

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2025): 4.49

Maths 2026; 11(5): 131-140

© 2025 Stats & Maths

<https://www.mathsjournal.com>

Received: 17-03-2026

Accepted: 19-04-2026

Kwame Danso Fenteng

Department of Statistics, North
Dakota State University, Fargo,
ND, USA

Rhonda Magel

Department of Statistics, North
Dakota State University, Fargo,
ND, USA

Millicent Anane Berko

Department of Statistics, North
Dakota State University, Fargo,
ND, USA

Corresponding Author:

Rhonda Magel

Department of Statistics, North
Dakota State University, Fargo,
ND, USA

Power analysis of combined nonparametric tests: Fligner-Wolfe and Wald-Wolfowitz approaches

Kwame Danso Fenteng, Rhonda Magel and Millicent Anane Berko

DOI: <https://www.doi.org/10.22271/math.2026.v11.i5b.2326>

Abstract

This paper provides a comprehensive power analysis of two nonparametric statistical tests: the Fligner-Wolfe (FW) rank test and the Wald-Wolfowitz (WW) runs test. Additionally, it evaluates a combined test proposed by Schukr and Magel (2024) to detect differences in location and scale. A simulation study in SAS is performed under various different distributions including a normal distribution, the Student t -distribution, and an exponential distribution. The simulated scenarios encompass purely location-based differences, scale-based differences, and hybrid alternatives (a combination of location and scale differences). Both balanced and unbalanced group sizes ranging from $n = 9$ to $n = 30$ are analyzed. The estimated powers and Type I error rates are used to evaluate these tests' performance under various scenarios.

Keywords: Location, Scale, Simple Tree Alternative, Control, Treatment Levels

1. Introduction

1.1 Background and Significance

Nonparametric statistical methods play a crucial role in modern data analysis, particularly when real-world data violate the strict assumptions required for parametric tests such as normality and homogeneity of variance. In practical settings like clinical trials, quality control, biomedical research, and engineering, data are often skewed, heavy-tailed, or contain outliers. Under such conditions, traditional methods like the t -test or ANOVA may yield misleading results. Nonparametric tests offer a robust alternative by relying on data ranks instead of raw values, making them less sensitive to distributional violations (Conover, 1999; Gibbons & Chakraborti, 2011).

Most nonparametric tests are designed to assess either location differences (i.e., shifts in central tendency) or scale differences (i.e., dispersion), but not both. For instance, the Mann-Whitney U and Wilcoxon rank-sum tests effectively detect location shifts, while the Ansari-Bradley and Fligner-Wolfe (FW) tests are more sensitive to scale differences (Sheskin, 2003). In contrast, real-world applications often involve hybrid alternatives, where location and scale parameters differ between groups. In such cases, relying on a single-aspect test may reduce statistical power.

This paper examines two established nonparametric procedures: the Fligner-Wolfe (FW) rank test and the Wald-Wolfowitz (WW) runs test. Building on these methods, it applies the combined test proposed by Schukr and Magel (2024) ^[1], which is designed to detect differences in both location (mean) and scale (variance).

1.2 The Fligner-Wolfe (FW) Test

The Fligner-Wolfe test is a rank-based, distribution-free method designed to detect location differences between a control group and one or more treatment groups (Fligner & Wolfe, 1982) ^[4]. It is especially valuable in ordered settings such as dose-response experiments. The test computes a statistic based on the sum of ranks from treatment groups, allowing robust comparison without assuming normality. However, its sensitivity is limited when differences in scale alone are present, as it was not designed to detect changes in variance.

1.3 The Wald-Wolfowitz (WW) Runs Test

The WW test assesses whether two samples originate from the same distribution by analyzing the sequence of group labels in pooled and sorted data (Wald & Wolfowitz, 1940)^[14]. It counts the number of "runs" — consecutive observations from the same group — and compares this count to the expected number under randomness. The WW test is beneficial for identifying scale or distributional differences and can flag broader deviations beyond simple location shifts. However, its power for pure location changes is typically lower than rank-based location-specific tests (Hollander *et al.*, 2013)^[17].

1.4 The Need for a Combined Approach

While the FW test performs well in identifying location differences and the WW test is adept at detecting broader distributional changes, neither is optimal in hybrid scenarios where location and scale differ simultaneously. This gap motivates the development of a combined test that merges the strengths of both FW and WW. The approach is conceptually aligned with existing methods, such as the Lepage test, which integrates the Wilcoxon and Ansari-Bradley statistics to simultaneously test for location and scale differences (Lepage, 1971)^[9].

1.5 Significance of the Study

This paper contributes to the field of nonparametric statistics in the following ways:

- Systematic evaluation of the FW and WW tests' performance under diverse distributional settings, including normal, Student t , and exponential distributions.
- Development of a novel combined test that enhances power for detecting hybrid alternatives.
- Practical guidance for applied researchers in biostatistics, reliability engineering, and social science, where data often deviate from standard parametric assumptions.

Through comprehensive Monte Carlo simulations implemented in SAS, this study assesses each test's Type I error control and statistical power across balanced and unbalanced group designs (sample sizes $n = 9$ to 30). The results aim to help researchers select more effective tools for real-world, non-normal data.

1.6 Problem Statement

In structured group comparisons such as those involving a control and multiple treatments, researchers often face the challenge of detecting differences that may affect central tendency (location) and dispersion (scale). Many nonparametric tests are tailored to detect either location or scale differences, limiting their utility when both are present — particularly in studies with complex treatment effects that influence both mean and variability.

The simple tree alternative framework provides a meaningful structure for modeling such comparisons, especially when treatments are logically ordered (e.g., increasing dosages or time-based interventions). While tests like the Fligner-Wolfe (FW) test and the Wald-Wolfowitz (WW) runs test are well established for detecting location and distributional changes, respectively, there is limited research evaluating their performance under joint location-scale alternatives within this framework. This gap is critical, as real-world data often deviate from theoretical assumptions, especially in small-sample settings or under skewed distributions.

To address this, the present study explores whether a combined testing approach — integrating the FW and WW tests — can offer improved sensitivity to location and scale differences by systematically comparing the performance of each test individually and in combination. For instance, a new drug might increase average response (location shift) while altering variability (scale shift). Traditional methods that apply the Fligner-Wolfe or Wald-Wolfowitz test individually may suffer from limited statistical power when location and scale differences are simultaneously present. The FW test primarily detects rank-based location shifts, while the WW test captures distributional irregularities through run patterns (Alsubie & Magel, 2020)^[11]. Their distinct focuses highlight the need for a unified test that harnesses the strengths of both approaches to improve sensitivity under complex alternatives. This study addresses two key gaps:

- **Performance Evaluation:** No systematic comparison exists for FW and WW under hybrid alternatives, especially in small samples ($n < 30$).
- **Methodological Integration:** The potential of combining FW (rank-based) and WW (runs-based) statistics remains unexplored despite their shared robustness to non-normality.

