Received: 3 July 2018 | Revised: 12 July 2019 | Accepted: 21 July 2019

DOI: 10.1002/sim.8343

RESEARCH ARTICLE

Statistics in Medicine WILEY

The nonparametric Behrens-Fisher problem with dependent replicates

Akash Roy[1] | **Solomon W. Harrar**[2] | **Frank Konietschke**[3,4]

[1]Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, Texas

[2]Department of Statistics, University of Kentucky, Lexington, Kentucky

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

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

Correspondence
Frank Konietschke, Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology, Berlin, Germany; or
Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Stra 2, 10178 Berlin, Germany.
Email: Frank.Konietschke@charite.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG KO 4680/3-2

Purely nonparametric methods are developed for general two-sample problems in which each experimental unit may have an individual number of possibly correlated replicates. In particular, equality of the variances, or higher moments, of the distributions of the data is not assumed, even under the null hypothesis of no treatment effect. Thus, a solution for the so-called *nonparametric Behrens-Fisher problem* is proposed for such models. The methods are valid for metric, count, ordered categorical, and even dichotomous data in a unified way. Point estimators of the treatment effects as well as their asymptotic distributions will be studied in detail. For small sample sizes, the distributions of the proposed test statistics are approximated using Satterthwaite-Welch-type *t*-approximations. Extensive simulation studies show favorable performance of the new methods, in particular, in small sample size situations. A real data set illustrates the application of the proposed methods.

KEYWORDS

asymptotics, clustered data, empirical distribution, nonparametric effects, ranks, two-sample problem

1 | INTRODUCTION

Statistical comparisons of two independent groups are one of the most frequently occurring inference problems in scientific research, eg, in biomedical or in social sciences. Many different statistical methods are available for making inferences, eg, *t*-test type statistics for testing the equality of the means of normal samples (assuming equal or unequal variances), χ^2- tests for binary data, Wilcoxon-Mann-Whitney (WMW) tests for testing $H_0^F : F_1 = F_2$, the equality of the two distribution functions of skewed or even ordered categorical data (assuming equal variances or shapes of the distributions) or the Brunner-Munzel tests for testing the hypothesis formulated in terms of the WMW effect

$$H_0 : p = \int F_1 dF_2 = P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2) = \frac{1}{2} \quad (1)$$

(allowing for different variances or shapes of the distributions).[1-5] Here, X_1 and X_2 denote two independent random variables having distribution functions F_1 and F_2, respectively. The correct method to use depends on the shapes of the data distributions and their scales. If $X_i \sim N(\mu_i, \sigma_i^2)$, then $p = \Phi\left(\frac{\mu_2-\mu_1}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$ and, thus, $p = \frac{1}{2}$ if $\mu_1 = \mu_2$ even if the variances σ_1^2 and σ_2^2 are different. Here, $\Phi(x)$ denotes the standard normal cumulative distribution function. Testing the hypothesis $H_0 : p = \frac{1}{2}$ is therefore called the *"nonparametric Behrens-Fisher problem"* because inference methods upon the relative effect p allow for heteroscedastic variances or shapes of the distributions even under the null hypothesis.[2,6,7] Statistical methods, which do not rely on the assumption of equal variances, are especially meaningful when the distribution under the alternative hypothesis of a statistic is important, eg, for the computation of confidence intervals for the effect of interest.

All of these methods, however, are not applicable when measurements are taken with dependent replicates, eg, when visual acuity or any blood parameters of mice sharing the same cage are measured. In all of these scenarios, the replicates (ie, the observations coming from all mice sharing the same cage) should neither be assumed to be independent nor be seen as observations coming from different subjects. Furthermore, using a summary measure (eg, means or medians) of the replicates as a single observation would decrease precision of the effect estimates and thus decrease the powers of the test procedures (see the illustrative simulation results in Section 6). Therefore, there is a need for statistical procedures that allow the specific modeling of the dependent replicates. Under normality assumption of the data, the dependent replicates can be modeled using a linear mixed model and the hypothesis of the equality of the means can be tested using appropriate F-test statistics, eg, using SAS PROC MIXED.[8,9] Replicated binary data can be analyzed using χ^2-square tests for $R \times C$ contingency tables with clustered data.[10,11] Dutta and Datta,[12] Rosner et al (RGL),[13] as well as Datta and Satten (DS)[14] generalized the WMW test to clustered data and their methods can also be used to analyze two independent groups with dependent replicates to test the hypothesis of the equality of the distribution functions $H_0^F : F_1 = F_2$ of the two groups. This formulation of the null hypothesis is rather strict because (1) variances are assumed to be identical under the null hypothesis and (2) the test statistics cannot be inverted into confidence intervals for the WMW effect p given in (1). The computation of confidence intervals, however, is a rather important task in practice and even required in clinical trials by regulatory authorities *"Estimates of treatment effects should be accompanied by confidence intervals, whenever possible..." (ICH E9 Guideline 1998, ch. 5.5, p25)*.[15] The only known available inference methods that can be used for the computation of confidence intervals for the relative effect p are the Brunner-Munzel test and its generalizations.[1,3-5,16] Therefore, it is the aim of the present paper to generalize the applicability of the Brunner-Munzel test to situations in which data is observed with (possibly) dependent replications.

When such data are observed, the numbers of the replications may or may not play an important role for the scientists. Therefore, weighted as well as unweighted versions of the estimators of the treatment effects will be investigated and their asymptotic distributions will be derived in a closed form. The results achieved in this paper generalize the ideas on previous attempts for testing the rather strict hypothesis $H_0 : F_1 = F_2$[7,17] or even for testing $H_0 : p = 1/2$.[7,18-20] In comparison to these pioneering works, differently weighted estimators of the treatment effect p as well as unbiased variance estimators will be proposed in the current paper. Furthermore, major attention will be given to the accuracy of the tests in terms of controlling the nominal type-I error level as well as their powers to detect alternatives when sample sizes are rather small. Here, it will be shown that the distributions of the tests can be approximated using t-distributions with approximated Satterthwaite-Welch degrees of freedom. The degrees of freedom are estimated in such a way that the new methods coincide with the Brunner-Munzel test when single measurements are observed. Recently, Larocque et al[21] developed asymptotic weighted and unweighted tests for the nonparametric Behrens-Fisher problem and proposed two different consistent estimators of the variance of the effect estimator that are either consistent (1) only under the null hypothesis or (2) also even under the alternative. However, extensive simulation studies show that the tests based upon them tend to be either way too conservative or liberal and, therefore, (3) the use of a linear combination of them is recommended by Larocque et al[21] in practical applications. Still and all, the resulting test cannot be inverted into confidence intervals for the underlying effect because the linearly combined variance estimator is only consistent under the null hypothesis. The test procedures proposed by Larocque et al[21] are explained in detail in Section 5.1 and rigorously compared with the new approach in extensive simulation studies.

The remainder of this paper is organized as follows. In Section 2, an example that motivated the research reported in this paper is described. The statistical model and the quantity of inferential interest (nonparametric effect) are formally introduced in Section 3. In Section 4, two estimators for the effect size are given and their asymptotic properties are derived. The theories developed in Section 4 are applied in Section 5 for deriving tests and confidence intervals. The

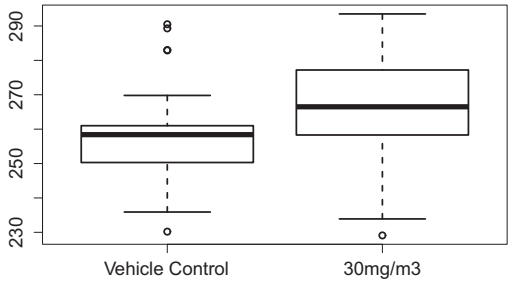

FIGURE 1 Boxplots of the body weights

finite-sample fidelity of the asymptotic theories is evaluated in Section 6 via simulation studies. In addition, in Section 6, the performance of the new methods are evaluated in comparison with existing methods. The analysis of the motivating data using the new methods is carried out in Section 7. Some remarks pertaining to data analysis strategies in light of the new methods are discussed in Section 8. All technical details and proofs are placed in the Appendix.

2 | MOTIVATING EXAMPLE

This research is motivated by a toxicological study involving small sample sizes and different numbers of dependent replicates per unit.

The data is obtained from the National Toxicological Program study number C20536, which investigates the effect of "specular hermatite" on body weights of male HSD rats.* Several rats share the same cage and thus, the cage is seen as the experimental unit with the replications being the body weights of the rats. We consider the two dose groups "vehicle control" and "30 mg/m3" of the active treatment and select the body weights of the rats after four weeks of treatment. In total, $n = 26$ cages are involved in the trial, where the vehicle control group consists of $n_1 = 13$ cages and the remaining $n_2 = 13$ cages are assigned to the active treatment group. We assume that the measurements obtained under the active treatment are independent from those in the vehicle control group. It is very evident from the boxplots displayed in Figure 1 (right) that the medians of the two groups are different. Therefore, it is of major interest to estimate the treatment effect and to test whether there is any significant difference between these two groups along with the computation of a confidence interval. The data also show slightly different variances. Furthermore, since sample sizes are very small, the data should be modeled with a "general" statistical model without restrictive assumptions. Note that the raw data (individual replicates) are displayed in the following boxplots.

Next, a general nonparametric model that allows for arbitrary distributions, different numbers of replicates per unit as well as arbitrary dependency patterns among the replications will be discussed. In particular, no linear relationship between the response variables and the treatment effects is not assumed. This will be explained in the next section.

3 | STATISTICAL MODEL AND HYPOTHESES

We consider two independent samples with replicated observations that can be modeled by independent random vectors

$$\mathbf{X}_{ik} = (X_{ik1}, \ldots, X_{ikm_{ik}})', \; i = 1, 2; k = 1, \ldots, n_i, \tag{2}$$

with distributions $X_{iks} \sim F_i, i = 1, 2$. Here, m_{ik} denotes the number of replicates of subject k under treatment i. The replicates X_{iks} may be arbitrarily correlated. We note that the numbers of replicates may not be under experimental control and may be different for each subject involved in the study. The total number of subjects (units) involved in the study is given by $n = n_1 + n_2$ and the total number of observations is given by $N = m_1 + m_2 = \sum_{i=1}^{2} \sum_{k=1}^{n_i} m_{ik}$. In order to allow for metric, discrete, dichotomous, as well as ordered categorical data in a unified way, we use the normalized version $F_i(x) = \frac{1}{2}[F_i^{(-)}(x) + F_i^{(+)}(x)]$ of the distribution function of X_{iks}, which is the average of the left-continuous, $F_i^{(-)}(x) = P(X_{iks} < x)$, and the right-continuous, $F_i^{(+)}(x) = P(X_{iks} \leq x)$, versions of the distribution function, respectively. The normalized version of the distribution function has first been used by Lèvy[22] and later by other works[20,23-25] to derive asymptotic results for rank statistics including the case of ties. The statistical model considered here does not entail any parameters by

*https://tools.niehs.nih.gov/cebs3/views/index.cfm?action=main.download&bin_id=1600&library_id=4877&fileIdsSelected=1de2c1a657894850015790801 6d60027 accessed on December 16, 2018.

which adequate treatment effects could be described. Therefore, the distributions F_1 and F_2 are used to define a treatment effect by

$$p = \int F_1 dF_2 = P(X_{111} < X_{211}) + \frac{1}{2}P(X_{111} = X_{211}), \tag{3}$$

which is the generalized WMW effect p introduced in (1) to dependent replicates. If $p < \frac{1}{2}$, then the observations coming from distribution F_1 tend to be smaller than those coming from F_2. If $p = \frac{1}{2}$, then the observations coming from these two distributions are expected to be almost similar. Thus, the effect p can be interpreted as a measure of tendency to larger or smaller values. As indicated in the introduction, the effect $p = \frac{1}{2}$ does not imply that the distributions F_1 and F_2 are identical; indeed, inference methods upon p allow for heteroscedastic variances, skewness or other shapes of the distributions even under the null hypothesis $H_0 : p = \frac{1}{2}$. The derivation of appropriate inference methods requires (1) consistent estimation of the effect in model (2) and (2) the computation of the asymptotic distribution of the estimates along with the consistent estimation of its parameter estimates. Unbiased and consistent estimators of p as well as their asymptotic normality will be established in the next section.

4 | POINT ESTIMATORS AND THEIR ASYMPTOTIC DISTRIBUTIONS

When no replicates were observed, the relative effect p can be estimated by plugging-in the empirical versions $\widehat{F}_1(x)$ and $\widehat{F}_2(x)$ of the distribution functions F_1 and F_2 into the integral representation of $p = \int F_1 dF_2$ given in (1). In our situation (2), however, replicates of the measurements per unit may be apparent, and, thus, the traditional estimators of the cumulative distribution functions may not be applicable in this situation. Furthermore, different weighting schemes to incorporate the information from the numbers of replicates may play an important role in the definition of a reasonable effect estimate. Here, we investigate two different versions of the empirical distribution functions and investigate their impact on the interpretation of the resulting estimator as well as their asymptotic behavior in detail. As weighting factors we use the sizes of the clusters and define estimators of the empirical distribution functions in a way that (1) larger clusters add more weight to the estimator than smaller ones and (2) each cluster adds the same weight to the estimator disregarding their sizes. Throughout this paper, the resulting estimators will be called *weighted* and *unweighted* estimators, respectively. Let $c(x) = 0, 1/2, 1$ according as $x < 0, = 0, > 0$ denote the normalized version of count function and consider two different versions of empirical distribution functions

$$\widehat{F}_g^{(u)}(x) = \frac{1}{n_g}\sum_{k=1}^{n_g}\frac{1}{m_{gk}}\sum_{s=1}^{m_{gk}} c(x - X_{gks}) \quad \text{and} \quad \widehat{F}_g^{(w)}(x) = \frac{1}{m_g}\sum_{k=1}^{n_g}\sum_{s=1}^{m_{gk}} c(x - X_{gks}),\ g = 1, 2. \tag{4}$$

Here, both $\widehat{F}_g^{(u)}(x)$ and $\widehat{F}_g^{(w)}(x)$ represent estimators of $F_g^{(c)}(x)$, where the sums of counts within each cluster are first averaged, and then the mean of these averages is computed in $\widehat{F}_g^{(u)}(x)$, while the sums of all counts obtained from all observations are averaged in $\widehat{F}_g^{(w)}(x)$ for all $x \in \mathbb{R}$. Thus, $\widehat{F}_g^{(w)}(x)$ basically is the standard version of the empirical distribution function of a random sample. The impact of these two different weighting versions becomes noticeable when they are plugged-in into the integral representation of p given in (1) to get the unweighted and weighted version of the estimators

$$\widehat{p}^{(u)} = \int \widehat{F}_1^{(u)} d\widehat{F}_2^{(u)} = \frac{1}{n_2}\sum_{k=1}^{n_2}\frac{1}{m_{2k}}\sum_{s=1}^{m_{2ks}}\frac{1}{n_1}\sum_{k'=1}^{n_1}\frac{1}{m_{1k'}}\sum_{s'=1}^{m_{1k's'}} c(X_{2ks} - X_{1k's'}), \quad \text{and} \tag{5}$$

$$\begin{aligned}\widehat{p}^{(w)} &= \int \widehat{F}_1^{(w)} d\widehat{F}_2^{(w)} = \frac{1}{m_1 m_2}\sum_{k=1}^{n_2}\sum_{s=1}^{m_{2ks}}\sum_{k'=1}^{n_1}\sum_{s'=1}^{m_{1k's'}} c(X_{2ks} - X_{1k's'}) \\ &= \frac{1}{N}(\overline{R}_{2\cdot\cdot} - \overline{R}_{1\cdot\cdot}) + \frac{1}{2},\end{aligned} \tag{6}$$

where $\overline{R}_{g\cdot\cdot} = m_g^{-1}\sum_{k=1}^{n_g}\sum_{s=1}^{m_{gk}} R_{gks}$ and R_{gks} is the (mid)rank of X_{gks} among all the N observations.

