase as the observed p-value moves further from the threshold, regardless of whether H_0 is rejected. The

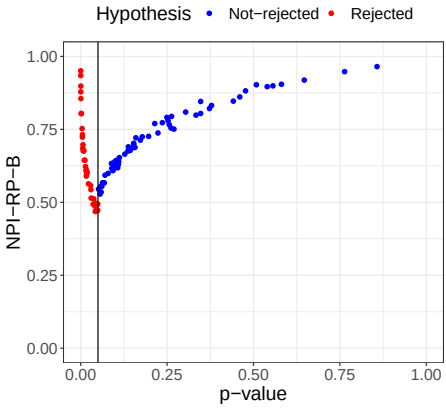


(a) MW, under H_0

(b) MW, under H_1



(c) EF, under H_0



(d) EF, under H_1

Fig. 1: NPI-RP-B for the MW test and EF test, $n = 20$, $\alpha = 0.05$

similarity between the median and mean of RP_i $(i = 1, \ldots, T)$ across simulated datasets suggests that the distribution of RP_i values is reasonably symmetric for each scenario.

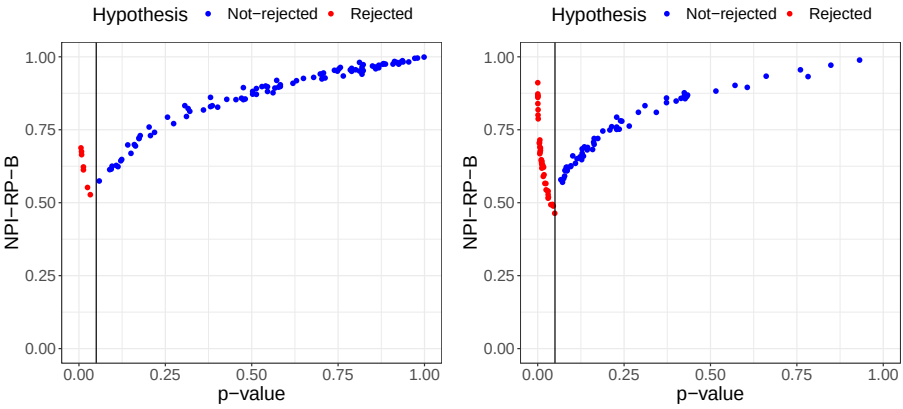
Example 4 (**NPI-RP-SO for MW and EF tests, real-world data with different assumptions for peak)** This example applies the NPI-RP-SO method to the MW test using the Telephone Communications data from [45], provided in Table 7. In this study, a firm seeks to enhance the cost-effectiveness of its communication strategies. Ten home office executives were randomly selected from the Sales, Production, and Research and Development departments to participate in the study.

Table 5: RP for the JT test under H_0, $n = 20$, $J_{0.0499} = 2271$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
2278	0.046	R	0.428	0.470	0.471	0.505	1935	0.655	NR	0.893	0.916	0.917	0.934
2261	0.057	NR	0.515	0.547	0.548	0.577	1902	0.725	NR	0.925	0.942	0.942	0.955
2228	0.083	NR	0.560	0.608	0.610	0.645	1892	0.745	NR	0.943	0.959	0.959	0.973
2195	0.119	NR	0.643	0.667	0.666	0.695	1850	0.820	NR	0.930	0.951	0.951	0.965
2157	0.171	NR	0.701	0.731	0.732	0.761	1814	0.871	NR	0.966	0.978	0.978	0.987
2148	0.185	NR	0.692	0.727	0.726	0.753	1736	0.946	NR	0.977	0.987	0.987	0.996
2114	0.245	NR	0.743	0.775	0.776	0.799	1660	0.981	NR	0.989	0.995	0.996	1
2093	0.287	NR	0.766	0.806	0.808	0.835	1609	0.992	NR	0.992	0.998	0.998	1
2071	0.334	NR	0.790	0.829	0.829	0.858	1556	0.997	NR	0.995	0.999	0.999	1
1988	0.530	NR	0.873	0.896	0.896	0.914	1397	1.000	NR	0.997	1.000	1	1