1.7 Research Objectives

The primary aim of this study is to evaluate and compare the performance of the Fligner-Wolfe (FW) rank test, the Wald-Wolfowitz (WW) runs test, and the combined test proposed by Schukr and Magel (2024)^[12] within the framework of the simple tree alternative. Specifically, the study seeks to determine whether these tests maintain appropriate Type I error control under the null hypothesis, to assess their relative statistical power under location-only, scale-only, and hybrid alternatives, and to develop and evaluate a unified test statistic that leverages the complementary strengths of FW and WW. Through extensive Monte Carlo simulations across varying distributional assumptions and sample size configurations, the study also aims to generate practical, evidence-based recommendations to assist applied researchers in selecting effective nonparametric methods for ordered treatment designs.

The specific objectives of the study are to:

- Examine Type I error control by evaluating whether the FW, WW, and combined tests maintain nominal significance levels under the null hypothesis of no treatment effect.
- Estimate statistical power under structured alternatives by comparing test performance in three scenarios: location-only differences (expected to favor FW), scale-only differences (expected to favor WW), and hybrid location-scale differences (to assess the benefit of combining).
- Develop and implement a combined test statistic that synthesizes the FW and WW procedures into a unified approach and evaluate its power advantages.
- Conduct simulations under the simple tree structure using both balanced and unbalanced group sizes (ranging from $n = 9$ to $n = 30$) across three distributional settings: the symmetric normal distribution, the heavy-tailed Student's t distribution with three degrees of freedom, and the right-skewed exponential distribution.
- Provide evidence-based recommendations to guide applied researchers in selecting appropriate nonparametric methods for studies involving ordered treatments and complex distributional shifts.

1.8 Scope of the Study

This study investigates the performance of three nonparametric statistical tests — the Fligner-Wolfe (FW) rank test, the Wald-Wolfowitz (WW) runs test, and a proposed combined test — within the simple tree alternative framework, which models structured comparisons between a control group and two or three treatment groups. The scope is limited to univariate, continuous data under one-sided hypothesis testing, particularly detecting location shifts, scale differences, and hybrid alternatives.

Simulations are conducted using SAS, generating data from normal, Student t ($df = 3$), and exponential distributions to reflect symmetric, heavy-tailed, and positively skewed distributions, respectively. Sample sizes range from $n = 9$ to $n = 30$ per group, with balanced and unbalanced group configurations considered. The performance of each test is evaluated based on Type I error rates and statistical power. Multivariate extensions and other nonparametric procedures beyond FW, WW, and their combination are beyond the scope of this work.

2. Materials and Methods

2.1 Research Design

This study evaluates the performance of three nonparametric tests within the simple tree alternative framework, a structured model used to compare a control group against two or more ordered treatment groups. The core objective is to assess how effectively each test detects location shifts, scale differences, or both under varied data and design conditions. The three tests considered are:

- **Fligner-Wolfe (FW) rank test** — a rank-based procedure designed to detect location differences between a control and multiple treatments (Fligner & Wolfe, 1982)^[4].
- **Wald-Wolfowitz (WW) runs test** — a nonparametric test that assesses scale and distributional differences through the pattern of label sequences in the pooled data (Wald & Wolfowitz, 1940)^[14].
- **Combined FW-WW test** — proposed by Schukr and Magel (2024)^[12], integrates FW and WW statistics to simultaneously capture location and scale shifts, drawing inspiration from composite approaches such as the Lepage test (Lepage, 1971)^[9].

The simple tree alternative offers a flexible framework for ordered comparisons. In this model, the null hypothesis assumes no differences in means or variances across groups, while the alternative hypothesis posits ordered increases in means and/or variances from the control to the treatments. This is particularly relevant in real-world applications such as clinical trials, industrial experimentation, or social intervention studies where increasing treatments may influence both central tendency and dispersion.

The simulation design adheres to the methodology proposed by Alsubie and Magel (2020)^[1] and implements the following components:

- **Multiple distributional conditions:** Data are simulated from three distinct continuous distributions — normal (symmetric, light-tailed), Student t with 3 degrees of freedom (symmetric, heavy-tailed), and exponential (skewed).
- **Balanced and unbalanced sample sizes:** Sample sizes per group range from $n = 9$ to $n = 30$, with both balanced and unbalanced configurations examined.

- **Controlled Type I error assessment:** Tests are evaluated under the null hypothesis to ensure that they maintain the nominal significance level ($\alpha = 0.05$).
- **Systematic power comparisons:** Each test's power is estimated across pure location differences (equal variances), pure scale differences (equal means), and hybrid alternatives (both mean and variance differ).

By simulating 10,000 replications per scenario, the study ensures stable and reliable estimates of Type I error and statistical power. SAS is used to implement the simulation algorithms, test procedures, and result aggregation.

2.2 Simulation Framework

2.2.1 Simple Tree Alternative Structure

The simple tree alternative is designed for hypothesis testing in situations where treatments are ordered or structured relative to a common control. The hypotheses are formulated as follows.

Null Hypothesis (H_0)

$$\mu_1 = \mu_2 = \dots = \mu_k \text{ and } \sigma_1 = \sigma_2 = \dots = \sigma_k \quad (2.1)$$

Alternative Hypothesis (H_1)

$$\mu_1 \leq [\mu_2, \dots, \mu_k] \text{ and/or } \sigma_1 \leq [\sigma_2, \dots, \sigma_k] \quad (2.2)$$

This structure reflects realistic scenarios such as increasing drug dosage effects or progressive treatment intensities, where both the central tendency and variability may increase across treatment groups.

2.2.2 Data Generation Process

To evaluate the performance of the tests under different conditions, data were generated using SAS 9.4. The study includes data from three continuous distributions that differ in symmetry, tail behavior, and skewness:

- **Normal:** $\text{rand}(\text{'Normal'})$, $\mu = 0$, $\sigma^2 = 1$
- **Student t ($df = 3$):** $\text{rand}(\text{'T'}, 3)$, scaled to $\sigma^2 \approx 1$
- **Exponential:** $\text{rand}(\text{'Expo'})$, $\theta = 1$ ($\mu = \sigma = 1$)

Each replication of the simulation generates samples for one control group and two or three treatment groups under the null condition (equal means and variances) and under alternative scenarios (location-only, scale-only, or hybrid).

2.3 Test Statistics

2.3.1 Fligner-Wolfe (FW) Rank Test

The FW test is a rank-based procedure developed to compare a control group with multiple treatment groups using a nonparametric framework. Let $N = n_1 + n_2$ be the total sample size, where n_1 is the size of the control group and n_2 is the total size of all treatment groups. Let r_j be the rank of the j -th observation from the control group.