Both estimators are means of the counts $c(X_{2ks} - X_{1k's'})$; however, $\widehat{p}^{(u)}$ is an unweighted and $\widehat{p}^{(w)}$ is a weighted mean of the normed placements $\widehat{F}_1^{(u)}(X_{2ks})$ and $\widehat{F}_1^{(w)}(X_{2ks})$, respectively. The main difference between these two estimators lies

in their interpretation when being applied to data, ie, subjects/units with larger clusters stack the estimator with more weight ($\widehat{p}^{(w)}$), while every cluster stucks the estimator with the same weight ($\widehat{p}^{(u)}$). The answer to "which estimator to use?" depends on the specific research question and experiment and, therefore, the choice has to be made on a case-by-case basis. Both estimators are identical in case of balanced clusters ($m_{1k} = m_{2k} = M$) and identical to the standard rank-based estimator of the relative treatment effect when single observations were measured ($m_{1k} = m_{2k} = 1$). Note that weighted means play an important role in statistical sciences and the weighted estimator is also often chosen in count data analysis with offset variables.[26]

Both of the estimators $\widehat{p}^{(u)}$ and $\widehat{p}^{(w)}$ are unbiased and strongly consistent for p, if sample sizes (ie, the number of experimental units) are reasonably large. The unbiasedness follows from $E(c(X_{2ks} - X_{1k's'})) = \int F_1 dF_2$, because X_{2ks} and $X_{1k's'}$ are independent. The consistency is outlined in the Appendix. Next, the asymptotic distributions of the estimators will be established. Note that each estimator is a sum of dependent random variables; thus, standard central limit theorems do not apply and their asymptotic normality is not obvious at first hand. Brunner et al[6] derived the asymptotic normality of the estimator in case of single observations by exploring the so-called asymptotic equivalence theorem that introduces sums of independent random variables which have the same asymptotic distribution as the estimator. Here, we will adapt their results to dependent replicates.

Define the unobservable random variables $Y_{1ks} = F_2(X_{1ks})$ and $Y_{2ks} = F_1(X_{2ks})$. Then, under mild conditions on the sample and cluster sizes, it can be shown that

$$\sqrt{n}(\widehat{p}^{(u)} - p) = \sqrt{n}\left\{\overline{Y}^{(u)}_{2\cdot\cdot} - \overline{Y}^{(u)}_{1\cdot\cdot} + (1-2p)\right\} + o_P(1) \tag{7}$$

$$\sqrt{N}(\widehat{p}^{(w)} - p) = \sqrt{N}\left\{\overline{Y}^{(w)}_{2\cdot\cdot} - \overline{Y}^{(w)}_{1\cdot\cdot} + (1-2p)\right\} + o_P(1), \tag{8}$$

where

$$\overline{Y}^{(u)}_{g\cdot\cdot} = \frac{1}{n_g}\sum_{k=1}^{n_g}\overline{Y}_{gk\cdot}, \quad \overline{Y}^{(w)}_{g\cdot\cdot} = \frac{1}{m_g}\sum_{k=1}^{n_g} m_{gk}\overline{Y}_{gk\cdot} \quad \text{and} \quad \overline{Y}_{gk\cdot} = \frac{1}{m_{gk}}\sum_{s=1}^{m_{gk}} Y_{gks}$$

denote unweighted and weighted means of the (unobservable random variables) Y_{gks} for $g = 1, 2$. This means that both quantities $\sqrt{n}(\widehat{p}^{(u)} - p)$ and $\sqrt{N}(\widehat{p}^{(w)} - p)$ have the same distribution as the sums of independent random variables given in the right-hand side of (7) and (8), respectively. However, note that $\overline{Y}_{gk\cdot}$ are independent but not identically distributed. This occurs because each variable $\overline{Y}_{gk\cdot}$ has variance $\sigma^2_{gk} = \mathrm{V}ar(\overline{Y}_{gk\cdot})$. For the derivation of the asymptotic normality, the following assumptions on the sample and replication sizes are necessary, which ensure that variance components of the limiting distributions exist. All of the following assumptions hold for $g = 1, 2$:

A1: $n \to \infty$ such that $\dfrac{n}{n_g} \to \lambda^{(u)}_g$ **A2:** $N \to \infty$ such that $\dfrac{N}{m_g} \to \lambda^{(w)}_g$

A3: $\sigma^{2(u)}_g = \lim\limits_{n\to\infty} n_g^{-1}\sum\limits_{k=1}^{n_g}\sigma^2_{gk} \in (0, \infty)$ **A4:** $\sigma^{2(w)}_g = \lim\limits_{N\to\infty} m_g^{-1}\sum\limits_{k=1}^{n_g} m^2_{gk}\sigma^2_{gk} \in (0, \infty)$

A5: $1 \le m_{ik} \le M_0 < \infty$.

Then, it follows that

$$\sqrt{n}(\widehat{p}^{(u)} - p) \xrightarrow{D} N(0, \sigma^{2(u)}) \text{ under } \mathbf{A1},\ \mathbf{A3} \text{ and } \mathbf{A5},\ \text{and} \tag{9}$$

$$\sqrt{N}(\widehat{p}^{(w)} - p) \xrightarrow{D} N(0, \sigma^{2(w)}) \text{ under } \mathbf{A2},\ \mathbf{A4} \text{ and } \mathbf{A5}, \tag{10}$$

respectively. Here,

$$\sigma^{2(u)} = \lambda^{(u)}_1\sigma^{2(u)}_1 + \lambda^{(u)}_2\sigma^{2(u)}_2 \quad \text{and} \quad \sigma^{2(w)} = \lambda^{(w)}_1\sigma^{2(w)}_1 + \lambda^{(w)}_2\sigma^{2(w)}_2, \tag{11}$$

denote the sums of the variance components, respectively. The variances are, however, unknown and must be estimated in real data applications. Consistent estimators will be developed in the next section.

4.1 | Estimation of the variances

In the previous section, the asymptotic normalities of the quantities $\sqrt{n}(\widehat{p}^{(u)} - p)$ and $\sqrt{N}(\widehat{p}^{(w)} - p)$ have been established. It turns out that limiting distributions of both random variables exist, the variances of which are given by $\sigma^{2(u)}$ and $\sigma^{2(w)}$ defined in (11), respectively. Both of the variances $\sigma^{2(u)}$ and $\sigma^{2(w)}$, however, do not only consist of a sum of two variance

constants, they both are rather a mean of variances. The upcoming arising task is the consistent estimation of these sums of variances in this nonparametric framework. One solution for this problem is first estimating the variances of the asymptotic equivalent sums in (7) and (8) using the unobservable random variables Y_{gks} and in a second step replacing them with observable random variables that are close enough to the Y_{gks} in an appropriate norm. We will first derive estimators for the variance $\sigma_g^{2(u)}$. Computing the variance of the mean $\overline{Y}_{gk\cdot}$ in the right-hand side of (7), we obtain by the independence of $\overline{Y}_{gk\cdot}$ and $\overline{Y}_{gk'\cdot}, k \neq k'$,

$$Var\left(\overline{Y}_{g\cdot\cdot}^{(u)}\right) = \frac{1}{n_g^2}\sum_{k=1}^{n_g} Var(\overline{Y}_{gk\cdot}) = \frac{1}{n_g^2}\sum_{k=1}^{n_g}\sigma_{gk}^{2(u)}.$$

Thus, an unbiased and consistent estimator of $\sigma_g^{2(u)} = \lim_{n\to\infty} n_g Var(\overline{Y}_{g\cdot\cdot}^{(u)})$ is given by the empirical variance

$$\widetilde{\sigma}_g^{2(u)} = \frac{1}{(n_g-1)}\sum_{k=1}^{n_g}\left(\overline{Y}_{gk\cdot} - \overline{Y}_{g\cdot\cdot}^{(u)}\right)^2 \tag{12}$$

for $g = 1, 2$. The variance estimation of the weighted estimator $\widehat{p}^{(w)}$ is a more challenging task. Analogous to the above, computing the variance of $\overline{Y}_{2\cdot\cdot}^{(w)}$ given in (8) yields

$$\mathrm{Var}\left(\overline{Y}_{g\cdot\cdot}^{(w)}\right) = \mathrm{Var}\left(\frac{1}{m_g}\sum_{k=1}^{n_g} m_{gk}\overline{Y}_{gk\cdot}\right) = \frac{1}{m_g^2}\sum_{k=1}^{n_g}\mathrm{Var}(m_{gk}\overline{Y}_{gk\cdot}) = \frac{1}{m_g^2}\sum_{k=1}^{n_g}\mathrm{Var}(Y_{gk\cdot}) = \frac{1}{m_g^2}\sum_{k=1}^{n_g}\sigma_{gk}^{2(w)},$$

where $Y_{gk\cdot} = \sum_{s=1}^{m_{gk}} Y_{gks}$. Thus, the variance components $\sigma_{gk}^{2(w)}$ represent variances of the sums of the variables Y_{gks}. In comparison to the investigations with respect to the variance of the unweighted estimator, here, the variables $Y_{gk\cdot}$ may have a different expectation when cluster sizes are different. Therefore, the variance estimator is derived by considering the squared deviation of $Y_{gk\cdot}$ to its estimated expectation $m_{gk}\overline{Y}_{g\cdot\cdot}^{(w)}$ along with a bias correction. To this end, define the known weight $K_g = m_g^{-1}\sum_{k=1}^{n_g} m_{gk}^2(m_g - 2m_{gk})^{-1}$ for $g = 1, 2$ and consider the estimator

$$\widetilde{\sigma}_g^{2(w)} = \frac{1}{(1+K_g)m_g}\sum_{k=1}^{n_g}\frac{m_g}{m_g - 2m_{gk}}\left(Y_{gk\cdot} - m_{gk}\overline{Y}_{g\cdot\cdot}^{(w)}\right)^2. \tag{13}$$

It is explained in the Appendix that $\widetilde{\sigma}_g^{2(w)}$ is an unbiased and consistent estimator of $\sigma_g^{2(w)}$. Both of the quantities $\widetilde{\sigma}_g^{2(u)}$ and $\widetilde{\sigma}_g^{2(w)}$ given in (12) and (13), are, however, not observable in real data applications. Therefore, the unobservable random variables $Y_{1ks} = F_2(X_{1ks})$ and $Y_{2ks} = F_1(X_{2ks})$ are replaced by the observable random variables

$$Z_{1ks}^{(c)} = \widehat{F}_2^{(c)}(X_{1ks}) \quad \text{and} \quad Z_{2ks}^{(c)} = \widehat{F}_1^{(c)}(X_{2ks}), \quad \text{for} \quad c \in \{u, w\},$$

where $F_g^{(c)}$ denotes the empirical distribution function of sample $g = 1, 2$ defined in (4), respectively. Finally, these variables replace the Y_{gks} used in $\widetilde{\sigma}_g^{2(u)}$ and $\widetilde{\sigma}_g^{2(w)}$, and thus, the estimators become

$$\widehat{\sigma}_g^{2(u)} = \frac{1}{(n_g-1)}\sum_{k=1}^{n_g}\left(\overline{Z}_{gk\cdot}^{(u)} - \overline{Z}_{g\cdot\cdot}^{(u)}\right)^2 \quad \text{and} \quad \widehat{\sigma}_g^{2(w)} = \frac{1}{(1+K_g)m_g}\sum_{k=1}^{n_g}\frac{m_g}{m_g - 2m_{gk}}\left(Z_{gk\cdot}^{(w)} - m_{gk}\overline{Z}_{g\cdot\cdot}^{(w)}\right)^2, \tag{14}$$

respectively. Combining these results, consistent estimators of the limiting variances $\sigma^{2(u)}$ and $\sigma^{2(w)}$ displayed in (11) are given by

$$\widehat{\sigma}^{2(u)} = \frac{n}{n_1}\widehat{\sigma}_1^{2(u)} + \frac{n}{n_2}\widehat{\sigma}_2^{2(u)} \quad \text{and} \quad \widehat{\sigma}^{2(w)} = \frac{N}{m_1}\widehat{\sigma}_1^{2(w)} + \frac{N}{m_2}\widehat{\sigma}_2^{2(w)}. \tag{15}$$

It is shown in the Appendix that both estimators $\widehat{\sigma}_g^{2(u)}$ and $\widehat{\sigma}_g^{2(w)}$ are consistent. Based on the asymptotic distribution of the effect estimators and their consistent variance estimation, test procedures for testing the hypothesis $H_0 : p = 1/2$ as well as confidence intervals for p can be derived. This will be explained in the next section.