Table 6: RP for the JT test under H_1, $n = 20$, $J_{0.0499} = 2271$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
2518	0.001	R	0.815	0.837	0.837	0.862	2206	0.106	NR	0.623	0.659	0.659	0.692
2469	0.002	R	0.755	0.780	0.780	0.810	2190	0.125	NR	0.628	0.652	0.653	0.683
2457	0.003	R	0.711	0.737	0.737	0.768	2151	0.180	NR	0.682	0.715	0.716	0.747
2427	0.005	R	0.673	0.706	0.707	0.736	2072	0.332	NR	0.764	0.798	0.798	0.828
2375	0.011	R	0.596	0.633	0.632	0.666	2029	0.431	NR	0.814	0.854	0.854	0.877
2330	0.022	R	0.533	0.567	0.570	0.605	2003	0.494	NR	0.886	0.912	0.912	0.932
2308	0.031	R	0.501	0.538	0.540	0.570	1964	0.588	NR	0.896	0.914	0.915	0.933
2286	0.041	R	0.462	0.496	0.497	0.528	1923	0.681	NR	0.917	0.932	0.932	0.948
2280	0.044	R	0.438	0.483	0.484	0.515	1842	0.832	NR	0.953	0.968	0.968	0.980
2242	0.071	NR	0.535	0.581	0.581	0.615	1789	0.901	NR	0.969	0.981	0.981	0.990



(a) under H_0 (b) under H_1

Fig. 2: NPI-RP-B for the JT test, with $g = 5$, $n = 20$, $\alpha = 0.05$

The hypothesis of interest is $H_0 : \mu_x = \mu_y = \mu_z$ against $H_1 : \mu_x \leq \mu_y \geq \mu_z$, with equal sample sizes of $n_x = n_y = n_z = 10$. The test is conducted at a significance level of $\alpha = 0.05$, with a threshold value of $A_{2,0.0498} = 138$, meaning the null hypothesis is rejected if $A_p \geq 138$. For the EF test, the null hypothesis is rejected if $\tilde{A}_p^* \geq 1.645$. Three different cases are considered, as summarised in Table 8. Since the EF test

Table 7: Telephone communications data

	Data										mean	std. dev.
Sales	343	495	602	666	796	813	894	920	960	1499	798.8	315.637
Production	126	156	216	291	345	488	516	542	546	1362	458.8	355.422
Research and Development	391	450	472	496	609	645	705	763	910	1309	675	273.985

Table 8: Test results for the three cases of Telephone Communications data

Cases	X	Y	Z	A_p	$\tilde{A}_p^*$	Reject H_0?
Case 1	Production	Sales	Research and Development	148	2.112	Reject H_0
Case 2	Sales	Research and Development	Production	110	0.440	Do not reject H_0
Case 3	Sales	Production	Research and Development	42	-2.552	Do not reject H_0

reaches the same conclusions as the MW test, its outcomes are shown in Table 8 but omitted from further analysis.

Table 9: NPI-RP-SO for the MW test with $H_1 : \mu_x \leq \mu_y \geq \mu_z$, $p = 2$, $\alpha = 0.05$, $A_{2,0.0498} = 138$