The test statistic is:

$$FW = \sum_{j=1}^{n_1} r_j \quad (2.3)$$

The expected value and variance of FW are:

$$E(FW) = \frac{n_1(N+1)}{2} \quad (2.4)$$

$$\text{Var}(FW) = \frac{n_1 n_2 (N+1)}{12} \quad (2.5)$$

The standardized test statistic is:

$$FW^* = \frac{FW - E(FW)}{\sqrt{\text{Var}(FW)}} \quad (2.6)$$

Decision Rule: Reject H_0 if $FW^* \leq -Z\alpha$.

2.3.2 Wald-Wolfowitz (WW) Runs Test

The WW test evaluates the number of runs in the combined data from two samples. Let n_1 be the size of group 1 (e.g., control), n_2 be the size of group 2 (e.g., treatment), and R be the observed number of runs. The null hypothesis is $H_0: FX(x) = FY(y)$.

The expected number of runs and its variance are:

$$E(WW) = \frac{2n_1n_2}{n_1+n_2} + 1 \quad (2.7)$$

$$\text{Var}(WW) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2 - 1)} \quad (2.8)$$

The standardized test statistic is:

$$WW^* = \frac{R - E(WW)}{\sqrt{\text{Var}(WW)}} \quad (2.9)$$

Decision Rule: Reject H_0 if $WW^* \leq -Z\alpha$.

2.3.3 Combined FW-WW Test Statistic

To improve sensitivity in hybrid alternatives where both location and scale vary, a combined test statistic is proposed. Under H_0 , both FW^* and WW^* follow a standard normal distribution $N(0,1)$. Assuming independence under the null, their sum has variance 2, and the combined statistic is:

$$T = \frac{FW^* + WW^*}{\sqrt{2}} \quad (2.10)$$

Therefore $T \sim N(0, 1)$ under the null hypothesis.

Decision Rule: Reject H_0 if $T \leq -Z\alpha$.

3. Results and Discussion

3.1 Normal Distribution: Three Populations

The simulation study under the normal distribution considered three populations and three scenarios that isolate or combine the effects of location and scale differences. Because the normal distribution is symmetric with light tails, it provides an ideal baseline for assessing the theoretical properties of each test. Results for the balanced case ($n = 9, 15$, and 30 per group) are supplemented by an unbalanced design with $n_1 = 9$, $n_2 = n_3 = 18$, in which the control group is deliberately undersized relative to the treatment groups. Since results for $k=3$ are similar to results for $k=4$, only the results for $k=4$ are given.

3.1 Normal Distribution: Four Populations

This section extends the analysis to four populations (one control and three treatment groups) under the normal distribution. The four-population setting increases the structural complexity of the simple tree alternative and generally provides more opportunities for the tests to accumulate information about treatment effects, particularly at moderate to large effect sizes.

3.2.1 Normal Distribution with Different Means and Equal Variances (Four Populations)

With four populations and equal variances, the FW test has the highest powers when the means differ. This was also the case when $k=3$ populations. One can see that the powers of the combined test are much greater than the powers of the WW test in this case.

In Table 1, the balanced $n = 15$ design shows that even at the smallest mean spacing considered, $\mu = (0, 0.25, 0.50, 0.75)$ FW achieves 0.4868. The WW test is again the weakest at 0.0964, while the combined test reaches 0.3851. At the larger spacing $\mu = (0, 0.75, 1.00, 1.25)$, $FW = 0.9447$, $WW = 0.3577$, and Combined = 0.8561, a difference of nearly 39 percentage points between FW and WW.

Table 1: Four samples with different means and equal variances ($n = 15$) - Normal distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	FW	WW	Combined
15	15	15	15	0	0.00	0.00	0.00	1	1	1	1	0.0490	0.0397	0.0561
15	15	15	15	0	0.25	0.50	0.75	1	1	1	1	0.4868	0.0964	0.3851
15	15	15	15	0	0.50	0.75	1.00	1	1	1	1	0.7766	0.1923	0.6468
15	15	15	15	0	0.75	1.00	1.25	1	1	1	1	0.9447	0.3577	0.8561
15	15	15	15	0	1.25	1.50	1.75	1	1	1	1	0.9988	0.7825	0.9903
15	15	15	15	0	1.25	1.75	1.75	1	1	1	1	0.9998	0.8178	0.9951
15	15	15	15	0	1.50	1.75	2.00	1	1	1	1	1.0000	0.9083	0.9994

Table 2 introduces the unbalanced design with $n_1 = 9$, $n_2 = n_3 = n_4 = 18$. The power profile qualitatively mirrors that of the balanced case but at lower levels. At $\mu = (0, 0.25, 0.50, 0.75)$, $FW = 0.3711$ versus 0.4868 in the balanced $n = 15$ case — a reduction of about 12 points from losing information in the small control group. The WW test is only modestly affected, with power changing from 0.0964 to 0.1281, suggesting that runs-based detection is less sensitive to the specific size of the control group when the total sample size is fixed. At the largest mean spacing, $\mu = (0, 1.50, 1.75, 2.00)$, the unbalanced design still achieves $FW = 0.9983$, $WW = 0.7631$, and Combined = 0.9862.

Table 2: Four samples with different means and equal variances ($n_1 = 9$, $n_2 = 18$, $n_3 = 18$, $n_4 = 18$) - Normal distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	FW	WW	Combined
9	18	18	18	0	0.00	0.00	0.00	1	1	1	1	0.0524	0.0775	0.0618
9	18	18	18	0	0.25	0.50	0.75	1	1	1	1	0.3711	0.1281	0.2994
9	18	18	18	0	0.50	0.75	1.00	1	1	1	1	0.6253	0.1927	0.5076
9	18	18	18	0	0.75	1.00	1.25	1	1	1	1	0.8439	0.2952	0.7185
9	18	18	18	0	1.25	1.50	1.75	1	1	1	1	0.9903	0.6047	0.9537
9	18	18	18	0	1.25	1.75	1.75	1	1	1	1	0.9943	0.6515	0.9690
9	18	18	18	0	1.50	1.75	2.00	1	1	1	1	0.9983	0.7631	0.9862

3.1.2 Normal Distribution with Equal Means and Different Variances (Four Populations)

Table 3 presents results for the balanced $n = 15$ four-population design. As the variance vector increases from $\sigma = (1, 1.50, 2.00, 2.25)$ to $\sigma = (1, 2.75, 3.25, 3.50)$, WW power climbs from 0.1061 to 0.3841. Notably, the combined test in this setting exhibits a pattern not seen in the three-population case: at $\sigma = (1, 1.50, 2.00, 2.25)$, Combined = 0.1084, which actually slightly exceeds WW = 0.1061. This occurs because the combined test's averaging of FW and WW occasionally benefits from a mild run structure that emerges when four groups have progressively different variances. However, the advantage is marginal and disappears at larger contrasts, where WW = 0.3841 versus Combined = 0.3002 at the largest variance configuration. FW power remains flat at 0.02 to 0.05 throughout, essentially at or below the nominal level.

