University of Wollongong

Research Online

Faculty of Engineering and Information
Sciences - Papers: Part B

Faculty of Engineering and Information
Sciences

2017

A Mann-Whitney type effect measure of interaction for factorial designs

Jan De Neve

Ghent University, Belgium, JanR.DeNeve@UGent.be

Olivier Thas

University of Wollongong, olivier@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/eispapers1>

Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

A Mann-Whitney type effect measure of interaction for factorial designs

Abstract

We propose a measure for interaction for factorial designs that is formulated in terms of a probability similar to the effect size of the Mann-Whitney test. It is shown how asymptotic confidence intervals can be obtained for the effect size and how a statistical test can be constructed. We further show how the test is related to the test proposed by Bhapkar and Gore [Sankhya A, 36:261-272 (1974)]. The results of a simulation study indicate that the test has good power properties and illustrate when the asymptotic approximations are adequate. The effect size is demonstrated on an example dataset.

Disciplines

Engineering | Science and Technology Studies

Publication Details

De Neve, J. & Thas, O. (2017). A Mann–Whitney type effect measure of interaction for factorial designs. *Communications in Statistics: Theory and Methods*, 46 (22), 11243-11260.

A Mann–Whitney type effect measure of interaction for factorial designs

Jan De Neve & Olivier Thas

To cite this article: Jan De Neve & Olivier Thas (2016): A Mann–Whitney type effect measure of interaction for factorial designs, Communications in Statistics - Theory and Methods, DOI: [10.1080/03610926.2016.1263739](https://doi.org/10.1080/03610926.2016.1263739)

To link to this article: <http://dx.doi.org/10.1080/03610926.2016.1263739>

Accepted author version posted online: 28 Nov 2016.

Submit your article to this journal [↗](#)

Article views: 43

View related articles [↗](#)

View Crossmark data [↗](#)

A Mann–Whitney type effect measure of interaction for factorial designs

Jan De Neve^{1,*} and Olivier Thas^{2,3}

Running head. A measure of interaction for factorial designs

Keywords. factorial designs; interaction; probability of superiority; rank test; Wilcoxon–Mann–Whitney.

Abstract. We propose a measure for interaction for factorial designs that is formulated in terms of a probability similar to the effect size of the Mann–Whitney test. It is shown how asymptotic confidence intervals can be obtained for the effect size and how a statistical test can be constructed. We further show how the test is related to the test proposed by Bhapkar and Gore [Sankhya A, 36:261–272 (1974)]. The results of a simulation study indicate that the test has good power properties and illustrate when the asymptotic approximations are adequate. The effect size is demonstrated on an example dataset.

1 Introduction

Over the years a variety of rank tests have been proposed for testing interaction in a two-way layout. For example, Patel and Hoel (1973) proposed a test based on the differences of two Mann–Whitney statistics, Conover and Iman (1981) considered a rank transform approach as a tool to develop nonparametric procedures, while Mansouri and Chang (1995), among others, constructed aligned rank tests. Akritas and Arnold (1994) considered a different approach by introducing a nonparametric hypothesis of interaction for which they constructed statistical tests. We refer to Gao and Alvo (2005) for an extensive literature overview on this topic in a two-way layout. In this paper we propose a rank test by starting from a particular summary measure for interaction.

More specifically, consider two factors, A and B , each with two levels, labeled 1 and 2, and let Y_{ab} denote the outcome associated with level $a = 1, 2$ of factor A and level $b = 1, 2$ of factor B . The conventional measure of interaction is defined in terms of the mean outcome:

$$\alpha := (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}), \quad (1)$$

where $\mu_{ab} = E(Y_{ab})$. Here α quantifies the difference in the effect of A (where the effect is defined as the difference between two means), between levels 1 and 2 of factor B . Equivalently $\alpha = (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$ expresses the difference in the effect of B between levels 1 and 2 of factor A . Now consider the transformed outcomes