5 | TEST STATISTICS

The asymptotic normality of the estimators $\widehat{p}^{(c)}$, where $c \in \{u, w\}$ along with the consistent estimators of their variances, can now be used for the derivation of appropriate test statistics for testing the null hypothesis $H_0 : p = 1/2$. To this end, define the quantities $f^{(u)} = \sqrt{n}$ and $f^{(w)} = \sqrt{N}$ and consider

$$T^{(c)} = f^{(c)} \frac{\widehat{p}^{(c)} - p}{\widehat{\sigma}^{(c)}},\ c \in \{u, w\}, \tag{16}$$

where the superscript (c) refers to the weighted and unweighted estimation approaches, respectively. It follows from the above that the variables $T^{(c)}$ follow, asymptotically, as $f^{(c)} \to \infty$, a standard normal distribution. Thus, under the hypothesis $H_0 : p = 1/2$,

$$T^{(c)} = f^{(c)} \frac{\widehat{p}^{(c)} - 1/2}{\widehat{\sigma}^{(c)}} \xrightarrow{\mathcal{D}} N(0, 1),\ c \in \{u, w\}. \tag{17}$$

For large sample sizes, the null hypothesis $H_0 : p = 1/2$ will be rejected at level α, if $|T^{(c)}| \geq z_{1-\alpha/2}$. One-sided test results can be achieved in the obvious way. Extensive simulation studies show, however, that the test tends to be liberal and to over reject the hypothesis when sample sizes are rather small. In order to provide an approximate version of the tests that control the nominal type-I error rate in small sample size situations, the idea from the work of Brunner et al[6] motivates us to approximate the distribution of $T^{(c)}$ by a central t_ν-distribution and estimate its approximate degree of freedom using Satterthwaite-Welch equations. This type of approximation is also known as Box-type approximation.[27] The problem that arises here is that each of the variables $\overline{Y}_{gk\cdot}$ represents a mean of the variables Y_{gks}; thus, each variable may have a different variance σ^2_{gk}. Computing the variance of the variance estimators $\widehat{\sigma}^{2(c)}_g$ involves sums of σ^4_{gk} and σ^3_{gk}, quantities rather difficult to estimate in this setup due to overfitting issues. Therefore, we define approximate degrees of freedom of the resulting t_ν-distribution such that the methods coincide with the Brunner-Munzel test when cluster sizes are equal to 1 and are given by

$$\nu^{(u)} = \frac{\left(\widehat{\sigma}^{2(u)}_1/n_1 + \widehat{\sigma}^{2(u)}_2/n_2\right)^2}{\widehat{\sigma}^{4(u)}_1/\left(n_1^2(n_1-1)\right) + \widehat{\sigma}^{4(u)}_2/\left(n_2^2(n_2-1)\right)} \quad \text{and} \quad \nu^{(w)} = \frac{\left(\widehat{\sigma}^{2(w)}_1/m_1 + \widehat{\sigma}^{2(w)}_2/m_2\right)^2}{\widehat{\sigma}^{4(w)}_1/\left(m_1^2(n_1-1)\right) + \widehat{\sigma}^{4(w)}_2/\left(m_2^2(n_2-1)\right)}. \tag{18}$$

For small sample sizes, the null hypothesis $H_0 : p = 1/2$ will be rejected at level α, if

$$|T^{(c)}| \geq t_{1-\alpha/2}(\nu^{(c)}),\ c \in \{u, w\}, \tag{19}$$

where $t_{1-\alpha/2}(\nu^{(c)})$ denotes the $(1-\alpha/2)$-quantile from the central $t_{\nu^{(c)}}$-distribution with estimated degree of freedom $\nu^{(c)}$ given in (18). Approximate $(1-\alpha)$-confidence intervals for p are given by

$$CI^{(c)} = \left[\widehat{p}^{(c)} - \frac{t_{\nu^{(c)},1-\alpha/2}}{\sqrt{f^{(c)}}}\widehat{\sigma}^{(c)};\ \widehat{p}^{(c)} + \frac{t_{\nu^{(c)},1-\alpha/2}}{\sqrt{f^{(c)}}}\widehat{\sigma}^{(c)}\right], c \in \{u, w\}. \tag{20}$$

One-sided confidence intervals and tests can be computed in the usual way by using $(1-\alpha)$-quantiles and setting the lower or upper bound of the confidence intervals to 0 or 1, depending on the direction. We note that $\nu^{(c)} \to \infty$ as $f^{(c)} \to \infty$ and, therefore, the approximation is asymptotically correct. Furthermore, $\nu^{(u)}$ and $\nu^{(w)}$ are identical when clusters are equally sized (ie, $m_{gk} \equiv M$). Furthermore, both of $\nu^{(u)}$ and $\nu^{(w)}$ are identical to the Brunner-Munzel degree of freedom when $m_{gk} \equiv 1$.

Remark 1. When the numbers of replicates of any unit is way larger than those of the others, it may happen that the weighted variance estimator $\widehat{\sigma}^{2(w)}_g$ given in (14) becomes negative and, thus, the test statistics $T^{(w)}$ cannot be computed. In this case, we propose to replace $\widehat{\sigma}^{2(w)}_g$ by the asymptotically unbiased version

$$\widehat{\tau}^{2(w)}_g = \frac{1}{(n_g - 1)} \sum_{k=1}^{n_g} \left(\frac{1}{m_{gk}} Z^{(w)}_{gk\cdot} - \overline{Z}^{(w)}_{g\cdot\cdot}\right)^2 \tag{21}$$

in $T^{(w)}$ and in the confidence intervals given in (20). Throughout this manuscript, the resulting test will be denoted as *TW*.

Remark 2. Note that the confidence intervals $CI^{(c)}$ as given in (20) may not necessarily be range-preserving, ie, the lower bound may be small than zero and/or the upper bound may be larger than one. Range-preserving confidence intervals for the effects p can be derived using the delta method and an appropriate transformation, eg, the $logit(x) = \log(x/(1-x))$ or $probit(x) = \Phi^{-1}(x)$ transformation function. For example, the logit-type confidence intervals for p are given by

$$\begin{aligned} CI^{(c)}_{\text{Logit}} &= \left[\text{expit}\left(CI^{(c)}_L\right), \text{expit}\left(CI^{(c)}_U\right)\right] \subseteq [0,1], \text{ where} \\ CI^{(c)}_L &= \text{logit}(\widehat{p}^{(c)}) - \frac{z_{1-\alpha/2}}{\sqrt{f^{(c)}}} \frac{\widehat{\sigma}^{(c)}}{\widehat{p}^{(c)}(1-\widehat{p}^{(c)})} \text{ and} \\ CI^{(c)}_U &= \text{logit}(\widehat{p}^{(c)}) + \frac{z_{1-\alpha/2}}{\sqrt{f^{(c)}}} \frac{\widehat{\sigma}^{(c)}}{\widehat{p}^{(c)}(1-\widehat{p}^{(c)})}. \end{aligned}$$

Here, $\text{expit}(y) = \exp(y)/(1+\exp(y))$ denote the inverse of the logit function.

The quality of the proposed tests in terms of controlling the nominal type-I error rate α and their powers to detect alternatives will be investigated in extensive simulation studies in the next section.

5.1 | Approach from the work of Larocque et al

Recently, Larocque et al[21] proposed solutions for the nonparametric Behrens-Fisher problem with clustered data, where clusters may contain observations from each group, respectively. Their methods are also valid in our model (2) and shall be briefly explained as follows: The methods are intended to test under the null hypothesis $H_0^{(L)} : E(s(X_{111} - X_{211})) = 0$, where $s(x)$ denotes the sign function, respectively. Note that testing $H_0^{(L)}$ is equivalent to testing $H_0^{(L)} : P(X_{111} < X_{211}) = P(X_{111} > X_{211})$, and thus, basically identical to testing $H_0 : p = 1/2$. In order to estimate the treatment effect $E(s(X_{111} - X_{211}))$, define the variables

$$S_{ik} = \sum_{s=1}^{m_{1k}} \sum_{l=1}^{m_{2k}} s(X_{1ks} - X_{2kl}),\ i = 1,2, k = 1, \ldots, n_i, \text{and } S = \frac{1}{N_1 N_2} \sum_{k=1}^{n_1} \sum_{k'=1}^{n_2} w_{ik} S_{ik},$$

where w_{ik} are nonnegative weights associated with S_{ik}. For the computation of the estimator of $Var(S)$, let

$$S^0_{ik} = w_{ik} S_{ik} \quad \text{and} \quad S^1_{ik} = w_{ik}(S_{ik} - m_{1k} m_{2k} S) \tag{22}$$

denote the noncentered (S^0_{ik}) and centered (S^1_{ik}) versions of the sums of signs S_{ik} given above. For an easy representation of the quite involved computation of the variance estimator, let $\mathbf{S} = (S^h_{ij})_{n_1 \times n_2}$ denote the matrices of the S^h_{ik} for $h = 0, 1$, respectively. Note that, in the third term of the variance estimator in the work of Larocque et al,[21p759] one of w_{ik} or w_{ri} must be zero in our setting. To see this, suppose i is an index value in the first group. Then $w_{ri} = 0$ for any r because $m_{i2} = 0$. Therefore, the expression of the variance estimator in the work of the aforementioned authors[21p759] can be written as

$$\begin{aligned} \widehat{\sigma}^2_{S_h} &= \frac{N}{(N_1 N_2)^2} \left[\text{Vec}(\mathbf{S}')' \left\{\mathbf{I}_{n_1} \otimes (\mathbf{J}_{n_2} - \mathbf{I}_{n_2})\right\} \text{Vec}(\mathbf{S}') + \text{Vec}(\mathbf{S})' \left\{\mathbf{I}_{n_2} \otimes (\mathbf{J}_{n_1} - \mathbf{I}_{n_1})\right\} \text{Vec}(\mathbf{S})\right] \\ &= \frac{N}{(N_1 N_2)^2} \left[\text{Vec}(\mathbf{S})' \left\{(\mathbf{J}_{n_2} - \mathbf{I}_{n_2}) \otimes \mathbf{I}_{n_1}\right\} \text{Vec}(\mathbf{S}) + \text{Vec}(\mathbf{S})' \left\{\mathbf{I}_{n_2} \otimes (\mathbf{J}_{n_1} - \mathbf{I}_{n_1})\right\} \text{Vec}(\mathbf{S})\right] \\ &= \frac{N}{(N_1 N_2)^2} \left[\text{tr}\left\{\mathbf{S}(\mathbf{J}_{n_2} - \mathbf{I}_{n_2})\mathbf{S}'\right\} + \text{tr}(\mathbf{S}'(\mathbf{J}_{n_1} - \mathbf{I}_{n_1})\mathbf{S}\right], \end{aligned}$$

where Vec(·) denotes the vector operator that stacks the columns of a matrix on top of each other and tr(·) denotes the trace of a matrix, respectively. Based on the previous calculations, Larocque et al[21] proposed three different estimators of

the variance $Var(S)$ as follows:

- $\widehat{\sigma}^2_{S_0}$ that uses S^0_{ik} given in (22) and is only consistent under the null hypothesis,
- $\widehat{\sigma}^2_{S_1}$ that uses S^1_{ik} given in (22) and is also consistent under the alternative hypothesis, and
- $\widehat{\sigma^2_S} = \frac{2}{3}\widehat{\sigma}^2_{S_0} + \frac{1}{3}\widehat{\sigma}^2_{S_1}$ as a linear combination of $\widehat{\sigma}^2_{S_0}$ and $\widehat{\sigma}^2_{S_1}$.

These three different consistent variance estimators lead to three different versions of test statistics for testing $H_0^{(L)}$ in

$$T_L^{(0)} = \sqrt{N}\frac{S}{\widehat{\sigma}_{S_0}}, \quad T_L^{(1)} = \sqrt{N}\frac{S}{\widehat{\sigma}_{S_1}}, \quad \text{and} \quad T_L^{(F)} = \sqrt{N}\frac{S}{\widehat{\sigma}_S}, \tag{23}$$

respectively. Under the null hypothesis $H_0^{(L)}$, all three versions have a standard normal distribution. Note that only the test statistic $T_L^{(1)}$ can be inverted into a confidence interval for the treatment effect, because the variance estimator used ($\widehat{\sigma}^2_{S_1}$) is consistent under the alternative hypothesis. However, all of the three versions of the tests will be used as competing procedures in the simulation studies. Those will be explained and discussed in detail in the next section.

6 | SIMULATIONS

All of the methods proposed in the previous sections are valid for large sample sizes. The arising questions are (1) "How accurate do they control the nominal type-I error rate under the null hypothesis?" and (2) "How much power do the procedures have to detect alternatives when sample sizes are small?" Extensive simulation studies were conducted to find answers to these questions in different scenarios involving very small and moderate sample sizes in balanced and unbalanced situations with different settings for the numbers of replications. Throughout the simulations, a two-sample design $\mathbf{X}_{ik} = (X_{1k1}, \ldots, X_{ikm_{ik}})', i = 1, 2; k = 1, \ldots, n_i$, was simulated with sample sizes $n_1, n_2 \in \{7, 10, 20\}$ and cluster sizes

- Setting 1: $m_{ik} = 1$,
- Setting 2: $m_{ik} = 2$,
- Setting 3: m_{ik} are realizations of independent $Binomial(4, 0.6) + 1$ variables, and
- Setting 4: m_{ik} are realizations of independent $Binomial(10, 0.3) + 1$ variables.

Thus, single observations are modeled in Setting 1 in which the new procedures $T^{(u)}$, $T^{(w)}$ and TW are all equivalent to the Brunner-Munzel test T_{BM}, equally sized numbers of replications are covered in Setting 2 where the weighted and unweighted estimators are identical but different to the Brunner-Munzel test, and different replication sizes are investigated in Settings 3 and 4 with sizes $m_{ik} \in \{1, \ldots, 4\}$ in Setting 3 and $m_{ik} \in \{1, \ldots, 10\}$ in Setting 4, respectively. To investigate the impact of the shape of data distributions on the quality of the procedures, three different types of distributions are considered in the simulation studies, namely,

- Multivariate normal: $\mathbf{X}_{ik} \sim N(\mathbf{0}, \mathbf{\Sigma}_{ik})$ with covariance matrix $\mathbf{\Sigma}_{ik} = \mathbf{I}_{m_{ik}\times m_{ik}}\sigma_i^2 + \rho(\mathbf{J}_{m_{ik}\times m_{ik}} - \mathbf{I}_{m_{ik}\times m_{ik}})$. Homoscedastic ($\sigma_1^2 = \sigma_2^2$) as well as two different heteroscedastic scenarios with $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$ and $\sigma_1^2 = 1$ and $\sigma_2^2 = 3$ with correlation values $\rho \in \{0, 0.5, 0.9\}$ were investigated. Thus, both positive (the larger sample has the larger variance) and negative (the larger sample has the smaller variance) pairing situations are covered within these settings;
- Multivariate lognormal: $\mathbf{X}_{ik} = (X_{1k1}, \ldots, X_{ikm_{ik}})'$ where $X_{iks} = \exp(Y_{iks})$ and $\mathbf{Y}_{ik} = (Y_{1k1}, \ldots, Y_{ikm_{ik}})' \sim N(\mathbf{0}, \mathbf{\Sigma}_{ik})$. Here, the covariance matrices were chosen from $\mathbf{\Sigma}_{ik} = \mathbf{I}_{m_{ik}\times m_{ik}} + \rho(\mathbf{J}_{m_{ik}\times m_{ik}} - \mathbf{I}_{m_{ik}\times m_{ik}})$ with correlation values $\rho \in \{0, 0.5, 0.9\}$ (note that the actual correlation coefficients of the resulting lognormal distributions are different[28]);
- Ordinal data: $\mathbf{X}_{ik} = (X_{1k1}, \ldots, X_{ikm_{ik}})'$ where $X_{iks} = [Y_{iks}]$ and $\mathbf{Y}_{ik} = (Y_{1k1}, \ldots, Y_{ikm_{ik}})' \sim N(\mathbf{0}, \mathbf{\Sigma}_{ik})$. Here, the covariance matrices were chosen from $\mathbf{\Sigma}_{ik} = \mathbf{I}_{m_{ik}\times m_{ik}} + \rho(\mathbf{J}_{m_{ik}\times m_{ik}} - \mathbf{I}_{m_{ik}\times m_{ik}})$ with correlation values $\rho \in \{0, 0.5, 0.9\}$ and the symbol $[\cdot]$ represents the rounding operator, respectively.

All simulations were conducted using R computational environment version 3.4.3 (www.r-project.org) each with $nsim = 10\,000$ simulation runs. Throughout the simulations, the newly developed tests $T^{(u)}$, $T^{(w)}$, and TW proposed in (16) and (21) were implemented using the corresponding $t_{\nu^{(c)}}$-approximation proposed in (19). They were compared with the methods $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ proposed by Larocque et al[21] given in (23). Another competitor is the Brunner-Munzel test T_{BM} for the application of which the first observation X_{ik1} within each cluster $\mathbf{X}_{ik}, i = 1, 2; k = 1, \ldots, n_i$, was used. The aim of simulating the Brunner-Munzel test is exploring its difference to the Larocque test $T_L^{(F)}$when single observations

(see Setting 1) were observed as well as investigating if the new methods increase its power when replicates were observed. Furthermore, we simulated the behavior of the rank-based methods for testing the hypothesis $H_0^F : F_1 = F_2$ with clustered data proposed by RGL[13] and DS.[14] These tests were computed using the *clusrank R*- package.[29] The simulation results are summarized for all the four settings $1 - 4$, the three different correlation values, and sample size configurations for each of the four different data distributions separately. Type-I errors are displayed by multiplying a factor 100 for the ease of visualization.