Case 1: Sales is the peak(Y)				
r^*	$\underline{\widehat{RP}}$	CI(95%)	$\widehat{\widehat{RP}}$	CI(95%)
10	0.400	(0.096, 0.704)	0.700	(0.416, 0.984)
100	0.350	(0.257, 0.443)	0.830	(0.756, 0.904)
500	0.284	(0.244, 0.324)	0.794	(0.759, 0.829)
1,000	0.364	(0.334, 0.394)	0.800	(0.775, 0.825)
5,000	0.331	(0.318, 0.344)	0.807	(0.796, 0.818)
10,000	0.327	(0.318, 0.336)	0.801	(0.793, 0.809)
50,000	0.320	(0.316, 0.324)	0.805	(0.802, 0.808)
100,000	0.322	(0.319, 0.325)	0.803	(0.801, 0.805)
150,000	0.320	(0.318, 0.322)	0.807	(0.805, 0.809)
Case 2: Research and development is the peak(Y)				
r^*	$\underline{\widehat{RP}}$	CI(95%)	$\widehat{\widehat{RP}}$	CI(95%)
10	0.800	(0.444, 0.975)	1	(0.692, 1)
100	0.650	(0.557, 0.743)	0.950	(0.907, 0.993)
500	0.680	(0.639, 0.721)	0.958	(0.940, 0.976)
1,000	0.622	(0.592, 0.652)	0.945	(0.931, 0.959)
5,000	0.656	(0.643, 0.669)	0.950	(0.944, 0.956)
10,000	0.658	(0.649, 0.667)	0.954	(0.950, 0.958)
50,000	0.661	(0.657, 0.665)	0.953	(0.951, 0.955)
100,000	0.663	(0.660, 0.666)	0.954	(0.953, 0.955)
150,000	0.664	(0.662, 0.666)	0.955	(0.954, 0.956)
Case 3: Production is the peak(Y)				
r^*	$\underline{\widehat{RP}}$	CI(95%)	$\widehat{\widehat{RP}}$	CI(95%)
10	1	(0.692, 1)	1	(0.692, 1)
100	0.970	(0.915, 0.994)	1	(0.964, 1)
500	0.978	(0.965, 0.991)	1	(0.993, 1)
1,000	0.976	(0.967, 0.985)	1	(0.996, 1)
5,000	0.976	(0.972, 0.980)	0.999	(0.999, 1)
10,000	0.980	(0.977, 0.983)	0.999	(1.000, 1)
50,000	0.979	(0.978, 0.980)	0.999	(1.000, 1)
100,000	0.979	(0.978, 0.980)	0.999	(1.000, 1)
150,000	0.978	(0.977, 0.979)	0.999	(1.000, 1)

For Case 1, the MW test is applied with Production as group X, Sales as group Y, and Research and Development as group Z, yielding $A_p = 148$, which exceeds the threshold of 138. Thus, the null hypothesis is rejected. In Case 2, Sales is considered as group X, Research and Development as group Y, and Production as group Z, leading to $A_p = 110 < 138$, so the null hypothesis is not rejected. In Case 3, Sales is group X, Production is group Y, and Research and Development is group Z, yielding $A_p = 42 < 138$, again resulting in non-rejection of the null hypothesis.

Since computing exact lower and upper reproducibility probabilities requires evaluating $\binom{20}{10}\binom{20}{10}\binom{20}{10} = 6.307 \times 10^{15}$ ordering combinations, it is computationally infeasible. Instead, the sampling of orderings method is used to approximate these probabilities. Random samples of orderings r are drawn for each group, and the sampled orderings are used to compute the minimum and maximum MW test statistics. This allows for approximate estimation of the lower and upper reproducibility probabilities, along with their corresponding 95% confidence intervals.

From Table 9, in Case 1, the lower reproducibility probability is relatively low since the test statistic $A_p = 148$ is close to the threshold. In contrast, for Case 3, $\widehat{\underline{RP}}$ is higher because $A_p = 42$ is far from the threshold, making the test outcome more stable under future replications.

The table also shows that increasing r has a minor impact on the estimates, with changes occurring only in the second decimal place. This suggests that reasonable approximations of the NPI lower and upper reproducibility probabilities can be obtained with $r \geq 10,000$, which is significantly smaller than the total number of orderings. The first application of NPI-RP-SO for test reproducibility, as carried out by [14] for the likelihood ratio test, suggests that sampling at least 2,000 orderings is sufficient for reliable estimates. Thus, NPI-RP-SO provides a computationally efficient way to approximate lower and upper reproducibility probabilities while avoiding the computational burden of exhaustive enumeration.

Example 5 (**NPI-RP-B for MW Test with Unknown Peak**) This example examines the NPI reproducibility probability for the MW test when the peak position is unknown and must be estimated from the data. The NPI-RP-B method is applied using Algorithm 1, with $B = 1000$ and $T = 1$. The peak is estimated by first computing standardised Mann–Whitney statistics for each group, and identifying the group(s) with the maximum value. The Mack–Wolfe statistic is then calculated based on this estimated peak position. This setup enables us to investigate how uncertainty in peak estimation impacts the reproducibility of test results.