$$Z := Y_{11} - Y_{21} \quad \text{and} \quad Z^* := Y_{12} - Y_{22}.$$

We refer to Z as the *effect outcome* of A at $B = 1$, for it considers the effect of A (now defined as a random variable and thus not restricted to the mean) when B is fixed at level 1. Similarly, Z^* denotes the effect outcome of A at $B = 2$. The distribution of the difference $Z - Z^*$ now describes the effect of the interaction on the entire outcome distribution. As an illustration, consider an experimental set-up where A denotes the treatment ($A = 1$ for the control group and $A = 2$ for the active treatment group) and B the gender ($B = 1$ for men and $B = 2$ for women). The treatment may affect different moments of the outcome distribution and these effects might be different for men and women. One way of summarizing the distribution of the difference $Z - Z^*$ is by considering a measure of location. For example the mean or median. Note that for the mean it follows $E(Z - Z^*) = E(Z) - E(Z^*) = \alpha$. Hence, summarizing the distribution of the differences by the mean results in the conventional measure of interaction (1). To emphasize that α summarizes the interaction in terms of the mean, we further refer to α as the *interaction average*.

If $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 , the top left panel of Figure 1 displays densities for $Y_{11} \stackrel{d}{=} Y_{12} \stackrel{d}{=} Y_{21} \stackrel{d}{=} N(0, 0.5)$ and $Y_{22} \stackrel{d}{=} N(-3, 0.5)$, for which the interaction average is $\alpha = -3$. The bottom left panel shows the densities of the effect outcomes $Z = Y_{11} - Y_{21} \stackrel{d}{=} N(0, 1)$ and $Z^* = Y_{12} - Y_{22} \stackrel{d}{=} N(3, 1)$: the density of the latter is shifted 3 units to the right as compared to the density of the former. Under this location-shift assumption, α captures all information on the difference between Z and Z^* . However, if location-shift does not hold, the average does not always capture all information. The top right panel of Figure 1 shows the densities when $Y_{11} \stackrel{d}{=} Y_{21} \stackrel{d}{=} N(0, 0.5)$, $Y_{12} \stackrel{d}{=} N(0, 16)$ and $Y_{22} \stackrel{d}{=} N(-3, 16)$. For this setting the interaction average is still $\alpha = -3$. However, as can be seen from the bottom right panel, the difference between the densities of Z and Z^* is now less pronounced as compared to the location-shift setting: changing B does not only alter the effect of A on average, it also affects the outcome variability resulting in a less pronounced interaction effect in terms of the full distribution. The interaction average,

however, does not take this change in variability into account.

To take this change in variability into account, we propose to quantify the interaction by

$$\beta := P(Z \leq Z^*) \equiv P(Z < Z^*) + 0.5P(Z = Z^*). \quad (2)$$

For a continuous outcome, β simplifies to $P(Z < Z^*)$, i.e. the probability that the effect outcome of A at $B = 2$ exceeds the effect outcome at $B = 1$. The general definition (2) allows for discrete outcomes as well. Probabilities of the form (2) have a long history. In a two-sample design, it corresponds to the summary measure associated with the Wilcoxon–Mann–Whitney (WMW) test (Wilcoxon, 1945; Mann and Whitney, 1947; Kruskal, 1952). Several authors have argued that this probability is well suited as a summary measure, mainly because 1) it often has an informative and intuitive interpretation, 2) it provides a general measure for the difference between two groups, and 3) it is robust. Bamber (1975) considered this quantity as a measure of the size of the difference between two populations, while Brumback et al. (2006) discussed its meaning as a treatment effect. For a more detailed discussion on this probability as an effect size see, for example, Laine and Davidoff (1996); Newcombe (2006); D’Agostino et al. (2006); Senn (2006); Zhou (2008); Tian (2008); Senn (2011); Thas et al. (2012); Kieser et al. (2013). Several names have been proposed for the probability that one outcome exceeds another: the non-parametric treatment effect, the Mann–Whitney functional, the individual exceedance probability, the stress-strength measure, measure of a generalized treatment effect, the relative effect or the probabilistic index (Wilcox, 2003; Acion et al., 2006; Senn, 2006; Kieser et al., 2013; Thas et al., 2012; Nussbaum, 2014). Note that the term ‘probabilistic index’ is not unambiguous since it may have a different meaning in other research disciplines; in ecology, for example, it is used to denote the water quality (Cordoba et al., 2010), while Billinton and Kuruganty (1980) use it in the context of transient stability. Similar as in Grissom and Kim (2005) we will use the term *Probability of Superiority* (PS) to denote the probability that one outcome exceeds another.