The type-I error simulation results using homogeneous normal distributions are displayed in Table 1, for heteroscedastic normal distributions having variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$ are displayed in Table 2, for heteroscedastic normal distributions where the two different groups have variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 3$ are displayed in Table 3, results for lognormally distributed data are given in Table 4 and empirical type-I error rates for ordered categorical data are displayed in Table 5, respectively. First, it can be seen from Tables 1 to 5 that the methods $T_L^{(0)}$ and $T_L^{(1)}$ given in (23) tend to be either way too conservative or too liberal throughout all of the investigated settings and data distributions. The test procedure $T_L^{(F)}$ tends to control the type-I error level conservatively when sample sizes are rather small. The conservatism decreases when sample sizes n_1 and n_2 get larger. The WMW-type test statistics tend to control the nominal type-I error rate reasonably well, even under heteroscedasticity. When sample sizes are unbalanced and variances are heteroscedastic, the methods tend to be slightly liberal or conservative, depending on size and variance allocations. However, neither the methods proposed by Larocque et al[21] nor the WMW-type tests can be inverted into confidence intervals for the effect p. The newly developed methods $T^{(w)}$ and $T^{(u)}$ tend to control the nominal type-I error reasonably well in all the investigated scenarios. When sample sizes are very small and correlation is very high the tests tend to be slightly liberal. The liberality decreases with increasing sample sizes. Furthermore, it can be seen from Tables 1 to 5 that the weighting scheme of the estimators does not impact the behavior of the tests. Sometimes, $T^{(w)}$ tends to be slightly more liberal than $T^{(u)}$. This occurs because the standard error (SE) of $\widehat{p}^{(w)}$ is usually a bit "harder" to estimate as variances of sums are estimated rather than variances of means. In all of the investigated scenarios, the weighted variance estimators $\widehat{\sigma}_i^{2(w)}$ did not become negative and all tests $T^{(w)}$ could be computed. However, to investigate the behavior of TW motivated in (21), the test was simulated in all scenarios. It turns out that the test controls the nominal size well and sometimes even better than $T^{(w)}$. However, when sample sizes are small and unbalanced, a conservative behavior of the test may become apparent. Summarizing the findings discussed above, the new methods seem to be pretty accurate and their usage is recommended when the sample sizes $n_i \geq 7$. In case of extreme small samples, an accurate control of the type-I error rate using asymptotic normal or t-quantiles for these rank-statistics cannot be expected. Simulation studies using negatively correlated data show very similar results to the reported above and are therefore omitted.

Next, the powers of all the methods was simulated to detect the alternative $H_1 : p \neq \frac{1}{2}$. Two different balanced designs were simulated, namely,

- **Design 1:** $\mathbf{X}_{ik} \sim N(\mu_i, \mathbf{\Sigma}_{ik})$ with $\mathbf{\Sigma}_{ik} = \mathbf{I}_{2\times 2} + \rho(\mathbf{J}_{2\times 2} - \mathbf{I}_{2\times 2})$ and $\rho \in \{-0.8, 0.8\}$. Here, $\mu_1 = \mathbf{0}$ and $\mu_2 = (\mu_1, \mu_2)'$ with $\mu_\ell = \Phi^{-1}(p)\sqrt{2}$ for various values of $p \in \{0.5, \ldots, 0.95\}$. Thus, the numbers of replicates $m_{ik} \equiv 2$ for all units and data have a predefined relative effect of p.
- **Design 2:** $\mathbf{X}_{ik} \sim N(\mu_i, \mathbf{\Sigma}_{ik})$ with $\mathbf{\Sigma}_{ik} = \mathbf{I}_{m_{ik}\times m_{ik}} + \rho(\mathbf{J}_{m_{ik}\times m_{ik}} - \mathbf{I}_{m_{ik}\times m_{ik}})$ and $\rho \in \{-0.8, 0.8\}$. Here, $\mu_1 = \mathbf{0}$ and $\mu_2 = (\mu_1, \ldots, \mu_{m_{2k}})'$ with $\mu_\ell = \Phi^{-1}(p)\sqrt{2}$ for various values of $p \in \{0.5, \ldots, 0.95\}$. The numbers of replications m_{ik} are realizations of independent $Binomial(4, 0.6) + 1$ variables. Thus, Design 2 represents a setting with different numbers of replications per unit. Note that the covariance matrix might be singular when $\rho = -0.8$.

The sample sizes were chosen to be moderately large ($n_i = 20$) for both of the designs. The power simulation results (multiplied by a factor 100) are displayed in Tables 6 and 7 with $\alpha = 5\%$. First, it can be readily seen that all of the methods that take the replications into account have a higher power than the Brunner-Munzel test. It should, however, be noted that the comparisons have to be confined to the newly proposed methods and the methods of Larocque et al,[21] ie, $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$. More specifically, RGL and DS are designed for the null hypothesis $H_0 : F_1 = F_2$. Hence, they should theoretically be sensitive to heteroscedastic settings as those setting are alternative points for the two tests. Therefore, strictly speaking, these two tests have low powers. When comparing the new method with that of the work of Larocque et al,[21] the power simulation (especially Design 2 in Table 7 as follows) clearly shows the advantage of the new methods when the within-cluster correlation is negative. Even for positive correlation, the new methods have slightly better power compared to the work of Larocque et al.[21] Furthermore, the type of weighting slightly impacts the powers of the methods $T^{(w)}$ or $T^{(u)}$ and seems to depend on the correlations within the clusters. Comparing the two weighted tests TW and $T^{(w)}$, it seems that the power of $T^{(w)}$ is slightly lower in some scenarios. This result may occur because the used

TABLE 1 Type-I error simulations ($\alpha = 5\%$) using homogeneous multivariate normal distributions for the repeated measurements in both groups. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

n1	n2	rho1	$T^{(u)}$	TW	$T^{(w)}$	T_{BM}	$T_L^{(0)}$	$T_L^{(1)}$	$T_L^{(F)}$	RGL	DS	Setting
7	7	0	5.5	5.5	5.5	5.5	5	18.0	5.5	5.2	5.2	1
7	7	0	5.3	5.3	5.3	5.5	0	7.1	1.5	4.7	5.7	2
7	7	0	4.9	4.7	5.1	5.7	1.3	3.8	7	4.6	5.4	3
7	7	0	4.4	3.9	4.8	5.2	7.8	5.7	3.5	3.8	5.0	4
7	7	0.5	5.9	5.9	5.9	5.9	0.5	18.0	5.9	5.3	5.3	1
7	7	0.5	6.1	6.1	6.1	5.7	0.2	11.6	3.4	4.8	5.7	2
7	7	0.5	5.9	5.7	6.1	6.0	0.1	6.7	1.7	4.4	5.4	3
7	7	0.5	6.1	6.2	6.7	5.8	0.1	4.6	1.1	3.9	5.7	4
7	7	0.9	6.0	6.0	6.0	6.0	0.7	18.0	6.1	5.5	5.5	1
7	7	0.9	5.5	5.5	5.5	5.4	0.5	8.3	2.5	4.4	5.3	2
7	7	0.9	6.2	6.8	7.5	6.2	0.5	5.4	1.5	4.5	5.9	3
7	7	0.9	6.4	6.8	7.8	6.0	0.4	4.6	1.6	3.4	5.7	4
10	10	0	6.0	6.0	6.0	6.0	2.5	13.9	6.2	5.7	5.7	1
10	10	0	5.2	5.2	5.2	5.5	1.2	5.8	1.9	4.7	5.4	2
10	10	0	5.1	4.4	5.1	5.2	3.8	4.3	1.9	4.6	5.4	3
10	10	0	5.1	3.3	5.1	5.5	6.8	6.0	3.7	4.0	5.2	4
10	10	0.5	5.4	5.4	5.4	5.4	2.3	13.3	5.5	5.1	5.1	1
10	10	0.5	5.3	5.3	5.3	5.0	1.7	8.8	3.8	4.5	5.1	2
10	10	0.5	5.2	5.6	6.0	5.6	1.5	6.4	3.0	4.4	5.1	3
10	10	0.5	5.8	6.0	6.4	5.7	1.5	5.4	2.6	4.4	5.7	4
10	10	0.9	5.4	5.4	5.4	5.4	2.1	13.4	5.6	4.9	4.9	1
10	10	0.9	5.5	5.5	5.5	5.5	2.1	7.1	3.7	4.7	5.3	2
10	10	0.9	5.9	6.2	6.3	5.5	1.8	5.3	2.9	4.6	5.6	3
10	10	0.9	5.5	6.4	6.8	5.1	1.9	3.7	2.6	4.0	5.2	4
10	20	0	5.9	5.9	5.9	5.9	3.2	11.1	5.9	5.0	5.5	1
10	20	0	4.9	4.9	4.9	5.7	2.3	5.7	2.3	4.6	4.8	2
10	20	0	5.1	2.8	5.2	5.0	3.9	5.2	2.4	4.4	4.8	3
10	20	0	5.1	3.7	4.9	5.8	4.9	4.5	2.7	4.6	5.0	4
10	20	0.5	5.5	5.5	5.5	5.5	3.2	11.0	5.6	4.7	5.1	1
10	20	0.5	5.4	5.4	5.4	5.4	2.6	8.1	4.3	4.8	5.1	2
10	20	0.5	5.9	6.0	6.1	5.7	2.8	6.4	4.0	5.2	5.4	3
10	20	0.5	5.6	5.9	6.0	5.4	2.8	4.8	3.5	4.6	5.2	4
10	20	0.9	5.8	5.8	5.8	5.8	3.2	11.2	5.8	5.1	5.4	1
10	20	0.9	5.4	5.4	5.4	5.4	2.9	6.4	4.0	4.6	5.0	2
10	20	0.9	5.3	6.0	6.0	5.1	2.6	4.6	3.3	4.6	4.7	3
10	20	0.9	5.6	6.1	5.9	5.3	2.8	4.4	3.4	4.8	5.3	4
20	10	0	5.8	5.8	5.8	5.8	3.0	11.1	5.9	5.2	5.3	1
20	10	0	5.5	5.5	5.5	5.5	2.5	5.6	2.7	5.0	5.3	2
20	10	0	5.2	4.4	5.2	5.8	3.1	4.2	2.1	4.6	5.1	3
20	10	0	5.0	4.4	5.0	5.6	3.1	3.8	1.9	4.6	5.0	4
20	10	0.5	5.4	5.4	5.4	5.4	2.8	10.6	5.4	4.6	5.0	1
20	10	0.5	5.5	5.5	5.5	5.4	2.6	8.3	4.4	4.8	5.2	2
20	10	0.5	5.6	5.6	5.7	5.6	2.7	5.5	3.6	4.6	5.2	3
20	10	0.5	5.9	6.0	6.1	5.7	2.7	5.5	3.5	4.4	5.6	4
20	10	0.9	5.5	5.5	5.5	5.5	3.3	11.2	5.6	4.9	5.2	1
20	10	0.9	5.2	5.2	5.2	5.3	2.8	6.2	4.0	4.5	4.8	2
20	10	0.9	5.3	5.6	5.6	5.3	2.9	4.2	3.4	4.5	4.9	3
20	10	0.9	5.6	6.5	6.5	5.5	3.0	4.7	3.5	4.5	5.1	4

variance estimators are based upon $\hat{\tau}_i^2$ given in (21) along with a bias correction that results in an estimator with larger variance.

All of the results and conclusions made here are, however, based on few selected designs of replications. Overall, it seems that all of the methods have a substantial power to detect departure from the null hypothesis $H_0 : p = 1/2$. A general conclusion cannot be made due to the abundance of possible designs and sample size configurations. Additional theoretical power and efficiency investigations of rank-tests are found in the work of Janssen[1] and references therein.

TABLE 2 Type-I error simulations using heteroscedastic multivariate normal distributions having variance 1 in group 1 and variance 2 in group 2. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

n1	n2	rho1	$T^{(u)}$	TW	$T^{(w)}$	T_{BM}	$T_L^{(0)}$	$T_L^{(1)}$	$T_L^{(F)}$	RGL	DS	Setting
7	7	0	5.5	5.5	5.5	5.5	0.8	17.3	5.5	5.5	5.5	1
7	7	0	5.5	5.5	5.5	5.9	0	6.9	1.3	5.1	6.1	2
7	7	0	5.2	4.3	5.9	5.8	6.0	5.7	3.1	4.3	6.0	3
7	7	0	5.3	3.7	5.5	5.7	7.9	7.8	4.2	3.4	5.5	4
7	7	0.5	5.5	5.5	5.5	5.5	0.6	17.2	5.5	5.4	5.4	1
7	7	0.5	5.3	5.3	5.3	5.6	0.1	11.9	2.7	4.3	5.2	2
7	7	0.5	5.8	5.9	6.1	6.2	0.5	8.5	1.8	4.0	5.7	3
7	7	0.5	5.7	5.5	6.1	5.9	1.0	7.8	2.1	3.6	5.2	4
7	7	0.9	5.6	5.6	5.6	5.6	0.6	17.9	5.7	5.4	5.4	1
7	7	0.9	5.6	5.6	5.6	5.5	0.2	10.2	3.1	4.2	5.1	2
7	7	0.9	6.6	6.6	7.5	5.8	0.2	5.0	1.2	3.7	5.7	3
7	7	0.9	6.1	6.4	6.9	5.7	0.1	4.4	1.2	4.1	5.2	4
10	10	0	6.0	6.0	6.0	6.0	2.3	14.1	6.2	6.0	6.0	1
10	10	0	5.4	5.4	5.4	5.7	1.3	6.0	2.1	5.2	6.0	2
10	10	0	4.8	4.3	5.2	5.8	4.7	4.9	2.2	4.7	5.4	3
10	10	0	4.9	4.3	5.1	5.6	5.6	4.8	2.6	4.2	5.4	4
10	10	0.5	5.2	5.2	5.2	5.2	2.1	13.4	5.4	5.3	5.3	1
10	10	0.5	5.3	5.3	5.3	5.6	1.4	9.5	3.9	4.7	5.2	2
10	10	0.5	5.7	4.8	5.6	5.3	1.4	7.3	3.0	4.0	5.3	3
10	10	0.5	5.7	5.8	6.2	5.4	1.5	6.8	3.2	4.1	5.4	4
10	10	0.9	5.0	5.0	5.0	5.0	2.0	12.6	5.2	5.1	5.1	1
10	10	0.9	5.4	5.4	5.4	5.2	1.5	8.2	3.5	4.4	5.1	2
10	10	0.9	6.1	6.5	6.4	5.4	1.8	5.1	2.7	4.7	5.6	3
10	10	0.9	5.7	6.7	7.3	5.9	1.7	5.5	2.5	3.6	5.1	4
10	20	0	5.5	5.5	5.5	5.5	2.9	10.9	5.5	3.8	5.2	1
10	20	0	5.1	5.1	5.1	5.7	2.4	5.7	2.6	3.9	5.1	2
10	20	0	4.8	3.8	5.0	5.2	3.4	4.2	2.2	3.4	4.6	3
10	20	0	5.3	2.8	5.7	5.6	7.3	5.6	4.2	4.0	5.3	4
10	20	0.5	5.4	5.4	5.4	5.4	2.7	11.0	5.4	3.8	5.1	1
10	20	0.5	5.2	5.2	5.2	5.3	2.2	8.5	4.0	3.9	4.6	2
10	20	0.5	5.3	5.3	5.5	5.4	2.2	6.2	3.3	4.1	4.6	3
10	20	0.5	5.3	5.6	5.7	5.4	2.4	5.5	3.6	4.4	4.9	4
10	20	0.9	5.6	5.6	5.6	5.6	2.8	11.2	5.5	3.9	5.2	1
10	20	0.9	5.7	5.7	5.7	5.8	2.8	7.5	4.4	4.7	5.2	2
10	20	0.9	6.2	6.1	6.3	5.4	2.9	4.9	3.5	4.9	5.3	3
10	20	0.9	5.2	5.9	6.1	5.2	2.4	4.6	3.0	4.8	4.5	4
20	10	0	5.5	5.5	5.5	5.5	3.1	11.5	5.9	6.6	5.7	1
20	10	0	5.2	5.2	5.2	5.6	2.5	5.8	2.8	6.5	5.6	2
20	10	0	5.3	3.7	5.6	5.9	2.3	4.1	1.8	6.3	5.6	3
20	10	0	5.6	4.7	5.5	5.5	5.1	4.7	3.0	5.8	5.7	4
20	10	0.5	5.4	5.4	5.4	5.4	3.2	11.3	5.6	6.4	5.5	1
20	10	0.5	5.6	5.6	5.6	5.5	2.8	9.0	4.7	5.4	5.4	2
20	10	0.5	5.0	4.8	5.1	5.2	2.3	5.7	3.3	4.4	4.8	3
20	10	0.5	5.4	5.5	5.8	5.3	2.5	5.5	3.4	4.7	5.0	4
20	10	0.9	5.7	5.7	5.7	5.7	3.4	10.9	5.9	6.4	5.7	1
20	10	0.9	5.4	5.4	5.4	5.2	2.8	7.3	4.3	4.7	5.1	2
20	10	0.9	6.0	5.8	6.5	5.5	3.2	5.8	3.9	3.7	4.7	3
20	10	0.9	5.6	5.4	6.3	5.9	3.0	5.2	3.8	3.8	4.6	4

