symmetry

MDPI

Article

Estimation and Testing of Wilcoxon–Mann–Whitney Effects in Factorial Clustered Data Designs

Kerstin Rubarth [1,2], Paavo Sattler [3], Hanna Gwendolyn Zimmermann [4] and Frank Konietschke [1,2,*]

[1] Institute of Biometry and Clinical Epidemiology, Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany; kerstin.rubarth@charite.de
[2] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany
[3] Department of Statistics, TU Dortmund University, TU Dortmund, 44221 Dortmund, Germany; paavo.sattler@tu-dortmund.de
[4] Experimental and Clinical Research Center, Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany; hanna.zimmermann@charite.de
* Correspondence: frank.konietschke@charite.de

Abstract: Clustered data arise frequently in many practical applications whenever units are repeatedly observed under a certain condition. One typical example for clustered data are animal experiments, where several animals share the same cage and should not be assumed to be completely independent. Standard methods for the analysis of such data are Linear Mixed Models and Generalized Estimating Equations—however, checking their assumptions is not easy, especially in scenarios with small sample sizes, highly skewed, count, and ordinal or binary data. In such situations, Wilcoxon–Mann–Whitney type effects are suitable alternatives to mean-based or other distributional approaches. Hence, no specific data distribution, symmetric or asymmetric, is required. Within this work, we will present different estimation techniques of such effects in clustered factorial designs and discuss quadratic- and multiple contrast type-testing procedures for hypotheses formulated in terms of Wilcoxon–Mann–Whitney effects. Additionally, the framework allows for the occurrence of missing data: estimation and testing hypotheses are based on all-available data instead of complete-cases. An extensive simulation study investigates the precision of the estimators and the behavior of the test procedures in terms of their type-I error control. One real world dataset exemplifies the applicability of the newly proposed procedures.

Keywords: clustered data; non-parametric statistics; rank-based procedures; repeated measures; missing data

check for updates

Citation: Rubarth, K.; Sattler, P.; Zimmermann, H.G.; Konietschke, F. Estimation and Testing of Wilcoxon–Mann–Whitney Effects in Factorial Clustered Data Designs. *Symmetry* **2022**, *14*, 244. https://doi.org/10.3390/sym14020244

Academic Editors: Jian-Qiang Wang and Florentin Smarandache

Received: 12 November 2021
Accepted: 16 December 2021
Published: 26 January 2022

1. Introduction

Clustered data are commonly encountered in medical research and various disciplines and occur whenever a subject is not only observed once under a certain condition, but multiple times. For instance, animals sharing the same cage, students in classes, skin irritations, etc., are different examples of clustered data. In these situations, subjects provide multiple possibly dependent observations (not necessarily equally sized). Standard methods for the analysis of independent observations (e.g., *t*-test, linear regression, Analysis of Variance (ANOVA), etc.) are not applicable in such scenarios. Ignoring that structure might result in bias (such as inflated type-I error rates and estimation bias) and therefore appropriate models are necessary for making inference. Reducing the clusters to a single point by, e.g., computing their mean or median, typically results in a loss of power and decreased precision of point estimates [1–3]. Furthermore, estimation of treatment effects becomes an issue because of presence of intra-cluster correlations and unequally sized clusters, see, e.g., Gao [4]. Under certain assumptions such as multivariate normality and linear relationships, Linear Mixed Models and Generalized Estimating Equations can be used. However, testing these assumptions is difficult in practice as noted by Fitzmaurice et al. [5] and Johnson and Wichern [6], especially when dealing with small sample sizes. Further, if count, ordinal, or highly skewed data are present, mean-based approaches are not applicable and thus, another type of measure is needed. On the contrary, Wilcoxon–Mann–Whitney-type effects $p = P(X < Y) + \frac{1}{2}P(X = Y)$ [7] are purely non-parametric quantities, which can be used for the definition of a treatment effect for metric, discrete, ordinal, and even dichotomous data in a unified way. Thus, the response variable is not assumed to be symmetrically distributed. Here X and Y represent two independent random variables coming from different populations. In the literature, p is also called *relative effect* [8] or *probabilistic index* [9,10], see Brunner et al. [11] for an overview. It is the aim of this paper to discuss different estimation techniques of such effects in factorial repeated measures designs with a clustered structure. We hereby differ between non-informative and informative cluster sizes by presenting weighted and unweighted estimators. Here, informative cluster sizes mean that the cluster sizes might be related to the outcome. In addition to estimation, we further introduce different statistical inference methods for testing hypotheses formulated in terms of these effects. All methods allow for the occurrence of missing data and take all-available data into consideration, which is a novelty since previous methods by Akritas et al. [12], Fong et al. [13], Domhof et al. [14] and Amro et al. [15] can only be used for testing hypotheses in terms of distribution functions in scenarios with missing data and do not allow for clustered data. Akritas and Brunner [16] and Brunner et al. [17] propose ranking procedures for testing hypotheses green formulated in terms of distribution functions in clustered data designs, see Brunner et al. [18] for an excellent overview. Thus, the aim of this work is to provide a framework for estimating Wilcoxon–Mann–Whitney-type effects, as well as to present test procedures for hypotheses formulated in terms of these effects in factorial repeated measures designs with clustered data. The paper is organized as follows. First, a real world example is introduced in Section 2 that motivates the development of the methods. Next, in Section 3 the factorial repeated measures model with a clustered data structure is introduced. Subsequently, point estimators and their asymptotic distributions are derived in Sections 4 and 5. Further, test procedures and multiple hypotheses in this framework are presented in Sections 6 and 7. The results of extensive simulation studies are presented in the following Section 8. Finally, the motivating example is analyzed by using the newly proposed methods in Section 9 and a discussion and conclusion about the findings is given in Section 10. All proofs can be found in the Appendix A.

2. Motivating Example

In order to motivate the development of methods for factorial repeated measures data with a clustered structure, we consider the secondary/exploratory outcome analysis of retinal thickness in the 'Sunphenon in progressive forms of multiple sclerosis' (SUPREMES) trial by Klumbies et al. [19], which was a relatively small clinical trial in progressive multiple sclerosis (MS) patients. MS is the most common autoimmune disorder of the central nervous system, affecting approximately 2.8 million people worldwide [20]. In MS, immune cells attack the myelin sheaths of nerve fibers, often resulting in neurodegeneration and thus permanent disability. In most cases, the disease starts with a relapsing–remitting course, followed by a progressive course around 15–20 years after diagnosis. Around 15% of patients feature a progressive course from onset of the disease [21]. MS is associated with visual impairment caused by optic nerve and posterior visual pathway damage, which can be quantified with optical coherence tomography based thickness measurements of the retinal nerve fiber layer, the combined ganglion cell, and inner plexiform layer (GCIP) and inner nuclear layer (INL) [22]. Retinal thickness analysis has been suggested as an outcome parameter in MS clinical trials [23]. In animal models of MS, epigallocatechin gallate (EGCG), which is an anti-inflammatory agent, indicated neuroprotective properties. The recent paper of Klumbies et al. [19] investigated the effect of EGCG on retinal thickness as an indicator for treatment response in progressive MS. For this motivating example, only the parameter peripapillary retinal nerve fiber layer (pRNFL) will be investigated.

Longitudinal OCT data were available from 31 patients, from which 15 patients were assigned to the intervention group and 16 to the control group, respectively. For most patients, both eyes were investigated, thus leading to possibly dependent observations, since assuming independence of the eyes from the same patient would be dubious. Further, missing values occur: From 61 patients in the SUPREMES trial, only 31 contributed to the final analysis and at 3-year follow-up, only 8 patients remained in the study. Table 1 displays the numbers of patients with pRNFL measurements in each group at each time point. Due to the extremely small number of measurements after 3 years, this time point was discarded from the analysis.

Table 1. Number of patients with pRNFL measurements in each group at baseline, 1-year follow-up, 2-year follow-up, and 3-year follow-up.

Group	First OCT	1-Year F/U	2-Year F/U	3-Year F/U
Verum	15	14	7	3
Placebo	16	16	11	5

Since the sample size was quite small and many missing values occurred, the authors tested—besides other hypotheses—whether there exists a statistical interaction of treatment and time using non-parametric analysis of longitudinal data in factorial experiments, as proposed by Brunner et al. [18]. One disadvantage of this procedure is that it cannot handle missing values and a clustered data structure. Therefore, Klumbies et al. [19] conducted a complete-case analysis and modeled the eyes as a second sub-plot factor besides the sub-plot factor time. In order to close this methodological gap, a general factorial model with repeated measurements, allowing for possibly correlated dependent replicates will be introduced in the next section.

3. The Factorial Repeated Measures Model with Missing Values

First, we study the general factorial repeated measures model without a clustered structure with independent random vectors

$$\begin{aligned} \boldsymbol{X}_{ik} &= ((\lambda_{i1k}, X_{i1k}), \dots, (\lambda_{idk}, X_{idk}))', i = 1, \dots, a; k = 1, \dots, n_i, \text{with} \\ \lambda_{isk} &= \begin{cases} 1, & X_{isk} \text{ is observed} \\ 0, & X_{isk} \text{ is missing.} \end{cases} \end{aligned} \tag{1}$$

Here, the random variable $X_{isk} \sim F_{is}, i = 1, \dots, a; s = 1, \dots, d; k = 1, \dots, n_i$ represents the s-th repeated measurement of the k-th subject in group i.

To account for metric, discrete, ordinal and ordered categorical data in a unified way, we use the *normalized version* of the distribution function

$$F_{is}(x) = P(X_{isk} < x) + \frac{1}{2} P(X_{isk} = x),$$

which is the average of the left and the right continuous version of the distribution function $F_{is}^{-} = P(X_{is1} < x)$ and $F_{is}^{+} = P(X_{is1} \leq x)$, which was first introduced by Ruymgaart [24].

In model (1), the numbers of non-missing observations under time-point s in group i, the overall sample size and the minimal number of observations over all groups and time points are given by

$$\lambda_{is} = \sum_{k=1}^{n_i} \lambda_{isk} \text{ , } N = \sum_{i=1}^{a} n_i \text{ and } \lambda_{min} = min(\lambda_{11}, \dots, \lambda_{ad}).$$

We propose to use *unweighted relative effects*

$$p_{is} = \int G dF_{is} = P(Z < X_{is1}) + \frac{1}{2}P(Z = X_{is1}), i = 1, \ldots, a; s = 1, \ldots, d, \tag{2}$$

with $G = \frac{1}{ad}\sum_{i=1}^{a}\sum_{s=1}^{d} F_{is}$ being the unweighted mean distribution function and $Z \sim G$, independent of X_{isk}. The relative effect p_{is} models the relationship of the distribution F_{is} to the average distribution G. If $p_{is} > p_{jt}$, then data coming from F_{is} tend to be larger than data coming from F_{jt}. If $p_{is} = p_{jt}$, then there is no tendency to greenlarger nor smaller values between the two distributions. For more information on unweighted and weighted relative effects, we refer to Brunner et al. [25] and Brunner et al. [26]. In the following, we always refer to unweighted relative effects, if relative effects are mentioned.

General Factorial Model with Clustered Data

In the following we introduce a general factorial longitudinal model with clustered data. In comparison with model (1), we observe random vectors $\boldsymbol{X}_{ik} = (\boldsymbol{X}_{i1k}, .., \boldsymbol{X}_{idk})'$ with

$$\begin{aligned} &\boldsymbol{X}_{isk} = (\lambda_{isk}, (X_{isk1}, \ldots, X_{iskm_{isk}}))', \text{ where} \\ &X_{isku} \sim F_{is}, \ u = 1, \ldots, m_{isk}, \end{aligned} \tag{3}$$

and m_{isk} denotes the number of possibly dependent replicates of subject k in group i at time s and $m_{is} = \sum_{k=1}^{n_i} m_{isk}\lambda_{isk}$ denotes the total number of possibly dependent replicates in group i at time s. Thus, the number of dependent replicates may vary for each subject and may not be under experimental control. Note, that we do not assume any correlation structure of the dependent replicates. Similarly as in (2), we define the relative effect as

$$p_{is} = \int G dF_{is} = P(Z < X_{is11}) + \frac{1}{2}P(Z = X_{is11})$$

with

$$F_{is}(x) = P(X_{isku} < x) + \frac{1}{2}P(X_{isku} = x).$$

Note, that model (1) is contained within this model as a special case with $m_{isk} \equiv 1$. In order to derive asymptotic results, we impose the following model assumptions:

Assumption 1.
- ***A1.1**: $\frac{n_i}{N} \to \kappa_i \in (0, 1]$;*
- ***A1.2**: $N \to \infty$ such that $\frac{N}{\lambda_{min}} < N_0$, N_0 being a fixed constant;*
- ***A1.3**: $N \to \infty$ such that $m_{is} < M_0$, M_0 being a fixed constant.*

Assumption **A1.1** ensures that none of the groups vanishes asymptotically, whereas Assumptions **A1.2** and **A1.3** ensure that the total sample size N and number of clustered observations m_{is} is bounded. In the following section, estimators for the relative effect $\boldsymbol{p}$ will be derived.

4. Estimators and Their Asymptotic Distributions

We will first study estimators of relative effects in factorial repeated measures designs without clustered data (i.e., $m_{isk} \equiv 1$). In order to account for possible missing values, we define the empirical distribution function of the data under time point s in group i as the average of the all-available data by

$$\widehat{F}_{isk}(x) = c(x - X_{isk})\lambda_{isk} \quad \text{resulting in} \quad \widehat{F}_{is}(x) = \frac{1}{\lambda_{is}}\sum_{k=1}^{n_i}\widehat{F}_{isk}(x).$$

Here, $c(u) = 0, \frac{1}{2}, 1$, if $u <, =, > 0$. By plugging in the empirical counterparts $\widehat{F}_{is}$, we obtain

$$\widehat{G}(x) = \frac{1}{ad} \sum_{i=1}^{a} \sum_{s=1}^{d} \widehat{F}_{is}(x),$$

$$\widehat{p}_{is} = \int \widehat{G} d\widehat{F}_{is} = \frac{1}{\lambda_{is}} \sum_{k=1}^{n_i} \frac{\lambda_{isk}}{ad} \sum_{j=1}^{a} \sum_{t=1}^{d} \sum_{\ell=1}^{n_j} \frac{\lambda_{jt\ell}}{\lambda_{jt}} c(X_{isk} - X_{jt\ell}) \tag{4}$$

and define

$$\widehat{\boldsymbol{p}} = (\widehat{p}_{11}, \ldots, \widehat{p}_{1d}, \widehat{p}_{21}, \ldots \widehat{p}_{ad})'.$$

4.1. Effect Estimation in Factorial Designs with Clustered Data

To generalize the estimation of empirical distribution functions and relative effects to the case of clustered data, we follow the idea of Roy et al. [1] who proposed two different approaches for estimating the distribution functions by using the cluster sizes as weighting schemes. In the first version of the estimator of the relative effect $\boldsymbol{p}$, larger clusters add more weight to the estimation than smaller ones and in the second version, all clusters add the same weight to the estimation, disregarding their size. Analogously to Roy et al. [1] the estimators are called *unweighted* and *weighted* estimators, respectively. Note that Obuchowski [27] also used the weighted version in the two-sample case.

The two different versions of the empirical distribution functions (unweighted and weighted) are defined as follows:

$$\widehat{F}_{is}^{(v_1)}(x) = \frac{1}{\lambda_{is}} \sum_{k=1}^{n_i} \frac{1}{m_{isk}} \sum_{u=1}^{m_{isk}} c(x - X_{isku})\lambda_{isk}$$
$$\widehat{F}_{is}^{(v_2)}(x) = \frac{1}{m_{is}} \sum_{k=1}^{n_i} \sum_{u=1}^{m_{isk}} c(x - X_{isku})\lambda_{isk}.$$

$\widehat{F}_{is}^{(v_1)}(x)$ is the unweighted estimator of $F_{is}(x)$, where the average of the count function is calculated separately for each cluster and these averages are then again averaged. $\widehat{F}_{is}^{(v_2)}(x)$ is the weighted estimator of $F_{is}(x)$ where the counts are averaged over all observations. In order to write the estimators in a unified way, we define weights

$$w_{isk}^{v_1} = \frac{1}{\lambda_{is} m_{isk}} \quad \text{and} \quad w_{isk}^{v_2} = \frac{1}{m_{is}},$$

then an estimator for $F_{is}(x)$ and $G(x)$ is given by

$$\widehat{F}_{is}^{*}(x) = \sum_{k=1}^{n_i} \sum_{u=1}^{m_{isk}} w_{isk}^{*} c(x - X_{isku})\lambda_{isk}, * \in \{v_1, v_2\}$$

and

$$\widehat{G}^{*} = \frac{1}{ad} \sum_{i=1}^{a} \sum_{s=1}^{d} \widehat{F}_{is}^{*} = \frac{1}{ad} \sum_{i=1}^{a} \sum_{s=1}^{d} \sum_{k=1}^{n_i} \lambda_{isk} \sum_{u=1}^{m_{isk}} w_{isk}^{*} c(x - X_{isku}).$$

It then follows that an estimator of p_{is} is given by

$$\widehat{p}_{is}^{*} = \int \widehat{G}^{*} d\widehat{F}_{is}^{*} = \frac{1}{ad} \sum_{j=1}^{a} \sum_{t=1}^{d} \sum_{k=1}^{n_i} \sum_{u=1}^{m_{isk}} \lambda_{isk} w_{isk}^{*} \widehat{F}_{jt}^{*}(X_{isku}) \tag{5}$$

$$= \frac{1}{ad} \sum_{k=1}^{n_i} \sum_{j=1}^{a} \sum_{t=1}^{d} \sum_{\ell=1}^{n_j} \sum_{u=1}^{m_{isk}} \sum_{v=1}^{m_{jt\ell}} \lambda_{isk} \lambda_{jt\ell} w_{isk}^{*} w_{jt\ell}^{*} c(X_{isku} - X_{jt\ell v}). \tag{6}$$

Note that in order to derive the theory of these estimators, the weights need to fulfill the following properties

Proposition 1. *If Assumption **A1.3** is fulfilled, it holds that*

- ***A2.1**:* $\lambda_{isk} w^*_{isk} m_{isk} \leq \mathcal{O}\left(\frac{1}{\lambda_{min}}\right)$;
- ***A2.2**:* $\sum_{k=1}^{n_i} \lambda_{isk} w^*_{isk} m_{isk} = 1$.