Table 3: Four samples with equal means and different variances ($n = 15$) - Normal distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	FW	WW	Combined
15	15	15	15	0	0	0	0	1	1.00	1.00	1.00	0.0530	0.0435	0.0619
15	15	15	15	0	0	0	0	1	1.50	2.00	2.25	0.0302	0.1061	0.1084
15	15	15	15	0	0	0	0	1	1.75	2.25	2.50	0.0252	0.1492	0.1334
15	15	15	15	0	0	0	0	1	2.00	2.50	2.75	0.0258	0.2057	0.1724
15	15	15	15	0	0	0	0	1	2.25	2.75	3.00	0.0228	0.2544	0.2106
15	15	15	15	0	0	0	0	1	2.50	3.00	3.25	0.0219	0.3253	0.2538
15	15	15	15	0	0	0	0	1	2.75	3.25	3.50	0.0227	0.3841	0.3002

Table 4 presents the unbalanced design ($n_1 = 9$, $n_2 = n_3 = n_4 = 18$). A striking feature is the behavior of FW power, which falls progressively below the nominal level as variance contrasts increase: from 0.0481 at H_0 to 0.0073 at the largest variance configuration. This sustained sub-nominal FW power in the four-population unbalanced scale scenario occurs because the ranking process is systematically distorted by the different variances across groups of very different sizes. With a small control group and large, high-variance treatment groups, treatment ranks tend to be extreme, and control ranks cluster in the middle, artificially deflating the FW test statistic. The combined test is similarly deflated, reaching only 0.1722 at the largest variance contrast considerably below the 0.3002 seen in the balanced design. WW power is relatively less affected, remaining at 0.3864 compared to 0.3841 in the balanced case, confirming WW's robustness to imbalance under pure scale alternatives.

Table 4: Four samples with equal means and different variances ($n_1 = 9$, $n_2 = 18$, $n_3 = 18$, $n_4 = 18$) - Normal distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	FW	WW	Combined
9	18	18	18	0	0	0	0	1	1.00	1.00	1.00	0.0481	0.0839	0.0571
9	18	18	18	0	0	0	0	1	1.50	2.00	2.25	0.0186	0.1775	0.0744
9	18	18	18	0	0	0	0	1	1.75	2.25	2.50	0.0138	0.2243	0.0903
9	18	18	18	0	0	0	0	1	2.00	2.50	2.75	0.0118	0.2761	0.1170
9	18	18	18	0	0	0	0	1	2.25	2.75	3.00	0.0102	0.3018	0.1307
9	18	18	18	0	0	0	0	1	2.50	3.00	3.25	0.0084	0.3551	0.1479
9	18	18	18	0	0	0	0	1	2.75	3.25	3.50	0.0073	0.3864	0.1722

3.1.3 Normal Distribution with Different Means and Different Variances (Four Populations)

Table 5 presents the balanced $n = 15$ hybrid results. The combined test consistently leads across all six non-null configurations. At $\mu = (0, 0.25, 0.50, 0.75)$, $\sigma = (1, 1.50, 2.00, 2.25)$, the combined test achieves 0.3119, compared to FW = 0.2349 and WW = 0.1122. The gap widens at larger effects: at $\mu = (0, 1.25, 1.50, 1.75)$, $\sigma = (1, 2.25, 2.75, 3.00)$, Combined = 0.8444 versus FW = 0.8114 and WW = 0.3846. At the highest effect level $\mu = (0, 1.50, 1.75, 2.00)$, $\sigma = (1, 2.75, 3.00, 3.25)$. Combined = 0.9048, FW = 0.8444, WW = 0.4858. The combined procedure provides a 6-point gain over FW and a 42-point gain over WW at this configuration, making a compelling case for its use in practical four-group hybrid settings under normality.

Table 6 presents an unbalanced large-sample design with $n_1 = 18$, $n_2 = n_3 = n_4 = 30$. The larger treatment group sizes amplify both FW and Combined power substantially. At $\mu = (0, 1.25, 1.50, 1.75)$, $\sigma = (1, 2.25, 2.75, 3.00)$, FW = 0.9333, and Combined = 0.9047, here FW exceeds Combined by nearly 3 points. This reversal is attributable to the dominant location signal at this configuration. With 30 observations per treatment group, the rank-based FW statistic has tremendous power, and incorporating WW dilutes rather than supplements the signal. At smaller effects $\mu = (0, 0.50, 0.75, 1.00)$, $\sigma = (1,$

1.75, 2.25, 2.50). Combined = 0.5426 edges FW = 0.4935, confirming the combined test's advantage when location and scale contributions are more balanced. These results suggest that, in large samples with strong location effects, FW alone may be slightly better, while the combined test is preferred when both location and scale effects are of moderate magnitude.

Table 5: Four samples with different means and different variances ($n = 15$) - Normal distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	FW	WW	Combined
15	15	15	15	0	0.00	0.00	0.00	1	1.00	1.00	1.00	0.0496	0.0405	0.0572
15	15	15	15	0	0.25	0.50	0.75	1	1.50	2.00	2.25	0.2349	0.1122	0.3119
15	15	15	15	0	0.50	0.75	1.00	1	1.75	2.25	2.50	0.3922	0.1786	0.4911
15	15	15	15	0	0.75	1.00	1.25	1	2.00	2.50	2.75	0.5380	0.2596	0.6357
15	15	15	15	0	1.25	1.50	1.75	1	2.25	2.75	3.00	0.8114	0.3846	0.8444
15	15	15	15	0	1.25	1.75	1.75	1	2.50	3.00	3.00	0.8098	0.4309	0.8724
15	15	15	15	0	1.50	1.75	2.00	1	2.75	3.00	3.25	0.8444	0.4858	0.9048

Table 6: Four samples with different means and different variances ($n_1 = 18$, $n_2 = 30$, $n_3 = 30$, $n_4 = 30$) - Normal distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	FW	WW	Combined
18	30	30	30	0	0.00	0.00	0.00	1	1.00	1.00	1.00	0.0494	0.0596	0.0599
18	30	30	30	0	0.25	0.50	0.75	1	1.50	2.00	2.25	0.2816	0.1553	0.3481
18	30	30	30	0	0.50	0.75	1.00	1	1.75	2.25	2.50	0.4935	0.2230	0.5426
18	30	30	30	0	0.75	1.00	1.25	1	2.00	2.50	2.75	0.6806	0.3168	0.7053
18	30	30	30	0	1.25	1.50	1.75	1	2.25	2.75	3.00	0.9333	0.4380	0.9047
18	30	30	30	0	1.25	1.75	1.75	1	2.50	3.00	3.00	0.9385	0.4973	0.9182
18	30	30	30	0	1.50	1.75	2.00	1	2.75	3.00	3.25	0.9613	0.5419	0.9458

3.2 Results for the T-distribution with 3 degrees of freedom

The results for the T distribution with 3 degrees of freedom, $t(3)$, were similar to the results for the normal distribution for both $k=3$ and $k=4$, as to which test statistic did better in the various cases. The combined test did better when both locations and scales changed. The FW did better with just the locations changed, but the combined test did have a relatively high power in comparison to the FW test in these cases. When just the scales changed, the WW test had the higher powers.