Remark 3. The question whether the weighted or unweighted estimator should be used has not been answered yet. A helpful selection criteria might be the asymptotic relative efficiency (ARE) of the corresponding weighted and unweighted estimators of p. It follows from the asymptotic normal distributions of $\sqrt{n}(\hat{p}^{(u)} - p)$ (7) and $\sqrt{N}(\hat{p}^{(w)} - p)$ (8) that we can define the ARE of these two sequences as

$$\text{ARE} = \text{ARE}\left(\sqrt{N}(\hat{p}^{(w)} - p), \sqrt{n}(\hat{p}^{(u)} - p)\right) = \frac{\sigma^{2(w)}/N}{\sigma^{2(u)}/n} = \frac{\sigma^{2(w)}}{\sigma^{2(u)}} \cdot \frac{n}{N}, \tag{24}$$

TABLE 3 Type-I error simulations using heteroscedastic multivariate normal distributions having variance 1 in group 1 and variance 3 in group 2. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

n1	n2	rho1	$T^{(u)}$	TW	$T^{(w)}$	T_{BM}	$T_L^{(0)}$	$T_L^{(1)}$	$T_L^{(F)}$	RGL	DS	Setting
7	7	0	5.6	5.6	5.6	5.6	0.8	17.6	5.6	5.8	5.8	1
7	7	0	5.7	5.7	5.7	6.0	0	7.0	1.5	5.6	6.8	2
7	7	0	5.1	3.1	5.4	5.8	6.3	7.9	4.1	0.3	5.8	3
7	7	0	4.9	4.0	5.5	5.7	6.6	6.5	2.9	4.9	6.0	4
7	7	0.5	5.5	5.5	5.5	5.5	0.7	17.9	5.4	5.7	5.7	1
7	7	0.5	5.5	5.5	5.5	5.6	0.1	11.6	3.0	4.8	5.6	2
7	7	0.5	5.3	5.3	5.4	5.7	0.4	7.6	1.6	4.1	5.1	3
7	7	0.5	5.4	4.8	5.9	5.7	1.4	8.8	2.2	3.1	5.1	4
7	7	0.9	5.3	5.3	5.3	5.3	0.6	17.4	5.1	5.5	5.5	1
7	7	0.9	5.8	5.8	5.8	5.9	0.2	11.7	3.3	4.5	5.6	2
7	7	0.9	6.1	6.5	7.4	6.0	0.2	10.3	2.1	2.8	5.1	3
7	7	0.9	5.5	5.5	6.4	5.5	0.3	6.2	1.6	3.7	4.5	4
10	10	0	5.8	5.8	5.8	5.8	2.5	14.1	5.9	6.2	6.2	1
10	10	0	5.3	5.3	5.3	5.6	1.6	5.8	2.1	5.1	5.8	2
10	10	0	5.2	3.6	5.5	5.7	4.1	5.3	2.4	6.3	6.1	3
10	10	0	4.9	3.0	5.7	5.7	8.3	5.4	4.3	4.3	5.8	4
10	10	0.5	5.8	5.8	5.8	5.8	2.4	14.1	5.9	5.9	5.9	1
10	10	0.5	5.2	5.2	5.2	5.4	1.4	9.3	3.7	4.6	5.3	2
10	10	0.5	5.5	5.0	5.3	5.4	1.2	6.6	2.7	4.0	5.1	3
10	10	0.5	5.3	5.2	5.6	5.5	1.2	6.2	2.5	3.7	4.9	4
10	10	0.9	5.7	5.7	5.7	5.7	2.1	13.4	5.8	6.0	6.0	1
10	10	0.9	5.6	5.6	5.6	5.7	1.6	8.8	3.8	4.7	5.3	2
10	10	0.9	5.2	5.4	5.3	6.0	1.4	4.8	2.3	3.8	4.4	3
10	10	0.9	6.0	5.4	6.1	5.6	1.7	5.2	2.7	3.9	5.3	4
10	20	0	5.4	5.4	5.4	5.4	2.6	11.1	5.2	3.2	5.3	1
10	20	0	5.4	5.4	5.4	5.3	2.2	5.7	2.6	3.6	5.5	2
10	20	0	5.0	4.3	4.9	5.6	2.0	3.7	1.8	4.3	5.6	3
10	20	0	5.2	2.8	5.1	5.2	5.1	5.2	3.2	3.9	5.9	4
10	20	0.5	4.8	4.8	4.8	4.8	2.6	10.2	4.8	3.1	5.0	1
10	20	0.5	5.3	5.3	5.3	5.5	2.2	8.6	4.1	3.8	5.0	2
10	20	0.5	5.1	5.1	5.6	5.0	2.1	6.3	3.3	4.2	4.7	3
10	20	0.5	5.5	5.0	5.3	5.8	2.3	6.6	3.6	4.0	4.8	4
10	20	0.9	5.0	5.0	5.0	5.0	2.4	10.6	4.8	3.2	5.0	1
10	20	0.9	5.6	5.6	5.6	5.4	2.5	7.3	4.2	4.2	4.8	2
10	20	0.9	5.2	5.3	5.6	5.5	2.5	4.6	3.1	4.0	4.3	3
10	20	0.9	5.2	5.6	5.7	4.9	2.2	4.8	2.9	3.7	4.3	4
20	10	0	5.6	5.6	5.6	5.6	3.5	11.2	5.9	7.8	6.0	1
20	10	0	5.5	5.5	5.5	5.9	3.0	6.0	3.0	7.6	6.0	2
20	10	0	5.2	3.5	5.5	5.6	3.7	4.5	2.3	7.6	6.0	3
20	10	0	5.0	4.6	5.6	5.6	3.6	4.2	2.1	6.4	5.7	4
20	10	0.5	5.8	5.8	5.8	5.8	3.6	11.6	6.2	7.9	6.3	1
20	10	0.5	5.8	5.8	5.8	5.9	3.0	9.6	5.0	6.3	5.7	2
20	10	0.5	5.7	5.2	5.7	5.6	2.8	7.8	4.4	4.7	5.2	3
20	10	0.5	5.7	4.9	6.2	5.8	2.9	7.7	4.5	4.4	5.2	4
20	10	0.9	5.7	5.7	5.7	5.7	3.4	11.2	5.9	7.6	6.0	1
20	10	0.9	5.6	5.6	5.6	5.9	3.2	8.2	4.6	5.3	5.4	2
20	10	0.9	5.4	5.3	5.7	5.7	2.7	5.7	3.7	3.7	4.6	3
20	10	0.9	5.7	5.6	6.0	5.9	2.9	5.4	3.7	3.6	4.6	4

(see, eg, Boos and Stefanski[30p14]). Thus, in our case, the ARE indicates which of the two estimators has a smaller SE. If ARE < 1, then the weighted estimator is more efficient; if ARE > 1, then the unweighted and both are of equal quality if ARE $= 1$, respectively. For example, let $n_1 = n_2 = n_0$ and assume that $\mathrm{Var}(Y_{iks}) = \sigma^2$ and $\mathrm{Cov}(Y_{iks}, Y_{iks'}) = \tau$. Since both the unweighted and weighted estimators are identical in case of equally sized clusters, consider a scenario

TABLE 4 Type-I error simulations using homogeneous lognormal distributions. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

n1	n2	rho1	$T^{(u)}$	TW	$T^{(w)}$	T_{BM}	$T_L^{(0)}$	$T_L^{(1)}$	$T_L^{(F)}$	RGL	DS	Setting
7	7	0	5.8	5.8	5.8	5.8	0.7	18.0	5.9	5.4	5.4	1
7	7	0	5.3	5.3	5.3	6.3	0	6.5	1.4	4.9	6.0	2
7	7	0	4.7	4.5	4.9	6.0	2.9	4.1	1.1	4.3	5.2	3
7	7	0	5.0	4.0	5.1	5.8	6.4	6.0	3.3	2.9	5.7	4
7	7	0.5	5.4	5.4	5.4	5.4	0.6	18.2	5.5	5.0	5.0	1
7	7	0.5	5.8	5.8	5.8	5.6	0.2	11.9	3.0	4.4	5.4	2
7	7	0.5	6.2	6.2	6.6	5.6	0.2	8.7	2.3	4.7	6.0	3
7	7	0.5	6.3	6.5	6.9	5.7	0.2	4.4	1.0	4.4	5.9	4
7	7	0.9	5.6	5.6	5.6	5.6	0.6	17.8	5.6	5.1	5.1	1
7	7	0.9	6.2	6.2	6.2	5.6	0.5	8.9	2.7	4.8	5.8	2
7	7	0.9	6.5	6.8	7.5	6.1	0.5	5.2	1.5	4.4	6.0	3
7	7	0.9	6.6	7.5	8.6	5.9	0.4	5.2	1.7	3.6	5.8	4
10	10	0	5.4	5.4	5.4	5.4	2.2	13.4	5.6	5.1	5.1	1
10	10	0	5.2	5.2	5.2	5.3	1.3	6.0	2.2	4.6	5.3	2
10	10	0	4.6	3.2	4.7	5.3	4.9	4.9	2.8	4.3	4.7	3
10	10	0	4.6	3.3	4.9	5.7	5.1	4.7	2.5	4.8	4.6	4
10	10	0.5	5.6	5.6	5.6	5.6	2.1	13.4	5.7	5.3	5.3	1
10	10	0.5	6.0	6.0	6.0	5.5	1.7	9.8	4.3	5.1	5.9	2
10	10	0.5	5.8	5.8	6.2	5.5	1.7	5.9	3.0	4.2	5.7	3
10	10	0.5	5.5	5.6	6.2	5.4	1.4	6.2	2.8	4.5	5.4	4
10	10	0.9	5.7	5.7	5.7	5.7	2.1	13.5	5.8	5.5	5.5	1
10	10	0.9	5.9	5.9	5.9	5.5	2.2	7.4	3.7	5.0	5.6	2
10	10	0.9	5.8	6.3	6.2	5.5	2.1	3.9	2.6	4.8	5.5	3
10	10	0.9	5.4	6.1	6.1	5.3	1.8	3.1	2.2	4.5	5.0	4
10	20	0	5.2	5.2	5.2	5.2	2.6	11.0	5.3	4.6	4.9	1
10	20	0	5.4	5.4	5.4	5.8	2.4	5.6	2.6	4.9	5.2	2
10	20	0	5.0	3.7	5.2	5.4	3.4	4.4	2.1	4.8	4.9	3
10	20	0	5.2	4.3	5.0	5.7	3.4	4.0	1.8	4.2	5.1	4
10	20	0.5	5.7	5.7	5.7	5.7	2.9	11.7	5.7	4.8	5.2	1
10	20	0.5	5.4	5.4	5.4	5.5	2.5	8.2	4.3	4.6	5.1	2
10	20	0.5	5.8	5.9	6.0	5.9	2.8	6.0	3.8	4.8	5.3	3
10	20	0.5	5.6	6.1	6.2	5.8	2.6	4.8	3.2	4.7	5.1	4
10	20	0.9	5.6	5.6	5.6	5.6	2.9	11.0	5.6	4.8	5.2	1
10	20	0.9	5.6	5.6	5.6	5.5	2.9	6.8	4.2	4.8	5.1	2
10	20	0.9	5.5	6.3	6.3	5.3	3.0	4.6	3.6	4.9	4.9	3
10	20	0.9	5.6	6.0	6.0	5.4	2.6	4.0	3.1	4.5	5.1	4
20	10	0	5.5	5.5	5.5	5.5	3.1	11.0	5.6	5.1	5.3	1
20	10	0	5.0	5.0	5.0	5.6	2.3	5.3	2.5	4.7	4.8	2
20	10	0	5.1	4.3	5.4	5.9	3.4	4.1	1.9	4.8	5.0	3
20	10	0	5.1	3.8	5.1	5.5	7.4	5.4	3.6	4.6	5.0	4
20	10	0.5	5.3	5.3	5.3	5.3	2.9	11.0	5.4	4.7	5.0	1
20	10	0.5	5.2	5.2	5.2	5.4	2.4	8.0	4.3	4.6	4.9	2
20	10	0.5	5.2	5.4	5.6	5.7	2.4	5.3	3.3	4.6	4.8	3
20	10	0.5	5.8	5.9	5.8	5.7	2.6	5.2	3.5	4.9	5.3	4
20	10	0.9	5.4	5.4	5.4	5.4	3.2	11.4	5.5	5.0	5.1	1
20	10	0.9	5.8	5.8	5.8	5.6	2.9	6.9	4.2	4.8	5.3	2
20	10	0.9	5.8	6.3	6.3	5.7	3.0	4.7	3.7	4.6	5.5	3
20	10	0.9	5.6	6.8	6.5	5.4	2.7	4.6	3.4	4.4	5.2	4

in which all cluster sizes are equal except one of them, ie, let $m_{11} = \ldots = m_{1n_1} = m_{21} = \ldots = m_{2,n_2-1} = 1$ and let $m_{2n_2} = m_0 > 1$. Routine calculations show that ARE $\gtreqless 1$ if

$$\tau \gtreqless \sigma^2 \frac{\frac{n_0-1+1/m_0}{n_0^2} - \frac{1}{m_2}}{\frac{m_0(m_0-1)}{m_2^2} - \frac{m_0-1}{m_0 n_0^2}}, \tag{25}$$