Furthermore note, that the application of all weights which fulfill both properties is theoretically possible. For example, Zou [28] developed an 'optimal' estimator, which incorporates information on cluster sizes and intra-cluster correlations by a mixed model approach.

First, we will study the asymptotic properties of general estimators for $\boldsymbol{p}$ in the following proposition.

Proposition 2. *The estimator* $\widehat{\boldsymbol{p}}^* = \left(\widehat{p}^*_{11}, \ldots, \widehat{p}^*_{1d}, \widehat{p}^*_{21}, \ldots, \widehat{p}^*_{ab}\right)'$ *is asymptotically unbiased and strongly consistent, i.e.,*

1. $E(\widehat{\boldsymbol{p}}^*) = \boldsymbol{p} + \mathcal{O}(\frac{1}{\lambda_{min}})$;
2. $\widehat{\boldsymbol{p}} - \boldsymbol{p} \xrightarrow{a.s.} 0, \lambda_{min} \to \infty$.

Subsequently, the asymptotic distribution of the statistic $\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p})$ will be derived. It will be indicated in the next theorem that $\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p})$ has asymptotically under **A1.1** and **A1.2**, the same distribution as the random vector $\sqrt{N}\boldsymbol{B}^*$, with

$$\sqrt{N}\boldsymbol{B}^* = \sum_{h=1}^{a} \sqrt{N}\boldsymbol{B}^*_h = \sum_{h=1}^{a} \left(\frac{\sqrt{N}}{n_h} \sum_{k=1}^{n_h} (\boldsymbol{\Psi}^*_{hk} - E(\boldsymbol{\Psi}^*_{hk})) \right)$$

based on random variables defined as

$$\Psi^*_{is,hk} := \begin{cases} -\frac{n_h}{ad} \sum_{t=1}^{d} \sum_{u=1}^{m_{htk}} \lambda_{htk} w^*_{htk} F_{is}(X_{htku}) \text{ , for } h \neq i \\ \frac{n_h}{ad} \sum_{j\neq i}^{a} \sum_{t=1}^{d} \sum_{u=1}^{m_{isk}} \lambda_{isk} w^*_{isk} F_{jt}(X_{isku}) \\ +\frac{n_h}{ad} \sum_{t=1}^{d} \left(\sum_{u=1}^{m_{isk}} \lambda_{isk} w^*_{isk} F_{it}(X_{isku}) \right. \\ \left. - \sum_{u=1}^{m_{itk}} \lambda_{itk} w^*_{itk} F_{is}(X_{itku}) \right) \text{, else.} \end{cases}$$

The expectation of $\boldsymbol{\Psi}^*_{hk}$ can be written as

$$\beta^*_{is,ik} := E(\Psi^*_{is,hk}) = \begin{cases} -\frac{n_h}{ad} \sum_{t=1}^{d} \lambda_{htk} m_{htk} w^*_{htk} p^{(is,ht)} \text{ , for } h \neq i \\ \frac{n_h}{ad} \sum_{j\neq i}^{a} \sum_{t=1}^{d} m_{isk} \lambda_{isk} w^*_{isk} p^{(jt,is)}) \\ +\frac{n_h}{ad} \sum_{t=1}^{d} \left(m_{isk} \lambda_{isk} w^*_{isk} p^{(it,is)} \right. \\ \left. - m_{itk} \lambda_{itk} w^*_{itk} p^{(is,it)} \right) \text{ , else,} \end{cases}$$

with $p^{(is,ht)} := \int F_{is} dF_{ht}$ denoting pairwise relative effects between groups i and h and time points s and t.

Theorem 1. *Let* $\sqrt{N}\boldsymbol{B}^* = \sum_{h=1}^{a} \sqrt{N}\boldsymbol{B}^*_h = \sum_{h=1}^{a} \sqrt{N}\left(B^*_{11,h}, \ldots, B^*_{ad,h}\right)'$ *be the vector of the random variables* $\sqrt{n}B_{is}$, $i = 1, \ldots, d$; $s = 1, \ldots, d$. *If **A1.1** and **A1.2** hold true, then*

$$||\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p}) - \sqrt{N}\boldsymbol{B}^*||_2^2 = \mathcal{O}\left(\frac{1}{N}\right).$$

It follows that the asymptotic covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p})$ is given by

$$\boldsymbol{V}^*_N = Cov(\sqrt{N}\boldsymbol{B}^*).$$

The asymptotic multivariate normality of the linear statistic $\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p})$ is given in the next theorem.

Theorem 2. *Under Assumptions **A1.1** and **A1.2**, the statistic $\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p})$ follows asymptotically, as $N \to \infty$, a multivariate normal distribution with expectation $\mathbf{0}$ and covariance matrix $\boldsymbol{V}_N^*$.*

However, this covariance matrix is mostly unknown in practical applications and must be estimated in order to be able to make statistical inferences. In Section 5 we will derive a consistent and positive semi-definite estimator of the covariance matrix.

4.2. Informative Cluster Sizes

In many applications, the cluster sizes m_{isk} (might) depend on the outcome of interest, i.e.,

Assumption 2.

$$E(X_{isku}) \neq E(X_{isku}|m_{isk}), i = 1, \ldots, a; s = 1, \ldots, d; k = 1, \ldots, n_i; u = 1, \ldots, m_{isk},$$

Which makes them *non-ignorable* or *informative* [29]. As an example, consider the periodontal disease (an inflammation of the gums and bone that surround and support the teeth) study [30]. Severe periodontitis ends in the falling out of teeth and, thus, cluster sizes (patient’s teeth) depend on the clinical outcome. Hoffmann et al. [29], among others, suggest a *Within-Cluster-Resampling* (WCR) method for the analysis of informative clustered binary data. This approach is also applicable in the rather general model considered here and will be described in the following:

A randomly chosen observation X_{isk}^q is sampled from cluster $\mathbf{X}_{isk}$. This is done for each of all $N * d$ clusters, resulting in a dataset involving single observations only. The latter is repeated Q times, e.g., Q = 10,000, and for each of the Q datasets, the vector of relative effects $\mathbf{p}$ is estimated by adapting Equation (4):

$$\widehat{p}_{is}^{\Delta,q} = \frac{1}{\lambda_{is}} \sum_{k=1}^{n_i} \frac{\lambda_{isk}}{ad} \sum_{j=1}^{a} \sum_{t=1}^{d} \sum_{\ell=1}^{n_j} \frac{\lambda_{jt\ell}}{\lambda_{jt}} c(X_{isk}^q - X_{jt\ell}^q).$$

An estimator and its asymptotic distribution is given in the following theorem.

Theorem 3. *Let*

$$\widehat{\boldsymbol{p}}^{\Delta} = \frac{1}{Q} \sum_{q=1}^{Q} \widehat{\boldsymbol{p}}^{\Delta,q}$$

denote the Within-Cluster-Resampling based estimator. If $N \to \infty$, then

$$\sqrt{N}\left(\widehat{\boldsymbol{p}}^{\Delta} - \boldsymbol{p}\right) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\Delta}),$$

where $\boldsymbol{\Sigma}^{\Delta}$ is finite.

A consistent variance is estimator is provided in the next theorem.

Theorem 4. *Let N and $\widehat{\boldsymbol{p}}^{\Delta}$ be defined as in Theorem 3. Then, an estimator of $\boldsymbol{\Sigma}^{\Delta}$ is given by*

$$\widehat{\boldsymbol{\Sigma}}^{\Delta} = \widehat{Var}\left(\sqrt{N}\left(\widehat{\boldsymbol{p}}^{\Delta} - \boldsymbol{p}\right)\right) = N\left(\frac{1}{Q} \sum_{q=1}^{Q} \widehat{\boldsymbol{\Sigma}}_q - \frac{Q-1}{Q} S_{\boldsymbol{p}}^2\right), \tag{7}$$

where $\widehat{\boldsymbol{\Sigma}}_q$ is the estimated covariance matrix from the q-th analysis (see the following chapter for the derivation of an estimator) and

$$S_{\boldsymbol{p}}^2 = \frac{1}{Q-1}\sum_{q=1}^{Q}\left(\widehat{\boldsymbol{p}}^{\Delta,q} - \widehat{\boldsymbol{p}}^{\Delta}\right)\left(\widehat{\boldsymbol{p}}^{\Delta,q} - \widehat{\boldsymbol{p}}^{\Delta}\right)'$$

is the estimated covariance matrix among the Q resample-based estimates $\widehat{\boldsymbol{p}}^{\Delta,q}$. Then, $\widehat{\boldsymbol{\Sigma}}^{\Delta}$ is consistent for $\boldsymbol{\Sigma} = Var\left(\sqrt{N}\left(\widehat{\boldsymbol{p}}^{\Delta} - \boldsymbol{p}\right)\right)$.

The proofs of Theorem 3 and 4 can be found in the appendix of Hoffmann et al. [29]. The WCR-approach proposed by Hoffmann et al. [29] is computationally intensive and could possibly lead to negative variance estimators due to the subtraction in the variance estimation in Equation (7). Hoffmann et al. [29] noted, that this occurs rarely and concluded that in these scenarios, the number of resampled datasets Q or the number of clusters N may be too small for making inferences. However, the WCR-based approach is equivalent to the unweighted estimation of the relative effects $\boldsymbol{p}$ as proposed by Roy et al. [1]—in both analysis all clusters are given equal weight, regardless of their size. Thus, the use of the unweighted estimator should be preferred over the WCR-based approach since its computation is less intensive and always leads to positive variance estimators. However, it should be noted that Assumption 2 of ignorable cluster-sizes is never imposed during the development of the theory in this work. Therefore, all weighting schemes that fulfill Assumptions **A2.1** and **A2.2** can be applied in case of non-ignorable cluster sizes-however, the resulting estimators have a different interpretation.

5. Estimation of the Covariance Matrix

Now, an estimator of the covariance matrix $\boldsymbol{V}_N^*$ is derived. Similarly, as in Rubarth et al. [31], the random variables $\Psi_{is,hk}^*$ are not observable. Otherwise, an estimator of $\boldsymbol{V}_N^*$ would be given by

$$\widetilde{\boldsymbol{V}}_N^* = \sum_{h=1}^{a}\frac{N}{n_h}\widetilde{\boldsymbol{V}}_{N,h}^*$$

with $\widetilde{\boldsymbol{V}}_{N,h}^* = \frac{1}{n_h-1}\sum_{k=1}^{n_h}(\boldsymbol{\Psi}_{hk}^* - \boldsymbol{\beta}_{hk}^*)(\boldsymbol{\Psi}_{hk}^* - \boldsymbol{\beta}_{hk}^*)'$. Therefore, we replace the unknown $\Psi_{is,hk}^*$ with observable random variables. Define the vectors $\widehat{\boldsymbol{\Psi}}_{hk}^* = \left(\widehat{\Psi}_{11,hk}^*, \ldots, \widehat{\Psi}_{ad,hk}^*\right)'$ with

$$\widehat{\Psi}_{is,hk}^* := \begin{cases} -\frac{n_h}{ad}\sum_{t=1}^{d}\sum_{u=1}^{m_{htk}}\lambda_{htk}w_{htk}^*\widehat{F}_{is}^*(X_{htku})\text{ , for } h \neq i \\ \frac{n_h}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d}\sum_{u=1}^{m_{isk}}\lambda_{isk}w_{isk}^*\widehat{F}_{jt}^*(X_{isku}) \\ +\frac{n_h}{ad}\sum_{t=1}^{d}\Big(\sum_{u=1}^{m_{isk}}\lambda_{isk}w_{isk}^*\widehat{F}_{it}^*(X_{isku}) \\ -\sum_{u=1}^{m_{itk}}\lambda_{itk}w_{itk}^*\widehat{F}_{is}^*(X_{itku})\Big)\text{ , else} \end{cases}$$

and expectation values

$$\widehat{\beta}_{is,hk}^* := E(\widehat{\Psi}_{is,hk}^*) = \begin{cases} -\frac{n_h}{ad}\sum_{t=1}^{d}\lambda_{htk}m_{htk}w_{htk}^*\widehat{p}^{*(is,ht)}\text{ , for } h \neq i \\ \frac{n_h}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d}m_{isk}\lambda_{isk}w_{isk}^*\widehat{p}^{*(jt,is)} \\ +\frac{n_h}{ad}\sum_{t=1}^{d}\Big(m_{isk}\lambda_{isk}w_{isk}^*\widehat{p}^{*(it,is)} \\ -m_{itk}\lambda_{itk}w_{itk}^*\widehat{p}^{*(is,it)}\Big)\text{ , else,} \end{cases}$$

where $\widehat{p}^{*(is,ht)} = \int \widehat{F}^*_{is} d\widehat{F}^*_{ht} = \sum_{k=1}^{n_h} \sum_{u=1}^{m_{htk}} \lambda_{htk} w^*_{htk} \widehat{F}^*_{is}(X_{htku})$ denote the estimators of the pairwise relative effects $p^{(is,ht)}$. Finally, an estimator for the unknown covariance matrix $\boldsymbol{V}^*_{N,h}$ is given by

$$\widehat{\boldsymbol{V}}^*_{N,h} = \frac{1}{n_h - 1} \sum_{k=1}^{n_h} \left(\widehat{\boldsymbol{\Upsilon}}^*_{hk} - \widehat{\boldsymbol{\beta}}^*_{hk} \right) \left(\widehat{\boldsymbol{\Upsilon}}^*_{hk} - \widehat{\boldsymbol{\beta}}^*_{hk} \right)'$$

and an estimator for $\boldsymbol{V}^*_N := \sum_{h=1}^{a} \kappa_h^{-1} \boldsymbol{V}^*_{N,h}$ is given by

$$\widehat{\boldsymbol{V}}^*_N := \sum_{h=1}^{a} \frac{N}{n_h} \widehat{\boldsymbol{V}}^*_{N,h}.$$

Its properties are presented in the next theorem.

Theorem 5. *For $N \to \infty$, such that **A1.1** and **A1.3** are fulfilled, it holds*

1. $\widehat{\boldsymbol{V}}^*_{N,h}$ *and* $\widehat{\boldsymbol{V}}^*_N$ *are positive semi-definite;*
2. $\boldsymbol{V}^*_{N,h} - \widehat{\boldsymbol{V}}^*_{N,h} \xrightarrow{a.s.} \mathbf{0}$;
3. $\boldsymbol{V}^*_N - \widehat{\boldsymbol{V}}^*_N \xrightarrow{a.s.} \mathbf{0}$.

6. Multiple Hypotheses

In this section, the formulation of hypotheses for main- and interaction effects in the factorial repeated measures framework will be outlined. Let $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q)' \in R^{q \times ad}$ be an arbitrary contrast matrix and let

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

be a family of hypotheses, where $\boldsymbol{c}_\ell$ denotes the ℓ-th row vector of $\boldsymbol{C}$. The decision which contrast matrix is appropriate depends on the specific research questions. Well known types of contrast matrices are the *Tukey*-type contrast matrix, used for all-pairwise comparisons or the *Dunnett*-type contrast matrix, which is used for the comparison of several groups to one control group. User-specified contrast matrices can also be applied, as long as they have the property of a contrast matrix, which is, that each row of the contrast matrix $\boldsymbol{C}$ sums up to 0 (i.e., $\sum_{m=1}^{ad} \boldsymbol{c}_{\ell,m} = 0 \forall \ell = 1, \ldots, q$).

Since the layout in this paper is multifactorial, it is briefly demonstrated how to define appropriate contrast matrices for testing *main effects* of group membership and time and *interaction effect* between group membership and time.

- *Main effect group membership G* In order to make comparisons in terms of group membership, it is necessary to center and average over the repeated measures. Thus, a contrast matrix to test for no group effect will be defined as

$$\boldsymbol{C}_G := \boldsymbol{C}_g \left(\boldsymbol{P}_a \otimes \frac{1}{d} \mathbf{1}'_d \right) \in \mathbb{R}^{q \times ad},$$

  with $\boldsymbol{C}_g$ being a contrast matrix for the group effect with a time structure.
- *Main effect time T* Similarly for the time effect, the measurements across the groups need to be centered and averaged, leading to a contrast matrix to test for no time effect as

$$\boldsymbol{C}_T := \boldsymbol{C}_t \left(\frac{1}{a} \mathbf{1}'_a \otimes \boldsymbol{P}_d \right) \in \mathbb{R}^{q \times ad}.$$

  Again, $\boldsymbol{C}_t$ denotes a contrast matrix for the effect over time without the group structure.
- *Interaction effect $G \times T$* For the test of no interaction between group membership and time, the centering matrix

$$\boldsymbol{C}_{GT} = \boldsymbol{P}_a \otimes \boldsymbol{P}_d \in \mathbb{R}^{ad \times ad}$$

  will be used.

7. Test Statistics

In this section, we will present different test procedures for testing global and multiple hypotheses concerning the null hypothesis $H_0^p : \boldsymbol{C}\boldsymbol{p} = \boldsymbol{0}$, with $\boldsymbol{C}$ being an appropriate contrast matrix tailored to the specific research question. First, we propose two quadratic test procedures, a Wald-type statistic (WTS) and an ANOVA-type statistic (ATS) as already described by Brunner et al. [17], Domhof et al. [14], and Rubarth et al. [31]. These procedures can only be used to test the global null hypothesis and cannot be inverted to obtain (simultaneous) confidence intervals. Therefore, we will present a Multiple Contrast Test Procedure (MCTP), which has been introduced by Konietschke et al. in a general non-parametric factorial framework [32] and Rubarth et al. [31] for the case of incompletely observed data. Using this procedure, multiple hypotheses can be tested simultaneously and adjusted confidence intervals and p-values are directly obtained.