In this example, the reproducibility probability is assessed for $g = 3$ groups, X, Y, and Z, each with sample size $n = 10$. The hypothesis of interest is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$ against $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_{p-1} \leq \mu_p \geq \mu_{p+1} \geq \cdots \geq \mu_g$, where the peak position p is unknown. The level of significance is set at $\alpha = 0.05$, leading to the critical value $A_{0.0498} = 2.112$, obtained from [28]. The null hypothesis is rejected if $A'_{\hat{p}} \geq 2.112$. For the MW test with an unknown peak and large sample sizes, the Monte Carlo Approximation is used to obtain the critical values.

In Table 10, the original data are generated from Normal distributions with means $\mu_x = 0$, $\mu_y = 1.5$, and $\mu_z = 1$, with a standard deviation of 1. The table presents the NPI-RP estimates for 10 original samples. For each sample, the probability of reaching the same test conclusion in the future is computed based on the $B = 1000$ bootstrap replications. Additionally, the contribution of each peak position to this probability is shown.

Table 10: NPI-RP-B for the MW test with unknown peak, $k = 3$, $X \sim N(0,1)$, $Y \sim N(1.5,1)$, $Z \sim N(1,1)$, $n = 10$, $\alpha = 0.05$, $A_{0.0498} = 2.112$, $B = 1000$, $T = 1$

	Test conclusion			Rejection				Non-rejection			
Samples	$\hat{p}$	$A'_{\hat{p}}$	H_0	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
1	2	4.267	R	0.992	0.000	0.984	0.008	0.008	0.000	0.006	0.002
2	3	3.384	R	0.812	0.000	0.041	0.771	0.188	0.007	0.021	0.160
3	3	3.004	R	0.867	0.000	0.257	0.610	0.133	0.008	0.053	0.072
4	2	2.948	R	0.746	0.001	0.618	0.127	0.254	0.012	0.133	0.109
5	2	2.772	R	0.855	0.002	0.587	0.266	0.145	0.008	0.092	0.045
6	2	2.640	R	0.782	0.004	0.561	0.217	0.218	0.013	0.152	0.053
7	2	2.552	R	0.888	0.001	0.436	0.451	0.112	0.002	0.051	0.059
8	2	2.376	R	0.560	0.013	0.477	0.070	0.440	0.059	0.261	0.120
9	3	2.281	R	0.587	0.001	0.113	0.473	0.413	0.023	0.098	0.292
10	2	1.628	NR	0.369	0.017	0.278	0.074	0.631	0.124	0.306	0.201

For example, in the first sample in Table 10, the estimated peak group in the original data is the second group ($\hat{p} = 2$), and the null hypothesis is rejected. The probability of rejecting H_0 in the 1000 bootstrap replications is 0.992. The majority of this probability comes from cases where the peak remains in the second group ($\hat{p} = 2$) in future samples, contributing 0.984 to the estimate. Cases where the peak shifts to the third group contribute 0.008, while no future samples have a peak at $\hat{p} = 1$.

In general, the estimated peak group in the original sample contributes the most to the reproducibility probability in future bootstrap samples. The NPI reproducibility probability is lower when the test statistic is close to the threshold, particularly when the null hypothesis is rejected. Conversely, when the test statistic is further from the threshold, NPI-RP-B estimates are higher, indicating greater reproducibility.

7 Concluding Remarks

This paper introduced NPI-based methods for assessing the reproducibility of the Mack-Wolfe (MW) test, the Esra and Fikri (EF) test, and the Jonckheere-Terpstra (JT) test. Exact lower and upper reproducibility probabilities were derived for the MW test with three groups; however, for larger sample sizes and more than three groups, exhaustive enumeration of all possible orderings becomes computationally infeasible. To address this, two NPI-based approaches were implemented: the sampling of orderings and the NPI-bootstrap technique. These methods provide feasible alternatives for computing reproducibility probabilities across all sample sizes. The computational efficiency and consistency of results obtained using NPI-B and sampling-based methods demonstrate their effectiveness in overcoming computational challenges in large samples. This work extends the development of NPI reproducibility, originally introduced by Coolen and BinHimd [11], and the findings align with previous NPI studies on test reproducibility. Notably, the results confirm that reproducibility probability tends to be low when the test statistic is close to the decision threshold. A close similarity between the reproducibility probability estimates for the MW and EF tests is observed, particularly in large-sample settings, as demonstrated in Example 2.