TABLE 5 Type-I error simulations using rounded normal distributions. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

n1	n2	rho1	$T^{(u)}$	TW	$T^{(w)}$	T_{BM}	$T_L^{(0)}$	$T_L^{(1)}$	$T_L^{(F)}$	RGL	DS	Setting
7	7	0	6.4	6.4	6.4	6.4	0.3	16.6	5.0	4.7	5.8	1
7	7	0	5.2	5.2	5.2	6.1	0	6.0	1.0	4.4	5.5	2
7	7	0	5.0	4.3	5.1	6.5	7.2	5.5	3.1	3.9	5.4	3
7	7	0	4.9	3.4	5.4	6.3	11.6	7.0	5.4	4.0	5.3	4
7	7	0.5	6.2	6.2	6.2	6.2	0.3	16.4	5.0	4.8	5.7	1
7	7	0.5	6.0	6.0	6.0	6.2	0.1	11.2	2.8	4.7	5.6	2
7	7	0.5	6.0	5.8	6.9	6.3	0.3	10.1	2.0	4.2	5.6	3
7	7	0.5	5.7	6.0	6.6	5.9	0.4	7.9	1.8	4.2	5.4	4
7	7	0.9	5.8	5.8	5.8	5.8	0.3	16.5	4.7	4.5	5.6	1
7	7	0.9	6.3	6.3	6.3	6.4	0.4	8.9	2.4	4.8	5.9	2
7	7	0.9	6.0	6.2	6.5	5.9	0.2	3.5	1.0	4.4	5.8	3
7	7	0.9	6.0	6.4	7.0	6.3	0.2	2.4	0.7	4.0	5.6	4
10	10	0	6.1	6.1	6.1	6.1	1.9	12.9	5.4	5.0	5.8	1
10	10	0	5.2	5.2	5.2	6.1	1.3	5.7	2.2	4.7	5.2	2
10	10	0	5.2	3.8	5.2	5.7	3.4	4.2	1.7	4.7	5.3	3
10	10	0	5.2	4.0	5.2	6.0	3.3	4.1	2.1	4.3	5.2	4
10	10	0.5	5.8	5.8	5.8	5.8	2.0	13.2	5.2	4.7	5.5	1
10	10	0.5	5.1	5.1	5.1	5.4	1.4	8.7	3.4	4.3	4.8	2
10	10	0.5	5.9	6.1	6.3	5.5	1.4	6.1	2.8	4.9	5.7	3
10	10	0.5	5.7	5.8	5.9	5.6	1.3	5.4	2.6	4.4	5.3	4
10	10	0.9	5.7	5.7	5.7	5.7	1.8	12.4	5.0	4.6	5.3	1
10	10	0.9	6.2	6.2	6.2	5.9	1.8	7.4	3.4	5.1	5.8	2
10	10	0.9	5.6	5.8	5.8	5.8	1.5	3.6	2.3	4.5	5.3	3
10	10	0.9	5.7	6.4	6.7	5.7	1.9	4.3	2.5	4.6	5.2	4
10	20	0	5.4	5.4	5.4	5.4	2.5	10.8	5.0	4.6	5.0	1
10	20	0	5.1	5.1	5.1	5.6	2.1	5.2	2.4	4.5	4.9	2
10	20	0	5.1	4.2	5.0	5.7	3.0	3.8	1.9	4.6	5.1	3
10	20	0	5.1	3.8	5.2	5.7	4.1	4.6	2.7	4.4	5.0	4
10	20	0.5	5.7	5.7	5.7	5.7	2.9	10.6	5.4	4.9	5.2	1
10	20	0.5	5.3	5.3	5.3	5.7	2.5	8.1	4.1	4.8	4.8	2
10	20	0.5	5.3	5.6	5.7	5.9	2.5	5.5	3.4	4.7	5.0	3
10	20	0.5	5.7	5.6	6.3	5.5	2.5	6.3	3.6	4.4	5.1	4
10	20	0.9	5.2	5.2	5.2	5.2	2.4	10.1	4.8	4.2	4.7	1
10	20	0.9	5.4	5.4	5.4	5.5	2.7	6.4	3.6	4.6	4.9	2
10	20	0.9	5.4	5.7	5.8	5.3	2.6	4.1	3.1	4.7	4.9	3
10	20	0.9	6.1	6.8	6.5	5.9	2.9	4.7	3.5	4.8	5.7	4
20	10	0	5.8	5.8	5.8	5.8	2.9	10.4	5.3	4.9	5.3	1
20	10	0	5.8	5.8	5.8	5.5	2.3	5.5	2.5	4.9	5.5	2
20	10	0	4.6	3.9	5.1	5.6	3.2	3.6	1.8	4.5	4.5	3
20	10	0	5.4	3.6	5.5	5.7	4.7	4.9	2.7	4.3	5.2	4
20	10	0.5	5.5	5.5	5.5	5.5	2.6	10.7	5.0	4.6	5.0	1
20	10	0.5	5.9	5.9	5.9	5.6	2.8	8.4	4.5	5.1	5.4	2
20	10	0.5	5.7	6.0	6.4	5.7	2.6	6.5	3.8	4.8	5.3	3
20	10	0.5	5.4	5.7	5.8	5.5	2.3	4.6	3.0	4.6	4.8	4
20	10	0.9	5.8	5.8	5.8	5.8	3.0	10.7	5.4	5.1	5.3	1
20	10	0.9	5.6	5.6	5.6	5.9	2.7	6.6	4.1	4.8	5.2	2
20	10	0.9	6.0	6.4	6.4	6.0	2.7	4.6	3.4	5.0	5.6	3
20	10	0.9	5.8	6.4	6.4	5.8	3.0	4.3	3.4	4.7	5.4	4

where $m_2 = n_0 - 1 + m_0$. For example, if $n_0 = 10$ and $m_0 = 2$, an intraclass correlation value of about $\tau = 0.35$ leads to ARE = 1 (see Figure 2). Thus, the powers of both the weighted and unweighted tests $T^{(w)}$ and $T^{(u)}$ are expected to be identical (for large sample sizes) in this specific scenario. If $\tau = 0$, it follows that $T^{(w)}$ has a higher power than $T^{(u)}$, and thus, the weighted estimator is preferred. Otherwise, in case of large correlations, the unweighted estimator should be used. The aforementioned findings are numerically justified in a simulation study with $n_1 = n_2 = 30$, $m_0 = 10$. Data $X_{11}, \ldots, X_{2,n_2-1}$ were generated from normal distributions with variance 1, whereas the only cluster $\mathbf{X}_{2n_2}$ was

TABLE 6 Power simulations ($\alpha = 5\%$) using homogeneous multivariate normal distributions with positive intracluster correlation $\rho = 0.8$ for the repeated measurements in both groups. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

p	$T^{(u)}$	TW	$T^{(w)}$	T_{BM}	$T_L^{(0)}$	$T_L^{(1)}$	$T_L^{(F)}$	RGL	DS
					Design 1				
0.5	5.7	5.7	5.7	5.3	3.8	6.4	4.6	5.3	5.6
0.55	9.6	9.6	9.6	8.6	6.8	10.7	8.1	8.9	9.4
0.60	21.8	21.8	21.8	20.3	16.7	23.5	19.1	20.7	21.5
0.65	41.7	41.7	41.7	38.1	35.9	44.1	38.7	40.2	41.1
0.70	67.0	67.0	67.0	62.1	60.0	69.2	63.5	65.4	66.4
0.75	86.6	86.6	86.6	82.5	82.4	88.0	84.8	85.9	86.3
0.80	97.2	97.2	97.2	95.3	96.1	97.6	96.7	97.1	97.2
0.85	99.7	99.7	99.5	99.2	99.5	99.7	99.6	99.0	99.7
0.90	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.95	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
					Design 2				
0.50	5.4	6.2	5.7	5.3	3.6	4.8	4.0	5.0	5.3
0.55	9.6	9.7	9.2	8.6	6.6	8.3	7.1	8.1	9.4
0.60	23.4	23.2	21.7	20.7	18.0	21.9	19.4	20.0	23.0
0.65	42.7	43.8	42.0	37.9	35.9	40.0	37.3	39.7	42.2
0.70	68.8	68.5	66.1	61.7	62.3	67.8	64.0	56.6	68.2
0.75	88.3	88.2	87.0	82.2	84.2	86.9	85.3	82.0	87.8
0.85	99.8	99.8	99.8	99.4	99.5	99.8	99.7	99.7	99.8
0.90	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.95	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

TABLE 7 Power simulations ($\alpha = 5\%$) using homogeneous multivariate normal distributions with negative intracluster correlation (ICC) $\rho = -0.8$ for the repeated measurements in both groups. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

p	$T^{(u)}$	TW	$T^{(w)}$	T_{BM}	$T_L^{(0)}$	$T_L^{(1)}$	$T_L^{(F)}$	RGL	DS
					Design 1				
0.50	4.1	4.1	4.1	5.4	2.8	0.0	0.6	4.8	5.1
0.55	35.2	35.2	35.2	9.4	29.9	0.0	12.7	37.2	38.2
0.60	89.4	89.4	89.4	19.3	86.3	0.1	63.0	90.1	90.6
0.65	99.8	99.8	99.8	38.4	99.6	1.2	94.9	99.8	99.8
0.70	100.0	100.0	100.0	61.7	100.0	2.8	98.7	100.0	100.0
0.75	100.0	100.0	100.0	82.8	100.0	5.7	99.0	100.0	100.0
0.80	100.0	100.0	100.0	95.1	100.0	8.2	99.2	100.0	100.0
0.85	100.0	100.0	100.0	99.4	100.0	11.4	99.4	100.0	100.0
0.90	100.0	100.0	100.0	100.0	100.0	13.3	99.3	100.0	100.0
0.95	100.0	100.0	100.0	100.0	100.0	16.2	99.3	100.0	100.0
					Design 2				
0.50	5.1	5.5	5.2	5.2	4.5	1.3	1.3	5.4	5.6
0.55	17.3	19.5	16.4	7.9	6.1	1.9	1.4	14.3	19.3
0.60	54.6	53.9	46.1	14.8	13.6	7.8	5.9	37.8	57.2
0.65	82.0	81.6	80.9	27.3	22.5	9.7	8.2	78.9	83.7
0.70	97.8	97.2	96.9	47.7	36.4	14.2	14.5	96.2	98.1
0.75	99.9	99.8	99.7	67.3	43.0	17.8	19.3	99.7	99.9
0.80	100.0	100.0	100.0	84.4	46.4	18.3	18.9	100.0	100.0
0.85	100.0	100.0	100.0	94.7	64.8	24.9	30.1	100.0	100.0
0.90	100.0	100.0	100.0	99.5	53.8	34.3	34.1	100.0	100.0
0.95	100.0	100.0	100.0	100.0	53.2	11.0	12.7	100.0	100.0

generated from $N(\mu, \mathbf{V})$, where $\mu = \sqrt{2}\phi^{-1}(p)(1, \ldots, 1)^\top$ and $\mathbf{V} = \mathbf{I} + \rho(\mathbf{J} - \mathbf{I})$. Here, ρ was chosen according to (25). The results are displayed in Table 8.

It can be seen from Table 8 that the powers of the three tests are almost identical if ARE = 1. Otherwise, if ARE < 1, the weighted test $T^{(w)}$ has a higher power than $T^{(u)}$ and vice versa if ARE > 1. Roughly speaking, $T^{(w)}$ is more efficient than $T^{(u)}$ if the clustered data are low or mildly correlated. Investigations of the ARE of the tests will be part of future research.[31,32]

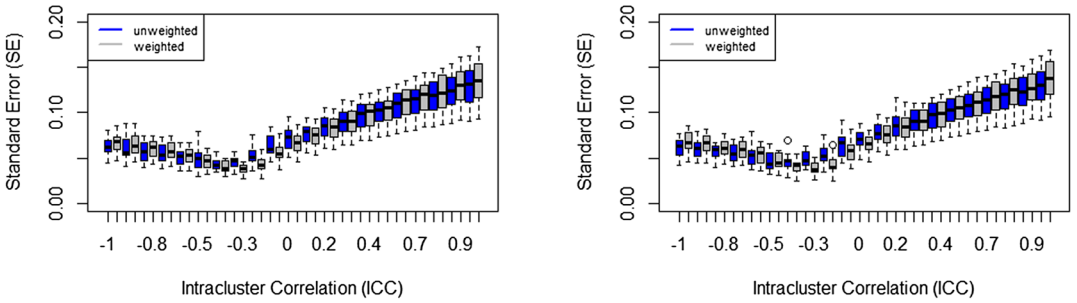

FIGURE 2 Boxplot of standard error of the relative effect size estimates. The true effect size is $p = 0.5$. A, Normal distribution; B, Lognormal distribution [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 8 Power simulations ($\alpha = 5\%$) of the unweighted and weighted tests $T^{(u)}$, TW and $T^{(w)}$ when the asymptotic relative efficiency (ARE) as given in (24) is smaller than, equal to, or larger than 1

	ARE < 1			ARE = 1			ARE > 1		
p	$T^{(u)}$	TW	$T^{(w)}$	$T^{(u)}$	TW	$T^{(w)}$	$T^{(u)}$	TW	$T^{(w)}$
0.50	0.0536	0.0331	0.0583	0.0521	0.0464	0.0537	0.0490	0.0362	0.0572
0.55	0.1083	0.0785	0.1152	0.1036	0.1002	0.1135	0.1021	0.0776	0.1067
0.60	0.2696	0.2274	0.2993	0.2793	0.2759	0.2927	0.2756	0.2110	0.2639
0.65	0.5449	0.5006	0.5896	0.5464	0.5458	0.5563	0.5389	0.4240	0.4883
0.70	0.7988	0.7825	0.8417	0.7998	0.8052	0.8057	0.7936	0.6595	0.7037
0.75	0.9466	0.9406	0.9627	0.9497	0.9518	0.9465	0.9480	0.8297	0.8535
0.80	0.9949	0.9955	0.9975	0.9924	0.9937	0.9913	0.9935	0.9292	0.9385
0.85	0.9999	1.0000	1.0000	0.9999	0.9999	0.9989	0.9997	0.9743	0.9764
0.90	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9931	0.9936

7 | DATA EVALUATIONS

In this section, the data set introduced in Section 2 will be analyzed. We will test the null hypothesis $H_0 : p = 1/2$ using the two new methods $T^{(u)}$, $T^{(w)}$ as given in (19), and also compute the modified version TW of $T^{(w)}$ as indicated in (21). We add the Brunner-Munzel test T_{BM} for comparative purposes for the computation of which the first observation within each cluster was used. The three methods $T_L^{(0)}$, $T_{(1)}$, and $T_L^{(F)}$ for testing $H_0 : E(S) = 0$ as given in (23) were also computed and compared with the aforementioned methods. Note that these differ only in the estimation of the asymptotic variance. Furthermore, for testing the null hypothesis $H_0 : F_1 = F_2$, the methods proposed by RGL[13] and DS[14] were computed. For all of the different methods, point estimators of the treatment effect, their SEs, test statistics, 95% confidence intervals, as well as p-values are reported. We used $\alpha = 0.05$ for all of the data evaluations and interpretations.