7.1. Quadratic Test Procedures

Following Konietschke et al. [33] and Rubarth et al. [31], we consider the Wald-type statistic (WTS)

$$Q_N^* = N\widehat{\boldsymbol{p}}'^*\boldsymbol{C}'\left[\boldsymbol{C}\widehat{\boldsymbol{V}}_N^*\boldsymbol{C}'\right]^+\boldsymbol{C}\widehat{\boldsymbol{p}}^*,$$

which can be approximated by a $\chi^2_{\widehat{f}}$ distribution with $\widehat{f} = rank(\boldsymbol{C}\widehat{\boldsymbol{V}}_N^*\boldsymbol{C}')$ degrees of freedom (see the discussion on further assumptions on $\boldsymbol{V}_N^*$ in Brunner et al. [25]). Here, $[.]^+$ denotes the Moore-Penrose inverse of a matrix. However, simulation studies by Konietschke et al. [33], Domhof et al. [14] and Rubarth et al. [31] indicate, that the WTS is very liberal in small or moderate sample size scenarios. Therefore, Akritas et al. [34] and Brunner et al. [25], among others, approximate the (asymptotic) distribution of

$$A_N^* = \frac{N}{tr(\boldsymbol{M}\boldsymbol{V}_N^*)}\widehat{\boldsymbol{p}}'^*\boldsymbol{M}\widehat{\boldsymbol{p}}^*$$

by a scaled χ^2_f/f distribution with

$$f = \frac{[tr(\boldsymbol{M}\boldsymbol{V}^*)]^2}{tr(\boldsymbol{M}\boldsymbol{V}_N^*\boldsymbol{M}\boldsymbol{V}_N^*)}$$

degrees of freedom. Here, $\boldsymbol{M} = \boldsymbol{C}'\left[\boldsymbol{C}\boldsymbol{C}'\right]^-\boldsymbol{C}$ and $\left[\boldsymbol{C}\boldsymbol{C}'\right]^-$ denotes a generalized inverse of $\boldsymbol{C}\boldsymbol{C}'$. Since $\boldsymbol{M}$ is a projection matrix, it holds that $\boldsymbol{M}\boldsymbol{p} = \boldsymbol{0} \iff \boldsymbol{C}\boldsymbol{p} = \boldsymbol{0}$. The unknown traces $tr(\boldsymbol{M}\boldsymbol{V}_N^*)$ and $tr(\boldsymbol{M}\boldsymbol{V}_N^*\boldsymbol{M}\boldsymbol{V}_N^*)$ are estimated by replacing $\boldsymbol{V}_N^*$ with $\widehat{\boldsymbol{V}}_N^*$, see Brunner et al. [17] for the derivation.

7.2. Multiple Contrast Test Procedure

To overcome the above outlined disadvantages of the quadratic test procedures, Konietschke et al. [32] proposed a rank-based MCTP for factorial designs, whereas Rubarth et al. [31] proposed a procedure for repeated measures designs with missing values.

Consider the ℓ-th individual null hypothesis $H_0^{(\ell)} : \boldsymbol{c}_\ell'\boldsymbol{p} = 0$ and the corresponding test statistic

$$T_\ell^* = \sqrt{N}\frac{\boldsymbol{c}_\ell'(\widehat{\boldsymbol{p}}^* - \boldsymbol{p})}{\sqrt{\boldsymbol{c}_\ell'\widehat{\boldsymbol{V}}_N^*\boldsymbol{c}_\ell}}$$

with $\boldsymbol{c}_\ell'$ being the ℓ-th row vector of $\boldsymbol{C}$. All test statistics are collected in the vector

$$\boldsymbol{T}^* = (T_1^*, \ldots, T_q^*)'.$$

Note that the test statistics T_ℓ^* and $T_m^*(\ell \neq m)$ are not necessarily independent depending on the chosen contrast and the repeated measures.

The distribution of T_ℓ^* is asymptotically standard normal. It follows then from Theorem 2 and Slutzky's theorem that $\boldsymbol{T}^*$ follows, asymptotically, as $N \to \infty$, a standard multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix

$$\boldsymbol{R}^* = \boldsymbol{D}^{*,-1/2}\boldsymbol{C}\boldsymbol{V}_N^*\boldsymbol{C}'\boldsymbol{D}^{*,-1/2},$$

with $\boldsymbol{D}^*$ being a diagonal matrix of the diagonal elements of $\boldsymbol{C}\boldsymbol{V}_N^*\boldsymbol{C}'$. For large samples, the local null hypothesis $H_0^{(\ell)} : \boldsymbol{c}_\ell'\boldsymbol{p} = 0$ will be rejected if $|T_\ell^*| \geq z_{1-\alpha,2,\boldsymbol{R}^*}$. Here, $z_{1-\alpha,2,\boldsymbol{R}^*}$ denotes the two-sided $(1-\alpha)$ equicoordinate quantile of the $\mathcal{N}(\mathbf{0}, \boldsymbol{R}^*)$ distribution [35]. By inverting the corresponding test statistic, simultaneous confidence intervals for the effects $\delta_\ell = \boldsymbol{c}_\ell'\boldsymbol{p}$ can be obtained by

$$CI_\ell = \left[\boldsymbol{c}_\ell'\widehat{\boldsymbol{p}}^* \mp \frac{z_{1-\alpha,2,\boldsymbol{R}^*}}{\sqrt{N}}\sqrt{\boldsymbol{c}_\ell'\widehat{\boldsymbol{V}}_N^*\boldsymbol{c}_\ell}\right].$$

It follows directly, that the global null hypothesis $H_0^p : \boldsymbol{C}\boldsymbol{p} = \mathbf{0}$ will be rejected, if $T_0 = max\{|T_1^*|, \ldots, |T_q^*|\} \geq z_{1-\alpha,2,\boldsymbol{R}^*}$. Analogously as in Konietschke et al. [36] and Rubarth et al. [31], the correlation matrix is unknown but can be consistently estimated by

$$\widehat{R}_N = \widehat{\boldsymbol{D}}^{*,-1/2}\boldsymbol{C}\widehat{\boldsymbol{V}}_N^*\boldsymbol{C}'\widehat{\boldsymbol{D}}^{*,-1/2}.$$

Again, $\widehat{\boldsymbol{D}}^*$ is denoted as the diagonal matrix obtained from the diagonal elements of $\boldsymbol{C}\widehat{\boldsymbol{V}}_N^*\boldsymbol{C}'$. We note that the method controls the family wise error rate α in the strong sense asymptotically. However, the proposed procedure is only valid for large sample sizes and the convergence of $\boldsymbol{T}^*$ to its asymptotic distribution is rather slow [32]. Therefore, we follow Konieschke et al. [32] who proposed a small sample approximation by using a central multivariate $T(\nu, \mathbf{0}, \widehat{\boldsymbol{R}}^*)$ distribution, with ν degrees of freedom and correlation matrix $\widehat{\boldsymbol{R}}^*$. We define for each linear contrast $\boldsymbol{c}_\ell' = (c_{\ell 11}, \ldots, c_{\ell ad})'$, $\ell = 1, .., q$ random variables $\Phi_{\ell hk}^* = \boldsymbol{c}_\ell'\boldsymbol{\Upsilon}_{hk}^*$. It can be directly seen that

$$\sqrt{N}\boldsymbol{c}_\ell'(\widehat{\boldsymbol{p}}^* - \boldsymbol{p}) \sim \sqrt{N}\sum_{h=1}^a \frac{1}{n_h}\sum_{k=1}^{n_h}[\Phi_{\ell hk}^* - E(\Phi_{\ell hk}^*)]$$

and by independence of $\Phi_{\ell hk}^*$ and $\Phi_{\ell hk'}^* (k \neq k')$ we obtain for the variance

$$Var\left(\sqrt{N}\sum_{h=1}^a \frac{1}{n_h}\sum_{k=1}^{n_h}[\Phi_{\ell hk}^* - E(\Phi_{\ell hk}^*)]\right) = N\sum_{h=1}^a \frac{1}{n_h}Var(\Phi_{\ell h1}^*) = N\sum_{h=1}^a \frac{1}{n_h}\omega_{\ell h}^{2,*}$$

with $\omega_{\ell h}^{2,*} = Var(\boldsymbol{c}_l'\boldsymbol{\Upsilon}_{h1}^*) = Var(\Phi_{\ell h1}^*)$. The unknown variances $\omega_{\ell h}^{2,*}$ can be consistently estimated by $\widehat{\omega}_{\ell h}^{2,*} = \frac{1}{n_h - 1}\sum_{k=1}^{n_h}(\Phi_{\ell hk}^* - \overline{\Phi}_{\ell h}^*)^2$ with $\overline{\Phi}_{\ell h}^* = \frac{1}{n_h}\sum_{k=1}^{n_h}\Phi_{\ell hk}^*$. We follow Gao et al. [37] and estimate the degree of freedom by

$$\nu = \max\{1, \min_{\ell=1,\ldots,q}\{\nu_1, \ldots, \nu_q\}\}$$

with

$$\nu_l = \frac{\left(\sum_{h=1}^a \widehat{\omega}_{\ell h}^{2,*}/n_h\right)^2}{\sum_{h=1}^a \widehat{\omega}_{\ell h}^{2,*}/(n_h^2(n_h-1))}, \ell = 1, \ldots, q.$$

8. Simulation Study

Within this section, the precision of the unweighted and weighted estimator and the behavior of the introduced test procedures in terms of their type-I error control are examined. The investigated metrics for the precision are Mean Squared Errors (MSEs) and biases, defined as

$$\begin{aligned}\text{bias} &= \frac{1}{n_{\text{sim}}}\sum_{i_{\text{sim}}=1}^{n_{\text{sim}}}\frac{1}{ad}\sum_{i=1}^{a}\sum_{s=1}^{d}\left(\widehat{p}_{is}^{*}-\frac{1}{2}\right)\\ \text{MSE} &= \frac{1}{n_{\text{sim}}}\sum_{i_{\text{sim}}=1}^{n_{\text{sim}}}\frac{1}{ad}\sum_{i=1}^{a}\sum_{s=1}^{d}\left(\widehat{p}_{is}^{*}-\frac{1}{2}\right)^{2}.\end{aligned}$$

As already pointed out by Domhof et al. [14], Konietschke et al. [33], and Rubarth et al. [31], the WTS requires large sample sizes to be able to maintain the type-I eror rate. Therefore, only the ATS and the MCTP will be examined.

8.1. Set-Up

The simulation study was conducted in *R* [38] version R 4.1.0 and for each scenario 10,000 simulation runs were performed. The complete simulation code can be found on https://github.com/KerstinRubarth/Clustered, last accessed on 10 November 2021. Due to the abundance of possible scenarios, the simulation study was restricted to the following parameter constellations: The number of independent groups was set to $a = 2$ and the number of repeated measures to $d = 3$. The sample sizes n_1 and n_2 were chosen to model balanced designs with $(n_1, n_2) \in \{(15, 15), (30, 30)\}$, as well as unbalanced designs with $(n_1, n_2) \in \{(20, 10), (40, 20)\}$. The number of dependent replicates of subject k at time s in group i (m_{isk}) were chosen to be

- $m_{isk} \equiv 1$ (no dependent replicates);
- $m_{isk} \equiv 2$ (two dependent replicates);
- m_{isk} realizations of a Binomial distribution with $Binom(5, 0.6) + 1$;
- m_{isk} realizations of a Binomial distribution with $Binom(10, 0.4) + 1$.

The correlation of the dependent replicates within a cluster was set to be

- $\rho_{isk} \equiv 0, \rho_{isk} \equiv 0.3, \rho_{isk} \equiv 0.9$ (same correlation within each cluster);
- ρ_{isk} realizations of a Binomial distribution with $Binom(10, 0.6)/10$ (different correlations within each cluster).

Data was generated by drawing from multivariate normal distributions having expectation $\boldsymbol{\mu}_{ik} = (\mu_{i1}, \ldots, \mu_{i1}, \ldots, \mu_{id}, \ldots, \mu_{id})' \in \mathbb{R}^{m_{ik}}$ ($m_{ik} = \sum_{s=1}^{d} m_{isk}$) and covariance matrices $\boldsymbol{\Sigma}_{ik} \in \mathbb{R}^{m_{ik} \times m_{ik}}$ with

$$\boldsymbol{\Sigma}_{ik} = \begin{pmatrix}
\sigma_{i1}^2 & \rho_{i1k} & \cdots & \rho_{i1k} & \sigma_{i12} & \cdots & \cdots & \sigma_{i12} & \sigma_{i13} & \cdots & \cdots & \sigma_{i13}\\
\rho_{i1k} & \ddots & \rho_{i1k} & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots\\
\vdots & \rho_{i1k} & \ddots & \rho_{i1k} & \vdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots\\
\rho_{i1k} & \cdots & \rho_{i1k} & \sigma_{i1}^2 & \sigma_{i12} & .. & \cdots & \sigma_{i12} & \sigma_{i13} & \cdots & \cdots & \sigma_{i13}\\
\sigma_{i21} & \cdots & \cdots & \sigma_{i21} & \sigma_{i2}^2 & \rho_{i2k} & \cdots & \rho_{i2k} & \sigma_{i23} & \cdots & \cdots & \sigma_{i23}\\
\vdots & \cdots & \cdots & \vdots & \rho_{i2k} & \ddots & \rho_{i2k} & \vdots & \vdots & \cdots & \cdots & \vdots\\
\vdots & \cdots & \cdots & \vdots & \vdots & \rho_{i2k} & \ddots & \rho_{i2k} & \vdots & \cdots & \cdots & \vdots\\
\sigma_{i21} & \vdots & \vdots & \sigma_{i21} & \rho_{i2k} & \cdots & \rho_{i2} & \sigma_{i2}^2 & \sigma_{i23} & \cdots & \cdots & \sigma_{i23}\\
\sigma_{i31} & \cdots & \cdots & \sigma_{i31} & \sigma_{i32} & \cdots & \cdots & \sigma_{i32} & \sigma_{i3}^2 & \rho_{i3k} & \cdots & \rho_{i3k}\\
\vdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots & \rho_{i3k} & \ddots & \rho_{i3k} & \vdots\\
\vdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots & \rho_{i3k} & \ddots & \rho_{i3k}\\
\sigma_{i31} & \cdots & \cdots & \sigma_{i31} & \sigma_{i32} & \cdots & \cdots & \sigma_{i32} & \rho_{i3k} & \cdots & \rho_{i3k} & \sigma_{i3}^2
\end{pmatrix}.$$

The components σ_{isk} are obtained from the following homo- and heteroscedastic covariance matrices of multivariate normal distributions:

$$\boldsymbol{\Sigma}_1 \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.1 & 0.2 \\ 0.1 & 1.2 & 0.3 \\ 0.2 & 0.3 & 1.5 \end{pmatrix}.$$

Since the simulation study of Rubarth et al. [31] indicated that the performance of the procedure is not dependent on the distribution, no other data generating distributions were considered. Analogously, we restricted our simulations to the case of Missing-Completely-At-Random (MCAR) scenarios, since no different behavior of the methods of Rubarth et al. [31] in MAR scenarios compared to MCAR scenarios could be observed. Thus, the indicators λ_{isk} greenwere generated by drawing from Binomial distributions $B(1-r)$ with r being the percentage of missing values $\boldsymbol{r} = (r_1, r_2) \in \{(0\%, 0\%), (0\%, 20\%), (10\%, 10\%), (30\%, 30\%)\}$. Since the power of the methods was already investigated in detail by Rubarth et al. [31], the simulation study green of the present paper focused solely on type-I error rates. Further, we additionally investigated the precision of the unweighted and weighted estimators.

8.2. Results—Type-I Error Rate

First, an overview of the impact of different **sample sizes** in scenarios with completely observed data is given in Figure 1 if no missing values occur. It can be readily seen that both procedures control the type-I error quite well even if the sample size is quite low with $n_1 = n_2 = 15$. Interestingly, the MCTP works better if sample sizes are unbalanced, whereas the ATS works better in case of balanced sample sizes.

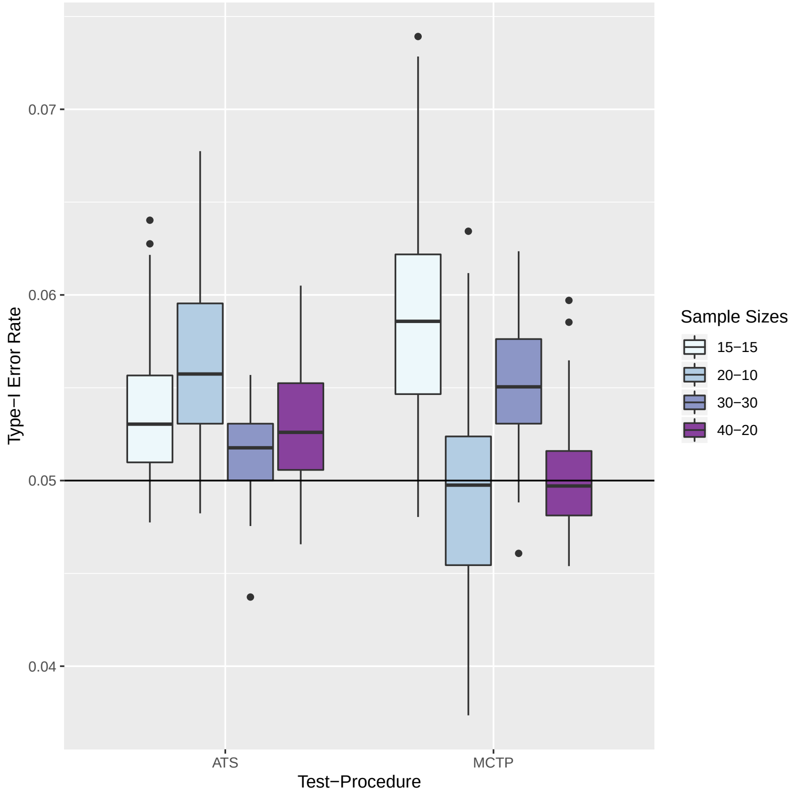

Figure 1. Boxplots of Type-I error rates in relation to sample sizes n_1 and n_2 in various settings without missing data.

Next, the impact of **missing values** will be inspected. In Figure 2, the sample sizes were fixed with $n_1 = n_2 = 30$. The empirical type-I error rates of both procedures increase, if missing values occur and the higher the relative frequency of missing values, the higher the type-I error rates. Furthermore, the simulation study indicates that the MCTP is more affected by the occurrence of missing values than the ATS, which was already noted by Rubarth et al. [31].