We refer to β as defined by (2) as the *interaction probability of superiority* (IPS). The major difference with the summary measure of the WMW test is that in the current approach transformations of the outcomes (i.e. $Y_{11} - Y_{21}$ and $Y_{12} - Y_{22}$) are modeled instead of the original outcomes Y_{ab} .

Under location-shift and when $\alpha = 0$, Z and Z^* are identically distributed so that $\beta = 0.5$. For the bottom left panel of Figure 1 we have $\beta = P(Z \leq Z^*) = 98\%$, i.e. with a 98% probability, the effect of

A is larger when $B = 2$ as compared to when $B = 1$. For the bottom right panel, when location-shift does not hold, this probability decreases to $\beta = P(Z \leq Z^*) = 70\%$. There is still an interaction effect, but it is less pronounced as compared to the left panel because of the increase in variability. Note that since $\beta = P(Y_{11} - Y_{21} \leq Y_{12} - Y_{22}) = P(Y_{11} - Y_{12} \leq Y_{21} - Y_{22})$, β also represents the probability that the effect outcome of B at $A = 2$ is greater than at $A = 1$. So the interaction effect can be interpreted in both directions.

The remainder of the paper is organized as follows. In Section 2 we provide an estimator for β and derive its asymptotic distribution; we propose a hypothesis test and its Pitman asymptotic relative efficiency is calculated relative to the ANOVA F -test. We further show how the test is related to the test of Bhapkar and Gore (1974). In Section 3 several interactions tests are discussed which are used as competitors in Section 4 to compare with the new test in a simulation study. Section 5 illustrates how the summary measure and test can be used in practice and Section 6 presents the conclusions and discussion.

2 Estimation and asymptotics

2.1 Two-by-two design

Let Y_{abi} , $a, b = 1, 2$, $i = 1, \dots, n_{ab}$ denote an i.i.d. sample. An unbiased estimator of β is given by

$$\hat{\beta} = \frac{1}{n_{11}n_{21}n_{12}n_{22}} \sum_{i=1}^{n_{11}} \sum_{j=1}^{n_{21}} \sum_{k=1}^{n_{12}} \sum_{l=1}^{n_{22}} I(Y_{11i} - Y_{21j} \leq Y_{12k} - Y_{22l}), \quad (3)$$

where $I(y_1 \leq y_2) := I(y_1 < y_2) + 0.5I(y_1 = y_2)$, with $I(\cdot)$ the indicator function. Instead of deriving the asymptotics for $\hat{\beta}$, we consider the asymptotics for $g(\hat{\beta})$ where $g(\cdot)$ is a smooth link function mapping the unit interval onto the real line, for example the logit link $g(x) = \log[x/(1 - x)]$ or the probit link $\Phi^{-1}(x)$ with $\Phi(\cdot)$ the standard normal distribution function. This will allow us to construct confidence intervals for β which are guaranteed to be within the unit interval. A sketch of the proof can be found in the Appendix.

Theorem 1. Let $N = \sum_{a=1}^2 \sum_{b=1}^2 n_{ab}$. As $N \rightarrow \infty$, assume $n_{ab}/N \rightarrow \lambda_{ab} \in (0, 1)$, $a, b = 1, 2$. Then, for a smooth function $g : [0, 1] \rightarrow \mathbb{R}$,

$$\sqrt{N}[g(\hat{\beta}) - g(\beta)] \xrightarrow{d} N(0, \sigma^2).$$

Furthermore, σ^2 can be consistently estimated by $N\hat{\sigma}_{\hat{\beta}}^2$, where

$$\begin{aligned}\hat{\sigma}_{\hat{\beta}}^2 &= \left(\frac{\dot{g}(\hat{\beta})}{n_{11}n_{21}n_{12}n_{22}} \right)^2 \\ &\times \left[\sum_i \left\{ \sum_{j,k,l} [\mathbb{I}(Y_{11i} - Y_{21j} \leq Y_{12k} - Y_{22l}) - \hat{\beta}] \right\}^2 + \sum_j \left\{ \sum_{i,k,l} [\mathbb{I}(Y_{11i} - Y_{21j} \leq Y_{12k} - Y_{22l}) - \hat{\beta}] \right\}^2 \right. \\ &\left. + \sum_k \left\{ \sum_{i,j,l} [\mathbb{I}(Y_{11i} - Y_{21j} \leq Y_{12k} - Y_{22l}) - \hat{\beta}] \right\}^2 + \sum_l \left\{ \sum_{i,j,k} [\mathbb{I}(Y_{11i} - Y_{21j} \leq Y_{12k} - Y_{22l}) - \hat{\beta}] \right\}^2 \right],\end{aligned}$$

where $\dot{g}(x) = dg(x)/dx$.