TABLE 9 Results for the body weight irritation study. Here, $T^{(u)}$ and $T^{(w)}$ represent the new methods given in (16); T_{BM} the Brunner-Munzel Test; $T_L^{(0)}$, $T_L^{(1)}$, and $T_L^{(F)}$ the methods from the work of Larocque et al[21] given in (23); the RGL test proposed by Rosner et al[13]; and DS the method proposed by Datta and Satten[14]

			Test	Confidence Interval		
Method	Point Estimators ($\hat{p}^{(c)}$, S)	SE($\hat{p}^{(c)}$)	Statistic	Lower Limit	Upper Limit	p-value
Results for testing $H_0 : p = 1/2$						
$T_L^{(u)}$	0.6607	0.0911	1.7646	0.4725	0.8489	0.0907
TW	0.6800	0.0913	1.9723	0.4913	0.8687	0.0606
$T_L^{(w)}$	0.6800	0.0903	1.9930	0.4918	0.8682	0.0599
T_{BM}	0.7160	0.1094	1.9742	0.4902	0.9418	0.0600
Results for testing $H_0 : E(S) = 0$ (Larocque et al (2010))						
$T_L^{(0)}$	0.1157	0.0336	1.7209	NA	NA	0.0853
$T_L^{(1)}$	0.1157	0.0292	1.9785	0.0011	0.2304	0.0479
$T_L^{(F)}$	0.1157	0.0322	1.7953	NA	NA	0.0726
Results for testing $H_0 : F_1 = F_2$						
RGL			-2.1565	NA	NA	0.0310
DS			-1.7423	NA	NA	0.0814

Here, $\hat{p}^{(c)}$ estimates the probability that the bodyweights from vehicle treated rats is smaller than from those under treatment.

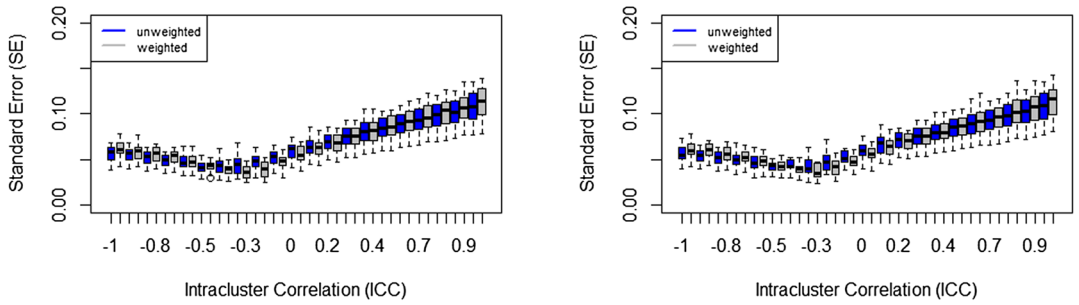

FIGURE 3 Boxplot of standard error of the relative effect size estimates. The true effect size is $p = 0.75$. A, Normal distribution; B, Lognormal distribution [Colour figure can be viewed at wileyonlinelibrary.com]

Since several rats share the same cage, the cage is assumed to be the experimental unit and the rats represent the replicates. Here, the numbers of cages in the two dose groups are identical ($n_1 = n_2 = 13$), while the numbers of replicates are different. The statistical analysis of the data is displayed in Table 9. First, it can be seen that both the unweighted $\widehat{p}^{(u)}$ as well as the weighted estimator $\widehat{p}^{(w)}$ are just slightly different. The estimated SEs are also about the same. This occurs, because the replicates have a medium correlation. Furthermore, the data do not provide the evidence to reject the null hypothesis $H_0 : p = 1/2$ at 5% level of significance. All of the results are, however, borderline and a remarkable improvement can be detected. The test decisions for testing $H_0 : F_1 = F_2$ differ slightly. However, both p-values indicate a difference in terms of the distributions.

8 | DISCUSSION

Dependent replications are observed in many experiments and there is a need for adequate statistical procedures that can be used for modeling them. Reducing the replications to single observations by either using their means, medians, or strictly using the first observation cannot be recommended because a lot of the information provided by the replications is not effectively used. Therefore, this strategy results in a loss of power and cannot be recommended for practitioners. Furthermore, data often follow a skewed distribution or are even observed on ordinal scales. Thus, there is a need for purely nonparametric flexible methods that can be used for analyzing such data in a unified way. Ranking procedures are known to be a robust and powerful statistical analysis tool for which parametric distributional assumptions are doubtful. Rosner et al[14] and DS[13] proposed rank-based WMW-type test procedures for testing $H_0^F : F_1 = F_2$ formulated in terms of the distribution functions of the data. This hypothesis implies that variances of the data across the two groups are identical. In particular, confidence intervals for the underlying effect cannot be computed in general nonparametric models. In this paper, purely nonparametric methods for testing hypotheses formulated in terms of the WMW-effect p given in (1) have been introduced. All of the methods neither imply that variances nor data distributional shapes are identical even under the null hypothesis. Thus, the nonparametric Behrens-Fisher problem with dependent replications has been investigated.

Different weighting schemes (weighted and unweighted) for estimating the treatment effect p have been investigated from a theoretical as well as empirical point of views. The choice of the weighting scheme to use depends on the specific study question and one cannot be recommended over the other for all situations. When, for example, exchangeable correlation structure can be assumed, the strength of the ICC present in the data could be used as a guiding factor. We conducted a small-scale simulation to shed some light on the effect of ICC on the precision in the estimation of the relative treatment effect. We examined the effects of ICC on the precision of the estimator for two distributions. Figures 2 and 3 show boxplots constructed from SEs computed from all possible sample sizes ($n_1, n_2 \in \{7, 10, 20\}$) and cluster sizes ($m_{ik}$s generated from Binomial$(10, 0.3) + 1$ and Binomial$(4, 0.3) + 1$) combinations (total of 18 numbers) for different values of ICC.

It can be seen from the figures that, for both distributions and effect sizes, the weighted analysis is preferred for low ICC (-0.4 to 0.4) and unweighted analysis is preferred otherwise. It is also interesting to see that high positive correlations show widely varying SE based on sample size and cluster size combinations compared to high negative correlations. As one would expect, the estimation is best in terms of low SEs for both weighted and unweighted analysis when the correlation is near zero.

In this paper, methods for specific clustered data were investigated. Clusters involving data from both treatment groups were not considered in this manuscript. The generalization of the estimating approaches to general clustered data designs will be part of future research. Furthermore, we restricted ourselves to two different weighting schemes. These can be generalized to "arbitrarily weighted" estimators by using a general weighting framework similar to the work of Larocque et al.[21] The motivating example represents a part of a complex study involving more than two treatment groups. The generalization of the methods to the several sample case will be considered in future investigations and appropriate global testing as well as multiple contrast test procedures for testing hypotheses formulated in terms of relative effects as in the work of Konietschke et al[33] will be explored theoretically as well as empirically. Resampling methods to approximate the distribution of the tests in both the two- and several sample case appears to be an intriguing option for small sample sizes. Since data is not exchangeable in the general setup considered here, studentized resampling (permutation or bootstrap) methods shall be explored. The aim of these methods is to mimic the asymptotic distribution of the test statistic (which turned out to be standard normal). A general theory for resampling methods is provided by other works[1,34-36] and will serve as an excellent basis for the development of such methods. Another important issue to tackle is allowing the cluster size to be informative.[12]

ACKNOWLEDGEMENTS

The authors are grateful to three expert referees, the associate editor, and the joint editor for their helpful comments which led to a considerable improvement of the original version of the paper. We would like to thank Paavo Sattler (Technical University of Dortmund) for valuable discussions. The research is supported by the Deutsche Forschungsgemeinschaft award number DFG KO 4680/3-2.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in https://tools.niehs.nih.gov/cebs3/views/index.cfm?action=main.download&bin_id=1600&library_id=4877&sfileIdsSelected=1de2c1a6578948500157908016d60027 accessed on December 16, 2018.

ORCID


Frank Konietschke https://orcid.org/0000-0002-5674-2076

REFERENCES

1. Janssen A. Testing nonparametric statistical functionals with applications to rank tests. *J Stat Plan Inference*. 1999;81(1):71-93.
2. Brunner E, Munzel U. The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biom J*. 2000;42(1):17-25.
3. Neubert K, Brunner E. A studentized permutation test for the non-parametric Behrens-Fisher problem. *Comput Stat Data Anal*. 2007;51(10):5192-5204.
4. Pauly M, Asendorf T, Konietschke F. Permutation-based inference for the AUC: a unified approach for continuous and discontinuous data. *Biom J*. 2016;58(6):1319-1337.
5. Fligner MA, Policello GE. Robust rank procedures for the Behrens-Fisher problem. *J Am Stat Assoc*. 1981;76:162-168.
6. Brunner E, Munzel U, Puri ML. The multivariate nonparametric Behrens-Fisher problem. *J Stat Plan Inference*. 2002;108(1-2):37-53.
7. Brunner E, Puri ML. Nonparametric methods in design and analysis of experiments. In: Ghosh S, Rao CR, eds. *Handbook of Statistics*. Vol. 13. 1996:631-703.
8. Verbeke G. *Linear Mixed Models for Longitudinal Data*. Vol. 126. New York, NY: Springer; 1997.
9. West BT, Welch KB, Galecki AT. *Linear Mixed Models: A Practical Guide using Statistical Software*. New York, NY: CRC; 2014.
10. Jung SH, Kang SH, Ahn C. Chi-square test for $R \times C$ contingency tables with clustered data. *J Biopharm Stat*. 2003;13(2):241-251.
11. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics*. 1992;1:577-585.
12. Dutta S, Datta S. A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics*. 2016;72(2):432-440.
13. Rosner B, Glynn RJ, Lee ML. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*. 2006;62(1):185-192.
14. Datta S, Satten GA. Rank-sum tests for clustered data. *J Am Stat Assoc*. 2005;100(471):908-915.
15. ICH Guideline on Statistical Principles for Clinical Trials 1998. Guideline. Available at http://private.ich.org
16. Tsimikas J, Bosch RJ, Coull BA, Barmi HE. Profile-likelihood inference for highly accurate diagnostic tests. *Biometrics*. 2002;58(4):946-956.

17. Akritas M, Brunner E. A unified approach to rank tests for mixed models. *J Stat Plan Inference*. 1997;61(2):249-277.
18. Domhof S. *Nichtparametrische Relative Effekte*. Göttingen, Germany: University of Göttingen; 2001.
19. Werner C. *Nichtparametrische Analyse Von Diagnostischen Tests* [PhD thesis]. Göttingen, Germany: University of Göttingen; 2006.
20. Konietschke F, Brunner E. Nonparametric analysis of clustered data in diagnostic trials: estimation problems in small sample sizes. *Comput Stat Data Anal*. 2009;53(3):730-741.
21. Larocque D, Haataja R, Nevalainen J, Oja H. Two sample tests for the nonparametric Behrens-Fisher problem with clustered data. *J Nonparametric Stat*. 2010;22(6):755-771.
22. Lèvy P. *Calcul Des Probabilitèes*. Paris, France: Gauthiers-Villars; 1925.
23. Ruymgaart FH. Unified approach to the asymptotic distribution theory of certain midrank statistics. In: Raoult JP, ed. *Statistique non Parametrique Asymptotique*. Springer: Berlin, Germany; 1980:1-18.
24. Munzel U. Linear rank score statistics when ties are present. *Stat Probab Lett*. 1999;41(4):389-395.
25. Werner C, Brunner E. Rank methods for the analysis of clustered data in diagnostic trials. *Comput Stat Data Anal*. 2007;51(10):5041-5054.
26. Konietschke F, Friede T, Pauly M. Semi-parametric analysis of overdispersed count and metric data with varying follow-up times: asymptotic theory and small sample approximations. *Biometrical Journal*. 2019;61(3):616-629.
27. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann Math Statist*. 1954;25(2):290-302.
28. Leonov S, Qaqish B. Correlated endpoints: simulation, modeling, and extreme correlations. *Statistical Papers*. 2017:1-26.
29. Jiang Y, He X, Lee MLT, Rosner B, Yan J. Wilcoxon rank-based tests for clustered data with R package clusrank. 2017. arXiv preprint 2017: arXiv:1706.03409.
30. Boos DD, Stefanski LA. *Essential Statistical Inference: Theory and Methods*. New York, NY: Springer; 2013.
31. Hájek J, Sidák Z, Sen PK. *Theory of Rank Tests*. Orlando, FL: Academic Press; 1999.
32. Janssen A. Asymptotic relative efficiency of tests at the boundary of regular statistical models. *J Stat Plan Inference*. 2004;126(2):461-477.
33. Konietschke F, Hothorn LA, Brunner E. Rank-based multiple test procedures and simultaneous confidence intervals. *Electron J Stat*. 2012;6:738-759.
34. Janssen A. Studentized permutation tests for non-iid hypotheses and the generalized Behrens-Fisher problem. *Stat Probab Lett*. 1997;36(1):9-21.
35. Janssen A, Pauls T. How do bootstrap and permutation tests work? *Ann Stat*. 2003;31(3):768-806.
36. Janssen A. Resampling student's t -type statistics. *Ann Inst Stat Math*. 2005;57(3):507-529.

How to cite this article: Roy A, Harrar SW, Konietschke F. The nonparametric Behrens-Fisher problem with dependent replicates. *Statistics in Medicine*. 2019;38:4939–4962. https://doi.org/10.1002/sim.8343

APPENDIX

PROOFS

A.1 | Underlying model

Given are two independent samples with dependent replicated data that can be modeled by independent random vectors

$$\mathbf{X}_{ik} = (X_{ik1}, \ldots, X_{ikm_{ik}})', \; i = 1, 2; k = 1, \ldots, n_i \tag{A1}$$

with distributions $X_{iks} \sim F_i, i = 1, 2$.

m_{ik} is the number of replicates of subject k under treatment i.

$N = \sum_{i=1}^{2} \sum_{k=1}^{n_i} m_{ik}$ is the total number of observations.