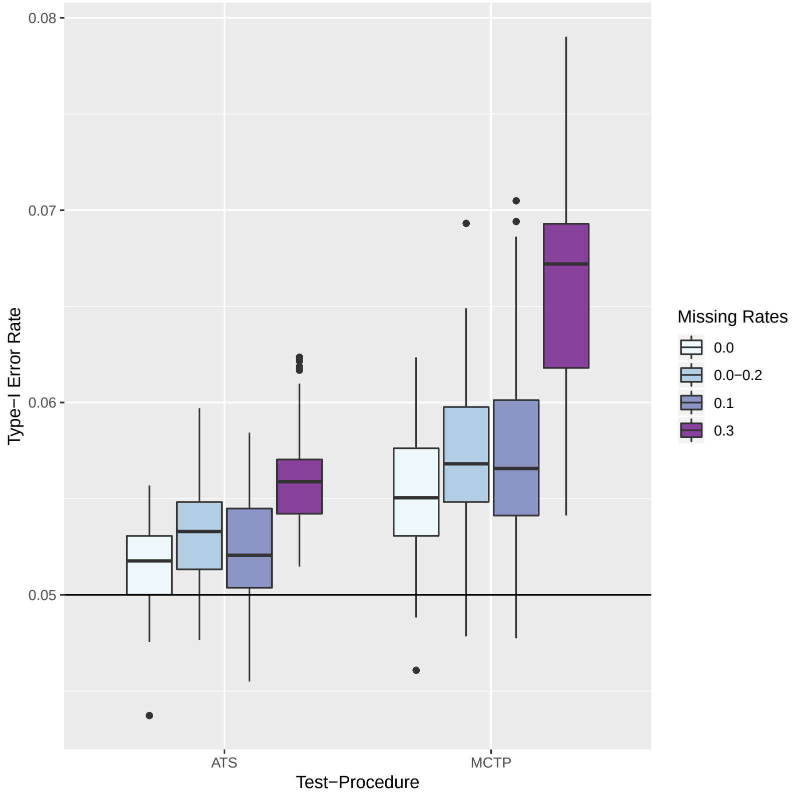

Figure 2. Boxplots of Type-I error rates in relation to missing rates r_1 and r_2 in various settings with $n_1 = n_2 = 30$.

The relationship between the **cluster sizes** m_{isk} and the type-I error rates is depicted in Figure 3. For this comparison, the sample sizes were again fixed with $n_1 = n_2 = 30$ and only completely observed data were investigated. It can be readily seen that type-I error rates decrease if two dependent replicates of each subject are present in comparison to a dataset without a clustered structure. However, type-I error rates of the ATS increases if the number of dependent replicates m_{isk} is arbitrary with an expected number of dependent replicates of 5 or 4, respectively. In contrast, the type-I error rates of the MCTP decrease on median in these scenarios.

Next, the influence of **intra-cluster correlations** ρ_{isk} on type-I error rates is investigated in scenarios with sample sizes $n_1 = n_2 = 30$ and green without missing data (Figure 4). The type-I error rates of the ATS decrease if non-arbitrary higher intra-cluster correlations are present, whereas the type-I error rates of the MCTP increase in case of higher (non-arbitrary) intra-cluster correlations ρ_{isk}. Interestingly, if arbitrary correlations ρ_{isk} are present with a mean correlation of 0.6, the type-I error rates of the ATS increase in comparison to scenarios with fixed correlations, whereas the type-I error rates of the MCTP are on the same level as in scenarios with a fixed correlation $\rho_{isk} = 0.9$.

Figure 5 depicts the impact of **homo- and heteroscedastic** covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in settings with $n_1 = n_2 = 30$ and green without missing data. Again, it can be seen that the type-I error rates of the ATS are on median smaller than those of the MCTP for both homo- and heteroscedastic covariance matrices. Type-I error rates of the ATS seem to be a bit smaller in case of homoscedasticity, whereas type-I error rates of the MCTP seem to be slightly larger in homoscedastic scenarios.

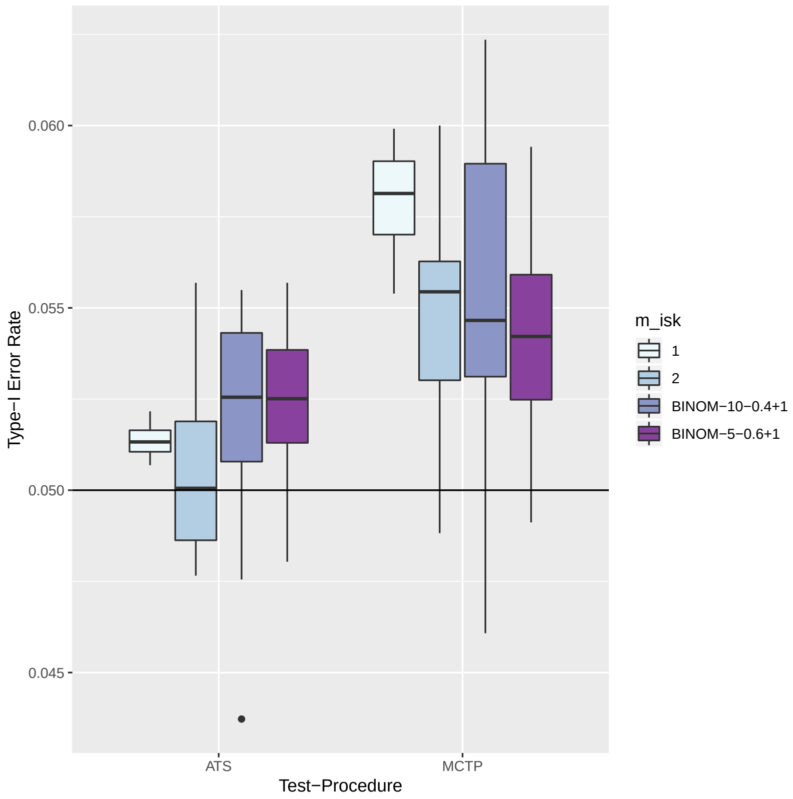

Figure 3. Boxplots of Type-I error rates in relation to cluster sizes m_{isk} in various settings with $n_1 = n_2 = 30$ and green without missing data.

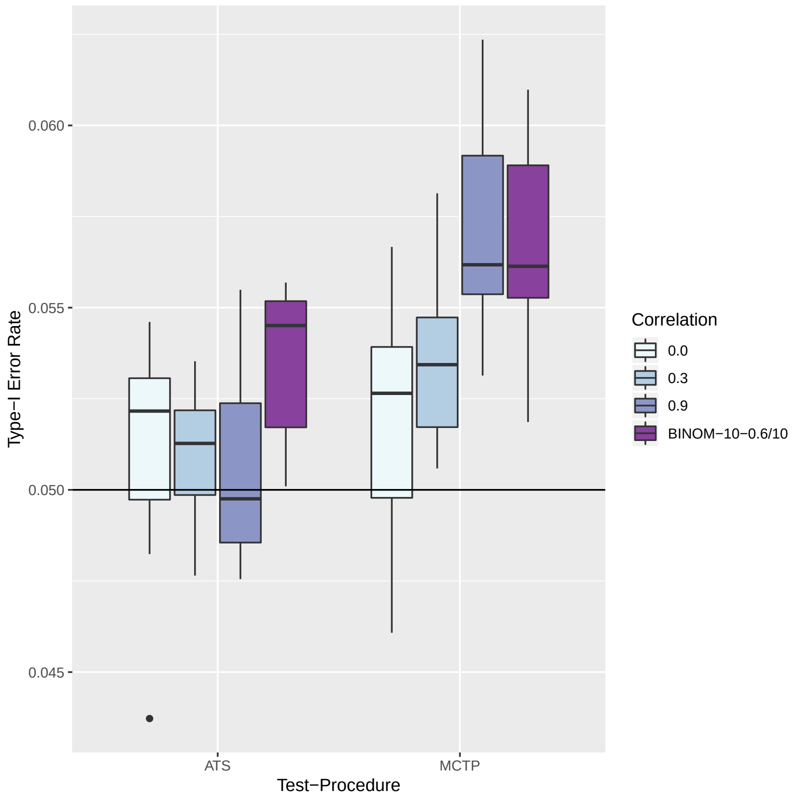

Figure 4. Boxplots of Type-I error rates in relation to intra-cluster correlation ρ_{isk} in various settings with $n_1 = n_2 = 30$ and green without missing data.

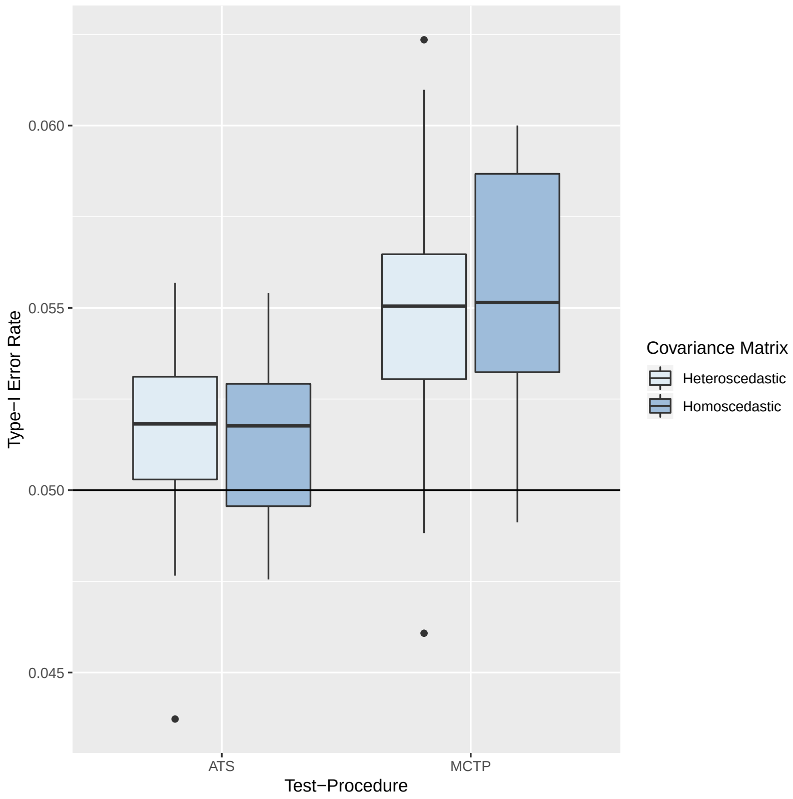

Figure 5. Boxplots of Type-I error rates in relation to covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ in various settings with $n_1 = n_2 = 30$ and without missing data.

Next, the relationship of **unweighted and weighted estimation** and type-I error rates will be inspected in scenarios with $n_1 = n_2 = 30$ and without missing data (Figure 6). It can be readily seen that the type-I error rates of the ATS do not differ on median in case of weighted and unweighted estimation of the relative effect $\boldsymbol{p}$, only the interquartile range is increased in case of weighted estimation. Contrary, the type-I error rates of the MCTP are on median smaller in case of unweighted estimation of the relative effect $\boldsymbol{p}$ but without an enlargement of the respective interquartile range.

To conclude, an analysis of the impact of **unweighted and weighted estimation** of the relative effect $\boldsymbol{p}$ and the fixed **intracluster correlation** ρ_{isk} is presented in Figure 7 (in scenarios with $n_1 = n_2 = 30$ and without missing data). The type-I error rates of the ATS decrease if the intra-cluster correlations ρ_{isk} increase, as already depicted in Figure 4. Interestingly, the medians of the type-I error rates are comparable in case of unweighted and weighted estimation if no intra-cluster correlation is present. However, in these scenarios, the interquartile range of the type-I error rates in case of weighted estimation is very enlarged in comparison to the case of unweighted estimation. If a medium intra-correlation is present, type-I error rates of the unweighted estimator are on median smaller than those of the weighted estimator. Here, the interquartile ranges are quite comparable. In scenarios with high intra-class correlations, the weighted estimator yields smaller type-I error rates on median; as well as a larger interquartile range.

Again, as already outlined in Figure 4, type-I error rates of the MCTP increase with higher intra-cluster correlations. The type-I error rates of the unweighted and weighted estimator are quite comparable if no intracluster-correlation is present. However, they are quite different if a medium correlation is present: type-I error rates by using the unweighted estimator tend to be smaller on median than those obtained from weighted estimation. If high correlations are present, the unweighted estimator yields smaller type-I error rates than the weighted version.

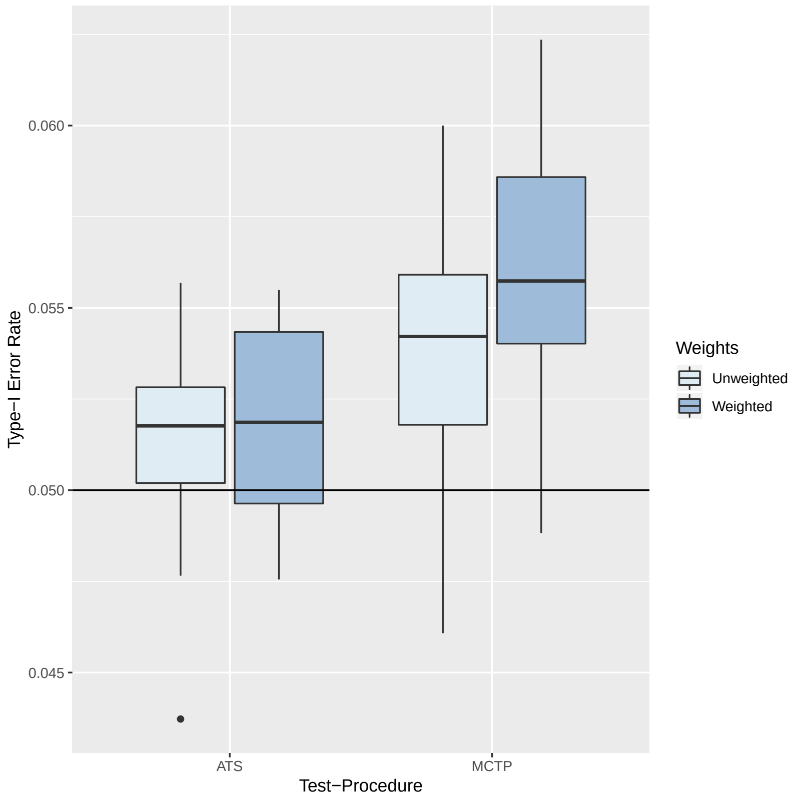

Figure 6. Boxplots of Type-I error rates in relation to unweighted and weighted estimation of the relative effect $\boldsymbol{p}$ in various settings with $n_1 = n_2 = 30$ and without missing data.

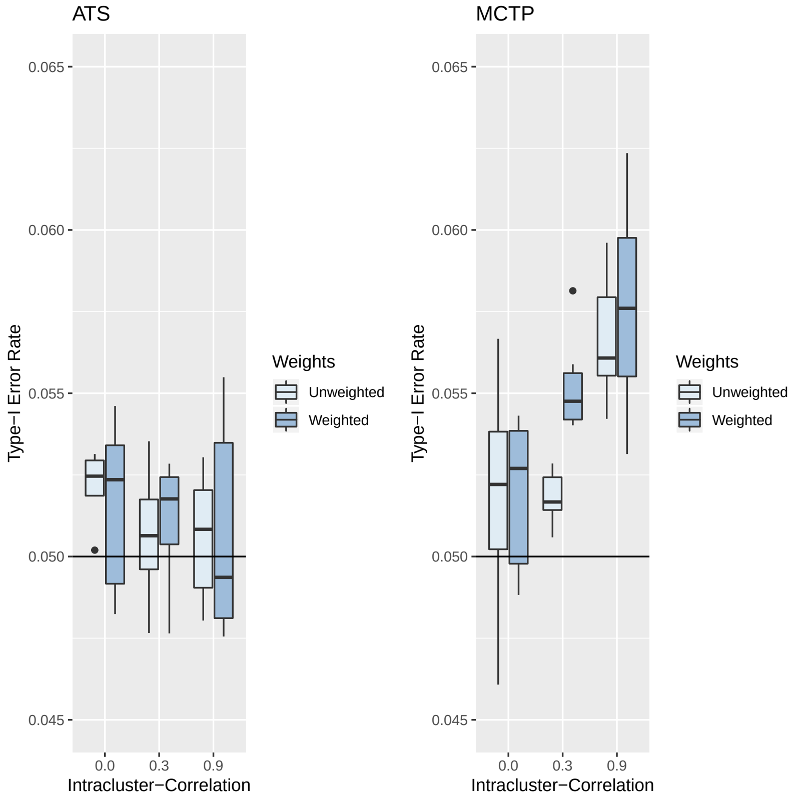

Figure 7. Boxplots of Type-I error rates in relation to unweighted and weighted estimation of the relative effect $\boldsymbol{p}$ and fixed intra-cluster correlations ρ_{isk} in various settings with $n_1 = n_2 = 30$ and without missing data.

8.3. Results—Precision

Analogously to the previous section, we will first explore the impact of the **sample size** on the precision of the unweighted and weighted estimators in scenarios with completely observed data. It can be readily seen in Figure 8 that the MSEs of the unweighted and weighted estimators are quite comparable. The MSEs decrease if sample size increase; balanced settings exhibit smaller MSEs than unbalanced settings. Regarding the bias of the estimators, scenarios with smaller sample sizes tend to exhibit biases in the negative direction, whereas scenarios with larger sample sizes exhibit biases in the positive direction. Interestingly, the interquartile range of biases is quite enlarged in scenarios with $n_1 = 40$ and $n_2 = 20$.