It is now straightforward to propose a Wald-type test for testing $H_0 : \beta = 0.5$ as well as to construct confidence intervals.

Corollary 1. *Under the conditions of Theorem 1 and under $H_0 : P(Z \leq Z^*) = 0.5$, as $N \rightarrow \infty$,*

$$\mathcal{IPS} = \frac{g(\hat{\beta}) - g(0.5)}{\hat{\sigma}_{\hat{\beta}}} \xrightarrow{d} N(0, 1). \quad (4)$$

Corollary 2. *Under the conditions of Theorem 1, an approximate $(1 - \alpha)$ confidence interval for β is given by*

$$\left[g^{-1} \left\{ g(\hat{\beta}) - \Phi^{-1}(1 - \alpha/2) \hat{\sigma}_{\hat{\beta}} \right\}, g^{-1} \left\{ g(\hat{\beta}) + \Phi^{-1}(1 - \alpha/2) \hat{\sigma}_{\hat{\beta}} \right\} \right].$$

2.2 General two-way layout

Consider the general two-way layout where factor A has $K \geq 2$ levels and factor B has $L \geq 2$ levels. Bhapkar and Gore (1974) proposed a score-type test statistic under the location-shift model

$$Y_{abi} = \mu + \delta_a + \zeta_b + \eta_{ab} + \varepsilon_{abi}, \quad (5)$$

where ε_{abi} are i.i.d. with median zero and common distribution F_ε and $\sum_a \delta_a = \sum_b \zeta_b = \sum_a \eta_{ab} = \sum_b \eta_{ab} = 0$. For testing $H_0 : \eta_{ab} = 0, a = 1, \dots, K, b = 1, \dots, L$, they proposed a test statistic based on Hoeffding's generalized U -statistics. In this section we show how their test can be expressed in terms of estimators similar to $\hat{\beta}$. For notational convenience we consider a balanced design $n = n_{ab} \forall a, b$; see Bhapkar and

Gore (1974) for the more general formulation when the cell frequencies are proportional to row and column marginal totals. Let

$$\hat{\beta}_{aa'bb'} = \frac{1}{n^4} \sum_{i,j,k,l} \mathbf{I}(Y_{abi} - Y_{a'b'j} \leq Y_{ab'k} - Y_{a'b'l}),$$

denote the estimator of the IPS when considering levels a and a' of factor A and levels b and b' of factor B .

The test statistic of Bhapkar and Gore can then be expressed as

$$IPS_{BG} = \frac{n}{K^2 L^2 (\hat{\nu} - 0.25)} \sum_{a=1}^K \sum_{b=1}^L \left[\sum_{a' \neq a}^K \sum_{b' \neq b}^L \hat{\beta}_{aa'bb'} - \frac{(K-1)(L-1)}{2} \right]^2, \quad (6)$$

with $\hat{\nu}$ an estimate of the nuisance parameter $\nu = \int F_{\varepsilon+\varepsilon'-\varepsilon''}^2(x) dF_{\varepsilon}(x)$ where $F_{\varepsilon+\varepsilon'-\varepsilon''}$ is the distribution of $\varepsilon + \varepsilon' - \varepsilon''$ for $\varepsilon, \varepsilon', \varepsilon''$ i.i.d. F_{ε} . They showed that IPS_{BG} has an asymptotic chi-squared null distribution with $(K-1)(L-1)$ degrees of freedom. Bhapkar and Gore (1974) provide a computationally intensive estimator for ν under model (5). However, Spurrier (2005) has shown that ν is bounded below by $239/840 \approx 0.28452$ and above by $(7 - o^2 + o/5)/24 \approx 0.29125$ where $o = (1 - \sqrt{2/3})/2$. Hence, instead of estimating ν , a conservative test can be obtained by replacing $\hat{\nu}$ in (6) by its upper bound. This test will only be slightly conservative because for most distributions ν is close to its upper bound (Hollander and Wolfe, 1999, p. 347).