3.3 Results for Exponential Distribution (Three Populations)

This section presents results for three independent samples drawn from the exponential distribution. Because the exponential distribution is inherently right-skewed, with $E[X] = \lambda$ and $\text{Var}[X] = \lambda^2$, the mean and variance are intrinsically linked. Three scenarios are examined: constant variance with changing means, constant mean with changing variances, and both changing simultaneously ($\sigma^2 = \mu^2$). Each test was performed using 10,000 Monte Carlo replications at $\alpha = 0.05$.

Table 7: Mean-Variance Mapping Used in Exponential Simulations

Mean (μ)	Variance (μ^2)
1.000	1.000
1.250	1.563
1.500	2.250
1.750	3.063
2.000	4.000
2.250	5.063
2.500	6.250
2.750	7.563

3.3.1 Exponential with Different Means and Equal Variances (Three Populations)

In the first scenario, the variance was held constant ($\sigma^2 = 1$) while the means differed among groups. Differences in means were created through additive shifts:

$$Y_j = X_j + \delta_j, X_j \sim \text{Exp}(1), E[Y_j] = 1 + \delta_j = \mu_j, \text{Var}[Y_j] = 1.$$

Under H_0 , empirical Type I error rates for all three tests were close to 0.05. As mean differences increased, both FW and the Combined test exhibited steadily increasing power, while WW was less sensitive to mean shifts.

Three power curves were drawn when the underlying distributions were exponential for 3 populations. All the

sample sizes were 9. Figure 1 shows the results of the power curve when the variances were constant and just the means changed. Figure 2 shows the results of the power curves when the variances change and the means were constant. Figure 3 shows the results of the power curves when both the mean and variance change. The results were the same for the other sample sizes considered, both equal and unequal. The combined test had higher powers when just the means changed (Figure 1). The Fligner-Wolfe test had higher powers when both the mean and variance changed (Figure 3). The WW test had higher powers when just the variance changed. (Figure 2).

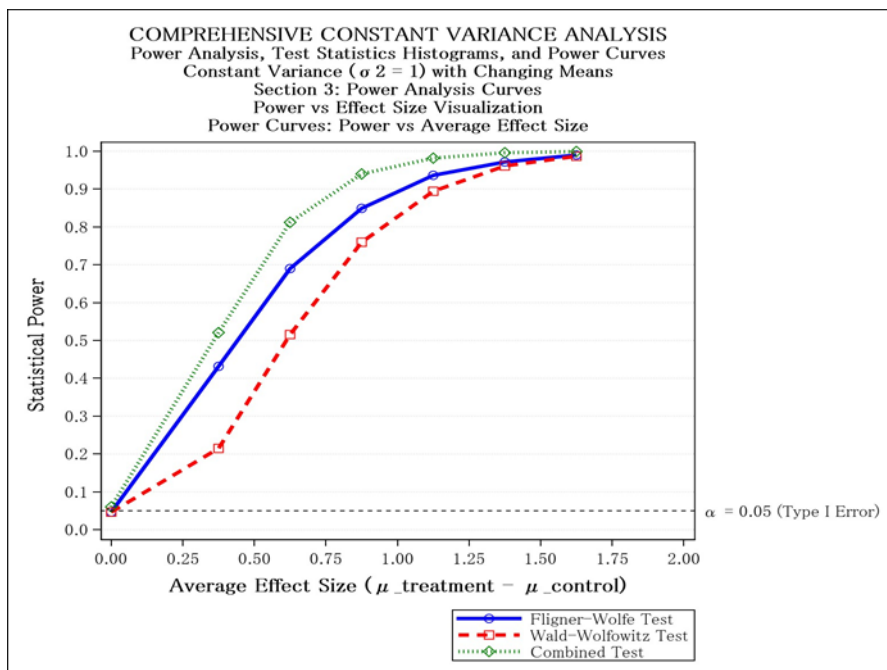


Fig 1: Power Analysis Curves of FW, WW and the Combined test under exponential distribution with same variances and different means (Power vs Average Effect Size) 3 populations

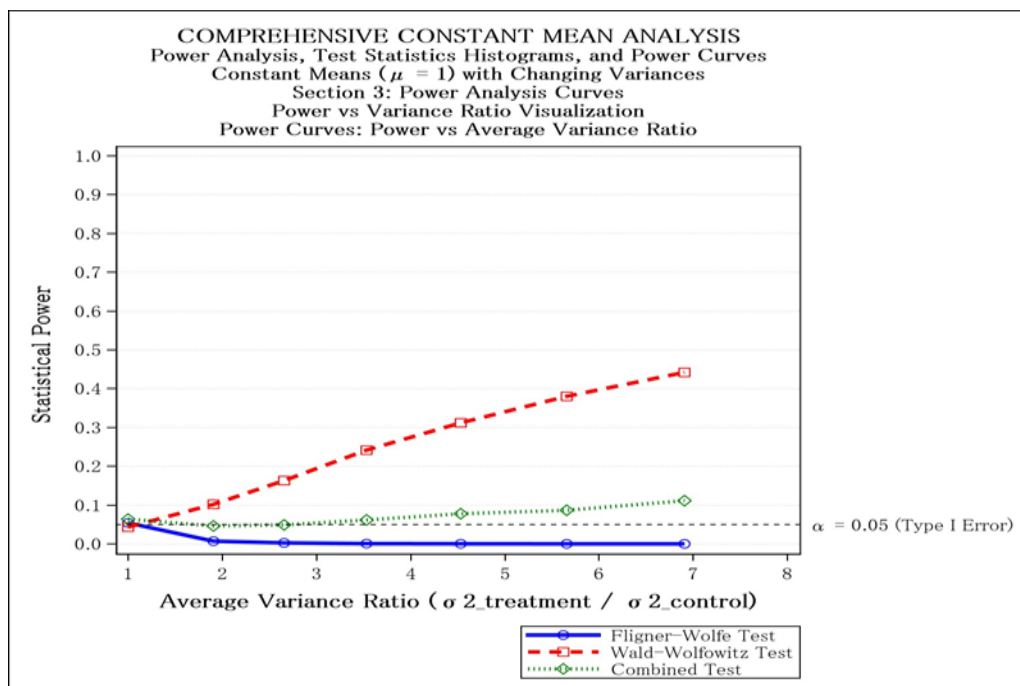


Fig 2: Power Analysis Curves of FW, WW and the Combined test under exponential .Three populations with equal sample sizes of size 9; exponential distributions with same means and different variances (Power vs Average Variance Ratio)