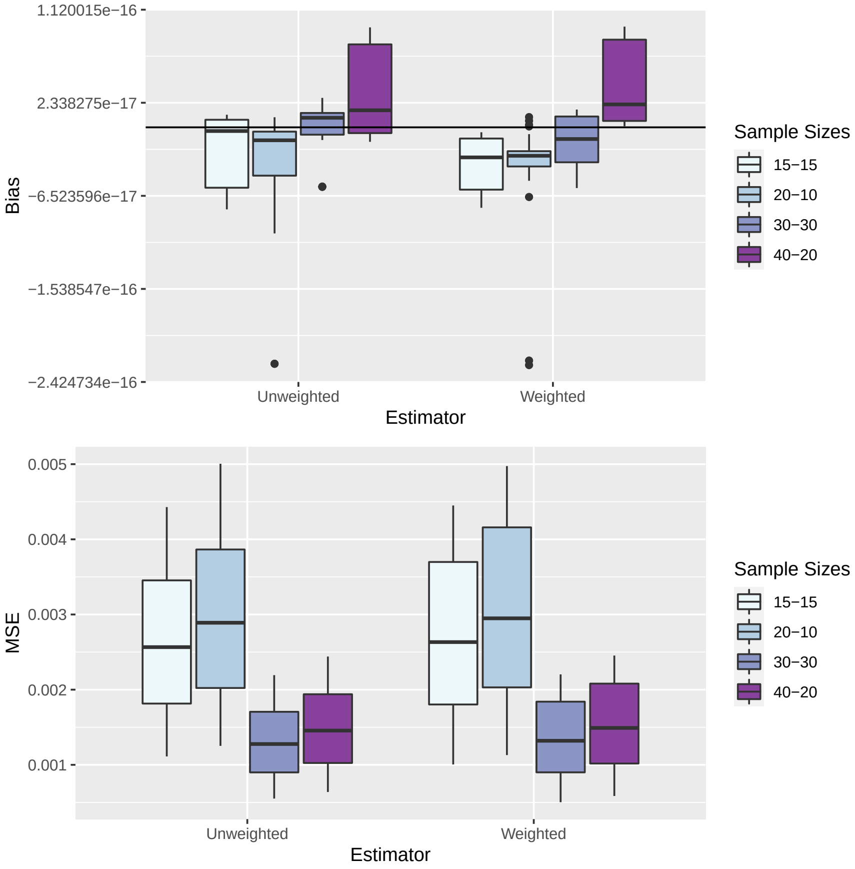

Figure 8. Boxplots of biases and MSEs of estimators $\widehat{p}^*_{is}$ in Equation (5) in relation to different sample sizes n_1 and n_2.

Next, the impact of **missing data** on the precision of the estimators is inspected in scenarios with $n_1 = n_2 = 30$ (see Figure 9). As seen before, the MSEs of unweighted and weighted estimators are quite comparable. The MSEs increase with an increasing missing rate. Interestingly, the interquartile ranges of the biases of the two different estimators are quite different; the weighted estimator exhibits a larger interquartile range (especially if 10% of data are missing) than the unweighted estimator. Further, biases of the unweighted estimator tend to be positive (except if the missing rate is 30%). Contrary, the biases of the weighted estimator tend to be negative.

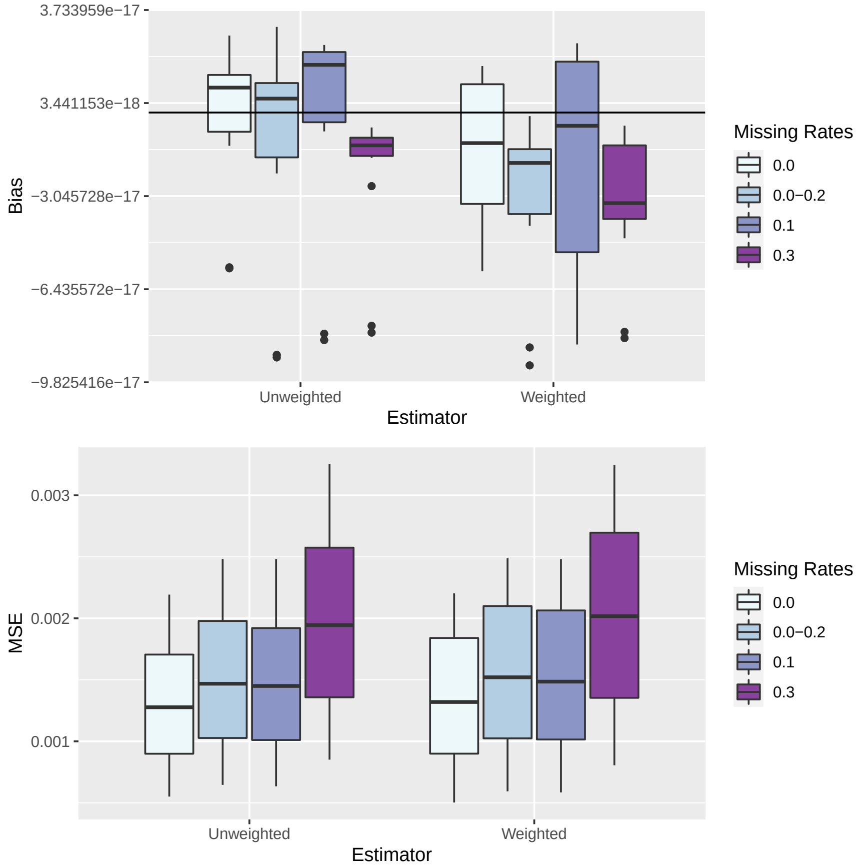

Figure 9. Boxplots of biases and MSEs of estimators $\widehat{p}_{is}^{*}$ in Equation (5) in relation to the amount of missing data in scenarios with $n_1 = n_2 = 30$.

In Figure 10, the influence of the **number of dependent replicates** m_{isk} on the precision is presented (again in scenarios with $n_1 = n_2 = 30$ and without missing data). As already pointed out, the distribution of MSEs of the unweighted and weighted estimator is very similar. Interestingly, there is very little variation if only one observation per subject and time point is available compared to scenarios with more possibly dependent replicates. Further, the MSEs decrease quite a lot if already two possibly dependent observations are available and the more clustered data are available, the smaller are the MSEs. The same observations can be made regarding the biases of both estimators.

Finally, the relationship between **intra-cluster correlations** will be inspected (see Figure 11, again scenarios with $n_1 = n_2 = 30$ and without missing data). Again, the MSEs of the unweighted and the weighted estimator do not differ much. It can be readily seen that in scenarios with fixed cluster correlations, the MSEs increase with increased intra-cluster correlations. Contrary, the biases of both estimators approach 0 with increasing fixed intra-cluster correlations but the two estimators exhibit a different behavior: the biases of the unweighted estimator tend to be positive whereas the biases of the weighted estimator tend to be negative.

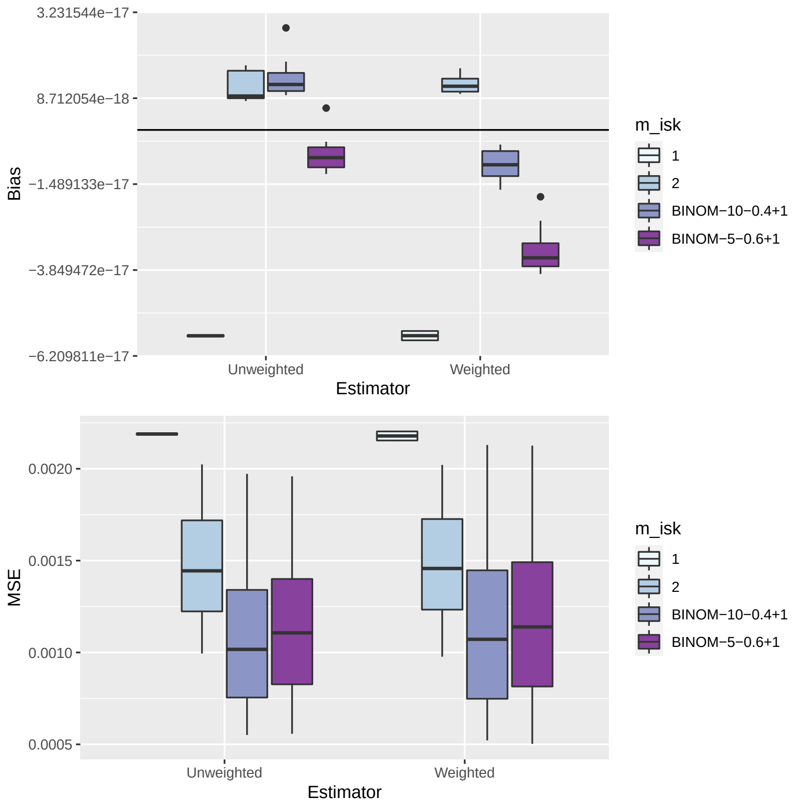

Figure 10. Boxplots of biases and MSEs of estimators $\widehat{p}_{is}^{*}$ in Equation (5) in relation to cluster sizes m_{isk} in scenarios with $n_1 = n_2 = 30$ and without missing data.

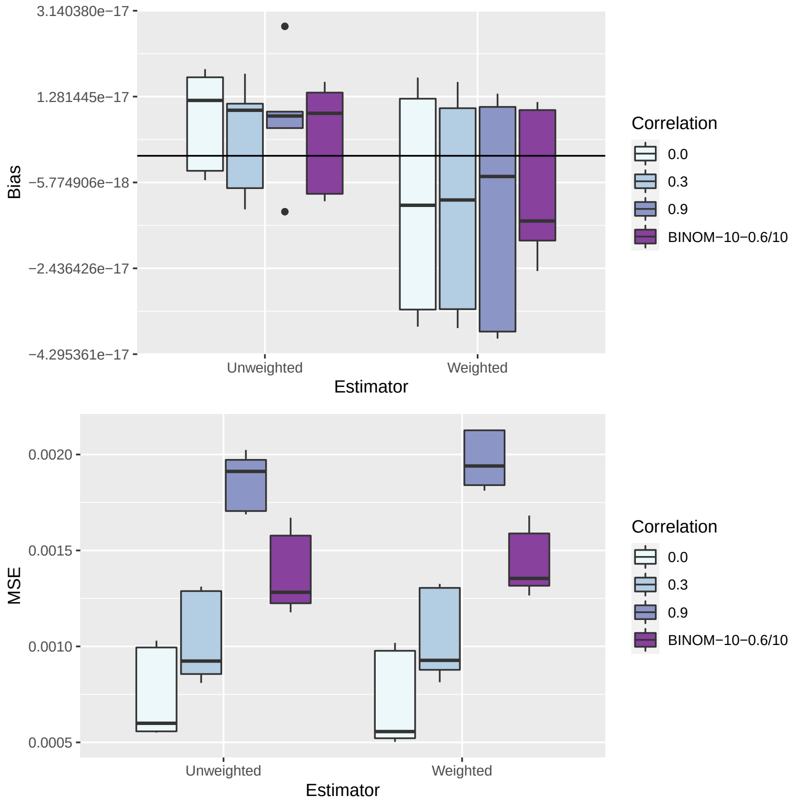

Figure 11. Boxplots of biases and MSEs of estimators $\widehat{p}_{is}^{*}$ in Equation (5) in relation to intra-cluster correlations ρ_{isk} in scenarios with $n_1 = n_2 = 30$ and without missing data.

9. Analysis of the Motivating Example

The parameter pRNFL from the SUPREMES study introduced in Section 2 can now be analyzed using the newly proposed methodology for factorial repeated measures designs with dependent replicates. It can be readily seen in Figures 12 and 13 that pRNFL baseline values in the placebo group tend to be smaller than those from the EGCG group.

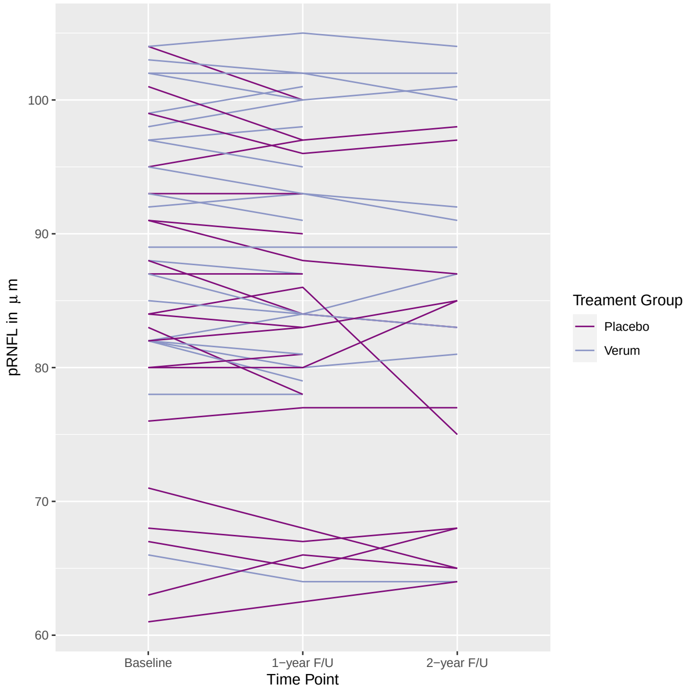

Figure 12. Lineplot of pRNFL values in both groups at baseline, 1-year follow-up and 2-year follow-up of the SUPREMES trial.

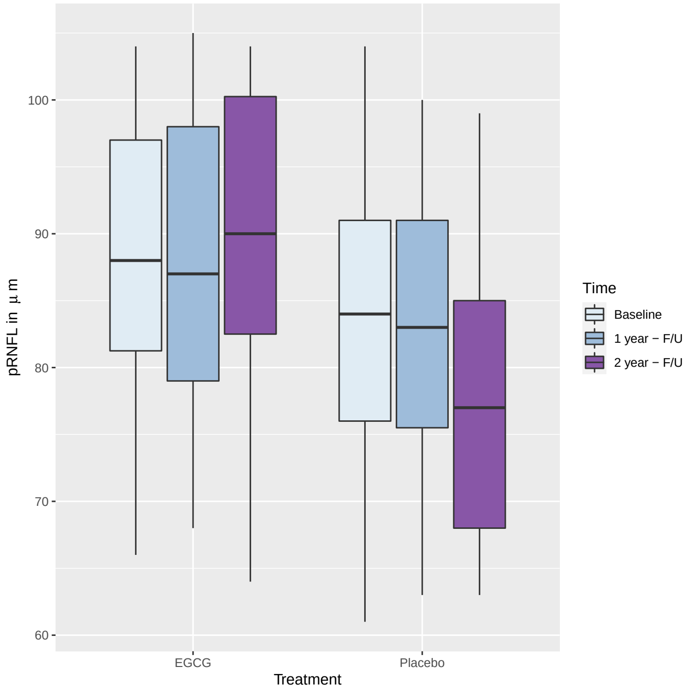

Figure 13. Boxplot of pRNFL values in both groups at baseline, 1-year follow-up and 2-year follow-up of the SUPREMES trial.

Further, Figure 13 indicates that pRNFL values seem to increase at 2-year follow-up; however, this needs to be interpreted with caution, since almost 50% of the patients could not be measured at this time point. These were mostly patients with smaller baseline pRNFL values. This could possibly violate the *Missing Completely at Random* (MCAR) assumption. However, the simulation study of Rubarth et al. [31] indicated that the proposed method is robust against violations of the MCAR assumption. In order to account for the baseline differences between the two groups following Klumbies et al. [19], the analysis is not based on raw pRNFL values but on differences to baseline. In contrast to the analysis of Klumbies et al. [19], the MCTP on baseline differences is applied. The Tukey contrast was chosen to compare all pairwise differences, thus, the contrast matrix for testing the null hypothesis $H_0^p : \boldsymbol{C}\boldsymbol{p} = \mathbf{0}$ is as follows:

$$\mathbf{C} = \begin{pmatrix} \boldsymbol{c}_1' \\ \boldsymbol{c}_2' \\ \boldsymbol{c}_3' \\ \boldsymbol{c}_4' \\ \boldsymbol{c}_5' \\ \boldsymbol{c}_6' \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

The relative effects were estimated by using the unweighted version of the estimator, leading to $\widehat{\boldsymbol{p}} = (\widehat{p}_{11}, \widehat{p}_{12}, \widehat{p}_{21}, \widehat{p}_{22})' = (0.525, 0.445, 0.505, 0.525)'$, indicating that the differences to baseline are the largest at 1-year follow-up simultaneously in group 1 (intervention group) and group 2 (Placebo group) and the smallest at 2-year follow-up in group 1. The results, including the values of the test statistics T_ℓ, p-values and 95% simultaneous confidence intervals, are displayed in Table 2.

Table 2. Point estimators of differences $p_{is} - p_{jt}$ (i, j: group, s, t: time point), simultaneous confidence intervals, t-values and p-values for Tukey-type contrasts in relative effects in the SUPREMES trial.

Comparison	Estimator	95%-Confidence Interval	*t*-Value	*p*-Value
$\widehat{p}_{12} - \widehat{p}_{11}$	−0.080	[−0.371, 0.212]	−0.777	0.850
$\widehat{p}_{21} - \widehat{p}_{11}$	−0.020	[−0.344, 0.303]	−0.178	0.998
$\widehat{p}_{22} - \widehat{p}_{11}$	0.000	[−0.468, 0.468]	0.000	1.000
$\widehat{p}_{21} - \widehat{p}_{12}$	0.059	[−0.335, 0.453]	0.429	0.969
$\widehat{p}_{22} - \widehat{p}_{12}$	0.080	[−0.442, 0.601]	0.435	0.968
$\widehat{p}_{22} - \widehat{p}_{21}$	0.020	[−0.306, 0.346]	0.177	0.998

Note that no adjustment for multiplicity was necessary, since the same critical value obtained from the MCTP was used for each comparison. It follows that no evidence exists to reject the global null hypothesis $H_0^p : \boldsymbol{C}\boldsymbol{p} = \mathbf{0}$, resulting in the same conclusion as already pointed out by Klumbies et al. [19]: Retinal thickness analysis did not reveal neuroprotective effects of EGCG, especially when considering the contrasts $\widehat{p}_{21} - \widehat{p}_{11} = 0$ and $\widehat{p}_{22} - \widehat{p}_{12} = 0$, which allow a direct comparison of the differences to baseline at 1-year and 2-year follow-up in the two groups.