Several research challenges remain for advancing NPI methods in reproducibility probability. This study examined the reproducibility of the JT test using the NPI-B approach, but deriving exact lower and upper reproducibility probabilities for the JT test remains an open problem that may require methodological developments. Similarly, while deriving exact lower and upper reproducibility probabilities for the MW test with more than three groups remains an open problem, the computational challenges involved may make it more practical to explore alternative methods for efficiently approximating these probabilities. Further investigations could also explore the reproducibility of other umbrella alternatives tests, such as the Modified Jonckheere-Terpstra test [46], the Page test [47], the Chen and Wolfe test [25], and the Hettmansperger and Norton test [24], contributing to a broader understanding of reproducibility in nonparametric hypothesis testing.

Acknowledgments. The first author acknowledges the financial support from the Saudi Arabian Cultural Bureau in London and King Faisal University in Saudi Arabia for funding her PhD studies at Durham University.

Appendix A

Proof of Equations (19) *and* (20) We consider the hypothesis test $H_0 : \mu_x = \mu_y = \mu_z$ against the alternative $H_1 : \mu_x \leq \mu_y \geq \mu_z$, where $p = 2$ refers to the second group, Y. In this case, the Mack-Wolfe test for three groups X, Y, and Z is based on the sum of the Mann-Whitney counts U_{XY} and U_{ZY}, expressed as:

$$A_p = U_{XY} + U_{ZY} = \left[R_{XY} - \frac{n_y(n_y+1)}{2}\right] + \left[R_{ZY} - \frac{n_y(n_y+1)}{2}\right],$$

where R_{XY} is the sum of the ranks of group Y when X and Y are combined, and R_{ZY} is the sum of the ranks of group Y when Y and Z are combined.

For each combination of orderings O_ℓ, the corresponding Mack-Wolfe test statistic is denoted by A_{p_ℓ}. Within the NPI framework, no assumptions are made about the exact locations of future observations within the intervals (x_{j-1}, x_j), (y_{i-1}, y_i), and (z_{k-1}, z_k). However, the number of future observations falling within each interval is known. As a result, an exact value of A_{p_ℓ} for a given ordering cannot be determined, but its minimum and maximum possible values, denoted by $\underline{A}_{p_\ell}$ and $\overline{A}_{p_\ell}$, respectively, can be derived. For simplicity, we omit the index ℓ in the following derivations.

To determine the minimum value, $\underline{A}_{p_\ell}$, all S_j^X future X observations in the interval (x_{j-1}, x_j) are set to x_j, all S_i^Y future Y observations in the interval (y_{i-1}, y_i) are set to y_{i-1}, and all S_k^Z future Z observations in the interval (z_{k-1}, z_k) are set to z_k. Under these conditions, the ranks of the S_i^Y future Y observations at y_{i-1} in the combined $X\&Y$ dataset are:

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + 1, \ldots, \sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + S_i^Y. \tag{A1}$$

Similarly, in the combined $Z\&Y$ dataset, the ranks are:

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z + 1, \ldots, \sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z + S_i^Y. \tag{A2}$$

Summing these ranks gives:

$$\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right] +$$
$$\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right]. \tag{A3}$$

Summing over all $i = 1, \ldots, n_y + 1$ and using $\sum_{i=1}^{n_y+1} S_i^Y = n_y$ leads to:

$$\underline{A}_{p_\ell} = \sum_{i=1}^{n_y+1} S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z\right]. \tag{A4}$$

To determine the maximum value, $\overline{A}_{p_\ell}$, all S_j^X future X observations in (x_{j-1}, x_j) are set to x_{j-1}, all S_i^Y future Y observations in (y_{i-1}, y_i) are set to y_i, and all S_k^Z future Z observations in (z_{k-1}, z_k) are set to z_{k-1}. The ranks of the S_i^Y future Y observations at y_i in the combined $X\&Y$ dataset are:

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + 1, \ldots, \sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + S_i^Y. \tag{A5}$$