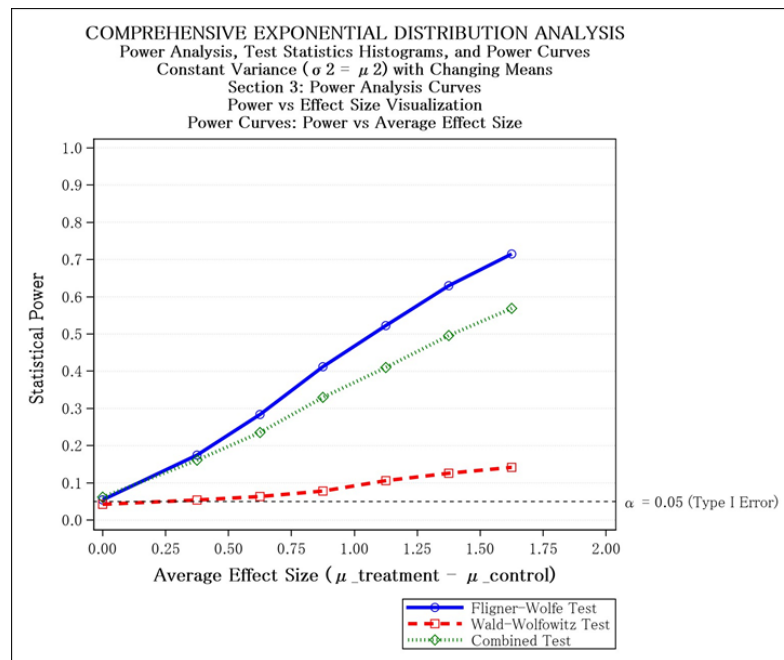


Fig 3: Power Analysis Curves of FW, WW and the Combined test under exponential distribution with different means and different variances (Power vs Average Effect Size) Three populations with equal sample sizes of 9.

3.4 Exponential Distribution: Four Populations

This section extends the analysis of the exponential distribution to four populations. The inclusion of a fourth group amplifies the joint location-scale signal available to each test, particularly under the hybrid scenario where both mean and variance increase monotonically across groups.

3.4.1 Exponential Distribution with Different Means and Equal Variances (Four Populations)

Table 8 shows results for the balanced $n = 15$ design. The combined test leads all three procedures across the entire range of mean spacings, a finding that distinguishes the four-

population exponential setting from the three-population case (Table 14), where Combined also led but by smaller margins. At $\mu = (1.00, 1.25, 1.50, 1.75)$, Combined = 0.742, FW = 0.689, WW = 0.445, a 5-point advantage for Combined over FW and a 30-point advantage over WW. At $\mu = (1.00, 1.50, 1.75, 2.00)$, these numbers grow to Combined = 0.950, FW = 0.905, WW = 0.794. The four-population structure amplifies WW's sensitivity to location shifts (by increasing skew in the pooled distribution), making the WW contribution more informative and, in turn, the combined test more powerful than FW alone.

Table 8: Four samples with different means and equal variances ($n = 15$) - Exponential distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ^2_1	σ^2_2	σ^2_3	σ^2_4	FW	WW	Combined
15	15	15	15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.048	0.053	0.058
15	15	15	15	1.00	1.25	1.50	1.75	1.00	1.00	1.00	1.00	0.689	0.445	0.742
15	15	15	15	1.00	1.50	1.75	2.00	1.00	1.00	1.00	1.00	0.905	0.794	0.950
15	15	15	15	1.00	1.75	2.00	2.25	1.00	1.00	1.00	1.00	0.975	0.934	0.991
15	15	15	15	1.00	2.00	2.25	2.50	1.00	1.00	1.00	1.00	0.995	0.985	0.999
15	15	15	15	1.00	2.25	2.50	2.75	1.00	1.00	1.00	1.00	0.999	0.996	1.000
15	15	15	15	1.00	2.50	2.75	3.00	1.00	1.00	1.00	1.00	1.000	0.999	1.000

Table 9 presents the unbalanced design ($n_1 = 9, n_2 = n_3 = n_4 = 18$). The combined test retains its leading position across all configurations. At $\mu = (1.00, 1.25, 1.50, 1.75)$, the unbalanced design yields FW = 0.578, WW = 0.412, and Combined = 0.665, compared to the balanced values of 0.689, 0.445, and 0.742, a consistent reduction of roughly 8-11 points due to the smaller control group. One notable feature of Table 9 is that

the WW test overtakes FW in the $n_1 = 9$ configuration at very large mean spacings: at $\mu = (1.00, 2.00, 2.25, 2.50)$, WW = 0.973 versus FW = 0.966. This crossover suggests that WW's run-based detection mechanism becomes more efficient when the control group is small, and the treatment group means are far from the control, a finding that could inform adaptive test selection strategies in practice.

Table 9: Four samples with different means and equal variances ($n_1 = 9, n_2 = 18, n_3 = 18, n_4 = 18$) - Exponential distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ^2_1	σ^2_2	σ^2_3	σ^2_4	FW	WW	Combined
9	18	18	18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.045	0.051	0.058
9	18	18	18	1.00	1.25	1.50	1.75	1.00	1.00	1.00	1.00	0.578	0.412	0.665
9	18	18	18	1.00	1.50	1.75	2.00	1.00	1.00	1.00	1.00	0.812	0.689	0.883
9	18	18	18	1.00	1.75	2.00	2.25	1.00	1.00	1.00	1.00	0.923	0.867	0.967
9	18	18	18	1.00	2.00	2.25	2.50	1.00	1.00	1.00	1.00	0.966	0.973	0.993
9	18	18	18	1.00	2.25	2.50	2.75	1.00	1.00	1.00	1.00	0.991	0.988	0.998
9	18	18	18	1.00	2.50	2.75	3.00	1.00	1.00	1.00	1.00	0.998	0.996	1.000

3.4.2 Exponential Distribution with Equal Means and Different Variances (Four Populations)

Table 10 presents the balanced $n = 15$ four-population pure scale results under the exponential distribution. FW power is effectively zero across all configurations. This further confirms FW's complete unsuitability for using to detect differences in location when the variances are not equal, especially in this skewed distribution setting. The alpha values are well below the stated alpha values. WW power rises sharply with the variance contrast, from 0.140 at $\sigma^2 = (1.000, 1.563, 2.250, 3.063)$ to 0.679 at $\sigma^2 = (1.000, 7.563, 9.000, 10.560)$. The combined test shows modest but real gains over FW, reaching 0.199 at the largest variance configuration, driven entirely by the WW component.