10. Discussion and Conclusions

In the present paper, we presented different estimation techniques for Wilcoxon–Mann–Whitney effects in factorial repeated measures designs with clustered data. In a first step, the information whether cluster sizes are informative or non-informative is key and plays a major role in precise effect size estimation. Furthermore, as indicated by Zou [28] the use of the intraclass correlation enlarges the precision of the estimators. Anyway, besides estimation, we furthermore discussed how to test global and multiple hypotheses in terms of Wilcoxon–Mann–Whitney effects using any of the aforementioned estimators. Here, no specific data distribution (symmetric or asymmetric) is required. The presented estimators and test procedures should be preferred over standard parametric methods such as Linear Mixed Models or Generalized Estimating Equations in scenarios with small sample sizes, heteroscedastic variances or count, ordinal outcomes. We recommend to use the MCTP instead of quadratic-type test procedures in most practical applications since testing global hypotheses does not usually answer the research questions. However, the MCTP presented in this paper does not precisely hold the nominal type-I error rate in case of very small sample sizes, high correlations, and strong heteroscedasticity between groups or time points. Recently, Friedrich et al. [39] proposed novel resampling methods in purely non-parametric designs. Extending these ideas to such designs will be part of future research to improve especially the MCTP in "extreme" scenarios.

Although an extensive simulation study was conducted to evaluate the precision and type-I error rates of the procedures in several scenarios, it is advised to conduct further simulation studies in practical applications, e.g., in the planning or data analysis phase of a study for a specific scenario. Further examinations indicate that the methods are applicable even in situations with rather small sample sizes, such as $n = 10$. The actual nominal level and accuracy depends, however, on the design of interest.

Furthermore, it is important to note that many scientists in applied research fields, e.g., biomedicine, have a misconception that the Wilcoxon–Mann–Whitney (WMW) test is a test for equality of means or medians, when outcomes are metric and distributions are skewed or ordinal and that this test is the non-parametric equivalent to a classic two-sample t-test [40]. As investigated by Fagerland et al. [41], the true significance level of the WMW test deviates enormously from the nominal level when the test is used for comparing means or medians in scenarios with deviations from a pure shift model (two populations having equal shapes and scales). In practical applications, the pure shift model is rarely present, since skewed distributions with different means have most likely also different variances. Especially in such scenarios, the Brunner–Munzel test [8] should be applied. Another disadvantage of the application of WMW tests as noted by Bergmann et al. [42] is that many versions and implementations in statistical software packages exist, e.g., large sample approximation, exact permutation form, versions with or without correction for continuity or ties and different algorithm variants, all leading to possibly different p-values and eventually to different conclusions. In this work we present a unified approach that does not need a correction for ties nor continuity. Furthermore, the WMW test and its p-value are rarely accompanied with their corresponding effect estimate, the Wilcoxon–Mann–Whitney parameter p and its corresponding confidence interval. As noted by Fay et al. [43], classic confidence interval procedures for the WMW parameter are not compatible with exact WMW tests, meaning, that the tests rejects a hypothesis at significance level α but the confidence interval for p includes $\frac{1}{2}$. Thus, Fay et al. [43] developed compatible confidence intervals for asymptotic WMW tests and for some exact WMW tests. Furthermore, Fay et al. [44] indicate that the WMW parameter p can be framed as a causal parameter (the probability that a randomly chosen subject from one population, e.g., the treatment group in a randomized-controlled trial, will have a larger response than one subject from the other population, e.g., the control group). However, this parameter is not equal to another closely related and non-identifiable causal effect, the probability that a randomly chosen subject will have a larger response under treatment than under control [44]. This paradox was first introduced by Hand et al. [45]. Therefore, caution must be given when interpreting effect estimates from non-parametric procedures. Thus, the literacy of non-parametric statistics of scientists working in applied fields should be fostered. This work aims to accomplish this by first introducing and explaining Wilcoxon–Mann–Whitney parameters p in special designs and then providing a flexible model for the analysis of factorial repeated measure designs with a clustered structure, allowing for missing values in a second step. For user friendly applications of the methods, it is planned to enrich the *R* software package nparLD [46].

Author Contributions: Conceptualization, K.R. and F.K.; methodology, K.R., P.S. and F.K.; formal analysis, K.R.; software, K.R. and P.S.; data curation, H.G.Z.; writing—original draft preparation, K.R. and P.S.; writing review and editing K.R., P.S., F.K. and H.G.Z.; funding acquisition, F.K. All authors have read and agreed to the published version of the manuscript.

Funding: The work of Kerstin Rubarth and Frank Konietschke was funded by Deutsche Forschungsgemeinschaft Grant/Award Number DFG KO 4680/3-2. The work of Paavo Sattler was funded by Deutsche Forschungsgemeinschaft Grant/Award Number DFG PA 2409/3-2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The R code of the simulation study is available at https://github.com/KerstinRubarth/Clustered, last accessed on 10 November 2021.

Conflicts of Interest: The authors declare no conflicts of interest. H.G.Z. received honoraria from Bayer Healthcare and Novartis, and research grants from Novartis, independent of the current study.

Appendix A

In this section, we present all the proofs of the theoretical results achieved.

Proof of Proposition 1. For the concrete weights, we calculate

$$\lambda_{jtk} w_{jtk}^{v_1} m_{jtk} = \lambda_{jtk} m_{jtk} \cdot \frac{1}{m_{jtk}\lambda_{jt}} \leq \frac{1}{\lambda_{\min}},$$

$$\lambda_{jtk} w_{jtk}^{v_2} m_{jtk} = \lambda_{jtk} m_{jtk} \cdot \frac{1}{m_{jt}} \leq \frac{M_0}{\lambda_{\min}}.$$

Finally, we calculate

$$\sum_{k=1}^{n_i} m_{isk} \lambda_{isk} w_{isk}^{v_1} = \sum_{k=1}^{n_i} m_{isk} \lambda_{isk} \cdot \frac{1}{\lambda_{is} m_{isk}} = \frac{1}{\lambda_{is}} \sum_{k=1}^{n_i} \lambda_{isk} = 1,$$

and

$$\sum_{k=1}^{n_i} m_{isk} \lambda_{isk} w_{isk}^{v_2} = \sum_{k=1}^{n_i} m_{isk} \lambda_{isk} \cdot \frac{1}{m_{is}} = 1.$$

□

Since these are all necessary properties of the weights, the following results hold for all kinds of weights fulfilling these properties.

Moreover, it becomes clear, that condition **A1.3** is only necessary for the first inequality and the weighted estimator. Since these inequalities are required for nearly all the following statements, it is therein claimed. Note that if the assumption of bounded cluster sizes is difficult to justify in practical applications, the unweighted estimator can be used without any restrictions.

Proposition A1. *For the empirical distributions, under condition **A1.3**, it holds*

$$||\widehat{F}_{is}^{*} - F_{is}||_{\infty} \xrightarrow{a.s.} 0,\ \lambda_{\min} \to \infty \quad \textit{and} \quad ||\widehat{G}^{*} - G||_{\infty} \xrightarrow{a.s.} 0 \quad \lambda_{\min} \to \infty \quad * \in \{v_1, v_2\}.$$

Proof. First, we demonstrate the pointwise almost sure convergence of the empirical distribution function $\widehat{F}_{is}^{*}$. Denote with $\Xi_{is} = \{k = 1, \ldots, n_i, : \lambda_{isk} > 0\}$, the amount containing the indices of subjects from group i, with existing observation in the s-th component. It is clear that $|\Xi_{is}| = \lambda_{is}$. Moreover, for fixed $x \in \mathbb{R}$ we define independent random variables $Y_{isk}^{*} := \lambda_{is} w_{isk}^{*} \sum_{u=1}^{m_{isk}} c(x - X_{isku})$. Then it holds that

$$\widehat{F}_{is}^{*}(x) = \frac{1}{\lambda_{is}} \sum_{k=1}^{n_i} \lambda_{isk} Y_{isk}^{*} = \frac{1}{\lambda_{is}} \sum_{\ell \in \Xi_{is}} Y_{is\ell}^{*}.$$

For the expectation of this sum, we calculate

$$\begin{aligned}
\mathbb{E}\left(\sum_{\ell\in\Xi_{is}} Y^*_{is\ell}\right) &= \sum_{\ell\in\Xi_{is}} \lambda_{is} w^*_{is\ell} \sum_{u=1}^{m_{is\ell}} \mathbb{E}(c(x - X_{is\ell u})) \\
&= \lambda_{is} \sum_{\ell\in\Xi_{is}} w^*_{is\ell} m_{is\ell} F_{is}(x) \\
&= \lambda_{is} F_{is}(x) \sum_{k=1}^{n_i} \lambda_{isk} w^*_{is\ell} m_{is\ell} \\
&= \lambda_{is} F_{is}(x).
\end{aligned}$$

Because of $|c(x)| < 1$ for the variance of $Y^*_{ik\ell}$ we obtain

$$Var(Y^*_{is\ell}) = \lambda^2_{is}(w^*_{is\ell})^2 Var\left(\sum_{u=1}^{m_{is\ell}} \mathbb{E}(c(x - X_{is\ell u}))\right) \leq \lambda^2_{is}(w^*_{is\ell})^2 {m_{is\ell}}^2.$$

For the application of the strong law of large numbers we finally consider

$$\begin{aligned}
\frac{1}{\lambda^2_{is}} \sum_{\ell\in\Xi_{is}} Var(Y^*_{ik\ell}) &\leq \sum_{\ell\in\Xi_{is}} (w^*_{is\ell})^2 {m_{is\ell}}^2 \\
&= \sum_{k=1}^{n_i} \lambda^2_{isk}(w^*_{isk})^2 {m_{isk}}^2 \\
&= \mathcal{O}(\lambda^{-1}_{\min}) \sum_{k=1}^{n_i} \lambda_{isk} w^*_{isk} m_{isk} \\
&= \mathcal{O}(\lambda^{-1}_{\min}) \to 0 \qquad\qquad \text{for } \lambda_{\min} \to \infty.
\end{aligned}$$

Thus, it holds that

$$\widehat{F}^*_{is}(x) \xrightarrow{a.s.} \frac{1}{\lambda_{is}} \mathbb{E}\left(\sum_{\ell\in\Xi_{is}} Y^*_{ik\ell}\right) = F_{is}(x).$$

Replacing $c(x)$ through $c^{(+)}(x)$ resp. $c^{(-)}(\cdot)$, this leads to the same convergence for the right-continuous resp. left-continuous versions of this distribution functions.

Now, this pointwise convergence has to be expanded for the supremum norm. This was already done by Domhof [47] for a similar setting and only requires the pointwise convergence proven above. The result for $\widehat{G}^*$ and G follows from this with the triangle inequality. □

Proof of Proposition 2. To prove the asymptotic unbiasedness, we consider the single components and calculate

$$\begin{aligned}
&\left|\mathbb{E}(\widehat{p}^*_{is}) - p_{is}\right| \\
=\ & \left|\tfrac{1}{ad}\sum_{k=1}^{n_i}\sum_{j=1}^{a}\sum_{t=1}^{d}\sum_{\ell=1}^{n_j}\sum_{u=1}^{m_{isk}}\sum_{v=1}^{m_{jt\ell}} \lambda_{isk}\lambda_{jt\ell}w^*_{jt\ell}w^*_{isk}\left[\mathbb{E}(c(X_{isku} - X_{jt\ell v})) - \int F_{jt} dF_{is}\right]\right| \\
\leq\ & \tfrac{1}{ad}\sum_{k=1}^{n_i}\sum_{j=1}^{a}\sum_{t=1}^{d}\sum_{\ell=1}^{n_j}\sum_{u=1}^{m_{isk}}\sum_{v=1}^{m_{jt\ell}} \lambda_{isk}\lambda_{jt\ell}w^*_{jt\ell}w^*_{isk}\left|\mathbb{E}(c(X_{isku} - X_{jt\ell v})) - \int F_{jt} dF_{is}\right| \\
=\ & \tfrac{1}{ad}\sum_{k=1}^{n_i}\sum_{j=1}^{a}\sum_{t=1}^{d} \lambda_{jtk}\lambda_{isk}w^*_{isk}w^*_{jtk}\sum_{u=1}^{m_{isk}}\sum_{v=1}^{m_{jtk}}\left|\mathbb{E}(c(X_{isku} - X_{jtkv})) - \int F_{jt} dF_{is}\right| \\
\leq\ & \tfrac{1}{ad}\sum_{k=1}^{n_i}\sum_{j=1}^{a}\sum_{t=1}^{d} \lambda_{jtk}\lambda_{isk}w^*_{isk}w^*_{jtk}m_{isk}m_{jtk} \\
\leq\ & \tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\sum_{k=1}^{n_i} \lambda_{isk}w^*_{isk}m_{isk}\cdot\mathcal{O}(\lambda^{-1}_{\min}) \\
=\ & \mathcal{O}(\lambda^{-1}_{\min}).
\end{aligned}$$

It is clear that through condition **A1.1** and **A1.2** we can also substitute this with $\mathcal{O}\big(N^{-1}\big)$.

For the second part we consider

$$\begin{aligned}
\left|\widehat{p}^*_{is} - p_{is}\right| &\leq \tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\left|\int \widehat{F}^*_{jt}\, d\widehat{F}^*_{is} - \int F_{jt}\, dF_{is}\right| \\
&= \tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\left|\int \widehat{F}_{jt}\, d\widehat{F}^*_{is} - \int F_{jt}\, d\widehat{F}^*_{is} + \int F_{jt}\, d\widehat{F}^*_{is} - \int F_{jt}\, dF_{is}\right| \\
&\leq \tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\left|\int (\widehat{F}^*_{jt} - F_{jt})\, d\widehat{F}^*_{is}\right| + \left|\int F_{jt}\, d(\widehat{F}^*_{is} - F_{is})\right| \\
&= \tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\left|\int (\widehat{F}^*_{jt} - F_{jt})\, d\widehat{F}_{is}\right| + \left|\int (F_{is} - \widehat{F}^*_{is})\, dF_{jt}\right| \\
&\leq \tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\Big(||\widehat{F}^*_{jt} - F_{jt}||_\infty + ||F_{is} - \widehat{F}^*_{is}||_\infty\Big).
\end{aligned}$$

From Proposition A1, we know that the differences between the distribution function and the empirical distribution function converge to zero, independent from the kind of weights. Therefore, as a finite sum of zero sequences, it holds that $\widehat{p}^*_{is} - p_{is} \xrightarrow{a.s.} 0$. From this, it directly follows that $\widehat{\boldsymbol{p}}^* - \boldsymbol{p} \xrightarrow{a.s.} 0$. □

With $\boldsymbol{\psi}^*_{hk} = (\Psi^*_{11,hk}, \Psi^*_{12,hk}, \ldots, \Psi^*_{ad,hk})$, $h = 1, \ldots, a$ $k = 1, \ldots, n_h$ consider the vector

$$\sqrt{N}\boldsymbol{B} = \left(\sum_{h=1}^{a}\sqrt{N}\cdot\frac{1}{n_h}\cdot\sum_{k=1}^{n_h}[\boldsymbol{\psi}^*_{hk} - \mathbb{E}(\boldsymbol{\psi}^*_{hk})]\right)$$

based on random variables defined as

$$\Psi^*_{is,hk} := \begin{cases} -\frac{n_h}{ad}\sum_{t=1}^{d}\sum_{u=1}^{m_{htk}}\lambda_{htk}w^*_{htk}F_{is}(X_{htku}) \text{ , for } h \neq i \\ [+2ex] \quad \frac{n_h}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d}\sum_{u=1}^{m_{isk}}\lambda_{isk}w^*_{isk}F_{jt}(X_{isku}) \\ [+2ex] + \frac{n_h}{ad}\sum_{t=1}^{d}\Big(\sum_{u=1}^{m_{isk}}\lambda_{isk}w^*_{isk}F_{it}(X_{isku}) \\ [+2ex] - \sum_{u=1}^{m_{itk}}\lambda_{itk}w^*_{itk}F_{is}(X_{itku})\Big) \text{ , else} \end{cases}$$

with expectation values

$$\beta^*_{is,hk} := \mathbb{E}(\Psi^*_{is,hk}) = \begin{cases} -\frac{n_h}{ad}\sum_{t=1}^{d}\lambda_{htk}m_{htk}w^*_{htk}p^{(is,ht)} \text{ , for } h \neq i \\ [+2ex] \quad \frac{n_h}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d}m_{isk}\lambda_{isk}w^*_{isk}p^{(jt,is)}) \\ [+2ex] + \frac{n_h}{ad}\sum_{t=1}^{d}\Big(m_{isk}\lambda_{isk}w^*_{isk}p^{(it,is)} \\ [+2ex] - m_{itk}\lambda_{itk}w^*_{itk}p^{(is,it)}\Big) \text{ , else.} \end{cases}$$

The term $\sqrt{N}\boldsymbol{B}$ can be used to calculate the asymptotic distribution of $\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p})$.