Similarly, in the combined $Z\&Y$ dataset, the ranks are:

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i)}-1} S_c^Z + 1, \ldots, \sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i)}-1} S_c^Z + S_i^Y. \tag{A6}$$

Summing these ranks gives:

$$\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right] +$$
$$\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i)}-1} S_c^Z\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right]. \tag{A7}$$

Summing over all $i = 1, \ldots, n_y + 1$ and using $\sum_{i=1}^{n_y+1} S_i^Y = n_y$ leads to:

$$\overline{A}_{p_\ell} = \sum_{i=1}^{n_y+1} S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + \sum_{c=1}^{k_{(i)}-1} S_c^Z\right]. \tag{A8}$$

□

References

[1] Atmanspacher, H., Maasen, S.: Reproducibility: Principles, Problems, Practices, and Prospects. Wiley, Hoboken (2016)

[2] Goodman, S.N.: A comment on replication, p-values, and evidence. Statistics in Medicine **11**(7), 875–879 (1992)

[3] Senn, S.: Comment on 'A comment on replication, p-values and evidence', by S. N. Goodman (letter to the editor). Statistics in Medicine **21**(15), 2437–2444 (2002). https://doi.org/10.1002/sim.1201

[4] Shao, J., Chow, S.C.: Reproducibility probability in clinical trials. Statistics in Medicine **21**, 1727–1742 (2002)

[5] De Martini, D.: Reproducibility probability estimation for testing statistical hypotheses. Statistics and Probability Letters **78**, 1056–1061 (2008)

[6] De Capitani, L., De Martini, D.: On stochastic orderings of the wilcoxon rank sum test statistic-with applications to reproducibility probability estimation testing. Statistics and Probability Letters **81**, 937–946 (2011)

[7] De Capitani, L., De Martini, D.: Reproducibility probability estimation and rp-testing for some nonparametric tests. Entropy **18** (2016)

[8] De Capitani, L.: An introduction to RP-testing. Epidemiology Biostatistics and Public Health **10**, 1–16 (2013)

[9] Boos, D.D., Stefanski, L.A.: P-value precision and reproducibility. American Statistician **65**, 213–221 (2011)

[10] BinHimd, S.: Nonparametric predictive methods for bootstrap and test reproducibility. PhD thesis, Durham University (2014)

[11] Coolen, F.P.A., BinHimd, S.: Nonparametric predictive inference for reproducibility of basic nonparametric tests. Journal of Statistical Theory and Practice **8**, 591–618 (2014)

[12] Coolen, F.P.A., Alqifari, H.N.: Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. REVSTAT: Statistical Journal. **16**, 167–185 (2018)

[13] Simkus, A., Coolen-Maturi, T., Coolen, F.P.A., Karp, N.A., Bendtsen, C.: Statistical reproducibility for multiple pairwise tests in pharmaceutical research. Statistical Methods in Medical Research **31**, 673–688 (2022)

[14] Marques, F.J., Coolen, F.P.A., Coolen-Maturi, T.: Introducing nonparametric predictive inference methods for reproducibility of likelihood ratio tests. Journal of Statistical Theory and Practice **13**, 15 (2019)

[15] Marques, F.J., Coolen, F.P.A.: Nonparametric predictive inference for test reproducibility by sampling future data orderings. Journal of Statistical Theory and Practice **14**, 62 (2020)

[16] Coolen, F.P.A., BinHimd, S.: Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov-Smirnov test. Journal of Statistical Theory and Practice **14**, 1–13 (2020)

[17] Mack, G.A., Wolfe, D.A.: K-sample rank tests for umbrella alternatives. Journal of the American Statistical Association **76**, 175–181 (1981)

[18] Hollander, M., Wolfe, D.A., Chicken, E.: Nonparametric Statistical Methods. Wiley, New Jersey (2013)

[19] Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 50–60 (1947)

[20] Millen, B.A., Wolfe, D.A.: A class of nonparametric tests for umbrella alternatives. Journal of Statistical Research **39**, 1–18 (2005)

[21] Esra, G., Fikri, G.: A modified mack–wolfe test for the umbrella alternative problem. Communications in Statistics-Theory and Methods **45**, 7226–7241 (2016)