For the special case where $K = L = 2$, it follows that $IPS_{BG} = (\hat{\beta} - 0.5)^2 / \hat{\sigma}_0^2$, where $\hat{\sigma}_0^2$ is an estimator for the variance of $\hat{\beta}$ under location-shift model (5). The test based on IPS (4), however, does not assume the location-shift model and its consistent variance estimator allows the construction of a confidence interval for β . Furthermore, the test based on IPS does not assume that the cell frequencies are proportional to row and column marginal totals.

2.3 Asymptotic relative efficiency

The Pitman ARE of the IPS_{BG} test versus the ANOVA F -test under location-shift and under a sequence of local alternatives $\eta_{ab} = \kappa_{ab} / \sqrt{N}$ with κ_{ab} constants, is given by

$$ARE = \frac{\tau^2 \sigma_{\varepsilon}^2}{\nu - 0.25}, \quad (7)$$

where $\sigma_{\varepsilon}^2 = \text{Var}(\varepsilon)$, $\tau = \int f_{\varepsilon-\varepsilon'}(x)^2 dx$ with $f_{\varepsilon-\varepsilon'}$ the density of $\varepsilon - \varepsilon'$; see Bhapkar and Gore (1974).

Since for $K = L = 2$ both IPS and IPS_{BG} are based on $\hat{\beta}$, expression (7) also gives the ARE's of the

IPS-test (4) versus the ANOVA *F*-test. Table 1 gives these ARE's for a variety of distributions F_ε , where τ is obtained with numerical integration and ν is approximated based on 10^8 Monte-Carlo simulations in R (R Core Team, 2016). For the uniform distribution, the *F*-test is more efficient, while for the normal distribution the efficiency of both tests is almost equal. For all other distribution, the *IPS* and *IPSBG*-tests are asymptotically more efficient than the *F*-test.

3 Other interaction tests

In this section we describe several interaction tests and briefly discuss their properties. These tests will be used in a simulation study in Section 4. For simplicity, we restrict the discussion to a balanced two-by-two design with $n = n_{11} = n_{12} = n_{21} = n_{22}$ replicates.

3.1 The ANOVA *F*-test

The ANOVA *F*-test statistic is equivalent to

$$\mathcal{F} = \frac{\hat{\alpha}^2}{\hat{\sigma}_{\hat{\alpha}}^2}, \quad (8)$$

where $\hat{\alpha}$ is an estimator of (1) obtained by replacing the population means μ_{ab} by their sample counterparts, say $\bar{Y}_{ab}$, and $\hat{\sigma}_{\hat{\alpha}}^2 = \sum_{a=1}^2 \sum_{b=1}^2 \sum_{i=1}^n (Y_{abi} - \bar{Y}_{ab})^2 / [n(n-1)]$. Under the null hypothesis $H_0 : \alpha = 0$, $\mathcal{F}$ follows an *F*-distribution with 1 and $4(n-1)$ degrees of freedom when $\varepsilon \stackrel{d}{=} N(0, \sigma_\varepsilon^2)$. When normality is not fulfilled and if n is large enough, the null distribution of $\mathcal{F}$ can be approximated by a chi-squared distribution with 1 degree of freedom. Similar as for *IPS*, the interpretation of $\mathcal{F}$ is clear since the statistic is constructed based on an estimator of a population parameter (here α). However, unlike *IPS*, $\mathcal{F}$ is sensitive to outliers. A robust version of the *F*-test can be constructed by estimating α based on rank regression; see e.g. McKean and Hettmansperger (1976).

3.2 The rank test of Patel and Hoel

Patel and Hoel (1973) proposed a difference between two PS's as a measure of interaction. More specifically, they defined

$$\gamma := P(Y_{11} \leq Y_{21}) - P(Y_{12} \leq Y_{22}) \quad \text{and} \quad \gamma' := P(Y_{11} \leq Y_{12}) - P(Y_{21} \leq Y_{22}). \quad (9)$$