Proof of Theorem 1. It is clear that

$$\int F_{is}d\widehat{G}^* = \frac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\int F_{is}d\widehat{F}^*_{jt} = \frac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\sum_{k=1}^{n_j}\lambda_{jtk}w^*_{jtk}\sum_{u=1}^{m_{htk}}F_{is}(X_{jtku}),$$

$$\int Gd\widehat{F}^*_{is} = \sum_{k=1}^{n_i}\lambda_{isk}w^*_{isk}\sum_{u=1}^{m_{isk}}G(X_{isku}) = \frac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\sum_{k=1}^{n_i}\lambda_{isk}w^*_{isk}\sum_{u=1}^{m_{isk}}F_{jt}(X_{isku}).$$

Analogously to Theorem 1 from Rubarth et al. [31] for the s-th component from the i-th group, it holds that

$$\begin{aligned}
& \sqrt{N}(\widehat{p}^*_{is} - p_{is}) \\
= \quad & \sqrt{N}\textstyle\int Gd\widehat{F}^*_{is} - \sqrt{N}\int \widehat{G}^*dF_{is} \\
= \quad & \sqrt{N}\textstyle\int \widehat{G}^*d(\widehat{F}_{is} - F_{is}) + \sqrt{N}\int \widehat{G}^*dF_{is} - \sqrt{N}p_{is} \\
= \quad & \sqrt{N}\textstyle\int Gd(\widehat{F}^*_{is} - F_{is}) + \sqrt{N}\int \widehat{G}^*dF_{is} - \sqrt{N}p_{is} + \mathcal{O}_P(1) \\
= \quad & \sqrt{N}\Big(\textstyle\int Gd\widehat{F}^*_{is} + \int \widehat{G}^*dF_{is} - 2p_{is}\Big) + \mathcal{O}_P(1) \\
= \quad & \sqrt{N}\Big(\textstyle\int Gd\widehat{F}^*_{is} + 1 - \int F_{is}d\widehat{G}^* - 2p_{is}\Big) + \mathcal{O}_P(1) \\
= \quad & \sqrt{N}\Bigg(\tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\sum_{k=1}^{n_i}\sum_{r=1}^{m_{isk}}\lambda_{isk}w^*_{isk}F_{jt}(X_{iskr}) - \tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\sum_{\ell=1}^{n_j}\sum_{u=1}^{m_{jt\ell}}\lambda_{jt\ell}w^*_{jt\ell}F_{is}(X_{jt\ell u}) \\
& +(1-2p_{is})) + \mathcal{O}_P(1) \\
= \quad & \sqrt{N}\Bigg(\tfrac{1}{ad}\sum_{j=1}^{a}\sum_{t=1}^{d}\Bigg[\sum_{k=1}^{n_i}\sum_{r=1}^{m_{isk}}\lambda_{isk}w^*_{isk}F_{jt}(X_{iskr}) - \sum_{\ell=1}^{n_j}\sum_{u=1}^{m_{jt\ell}}\lambda_{jt\ell}w^*_{jt\ell}F_{is}(X_{jt\ell u})\Bigg] \\
& +(1-2p_{is})) + \mathcal{O}_P(1)
\end{aligned}$$

$$
\begin{aligned}
= \quad & \sqrt{N}\Bigg(\tfrac{1}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d}\sum_{k=1}^{n_i}\sum_{u=1}^{m_{isk}} \lambda_{isk} w^*_{isk} F_{jt}(X_{isku}) \\
& + \tfrac{1}{ad}\sum_{k=1}^{n_i}\sum_{t=1}^{d}\Bigg[\sum_{u=1}^{m_{isk}}\lambda_{isk}w^*_{isk}F_{it}(X_{isku}) - \sum_{u=1}^{m_{itk}}\lambda_{itk}w^*_{itk}F_{is}(X_{itku})\Bigg] \\
& - \tfrac{1}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d}\sum_{\ell=1}^{n_j}\sum_{u=1}^{m_{jt\ell}}\lambda_{jt\ell}w^*_{jt\ell}F_{is}(X_{jt\ell u}) + (1-2p_{is})\Bigg) + \mathcal{O}_P(1) \\
= \quad & \sqrt{N}\Bigg(\sum_{h=1}^{a} \frac{1}{n_h}\sum_{k=1}^{n_h}\Psi^*_{is,hk} + (1-2p_{is})\Bigg) + \mathcal{O}_P(1),
\end{aligned}
$$

where the stochastic convergence to zero, denoted by $\mathcal{O}_P(1)$, holds regarding $\lambda_{\min} \to \infty$.

Considering now the expectations, this leads to

$$\beta^*_{is,hk} = -\frac{n_h}{ad}\sum_{t=1}^{d}\lambda_{htk}m_{htk}w^*_{htk}p^{(is,ht)}$$

for $h \neq i$ and else to

$$
\begin{aligned}
\beta^*_{is,ik} = \quad & \tfrac{n_i}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d} m_{isk}\lambda_{isk}w^*_{isk}p^{(jt,is)} \\
& + \tfrac{n_i}{ad}\sum_{t=1}^{d}\Big[m_{isk}\lambda_{isk}w^*_{isk}p^{(it,is)} - m_{itk}\lambda_{itk}w^*_{itk}p^{(is,it)}\Big]
\end{aligned}
$$

with $p^{(is,ht)} := \int F_{is}\, dF_{ht}$. Therefore, we can calculate

$$
\begin{aligned}
\sum_{k=1}^{n_i}\mathbb{E}(\Psi^*_{is,ik}) \quad = \quad & \sum_{k=1}^{n_i}\Bigg(\tfrac{n_i}{ad}\sum_{j\neq i}^{a}\sum_{t=1}^{d} m_{isk}\lambda_{isk}w^*_{isk}p^{(jt,is)} \\
& + \tfrac{n_i}{ad}\sum_{t=1}^{d}\Big[m_{isk}\lambda_{isk}w^*_{isk}p^{(it,is)} - m_{itk}\lambda_{itk}w^*_{itk}p^{(is,it)}\Big]\Bigg) \\
= \quad & \sum_{j\neq i}^{a}\tfrac{n_i}{ad}\sum_{t=1}^{d}p^{(jt,is)} + \tfrac{n_i}{ad}\sum_{t=1}^{d}\Big[p^{(it,is)} - p^{(is,it)}\Big]
\end{aligned}
$$

as well as, for $h \neq i$,

$$\sum_{k=1}^{n_i}\mathbb{E}(\Psi^*_{is,hk}) \quad = -\tfrac{n_h}{ad}\sum_{k=1}^{n_i}\sum_{t=1}^{d}\lambda_{htk}m_{htk}w^*_{htk}p^{(is,ht)} = -\tfrac{n_h}{ad}\sum_{t=1}^{d}p^{(is,ht)}.$$

In total, we obtain

$$
\begin{aligned}
& \mathbb{E}\Bigg(\sum_{h=1}^{a}\frac{1}{n_h}\sum_{k=1}^{n_h}\Psi^*_{is,hk}\Bigg) \\
= \quad & \Bigg(\sum_{h\neq i}^{a}\frac{1}{n_h}\sum_{k=1}^{n_h}\mathbb{E}(\Psi^*_{is,hk}) + \frac{1}{n_i}\sum_{k=1}^{n_g}\mathbb{E}(\Psi^*_{is,ik})\Bigg)
\end{aligned}
$$

$$\begin{aligned}
&= \tfrac{1}{ad}\left(- \sum_{h\neq i}^{a} \sum_{t=1}^{d} p^{(is,ht)} + \sum_{j\neq i}^{a} \sum_{t=1}^{d} p^{(jt,is)} + \sum_{t=1}^{d} \left[p^{(it,is)} - p^{(is,it)} \right] \right) \\
&= \tfrac{1}{ad}\left(- \sum_{h=1}^{a} \sum_{t=1}^{d} p^{(is,ht)} + \sum_{j=1}^{a} \sum_{t=1}^{d} p^{(jt,is)} \right) \\
&= -(1 - p_{is}) + p_{is} \\
&= -(1 - 2p_{is}).
\end{aligned}$$

Together with the other equation, we obtain

$$\sqrt{N}(\widehat{p}^*_{is} - p_{is}) = \sqrt{N}\left(\sum_{h=1}^{a} \frac{1}{n_h} \cdot \sum_{k=1}^{n_h} \left[\Psi^*_{is,hk} - \mathbb{E}(\Psi^*_{is,hk}) \right] \right) + \mathcal{O}_P(1)$$

resp.

$$\sqrt{N}(\widehat{\boldsymbol{p}}^* - \boldsymbol{p}) = \sqrt{N}\left(\sum_{h=1}^{a} \frac{1}{n_h} \cdot \sum_{k=1}^{n_h} [\boldsymbol{\psi}^*_{hk} - \mathbb{E}(\boldsymbol{\psi}^*_{hk})] \right) + \mathcal{O}_P(1).$$

□

Proof of Theorem 2. Through the construction of $\Psi^*_{is,hk}$ and **A1.2**, it holds that $|\Psi^*_{is,hk}| \leq \mathcal{O}(N_0)$, thus, these random variables have finite moments. Due to the independence of all $\boldsymbol{\psi}^*_{hk}$ by Lindeberg–Feller Theorem, we obtain for $\boldsymbol{B}^*_h = \frac{1}{n_h} \sum_{k=1}^{n_h} \left[\boldsymbol{\psi}^*_{hk} - \mathbb{E}(\boldsymbol{\psi}^*_{hk}) \right]$ that $\sqrt{n_h}\boldsymbol{B}^*_h$ is asymptotically distributed like a normal distributed random vector with expectation vector $\mathbf{0}$ and covariance matrix $\boldsymbol{V}^*_{N,h}$. Hereby, two of the requirements of the the Lindeberg–Feller Theorem are obviously fulfilled, since the random variables are uniformly bounded and centered. However, it is not sure that $\boldsymbol{V}^*_{N,h}$, which depends on N, converge to a fixed matrix. However, through the fact that all random variables are bounded, it follows that the covariance matrix is also bounded, and for each sequence we can find a subsequence where the covariance matrix converges. For each of this subsequences the asymptotic distribution of $\boldsymbol{B}^*_h$ matches with the actual normal distribution. Therefore, the result holds in general. Through the independence of the $\boldsymbol{B}^*_h$ and $\sqrt{N}\boldsymbol{B}^* = \sum_{h=1}^{a} \frac{\sqrt{N}}{\sqrt{n_h}} \sqrt{n_h}\boldsymbol{B}^*_h$ together with condition **A1.2** the result follows. □

Since the covariance matrix is unknown and the random variables are not observable, we use estimated versions of these random variables to estimate the covariance matrix. They are defined as

$$\widehat{\Psi}^*_{is,hk} := \begin{cases} -\frac{n_h}{ad} \sum_{t=1}^{d} \sum_{u=1}^{m_{htk}} \lambda_{htk} w^*_{htk} \widehat{F}^*_{is}(X_{htku}) \text{ , for } h \neq i \\ \frac{n_h}{ad} \sum_{j\neq i}^{a} \sum_{t=1}^{d} \sum_{u=1}^{m_{isk}} \lambda_{isk} w^*_{isk} \widehat{F}^*_{jt}(X_{isku}) \\ + \frac{n_h}{ad} \sum_{t=1}^{d} \Big(\sum_{u=1}^{m_{isk}} \lambda_{isk} w^*_{isk} \widehat{F}^*_{it}(X_{isku}) \\ - \sum_{u=1}^{m_{itk}} \lambda_{itk} w^*_{itk} \widehat{F}^*_{is}(X_{itku}) \Big) \text{ , else} \end{cases}$$

and for the expectation

$$\widehat{\beta}^*_{is,ik} := E(\widehat{\Psi}^*_{is,hk}) = \begin{cases} -\frac{n_h}{ad} \sum_{t=1}^{d} \lambda_{htk} m_{htk} w^*_{htk} \widehat{p}^{*(is,ht)} \text{ , for } h \neq i \\ \frac{n_h}{ad} \sum_{j\neq i}^{a} \sum_{t=1}^{d} m_{isk} \lambda_{isk} w^*_{isk} \widehat{p}^{*(jt,is)}) \\ + \frac{nh}{ad} \sum_{t=1}^{d} \Big(m_{isk} \lambda_{isk} w^*_{isk} \widehat{p}^{*(it,is)} \\ - m_{itk} \lambda_{itk} w^*_{itk} \widehat{p}^{(is,it)} \Big) \text{ , else.} \end{cases}$$

Based on these variables, we define an estimator for the unknown covariance matrix $\boldsymbol{V}_{N,h}$ through

$$\widehat{\boldsymbol{V}}^* = \frac{1}{n_h - 1}\sum_{k=1}^{n_h}(\widehat{\boldsymbol{p}}\boldsymbol{si}^*_{hk} - \widehat{\boldsymbol{\beta}}^*_{hk})(\widehat{\boldsymbol{p}}\boldsymbol{si}^*_{hk} - \widehat{\boldsymbol{\beta}}^*_{hk})',$$

whereby an estimator for $\boldsymbol{V}^*_N := \sum_{h=1}^{a} \kappa_h^{-1} \cdot \boldsymbol{V}^*_{N,h}$ is given by $\widehat{\boldsymbol{V}}^*_N := \sum_{h=1}^{a} \frac{N}{n_h} \cdot \widehat{\boldsymbol{V}}^*_{N,h}$.

Proof of Theorem 5. 1. Let $\boldsymbol{y} = (y_1, \ldots, y_{ad})$ be an arbitrary vector. Then, it holds that

$$\begin{aligned} \boldsymbol{y}'\widehat{\boldsymbol{V}}^*_h\boldsymbol{y} &= \tfrac{1}{n_h-1}\sum_{k=1}^{n_h}\Big[\boldsymbol{y}'(\widehat{\boldsymbol{p}}\boldsymbol{si}^*_{hk} - \widehat{\boldsymbol{\beta}}^*_{hk})\Big]\Big[\boldsymbol{y}'(\widehat{\boldsymbol{p}}\boldsymbol{si}^*_{hk} - \widehat{\boldsymbol{\beta}}^*_{hk})\Big]' \\ &= \tfrac{1}{n_h-1}\sum_{k=1}^{n_h}\Big[\boldsymbol{y}'(\widehat{\boldsymbol{p}}\boldsymbol{si}^*_{hk} - \widehat{\boldsymbol{\beta}}^*_{hk})\Big]^2 \geq 0. \end{aligned}$$

Since $\widehat{\boldsymbol{V}}^*_N$ is a convex combination of positive semi-definite matrices, it is also positive semi-definite.

2. It is clear that

$$\widetilde{\boldsymbol{V}}^*_{N,h} = \frac{1}{n_h - 1}\sum_{k=1}^{n_h}(\boldsymbol{\psi}^*_{hk} - \boldsymbol{\beta}^*_{hk})(\boldsymbol{\psi}^*_{hk} - \boldsymbol{\beta}^*_{hk})'$$

is a consistent estimator for $\boldsymbol{V}^*_{N,h}$. To demonstrate consistency, we use the triangle inequality and prove that $\widehat{\boldsymbol{V}}^*_{N,h} - \widetilde{\boldsymbol{V}}^*_{N,h} \xrightarrow{a.s} \boldsymbol{0}$. First, we remember that $|\Psi^*_{hk}| \leq \mathcal{O}(N_0)$ and with the same arguments it follows that $|\widehat{\Psi}^*_{hk}| \leq \mathcal{O}(N_0)$, $|\beta^*_{hk}| \leq \mathcal{O}(N_0)$ and $|\widehat{\beta}^*_{hk}| \leq \mathcal{O}(N_0)$.

Then, we consider the single components, where for the diagonal elements it holds that

$$\begin{aligned} &\Big|\tfrac{1}{n_h-1}\sum_{k=1}^{n_h}\Big(\widehat{\Psi}^*_{is,hk} - \widehat{\beta}^*_{is,hk}\Big)^2\Big(\Psi^*_{is,hk} - \beta^*_{is,hk}\Big)^2\Big| \\ \leq\ & 2 \cdot \max_{k=1,\ldots,n_h}\Big|\Big(\widehat{\Psi}^*_{is,hk} - \widehat{\beta}^*_{is,hk}\Big)^2\Big(\Psi^*_{is,hk} - \beta^*_{is,hk}\Big)^2\Big| \\ \leq\ & 2 \cdot \max_{k=1,\ldots,n_h}\Big|\Big(\widehat{\Psi}^*_{is,hk} - \widehat{\beta}^*_{is,hk} + \Psi^*_{is,hk} - \beta^*_{is,hk}\Big)\Big(\widehat{\Psi}^*_{is,hk} - \widehat{\beta}^*_{is,hk} - \Psi^*_{is,hk} + \beta^*_{is,hk}\Big)\Big| \\ \leq\ & \mathcal{O}(N_0) \cdot \max_{k=1,\ldots,n_h}|\widehat{\Psi}^*_{is,hk} - \widehat{\beta}^*_{is,hk} - \Psi^*_{is,hk} + \beta^*_{is,hk}| \\ \leq\ & \mathcal{O}(N_0) \cdot \max_{k=1,\ldots,n_h}\big|\widehat{\Psi}^*_{is,hk} - \Psi^*_{is,hk}\big| + \mathcal{O}(N_0) \cdot \max_{k=1,\ldots,n_h}\big|\beta^*_{is,hk} - \widehat{\beta}^*_{is,hk}\big|. \end{aligned}$$

First, we consider for $h \neq i$

$$\begin{aligned} |\Psi^*_{is,hk} - \widehat{\Psi}^*_{is,hk}| &= \tfrac{n_h}{ad}\Big|\sum_{t=1}^{d}\sum_{u=1}^{m_{htk}}\lambda_{htk}w^*_{htk}\Big(F_{is}(X_{htku}) - \widehat{F}^*_{is}(X_{htku})\Big)\Big| \\ &\leq \tfrac{n_h}{ad}\sum_{t=1}^{d} m_{ikt}\lambda_{htk}w^*_{htk}||F_{is} - \widehat{F}^*_{is}||_\infty \\ &\leq n_h \cdot \mathcal{O}(\lambda_{\min}^{-1}) \cdot ||F_{is} - \widehat{F}^*_{is}||_\infty \xrightarrow{a.s} 0 \end{aligned}$$

and similar

$$
\begin{aligned}
& |\boldsymbol{\Psi}^*_{is,ik} - \widehat{\boldsymbol{\Psi}}^*_{is,ik}| \\
= \quad & \tfrac{n_h}{ad}\Big| \sum_{j\neq i}^{a}\sum_{t=1}^{d}\sum_{u=1}^{m_{isk}} \lambda_{isk} w^*_{isk} \cdot \Big(F_{jt}(X_{isku}) - \widehat{F}^*_{jt}(X_{isku})\Big) \\
& + \sum_{t=1}^{d}\sum_{u=1}^{m_{isk}} \lambda_{isk} w^*_{isk} \cdot \Big(F_{it}(X_{isku}) - \widehat{F}^*_{it}(X_{isku})\Big) \\
& - \sum_{t=1}^{d}\sum_{u=1}^{m_{itk}} \lambda_{itk} w^*_{itk} \cdot \Big(F_{is}(X_{itku}) - \widehat{F}^*_{is}(X_{itku})\Big)\Big| \\
\leq \quad & \tfrac{n_h}{ad}\Bigg(\sum_{j\neq i}^{a}\sum_{t=1}^{d} m_{isk}\lambda_{isk} w^*_{isk} \cdot \big|\big|F_{jt} - \widehat{F}^*_{jt}\big|\big|_\infty \\
& + \sum_{t=1}^{d} m_{isk}\lambda_{isk} w^*_{isk} \cdot \big|\big|F_{it} - \widehat{F}^*_{it}\big|\big|_\infty + \sum_{t=1}^{d} m_{itk}\lambda_{itk} w^*_{itk} \cdot \big|\big|F_{is} - \widehat{F}^*_{is}\big|\big|_\infty \Bigg) \\
\leq \quad & \tfrac{n_h}{ad}\Bigg(\sum_{j\neq i}^{a}\sum_{t=1}^{d} \mathcal{O}(\lambda^{-1}_{\min}) \cdot \big|\big|F_{jt} - \widehat{F}^*_{jt}\big|\big|_\infty \\
& + \sum_{t=1}^{d} \mathcal{O}(\lambda^{-1}_{\min}) \cdot \big|\big|F_{it} - \widehat{F}^*_{it}\big|\big|_\infty + \sum_{t=1}^{d} \mathcal{O}(\lambda^{-1}_{\min}) \cdot \big|\big|F_{is} - \widehat{F}^*_{is}\big|\big|_\infty \Bigg) \\
\leq \quad & \tfrac{\mathcal{O}(N_0)}{ad}\Bigg(\sum_{j\neq i}^{a}\sum_{t=1}^{d} \big|\big|F_{jt} - \widehat{F}^*_{jt}\big|\big|_\infty + \sum_{t=1}^{d} \big|\big|F_{it} - \widehat{F}^*_{it}\big|\big|_\infty + \sum_{t=1}^{d} \big|\big|F_{is} - \widehat{F}^*_{is}\big|\big|_\infty \Bigg) \\
& \xrightarrow{a.s} 0.
\end{aligned}
$$