[22] Bhat, S.V.: Simple k-sample rank tests for umbrella alternatives. Research Journal of Mathematics and Statistics **1**, 27–29 (2009)

[23] Basso, D., Salmaso, L.: A permutation test for umbrella alternatives. Statistics and Computing **21**, 45–54 (2011)

[24] Hettmansperger, T.P., Norton, R.M.: Tests for patterned alternatives in k-sample problems. Journal of the American Statistical Association **82**, 292–299 (1987)

[25] Chen, Y.I., Wolfe, D.A.: A study of distribution-free tests for umbrella alternatives. Biometrical Journal **32**, 47–57 (1990)

[26] Magel, R., Qin, L.: A non-parametric test for umbrella alternatives based on ranked-set sampling. Journal of Applied Statistics **30**, 925–937 (2003)

[27] Grant, S., Eric, C., Rachel, B.: NSM3: Functions and Datasets to Accompany Hollander, Wolfe, and Chicken - Nonparametric Statistical Methods,

Third Edition. (2020). R package version 1.14. https://CRAN.R-project.org/package=NSM3

[28] Bonnini, S., Corain, L., Marozzi, M., Salmaso, L.: Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R. Wiley, Chichester (2014)

[29] Jonckheere, A.R.: A test of significance for the relation between m rankings and k ranked categories. British Journal of Statistical Psychology **7**, 93–100 (1954)

[30] Terpstra, T.J.: The asymptotic normality and consistency of kendall's test against trend, when ties are present in one ranking. Indagationes Mathematicae **14**, 327–333 (1952)

[31] Gibbons, J., Chakraborti, S.: Nonparametric Statistical Inference: Revised and Expanded. Statistics: A Series of Textbooks and Monographs. Marcel Dekker, Inc, New York (2003)

[32] Hill, B.M.: Posterior distribution of percentiles: Bayes' theorem for sampling from a population. Journal of the American Statistical Association **63**, 677–691 (1968)

[33] Hill, B.M.: De Finetti's Theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). Bayesian Statistics **3**, 211–241 (1988)

[34] Geisser, S.: Predictive Inference: An Introduction. Chapman and Hall, London (1993)

[35] De Finetti, B.: Theory of Probability: a Critical Introductory Treatment. Wiley, London (1974)

[36] Augustin, T., Coolen, F.P.A., de Cooman, G., Troffaes, M.C.M.: Introduction to Imprecise Probabilities. Wiley, Chichester (2014)

[37] Coolen, F.P.A.: On the Use of Imprecise Probabilities in Reliability. Quality and Reliability Engineering International **20**, 193–202 (2004)

[38] Coolen, F.P.A.: Low structure imprecise predictive inference for Bayes' problem. Statistics & Probability Letters **36**, 349–357 (1998)

[39] Coolen, F.P.A., Augustin, T.: A nonparametric predictive alternative to the imprecise Dirichlet model: the case of a known number of categories. International Journal of Approximate Reasoning **50**, 217–230 (2009)

[40] Coolen, F.P.A.: On nonparametric predictive inference and objective Bayesianism. Journal of Logic, Language and Information **15**, 21–47

(2006)

[41] Morisette, J.T., Khorram, S.: Exact binomial confidence interval for proportions. Photogrammetric Engineering and Remote Sensing **64**, 281–282 (1998)

[42] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC, New York (1993)

[43] Aldawsari, A.: Parametric predictive bootstrap and test reproducibility. PhD thesis, Durham University (2023). http://etheses.dur.ac.uk/14970/

[44] Simkus, A.: Contributions to statistical reproducibility and small-sample bootstrap. PhD thesis, Durham University (2023). http://etheses.dur.ac.uk/15294/

[45] Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models. McGraw-Hill/Irwin, New York (2005)

[46] Tryon, P.V., Hettmansperger, T.P.: A class of nonparametric tests for homogeneity against ordered alternatives. The Annals of Statistics, 1061–1070 (1973)

[47] Page, E.B.: Ordered hypotheses for multiple treatments: a significance test for linear ranks. Journal of the American Statistical Association **58**, 216–230 (1963)