Hence, γ gives the difference in effect of A (in terms of the PS) for the two levels of B , and γ' the difference in effect of B for the levels of A . Note that, in general, $\gamma \neq \gamma'$. To test the null hypothesis of no interaction $H_0 : \gamma = 0$, their test statistic is given by

$$\mathcal{PH} = \frac{\hat{\gamma}_{1121} - \hat{\gamma}_{1222}}{\sqrt{\hat{\sigma}_{1121}^2 + \hat{\sigma}_{1222}^2}}, \quad (10)$$

where $\hat{\gamma}_{aba'b'} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n I(Y_{abi} \leq Y_{a'b'j})$, and

$$\hat{\sigma}_{aba'b'}^2 = n^{-1} \hat{\gamma}_{aba'b'} (1 - \hat{\gamma}_{aba'b'}) [\phi_{aba'b'} + \phi_{a'b'ab} + 2(n-1)^{-1}],$$

with

$$\phi_{aba'b'} = \frac{1}{n^2(n-1)} \sum_{i=1}^n \sum_{j=1}^n \sum_{\substack{k=1 \\ k \neq j}}^n I(Y_{abi} \leq Y_{a'b'j}) I(Y_{abi} \leq Y_{a'b'k}),$$

equals a variance estimator due to Sen (1967).

Asymptotically, $\mathcal{PH}$ has a standard normal null distribution. For more information on this test statistic, we refer to Patel and Hoel (1973); Marden and Muyot (1995); Wilcox (1999).

3.3 The rank transform test

Conover and Iman (1981) proposed the rank transform method to construct rank tests for a variety of designs. In its simplest form, a test for testing interaction can be obtained by applying the ANOVA interaction F -test on the ranks of the outcomes. Let R_{abi} denote the rank associated with Y_{abi} where the ranking is performed within the pooled sample. Let $\bar{R}_{ab}$ denote the sample average of the ranks of group $A = a$ and $B = b$, and $\hat{\alpha}_R = (\bar{R}_{11} - \bar{R}_{21}) - (\bar{R}_{12} - \bar{R}_{22})$. The rank transform test statistic is given by

$$\mathcal{RT} = \frac{\hat{\alpha}_R^2}{\hat{\sigma}_{RT}^2}, \quad (11)$$

with $\hat{\sigma}_{RT}^2 = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^n (R_{ijk} - \bar{R}_{ij})^2 / [n(n-1)]$. The interpretation of the test statistic, however, is not always clear. Furthermore, the rank transform approach is not always suited for testing interaction. For example, Brunner and Neumann (1986) and later Thompson (1991) have shown that the expected value of the rank transform test statistic can tend to infinity with increasing sample size in the absence of interaction. Since the introduction of the rank transform approach, the properties of the related statistics have been studied in more detail; see, for example, Akritas (1990). Thompson (1991) has shown that for two-by-two designs, $\mathcal{RT}$ asymptotically follows a χ_1^2 -distribution. However, for other two-way layouts with main effects of both factors, this does not longer hold. Several authors have worked out the correct hypotheses and asymptotics for the rank transform method, see for example Akritas (1990); Akritas and Arnold (1994); Akritas et al. (1997); Brunner and Puri (2001); Fan and Zhang (2014).

3.4 The aligned rank test

Since the rank transform method may not always be suitable for testing interaction, Mansouri and Chang (1995), among others, proposed an aligned rank test. They assume an ANOVA decomposition of the population means

$$\mu_{ab} = \mu + \delta_a + \zeta_b + \eta_{ab},$$

and estimate the parameters by means of least-squares. We denote these estimators as $\hat{\mu}$, $\hat{\delta}_a$, $\hat{\zeta}_b$, and $\hat{\eta}_{ab}$. Instead of ranking the outcomes, as in the rank transform approach, they rank the aligned outcomes: $Y_{abi} - \hat{\delta}_a - \hat{\zeta}_b$. Let AR_{abi} denote the corresponding rank within the pooled sample of the aligned outcomes and $\bar{AR}_{ab}$ the sample average of these ranks for $A = a$ and $B = b$. Let $\hat{\alpha}_{AR} = (\bar{AR}_{11} - \bar{AR}_{21}) - (\bar{AR}_{12} - \bar{AR}_{22})$. The aligned rank transform test statistics is given by

$$\mathcal{ART} = \frac{\hat{\alpha}_{AR}^2}{\hat{\sigma}_{ART}^2}, \quad (12)$$

with $\hat{\sigma}_{ART}^2 = \sum_{a=1}^2 \sum_{b=1}^2 \sum_{i=1}^n (AR_{abi} - \bar{AR}_{ab})^2 / [n(n-1)]$. Note that Mansouri and Chang (1995) propose more sophisticated aligned rank tests as compared to (12). Instead of least squares, robust rank regression estimators for δ and ζ can be used; see, for example, McKean and Hettmansperger (1976). Note that, unlike the *IPS*-test, the aligned rank test is restricted to linear models.