Table 10: Four samples with equal means and different variances ($n = 15$) - Exponential distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ^2_1	σ^2_2	σ^2_3	σ^2_4	FW	WW	Combined
15	15	15	15	1.00	1.00	1.00	1.00	1.000	1.563	2.250	3.063	0.002	0.140	0.055
15	15	15	15	1.00	1.00	1.00	1.00	1.000	2.813	3.938	5.063	0.001	0.289	0.089
15	15	15	15	1.00	1.00	1.00	1.00	1.000	4.063	5.625	7.188	0.001	0.423	0.123
15	15	15	15	1.00	1.00	1.00	1.00	1.000	5.313	7.313	9.313	0.001	0.534	0.156
15	15	15	15	1.00	1.00	1.00	1.00	1.000	6.188	8.438	10.69	0.001	0.623	0.178
15	15	15	15	1.00	1.00	1.00	1.00	1.000	6.938	9.438	11.94	0.001	0.678	0.189
15	15	15	15	1.00	1.00	1.00	1.00	1.000	7.563	9.000	10.56	0.001	0.679	0.199

Table 11 presents the unbalanced design ($n_1 = 9, n_2 = n_3 = n_4 = 18$). FW power drops to 0.000 in several rows below what was observed in the balanced case, again reflecting the distorting effect of the small control group on the rank structure. WW power remains substantial: 0.606 at the largest variance contrast, compared to 0.679 in the balanced case, a smaller reduction than might be expected given the reduced control group size. This WW robustness to imbalance in the

four-population scale scenario likely reflects the fact that run patterns are determined by the full ordering of the combined sample rather than just the control-group contribution. The combined test, despite FW's near-zero contribution, provides mild power of 0.130 at the highest contrast, reflecting the WW component's signal being halved by averaging with FW's near-zero value under the combined statistic's $\sqrt{2}$ normalization.

Table 11: Four samples with equal means and different variances ($n_1 = 9, n_2 = 18, n_3 = 18, n_4 = 18$) - Exponential distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ^2_1	σ^2_2	σ^2_3	σ^2_4	FW	WW	Combined
9	18	18	18	1.00	1.00	1.00	1.00	1.000	1.563	2.250	3.063	0.001	0.198	0.046
9	18	18	18	1.00	1.00	1.00	1.00	1.000	2.813	3.938	5.063	0.001	0.312	0.063
9	18	18	18	1.00	1.00	1.00	1.00	1.000	4.063	5.625	7.188	0.000	0.401	0.082
9	18	18	18	1.00	1.00	1.00	1.00	1.000	5.313	7.313	9.313	0.000	0.478	0.098
9	18	18	18	1.00	1.00	1.00	1.00	1.000	6.188	8.438	10.69	0.000	0.545	0.112
9	18	18	18	1.00	1.00	1.00	1.00	1.000	6.938	9.438	11.94	0.000	0.589	0.123
9	18	18	18	1.00	1.00	1.00	1.00	1.000	7.563	9.000	10.56	0.000	0.606	0.130

3.4.3 Exponential Distribution with Different Means and Different Variances (Four Populations)

The final simulation scenario combines exponential changes in mean and variance across four populations, exploiting the natural μ - σ^2 linkage of the exponential distribution.

Table 12 presents the balanced $n = 15$ results for the four-population exponential hybrid scenario. The FW > Combined > WW hierarchy observed in the three-population exponential hybrid (Figure 3) is maintained here with an important difference: the power gap between FW and Combined has widened at large effect sizes. At $\mu = (1.00, 2.50, 2.75, 3.00)$, $\sigma^2 = (1.00, 6.25, 7.56, 9.00)$, FW = 0.904, and Combined = 0.775, a gap of 12.9 percentage points. In the three-population case, the largest gap was 4.7 points. The larger FW-Combined

gap in the four-population case suggests that with more treatment groups, the exponential distribution's skew creates stronger rank-based location signals that FW can exploit more fully. In contrast, the added WW dilution in the combined statistic becomes more costly.

WW power remains the lowest across all configurations, ranging from 0.060 for the smallest effect to 0.229 for the largest. The modest WW power under the exponential hybrid, even with four populations, stands in stark contrast to the WW power of 0.679 under the pure scale exponential scenario (Table 12). This confirms that WW's strength in exponential settings is specifically tied to scale-only differences; when location and scale increase together, the run-based signal is complicated by mean differences that favor FW.

Table 12: Four samples with different means and different variances ($n = 15$) - Exponential distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ^2_1	σ^2_2	σ^2_3	σ^2_4	FW	WW	Combined
15	15	15	15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.048	0.053	0.058
15	15	15	15	1.00	1.25	1.50	1.75	1.00	1.56	2.25	3.06	0.294	0.060	0.239
15	15	15	15	1.00	1.50	1.75	2.00	1.00	2.25	3.06	3.94	0.478	0.089	0.389
15	15	15	15	1.00	1.75	2.00	2.25	1.00	3.06	3.94	4.81	0.645	0.123	0.534
15	15	15	15	1.00	2.00	2.25	2.50	1.00	3.94	4.81	5.69	0.778	0.156	0.656
15	15	15	15	1.00	2.25	2.50	2.75	1.00	4.81	5.69	6.56	0.856	0.189	0.734
15	15	15	15	1.00	2.50	2.75	3.00	1.00	6.25	7.56	9.00	0.904	0.229	0.775

Table 13 presents the unbalanced four-population hybrid results ($n_1 = 9$, $n_2 = n_3 = n_4 = 18$). The FW > Combined > WW ordering persists throughout. Power reductions relative to the balanced case are visible but moderate: at $\mu = (1.00, 2.50, 2.75, 3.00)$, $\sigma^2 = (1.00, 6.25, 7.56, 9.00)$, FW = 0.778 (versus 0.904 balanced), Combined = 0.595 (versus 0.775 balanced), WW = 0.220 (versus 0.229 balanced). The WW test is least affected by the unbalanced design in this scenario,

a pattern consistent across all exponential scenarios, suggesting that runs-based detection is inherently less sensitive to control group size than rank-based detection in skewed distributions. The combined test is substantially reduced in the unbalanced case, primarily because FW power drops more steeply than WW when the control group is small under the hybrid exponential scenario.