For the expectation, we calculate for $h \neq i$

$$
\begin{aligned}
|\beta^*_{is,hk} - \widehat{\beta}^*_{is,hk}| &= \tfrac{n_h}{ad}\Big| \sum_{t=1}^{d} \lambda_{htk} m_{htk} w_{htk} \Big(p^{(is,ht)} - \widehat{p}^{*(is,ht)}\Big)\Big| \\
&\leq \tfrac{n_h}{ad} \sum_{t=1}^{d} \mathcal{O}(\lambda^{-1}_{\min}) \cdot \Big|p^{(is,ht)} - \widehat{p}^{*(is,ht)}\Big| \xrightarrow{a.s} 0
\end{aligned}
$$

and

$$\begin{aligned}
&|\beta^*_{is,ik} - \widehat{\beta}^*_{is,ik}| \\
\leq\;& \tfrac{n_h}{ad} \sum_{j\neq i}^{a} \sum_{t=1}^{d} m_{isk}\lambda_{isk}w^*_{isk}\Big(p^{(jt,is)} - \widehat{p}^{*(jt,is)}\Big) \\
&+ \tfrac{n_h}{ad} \sum_{t=1}^{d} \Big[m_{isk}\lambda_{isk}w^*_{isk}\Big(p^{(it,is)} - \widehat{p}^{*(it,is)}\Big) - m_{itk}\lambda_{itk}w^*_{itk}\Big(p^{(is,it)} - \widehat{p}^{*(is,it)}\Big)\Big] \\
\leq\;& \tfrac{n_h}{ad} \sum_{j\neq i}^{a} \sum_{t=1}^{d} m_{isk}\lambda_{isk}w^*_{isk}\big|p^{(jt,is)} - \widehat{p}^{*(jt,is)}\big| \\
&+ \tfrac{n_h}{ad} \sum_{t=1}^{d} \Big[m_{isk}\lambda_{isk}w^*_{isk}\big|p^{(it,is)} - \widehat{p}^{*(it,is)}\big| + m_{itk}\lambda_{itk}w^*_{itk}\big|p^{(is,it)} - \widehat{p}^{*(is,it)}\big|\Big] \\
\leq\;& \tfrac{\mathcal{O}(N_0)}{ad}\left(\sum_{j\neq i}^{a} \sum_{t=1}^{d} \big|p^{(jt,is)} - \widehat{p}^{*(jt,is)}\big| + \sum_{t=1}^{d}\Big[\big|p^{(it,is)} - \widehat{p}^{*(it,is)}\big| + \big|p^{(is,it)} - \widehat{p}^{*(is,it)}\big|\Big]\right) \\
&\xrightarrow{a.s} 0.
\end{aligned}$$

For the off diagonal elements we can use the same convergences, but first we define $\widehat{\Delta}^*_{is,hk} := \Big(\widehat{\Psi}^*_{is,hk} - \widehat{\beta}^*_{is,hk}\Big)$ and $\Delta^*_{is,hk} := \Big(\Psi^*_{is,hk} - \beta^*_{is,hk}\Big)$ which fulfill $|\widehat{\Delta}^*_{is,hk}| \leq \mathcal{O}(N_0)$ and $|\Delta^*_{is,hk}| \leq \mathcal{O}(N_0)$. Then, for the elements which correspond to $\Psi^*_{is,hk}$ and $\Psi^*_{jt,hk}$ we obtain

$$\begin{aligned}
&\Big|\tfrac{1}{n_h-1}\sum_{k=1}^{n_h} \widehat{\Delta}^*_{is,hk}\widehat{\Delta}^*_{jt,hk} - \Delta^*_{is,hk}\Delta^*_{jt,hk}\Big| \\
=\;& \Big|\tfrac{1}{n_h-1}\sum_{k=1}^{n_h} \widehat{\Delta}^*_{is,hk}\widehat{\Delta}^*_{jt,hk} - \Delta^*_{is,hk}\Delta^*_{jt,hk} + \widehat{\Delta}^*_{is,hk}\Delta^*_{jt,hk} - \widehat{\Delta}^*_{is,hk}\Delta^*_{jt,hk}\Big| \\
=\;& \Big|\tfrac{1}{n_h-1}\sum_{k=1}^{n_h} \widehat{\Delta}^*_{is,hk}\Big(\widehat{\Delta}^*_{jt,hk} - \Delta^*_{jt,hk}\Big) + \Delta^*_{jt,hk}\Big(\widehat{\Delta}^*_{is,hk} - \Delta^*_{is,hk}\Big)\Big| \\
\leq\;& \tfrac{1}{n_h-1}\sum_{k=1}^{n_h}\Big[\big|\widehat{\Delta}_{gi,hk}\Big(\widehat{\Delta}^*_{jt,hk} - \Delta^*_{jt,hk}\Big)\big| + \big|\Delta^*_{jt,hk}\Big(\widehat{\Delta}^*_{is,hk} - \Delta^*_{is,hk}\Big)\big|\Big] \\
\leq\;& \tfrac{1}{n_h-1}\sum_{k=1}^{n_h}\Big[\mathcal{O}(N_0)\big|\Big(\widehat{\Delta}^*_{jt,hk} - \Delta^*_{jt,hk}\Big)\big| + \mathcal{O}(N_0)\big|\Big(\widehat{\Delta}^*_{is,hk} - \Delta^*_{is,hk}\Big)\big|\Big] \\
\leq\;& \tfrac{2}{n_h}\sum_{k=1}^{n_h}\Big[\mathcal{O}(N_0)\big|\Big(\widehat{\Delta}^*_{jt,hk} - \Delta^*_{jt,hk}\Big)\big| + \mathcal{O}(N_0)\big|\Big(\widehat{\Delta}^*_{is,hk} - \Delta^*_{is,hk}\Big)\big|\Big] \\
\leq\;& \mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\Big(\widehat{\Delta}^*_{jt,hk} - \Delta^*_{jt,hk}\Big)\big| + \mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\Big(\widehat{\Delta}^*_{is,hk} - \Delta^*_{is,hk}\Big)\big| \\
=\;& \mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\Big(\widehat{\Psi}^*_{jt,hk} - \widehat{\beta}^*_{jt,hk}\Big) - \Big(\Psi^*_{jt,hk} - \beta^*_{jt,hk}\Big)\big| \\
&+\mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\Big(\widehat{\Psi}^*_{is,hk} - \widehat{\beta}^*_{is,hk}\Big) - \Big(\Psi^*_{is,hk} - \beta^*_{is,hk}\Big)\big| \\
\leq\;& \mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\widehat{\Psi}^*_{jt,hk} - \Psi^*_{jt,hk}\big| + \mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\beta^*_{jt,hk} - \widehat{\beta}^*_{jt,hk}\big| \\
&+\mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\widehat{\Psi}^*_{is,hk} - \Psi^*_{is,hk}\big| + \mathcal{O}(N_0)\max_{k=1,\dots,n_h}\big|\beta^*_{is,hk} - \widehat{\beta}^*_{is,hk}\big| \xrightarrow{a.s.} 0
\end{aligned}$$

Since we demonstrated consistency for the diagonal elements and the off diagonal elements, it follows that $\widehat{\boldsymbol{V}}^*_{N,h} - \widetilde{\boldsymbol{V}}^*_{N,h} \xrightarrow{a.s.} 0$ and consequently $\widehat{\boldsymbol{V}}^*_{N,h} - \boldsymbol{V}^*_{N,h} \xrightarrow{a.s.} 0$.

3. Follows directly from 2. with Slutzky's theorem. □

References

1. Roy, A.; Harrar, S.W.; Konietschke, F. The nonparametric Behrens-Fisher problem with dependent replicates. *Stat. Med.* **2019**, *38*, 4939–4962. doi: doi: 10.1002/sim.8343. [CrossRef] [PubMed]
2. 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*, 755–771. [CrossRef]
3. Cui, Y.; Konietschke, F.; Harrar, S.W. The nonparametric Behrens–Fisher problem in partially complete clustered data. *Biom. J.* **2021**, *63*, 148–167. [CrossRef] [PubMed]
4. Gao, X. A Nonparametric Procedure for the Two-Factor Mixed Model with Missing Data. *Biom. J.* **2007**, *49*, 774–788. [CrossRef] [PubMed]
5. Fitzmaurice, G.; Laird, N.; Ware, J. *Applied Longitudinal Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
6. Johnson, R.A.; Wichern, D. *Applied Multivariate Statistical Analysis*; Pearson Education Limited: London, UK, 2007.
7. Mann, H.B.; Whitney, D.R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60. doi: 10.1214/aoms/1177730491. [CrossRef]
8. Brunner, E.; Munzel, U. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biom. J.* **2000**, *42*, 17–25. [CrossRef]
9. Thas, O.; Neve, J.D.; Clement, L.; Ottoy, J.P. Probabilistic index models. *J. R. Stat. Soc. Ser. B* **2012**, *74*, 623–671. [CrossRef]
10. Acion, L.; Peterson, J.J.; Temple, S.; Arndt, S. Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Stat. Med.* **2006**, *25*, 591–602. [CrossRef]
11. Brunner, E.; Bathke, A.C.; Konietschke, F. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*; Springer: Berlin/Heidelberg, Germany, 2018.
12. Akritas, M.; Kuha, J.; Osgood, W. A Nonparametric Approach to Matched Pairs with Missing Data. *Sociol. Methods Res.* **2002**, *30*, 425–454. doi: 10.1177/0049124102030003006. [CrossRef]
13. Fong, Y.; Huang, Y.; Lemos, M.; Mcelrath, J. Rank-based two-sample tests for paired data with missing values. *Biostatistics* **2018**, *19*, 281–294. [CrossRef]
14. Domhof, S.; Brunner, E.; Osgood, W. Rank Procedures for Repeated Measures with Missing Values. *Sociol. Methods Res.* **2002**, *30*, 367–393. [CrossRef]
15. Amro, L.; Konietschke, F.; Pauly, M. Incompletely observed nonparametric factorial designs with repeated measurements: A wild bootstrap approach. *arXiv* **2021**, arXiv:2102.02871.
16. Akritas, M.; Brunner, E. A unified approach to rank tests for mixed models. *J. Stat. Plan. Inference* **1997**, *61*, 249–277. [CrossRef]
17. Brunner, E.; Munzel, U.; Puri, M.L. Rank-Score Tests in Factorial Designs with Repeated Measures. *J. Multivar. Anal.* **1999**, *70*, 286–317. [CrossRef]
18. Brunner, E.; Domhof, S.; Langer, F. *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*; Wiley-Interscience: Hoboken, NJ, USA, 2002; Volume 373.
19. Klumbies, K.; Rust, R.; Dörr, J.; Konietschke, F.; Paul, F.; Bellmann-Strobl, J.; Brandt, A.; Zimmermann, H.G. Retinal Thickness Analysis in Progressive Multiple Sclerosis Patients Treated With Epigallocatechin Gallate: Optical Coherence Tomography Results From the SUPREMES Study. *Front. Neurol.* **2021**, *12*, 615790. doi: 10.3389/fneur.2021.615790. [CrossRef]
20. Walton, C.; King, R.; Rechtman, L.; Kaye, W.; Leray, E.; Marrie, R.A.; Robertson, N.; Rocca, N.L.; Uitdehaag, B.; van der Mei, I.; et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult. Scler. J.* **2020**, *26*, 1816–1821.
21. Reich, D.S.; Lucchinetti, C.F.; Calabresi, P.A. Multiple Sclerosis. *N. Engl. J. Med.* **2018**, *378*, 169–180. NEJMra1401483. [CrossRef]
22. Petzold, A.; Balcer, L.J.; Calabresi, P.A.; Costello, F.; Frohman, T.C.; Frohman, E.M.; Martinez-Lapiscina, E.H.; Green, A.J.; Kardon, R.; Outteryck, O.; et al. Retinal layer segmentation in multiple sclerosis: a systematic review and meta-analysis. *Lancet Neurol.* **2017**, *16*, 797–812. [CrossRef]
23. Oertel, F.C.; Zimmermann, H.G.; Brandt, A.U.; Paul, F. Optical coherence tomography in neuromyelitis optica spectrum disorders: potential advantages for individualized monitoring of progression and therapy *Expert Rev. Neurother.* **2019**, *19*, 31–43.
24. Ruymgaart, F. *A Unified Approach to the Asymptotic Distribution Theory of Certain Midrank Statistics*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 821, pp. 1–18.
25. Brunner, E.; Konietschke, F.; Pauly, M.; Puri, M. Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects. *J. R. Stat. Soc. Ser. B* **2016**, *79*, 1463–1485. [CrossRef]
26. Brunner, E.; Konietschke, F.; Bathke, A.C.; Pauly, M. Ranks and Pseudo-ranks—Surprising Results of Certain Rank Tests in Unbalanced Designs. *Int. Stat. Rev.* **2020**, 89, 349–366. [CrossRef]
27. Obuchowski, N.A. Nonparametric analysis of clustered ROC curve data. *Biometrics* **1997**, *53*, 567–578. [CrossRef]
28. Zou, G. Confidence interval estimation for treatment effects in cluster randomization trials based on ranks. *Stat. Med.* **2021**, *40*, 3227–3250. doi: doi: 10.1002/sim.8918. [CrossRef]

29. Hoffman, E.; Sen, P.; Weinberg, C. Within-Cluster Resampling. *Biometrika* **2001**, *88*, 1121–1134. [CrossRef]
30. Williamson, J.; Datta, S.; Satten, G. Marginal Analyses of Clustered Data When Cluster Size Is Informative. *Biometrics* **2003**, *59*, 36–42. doi: 10.1111/1541-0420.00005. [CrossRef]
31. Rubarth, K.; Pauly, M.; Konietschke, F. Ranking Procedures for Repeated Measures Designs with Missing Data: Estimation, Testing and Asymptotic Theory. *Stat. Methods Med. Res.* **2022**, *31*, 105–118. [CrossRef] [PubMed]
32. Konietschke, F.; Hothorn, L.; Brunner, E. Rank-based multiple test procedures and simultaneous confidence intervals. *Electron. J. Stat.* **2012**, *6*, 738–759. [CrossRef]
33. Konietschke, F.; Bathke, A.; Hothorn, L.; Brunner, E. Testing and estimation of purely nonparametric effects in repeated measures designs. *Comput. Stat. Data Anal.* **2010**, *54*, 1895–1905. [CrossRef]
34. Akritas, M.; Arnold, S.; Brunner, E. Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs. *J. Am. Stat. Assoc.* **1997**, *92*, 258–265. [CrossRef]
35. Bretz, F.; Genz, A.; Hothorn, L. On the Numerical Availability of Multiple Comparison Procedures. *Biom. J.* **2001**, *43*, 645–656. [CrossRef]
36. Konietschke, F.; Harrar, S.W.; Lange, K.; Brunner, E. Ranking procedures for matched pairs with missing data—Asymptotic theory and a small sample approximation. *Comput. Stat. Data Anal.* **2012**, *56*, 1090–1102.
37. Gao, X.; Alvo, M.; Chen, J.; Li, G. Nonparametric multiple comparison procedures for unbalanced one-way factorial designs. *J. Stat. Plan. Inference* **2008**, *138*, 2574–2591. [CrossRef]
38. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
39. Friedrich, S.; Konietschke, F.; Pauly, M. A wild bootstrap approach for nonparametric repeated measurements. *Comput. Stat. Data Anal.* **2017**, *113*, 38–52. [CrossRef]
40. Fay, M.P.; Proschan, M.A. Wilcoxon-Mann-Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat. Surv.* **2010**, *4*, 1–39. doi: 10.1214/09-SS051. [CrossRef]
41. Fagerland, M.W.; Sandvik, L. The Wilcoxon–Mann–Whitney test under scrutiny. *Stat. Med.* **2009**, *28*, 1487–1497. doi: 10.1002/sim.3561. [CrossRef]
42. Bergmann, R.; Ludbrook, J.; Spooren, W.P.J.M. Different Outcomes of the Wilcoxon-Mann-Whitney Test from Different Statistics Packages. *Am. Stat.* **2000**, *54*, 72–77.
43. Fay, M.; Malinovsky, Y. Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test: Confidence Intervals on the Mann-Whitney Parameter. *Stat. Med.* **2018**, *37*, 3991–4006. doi: 10.1002/sim.7890. [CrossRef]
44. Fay, M.; Brittain, E.; Shih, J.; Follmann, D.; Gabriel, E. Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Stat. Med.* **2017**, *37*, 2923–2937. doi: 10.1002/sim.7799. [CrossRef]
45. Hand, D.J. On Comparing Two Treatments. *Am. Stat.* **1992**, *46*, 190–192. doi: 10.1080/00031305.1992.10475881. [CrossRef]
46. Noguchi, K.; Gel, Y.R.; Brunner, E.; Konietschke, F. nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *J. Stat. Softw.* **2012**, *50*, 12. [CrossRef]
47. Domhof, S. Nichtparametrische Relative Effekte. Doctoral Dissertation, Niedersächsische Staats-und Universitätsbibliothek Göttingen, Göttingen, Germany, 2001.