$m_i = \sum_{j=1}^{n_i} m_{ij} \; i = 1, 2.$

A.2 | Proof of strong consistency of relative treatment effect estimates

$$\begin{aligned}
\left|\hat{p}^{(c)} - p\right| &= \left|\int \hat{F}_1^{(c)} d\hat{F}_2^{(c)} - \int F_1 dF_2\right| = \left|\int \hat{F}_1^{(c)} d\hat{F}_2^{(c)} - \int F_1 dF_2 - \int F_1 d\hat{F}_2^{(c)} + \int F_1 d\hat{F}_2^{(c)}\right| \\
&= \left|\int \left(\hat{F}_1^{(c)} - F_1\right) d\hat{F}_2^{(c)} + \int F_1 d\left(\hat{F}_2^{(c)} - F_2\right)\right| \\
&\leq \left|\int \left(\hat{F}_1^{(c)} - F_1\right) d\hat{F}_2^{(c)}\right| + \left|\int F_1 d\left(\hat{F}_2^{(c)} - F_2\right)\right| \\
&= \left|\int \left(\hat{F}_1^{(c)} - F_1\right) d\hat{F}_2^{(c)}\right| + \left|\int \left(F_2 - \hat{F}_2^{(c)}\right) dF_1\right| \\
&\leq \left\|\hat{F}_1^{(c)} - F_1\right\|_\infty + \left\|\hat{F}_2^{(c)} - F_2\right\|_\infty \xrightarrow{as} 0; \quad \text{whenever } n_1, n_2 \text{ are sufficiently large and } m_{gk} \text{ are bounded.}
\end{aligned}$$

A.3 | Proof of asymptotic equivalence for the unweighted estimator in (5)

In a first step, we decompose the estimator as follows:

$$\begin{aligned}
\hat{p}^{(u)} - p &= \int \hat{F}_1^{(u)} d\hat{F}_2^{(u)} - \int F_1 dF_2 \\
&= \int \left(\hat{F}_1^{(u)} - F_1\right) dF_2 + \int F_1 d\left(\hat{F}_2^{(u)} - F_2\right) + \int \left(\hat{F}_1^{(u)} - F_1\right) d\left(\hat{F}_2^{(u)} - F_2\right) \\
&= \overline{Y}_{2\cdot\cdot}^{(u)} - \overline{Y}_{1\cdot\cdot}^{(u)} + (1-2p) + \underbrace{\int \left(\hat{F}_1^{(u)} - F_1\right) d\left(\hat{F}_2^{(u)} - F_2\right)}_{=:A_n}.
\end{aligned}$$

Now,

$$A_n = \int \left(\hat{F}_1^{(u)} - F_1\right) d\left(\hat{F}_2^{(u)} - F_2\right) = \frac{1}{n_1}\frac{1}{n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} \int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right)$$

and, thus,

$$A_n^2 = \frac{1}{n_1^2}\frac{1}{n_2^2}\sum_{k=1}^{n_1}\sum_{k'=1}^{n_1}\sum_{l=1}^{n_2}\sum_{l'=1}^{n_2} \int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right) \int \left(\hat{F}_{1k'} - F_1\right) d\left(\hat{F}_{2l'} - F_2\right).$$

Notice that

$$E(A_n) = \frac{1}{n_1}\frac{1}{n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} E\left[\int (\hat{F}_{1k} - F_1) d(\hat{F}_{2l} - F_2)\right] = 0,$$

which follows by applying Fubini's theorem and that $\hat{F}_{1k}$ is an unbiased estimator of F_1. Then, to complete the proof, it suffices to show that $E(nA_n^2) = o(1)$ as n tends to infinity. To show this, we consider two cases.

Case 1: $k \neq k'$ or $l \neq l'$. For example, if $k \neq k'$, we know that $\mathbf{X}_{1k}$ and $\mathbf{X}_{1k'}$ are independent

$$\begin{aligned}
&E\left[\int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right) \int \left(\hat{F}_{1k'} - F_1\right) d\left(\hat{F}_{2l'} - F_2\right)\right] \\
&= E\left[\int \left(\hat{F}_{1k'} - F_1\right) d\left(\hat{F}_{2l'} - F_2\right) E\left[\int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right) \middle| \mathbf{X}_{1k'}, \mathbf{X}_{2l}, \mathbf{X}_{2l'}\right]\right].
\end{aligned}$$

Again, by applying Fubini's theorem and the fact that $\hat{F}_{1k}$ is an unbiased estimator of F_1, it can be seen that the inner expectation is zero. Therefore,

$$E\left[\int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right) \int \left(\hat{F}_{1k'} - F_1\right) d\left(\hat{F}_{2l'} - F_2\right)\right] = 0.$$

By symmetry, we get the same expectation when $l \neq l'$.

Case 2: $k = k'$ and $l = l'$. Since both distribution functions are bounded in $[0, 1]$,

$$\left|\int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right)\right| \leq 2 \;\Rightarrow\; \left(\int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right)\right)^2 \leq 4$$

and, thus,

$$\frac{1}{n_1^2}\frac{1}{n_2^2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} E\left[\int \left(\hat{F}_{1k} - F_1\right) d\left(\hat{F}_{2l} - F_2\right)\right]^2 = O\left(\frac{1}{n_1 n_2}\right).$$

From the two cases we have,

$$E\left(nA_n^2\right) = O\left(\frac{n}{n_1 n_2}\right) = O\left(n^{-1}\right)$$

since, by Assumption **A1**, $n/n_i = O(1)$ as $n \to \infty$ for $i = 1, 2$.

A.4 | Proof of Asymptotic Equivalence for the Weighted Estimator in (6)

With the same arguments as above, we first decompose the random variable $\sqrt{N}(\hat{p}^{(w)} - p)$ in the following way:

$$\begin{aligned}
\hat{p}^{(w)} - p &= \int \hat{F}_1^{(w)} dF_2^{(w)} - \int F_1 dF_2 \\
&= \int \left(\hat{F}_1^{(w)} - F_1\right) dF_2 + \int F_1 d\left(\hat{F}_2^{(w)} - F_2\right) + \int \left(\hat{F}_1^{(w)} - F_1\right) d\left(\hat{F}_2^{(w)} - F_2\right) \\
&= \overline{Y}_{2\cdot\cdot}^{(w)} - \overline{Y}_{1\cdot\cdot}^{(w)} + (1 - 2p) + A_N,
\end{aligned}$$

where

$$\begin{aligned}
A_N &= \int \left(\hat{F}_1^{(w)} - F_1\right) d\left(\hat{F}_2^{(w)} - F_2\right) \\
&= \frac{1}{m_1}\frac{1}{m_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} \int \left(m_{1k}\hat{F}_{1k} - m_{1k}F_1\right) d\left(m_{2l}\hat{F}_{2l} - m_{2l}F_2\right).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
A_N^2 = \frac{1}{m_1^2}\frac{1}{m_2^2}\sum_{k=1}^{n_1}\sum_{k'=1}^{n_1}\sum_{l=1}^{n_2}\sum_{l'=1}^{n_2} &\int \left(m_{1k}\hat{F}_{1k} - m_{1k}F_1\right) d\left(m_{2l}\hat{F}_{2l} - m_{2l}F_2\right) \\
&\cdot \left(m_{1k'}\hat{F}_{1k'} - m_{1k'}F_1\right) d\left(m_{2l'}\hat{F}_{2l'} - m_{2l'}F_2\right).
\end{aligned}$$

Here, it can be similarly shown that $E(A_N) = 0$ and, thus, it remains to show that $NE(A_N^2) = o(1)$.

Case 1: $k \neq k'$ or $l \neq l'$. The proof in this case is similar the one for unweighted estimator.

Case 2: $k = k'$ and $l = l'$. Here also, assuming the cluster sizes are uniformly bounded, ie, $m_{ik} \leq M_0 < \infty$ for all $i = 1, 2$ and $k = 1, \ldots, n_i$,

$$\left|\int \left(m_{1k}\hat{F}_{1k} - m_{m_{1k}}F_1\right) d\left(m_{2l}\hat{F}_{2l} - m_{2l}F_2\right)\right| \leq 2M_0^2 \Rightarrow \left(\int \left(m_{1k}\hat{F}_{1k} - m_{1k}F_1\right) d\left(m_{2l}\hat{F}_{2l} - m_{2l}F_2\right)\right)^2 \leq 4M_0^4.$$

Combining the two cases,

$$E\left(A_N^2\right) = O\left(\frac{n_1 n_2 M_0^2}{m_1^2 m_2^2}\right) = O\left(N^{-2}\right)$$

by Assumption **A2**. Therefore, $NE(A_N^2) = O(N^{-1})$.

A.5 | Proof of the consistency of unweighted variance estimator in (14)

First, note that $E[\widetilde{\sigma}_g^{2(u)} - \sigma_g^{2(u)}]^2 \to 0$. Therefore, it is enough to show that $E[\widehat{\sigma}_g^{2(u)} - \widetilde{\sigma}_g^{2(u)}]^2 \to 0$. To that end, note that[17]

$$E\left(\widehat{F}_s^{(u)}(X_{gk\ell}) - F_s(X_{gk\ell})\right)^2 = O(n^{-1}),$$

for $s \neq g$. Now, we are ready to prove the consistency

$$\begin{aligned}
E[\widehat{\sigma}_g^{2(u)} - \widetilde{\sigma}_g^{2(u)}]^2 &= \frac{1}{(n_g-1)^2} E\left[\sum_{k=1}^{n_g}\left\{(\overline{Z}_{gk\cdot}^{(u)} - \overline{Z}_{g\cdot\cdot}^{(u)})^2 - (\overline{Y}_{gk\cdot} - \overline{Y}_{g\cdot\cdot}^{(u)})^2\right\}\right]^2 \\
&= \frac{1}{(n_g-1)^2} E\left[\sum_{k=1}^{n_g}\left(\overline{Z}_{gk\cdot}^{(u)} - \overline{Z}_{g\cdot\cdot}^{(u)} - \overline{Y}_{gk\cdot} + \overline{Y}_{g\cdot\cdot}^{(u)}\right)\left(\overline{Z}_{gk\cdot}^{(u)} - \overline{Z}_{g\cdot\cdot}^{(u)} + \overline{Y}_{gk\cdot} - \overline{Y}_{g\cdot\cdot}^{(u)}\right)\right]^2 \\
&\leq \frac{1}{(n_g-1)^2} E\left[\sum_{k=1}^{n_g}\left(\overline{Z}_{gk\cdot}^{(u)} - \overline{Z}_{g\cdot\cdot}^{(u)} - \overline{Y}_{gk\cdot} + \overline{Y}_{g\cdot\cdot}^{(u)}\right)^2 \sum_{k=1}^{n_g}\left(\underbrace{\underline{Z}_{gk\cdot}^{(u)} - \underline{Z}_{g\cdot\cdot}^{(u)} + \underline{Y}_{gk\cdot} - \underline{Y}_{g\cdot\cdot}^{(u)}}_{\leq 2}\right)^2\right] \\
&\leq \frac{4n_g}{(n_g-1)^2}\sum_{k=1}^{n_g} E\left(\overline{Z}_{gk\cdot}^{(u)} - \overline{Z}_{g\cdot\cdot}^{(u)} - \overline{Y}_{gk\cdot} + \overline{Y}_{g\cdot\cdot}^{(u)}\right)^2 \\
&= \frac{4n_g}{(n_g-1)^2}\sum_{k=1}^{n_g} E\left\{\left(\overline{Z}_{gk\cdot}^{(u)} - \overline{Y}_{gk\cdot}\right)^2 - \underbrace{n_g\left(\underline{Z}_{g\cdot\cdot}^{(u)} - \underline{Y}_{g\cdot\cdot}^{(u)}\right)^2}_{\geq 0}\right\} \\
&\leq \frac{4n_g}{(n_g-1)^2}\sum_{k=1}^{n_g} E\left(\overline{Z}_{gk\cdot}^{(u)} - \overline{Y}_{gk\cdot}\right)^2 \\
&\leq \frac{4n_g M_0^2}{(n_g-1)^2}\sum_{k=1}^{n_g}\sum_{\ell=1}^{m_{gk}} E\left(\widehat{F}_s^{(u)}(X_{gk\ell}) - F_s(X_{gk\ell})\right)^2, \quad s \neq g \\
&= \frac{4n_g M_0^2}{(n_g-1)^2} O(n_g n^{-1}).
\end{aligned}$$

The first inequality is by Cauchy-Schwartz, while the last one is using the inequality $(\sum_{i=1}^q a_i)^2 \leq q^2 \sum_{i=1}^q a_i^2$ and that $1 \leq m_{gk} \leq M_0 < \infty$ for all $g = 1, 2$ and $k = 1, \ldots, n_g$.

A.6 | Proof of the consistency of weighted variance estimator in (14)

Note that $E[\widetilde{\sigma}_g^{2(w)} - \sigma_g^{2(w)}]^2 \to 0$ under the aforementioned assumptions. Therefore, it is sufficient to show that $E[\widehat{\sigma}_g^{2(w)} - \widetilde{\sigma}_g^{2(w)}]^2 \to 0$. Again, it can also be proved that[17]

$$E(\widehat{F}_s^{(w)}(X_{gk\ell}) - F_s(X_{gk\ell}))^2 = O(N^{-1}),$$

for $s \neq g$. Since $(1+K_g)^{-1} = 1 + O(m_g^{-1})$ and $(m_g - 2m_{gk})^{-1} = m_g^{-1} + O(m_g^{-2})$, we only need to show that $m_g^{-2} E[\sum_{k=1}^{n_g}(\tilde{Z}_{gk\cdot}^2 - \tilde{Y}_{gk\cdot}^2)]^2 \to 0$, where $\tilde{Z}_{gk} = Z_{gk\cdot}^{(w)} - m_{gk}\overline{Z}_{g\cdot\cdot}^{(w)}$ and $\tilde{Y}_{gk} = Y_{gk\cdot} - m_{gk}\overline{Y}_{g\cdot\cdot}^{(w)}$. Along the same lines of manipulations

as in the proof of the unweighted variance estimator,

$$
\begin{aligned}
E\left[\sum_{k=1}^{n_g}\left(\tilde{Z}_{gk\cdot}^2-\tilde{Y}_{gk\cdot}^2\right)\right]^2 &\le \frac{m_g(M_0+1)^2}{m_g^2}\sum_{k=1}^{n_g}m_{gk}^2E\left(Z_{gk\cdot}^{(w)}-\overline{Z}_{g\cdot\cdot}^{(w)}-Y_{gk\cdot}+\overline{Y}_{g\cdot\cdot}^{(w)}\right)^2\\
&\le \frac{(M_0+1)^2M_0}{m_g}\sum_{k=1}^{n_g}m_{gk}E\left[\left(Z_{gk\cdot}^{(w)}-Y_{gk\cdot}\right)-\left(\overline{Z}_{g\cdot\cdot}^{(w)}-\overline{Y}_{g\cdot\cdot}^{(w)}\right)\right]^2\\
&= \frac{(M_0+1)^2M_0}{m_g}E\left[\sum_{k=1}^{n_g}m_{gk}\left(Z_{gk\cdot}^{(w)}-Y_{gk\cdot}\right)^2-m_g\left(\overline{Z}_{g\cdot\cdot}^{(w)}-\overline{Y}_{g\cdot\cdot}^{(w)}\right)^2\right]\\
&\le \frac{(M_0+1)^2M_0}{m_g}\sum_{k=1}^{n_g}m_{gk}E\left(Z_{gk\cdot}^{(w)}-Y_{gk\cdot}\right)^2\\
&\le \frac{(M_0+1)^2M_0^4}{m_g}\sum_{k=1}^{n_g}\sum_{\ell=1}^{m_{ik}}E\left[F_s^{(w)}(X_{gk\ell})-F_s(X_{gk\ell})\right]^2,\quad \text{for } s\neq g\\
&= \frac{(M_0+1)^2M_0^4}{m_g}O(m_gN^{-1}).
\end{aligned}
$$