Table 13: Four samples with different means and different variances ($n_1 = 9$, $n_2 = 18$, $n_3 = 18$, $n_4 = 18$) - Exponential distribution

n_1	n_2	n_3	n_4	μ_1	μ_2	μ_3	μ_4	σ^2_1	σ^2_2	σ^2_3	σ^2_4	FW	WW	Combined
9	18	18	18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.045	0.051	0.058
9	18	18	18	1.00	1.25	1.50	1.75	1.00	1.56	2.25	3.06	0.229	0.107	0.194
9	18	18	18	1.00	1.50	1.75	2.00	1.00	2.25	3.06	3.94	0.367	0.145	0.301
9	18	18	18	1.00	1.75	2.00	2.25	1.00	3.06	3.94	4.81	0.501	0.178	0.412
9	18	18	18	1.00	2.00	2.25	2.50	1.00	3.94	4.81	5.69	0.623	0.198	0.501
9	18	18	18	1.00	2.25	2.50	2.75	1.00	4.81	5.69	6.56	0.712	0.211	0.556
9	18	18	18	1.00	2.50	2.75	3.00	1.00	6.25	7.56	9.00	0.778	0.220	0.595

4. Conclusion

This paper provides a conclusive summary of the study, synthesizing the key findings from extensive simulation experiments that evaluated the performance of the Fligner-Wolfe (FW) rank test, the Wald-Wolfowitz (WW) runs test, and the proposed combined FW-WW test across various distributional shapes (Normal, Student's t , Exponential), sample sizes ($n = 9$ to $n = 30$), and alternative scenarios (location-shift, scale-shift, and hybrid).

4.1 Summary of Key Findings

4.1.1 Type I Error Control

All three tests — FW, WW, and the Combined test — demonstrated robust control of the Type I error rate, maintaining it close to the nominal α level of 0.05 across all distributional assumptions (normal, heavy-tailed, skewed) and both balanced and unbalanced sample size configurations except for the WW test at times when the sample sizes were unequal with the control group having a sample size of 9. (It is recommended that a sample size of at least 15 be used for the control group). This foundational result confirms the validity of the tests for use in the simple tree alternative framework.

4.1.2 Performance Under Pure Location Alternatives

The FW test was the most powerful procedure when the alternative hypothesis involved purely location-based shifts with homogeneous variances for symmetric populations. The WW test exhibited low to moderate power in this scenario. The Combined test closely tracked the performance of the FW test, consistently and significantly outperforming the WW test, establishing itself as a robust alternative when a location shift is the primary concern. In the case of the exponential distribution, the Combined test had higher powers than the FW test for location shifts and homogeneous variances.

4.1.3 Performance Under Pure Scale Alternatives

The WW test was unequivocally the dominant method for detecting differences in scale when group means were equal among these three tests. The Combined test provided a valuable compromise, achieving intermediate power and capturing a significant portion of the dispersion effect. This was true for both symmetric and asymmetric distributions.

4.1.4 Performance Under Hybrid Alternatives

The hybrid scenario was the setting in which the proposed Combined test demonstrated its greatest utility for symmetric

distributions. Under normality and the Student's t -distribution, the Combined test consistently achieved higher statistical power than either individual test. For instance, in a typical hybrid case with $n = 30$ under normality, where FW power was 0.91 and WW power was 0.75, the Combined test reached a power of 0.97. However, this advantage did not extend to skewed distributions. Under the exponential distribution, the FW test alone demonstrated better power than the Combined test for detecting hybrid alternatives.

4.2 Implications for Research and Practice

This research provides a clear, evidence-based guide for selecting nonparametric tests within the simple tree alternative framework:

- **For ordered location shifts with equal variance:** The Fligner-Wolfe (FW) test is the most powerful for symmetric distributions and the Combined test for the exponential distributions.
- **For detecting variance heterogeneity with equal locations:** The Wald-Wolfowitz (WW) test is the most effective tool.
- **For general-purpose use or when the nature of the alternative is unknown:** The Combined FW-WW test is the recommended procedure under symmetric distributions. It has higher powers when both location and scale changed. Its powers were relatively high for location only changes in relation to the FW test. The powers were in the middle for scale only changes in the symmetric case. The combined test also did well in the asymmetric case. It had high powers for location only changes, relatively high powers for both changes, and powers somewhere in the middle for scale only changes. Its performance ensures high sensitivity across a wide range of scenarios, making it a robust and reliable default option.

References

1. Alsubie A, Magel R. Proposed nonparametric tests using Moses test for location and scale testing. *Journal of Progressive Research in Mathematics*. 2020;16(2):2877-2887.
2. Barlow RE, Brunk HD. The isotonic regression problem and its dual. *Journal of the American Statistical Association*. 1972;67(337):140-147. DOI:10.2307/2284712.

3. Conover WJ. Practical nonparametric statistics. 3rd ed. New York: John Wiley & Sons; 1999.
4. Fligner M, Wolfe D. Distribution-free tests for comparing several treatments with a control. *Statistica Neerlandica*. 1982;36(3):119-127. DOI:10.1111/j.1467-9574.1982.tb00783.x.
5. Gibbons J, Chakraborti S. Nonparametric statistical inference. 5th ed. Boca Raton: Taylor & Francis; 2011.
6. Hettmansperger TP, McKean JW. Robust nonparametric statistical methods. Boca Raton: CRC Press; 2010. DOI:10.1201/b10451.
7. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. 3rd ed. New York: John Wiley & Sons; 2013.
8. Khalawi FA. Proposed nonparametric tests for the simple tree alternative for location and scale testing in a mixed design [dissertation]. Fargo: North Dakota State University; 2022. DOI:10.9734/ajpas/2022/v20i4446.
9. Lepage Y. A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika*. 1971;58(1):213-217. DOI:10.2307/2334333.
10. Magel R, Qin L. A non-parametric test for umbrella alternatives based on ranked-set sampling. *Journal of Applied Statistics*. 2003;30(8):925-937. DOI:10.1080/0266476032000075994.
11. Magel R, Terpstra J, Canonizado K, Park JI. Nonparametric tests for mixed designs. *Communications in Statistics - Simulation and Computation*. 2010;39(6):1228-1250. DOI:10.1080/03610918.2010.490315.
12. Schukr B, Magel R. Nonparametric tests for testing the control versus effects of treatment levels. *International Journal of Statistics and Applied Mathematics*. 2024;9(6):185-190. DOI:10.22271/math.2024.v9.i6c.1916.
13. Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. Boca Raton: Chapman & Hall/CRC; 2003.
14. Wald A, Wolfowitz J. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*. 1940;11(2):147-162. DOI:10.1214/aoms/1177731909.