3.5 Row and column rank test

Gao and Alvo (2005) proposed a rank test which is valid under the location-shift model (5). If R_{abi}^A denotes the ranking of Y_{abi} among outcomes for which $A = a$ and R_{abi}^B denotes the ranking of Y_{abi} among outcomes for which $B = b$, then their test statistic consists of a generalized quadratic form of a linear combination of the rankings R_{abi}^A and R_{abi}^B . We refer to Gao and Alvo (2005) for details on the construction of the test.

4 Simulation study

4.1 Estimation

To empirically evaluate the asymptotic approximations of Theorem 1, we set up a simulation study where we simulate data according to a 2^2 full factorial design with n replicates, where

$$Y = \theta_1 + \theta_2 X_A + \theta_3 X_B + \theta_4 X_A X_B + \varepsilon, \quad (13)$$

with $X_A \in \{-1, 1\}$ and $X_B \in \{-1, 1\}$ denoting the groups, and for several distribution functions F_ε : the standard normal distribution $N(0, 1)$, the t -distribution with 3 degrees of freedom t_3 , the exponential distribution with rate 1 $\text{Exp}(1)$, and the logistic distribution $\text{Logistic}(0,1)$. All distributions are centred to have a mean of zero and scaled to have a variance of one. Furthermore, $\theta_1 = \theta_2 = 1$, $\theta_3 = -2$, and several choices of θ_4 are considered, resulting in different values of β in equation (2).

Table 2 gives the results based on 10000 Monte-Carlo simulations. All simulations were performed in R (R Core Team, 2016).

The results confirm that for all choices of n and F_ε , β is unbiasedly estimated. For $n = 5$, $\hat{\sigma}_\beta^2$ underestimates the true variance, but this underestimation is less pronounced when β tends to 0.5. The empirical coverage is close to 95% for $\beta = 0.5$, but anti-conservative for the other choices of β . As n increases the bias of $\hat{\sigma}_\beta^2$ decreases, particularly for $\beta = 0.5$ for which the true coverage is close to 95% for $n = 10$. For $n = 20$ the coverage of the 95% approaches the nominal level except for $\beta = 0.903$ and $\beta = 0.884$ for which the coverage is less than 95%. Note that these are the β -values closest to the boundaries.

4.2 Hypothesis testing

4.2.1 Two-by-two design

To study the empirical properties of the *IPS*-test (4), data are simulated for a two-by-two design according to model (13). The following choices of F_ε are considered: $N(0, 4)$, t_3 , $LN(0, 4.67)$ the centred log-normal distribution with mean zero and variance 4.67, the normal distribution with mean 0 and variance 4 but with the first observation in $(A, B) = (1, 1)$ replaced by an outlier (the value 1000) which is denoted by $N(0, 4) + \text{outlier}$, and a heteroscedastic mean-zero normal distribution with variance $\sigma^2(X_A, X_B) = (1.5 + 1X_A - 1.3X_B + 0.4X_AX_B)^2$ and denoted by $N[0, \sigma^2(X_A, X_B)]$. Table 3 gives an overview of the values of $\theta^T = (\theta_1, \theta_2, \theta_3, \theta_4)$ and the different error distributions used in the simulation set-up.

Balanced design

Table 4 displays the empirical rejection rates at the 5% level of significance based on 1000 Monte-Carlo simulations. We consider 7 tests: the *IPS*-test (4) with logit link, the ANOVA F -tests ($\mathcal{F}$); the rank transform test ($\mathcal{RT}$); the aligned rank test ($\mathcal{ART}$); a robust version of the aligned rank test based on rank regression ($\mathcal{RART}$) using the `Rfit` package (Kloke and McKean, 2013); the test of Patel and Hoel ($\mathcal{PH}$); and the test of Gao and Alvo ($\mathcal{GA}$) as implemented in the `StatMethRank` package (Li, 2015). The results of the *IPSBG*-test were similar to the results of the *IPS*-test and are therefore not included.

Overall, the empirical type I error rate of the *IPS*-test is close to its nominal level for all choices of n , error distributions and independent of the main effects. The $\mathcal{RT}$ -test does not correctly control the type I error when there are moderate to large main effects, while both the $\mathcal{ART}$ and F -test are sensitive to the outlier. Both $\mathcal{ART}$ and $\mathcal{RART}$ -tests are anti-conservative for heteroscedastic data, but the latter is robust against the outlier. The $\mathcal{PH}$ -test is anti-conservative for $n = 10$ for heteroscedastic data and moderate main effects, while it is biased for all F_ε when there are large main effects. The $\mathcal{GA}$ -test does not correctly control the type I error for $n = 5$. For $n = 10$ and in the absence of main effects, the type I error of the $\mathcal{GA}$ -test is close to its nominal level, except for heteroscedastic data. For moderate main effects the test is anti-conservative except for the t -distributed error for which it is slightly conservative and for the heteroscedastic data for which there is a substantial inflation of the type I error. For large main effects, the test is biased for

all error distributions.

Overall the IPS -test has a stable power. The $\mathcal{RT}$ -test has no power in the presence of large main effects. This is not surprising since the $\mathcal{RT}$ -test does not test the hypothesis $H_0 : \theta_4 = 0$ (Akritas, 1990). One can show that the $\mathcal{RT}$ -test is equivalent to testing $H_0 : \vartheta_4 = 0$ in the model

$$E(F(Y) | X_A, X_B) = \vartheta_1 + \vartheta_2 X_A + \vartheta_3 X_B + \vartheta_4 X_A X_B,$$

where $F(y)$ denotes the weighted average of all conditional distributions $P(Y \leq y | X_A, X_B)$, see e.g. Akritas (1990); Akritas and Arnold (1994); Brunner and Puri (2001); Shah and Madden (2004); Fan and Zhang (2014); De Neve and Thas (2015) for more details. $E(F(Y) | X_A, X_B)$ is also referred to as the relative treatment effect (Brunner and Puri, 2001; Shah and Madden, 2004).

For F_ε equal to the normal distribution with mean zero and variance 4 and $\theta^T = (1, 1, -2, 0.7)$, it follows that approximately $\vartheta^T = (0.5, 0.09, -0.19, 0.06)$ while for $\theta^T = (1, 10, -20, 0.7)$ this becomes $\vartheta^T = (0.5, 0.125, -0.25, 0)$, explaining the (drastic) decrease in power as compared to some of the other tests. This is in line with the findings of Sawilowsky (1990). Hence, interaction defined in terms of the expected outcome is not equivalent to interaction defined in terms of relative treatment effects.

The $\mathcal{RART}$ -test has good power properties and it outperforms the IPS -test when $F_\varepsilon = t_3$ and $F_\varepsilon = \text{LN}(0, 4.67)$. The IPS -test has a higher power than the $\mathcal{RART}$ -test when there is an outlier. For a normally distributed error with $n = 5$ the IPS -test is slightly more powerful as compared to the $\mathcal{RART}$ -test, while for $n = 10$ the performances are similar. As mentioned earlier, the $\mathcal{RART}$ -test does not correctly control the type I error for heteroscedastic data, making comparisons of powers impossible. For $n = 10$ and in the absence of main effects, the $\mathcal{GA}$ -test has a similar to superior performance over the IPS -test. For $n = 10$, moderate main effects and a t -distributed error, the IPS -test is superior over the $\mathcal{GA}$ -test, while for large main effects and all error distributions, the $\mathcal{GA}$ -test has no power.

Similar conclusions hold for smaller ($\theta_4 = 0.3$) and larger interaction effects ($\theta_4 = 1.1$); see the Appendix for more simulation results.

Unbalanced design

Table 5 gives the simulation results for an unbalanced design where $n_{[i,j]}$ denotes the sample size for cell

$X_A = i$ and $X_B = j$, $i, j \in \{-1, 1\}$. For the heteroscedastic error $E = 5$ in Table 3, we consider two settings: for $E = 5a$ the sample sizes are inversely proportional to the variances, while for $E = 5b$ the sample sizes are proportional to the variances. For setting $E = 5a$, all tests except the $\mathcal{PH}$ -test have an inflated type I error, while for the setting $E = 5b$ the type I error of the IPS -test is close to its nominal level. For a total sample size of 20, the IPS -test has an inflated type I error in the presence of an outlier, while for a total sample size of 40, the type I error is closer to its nominal level. Note that the $\mathcal{RT}$ -test has different properties as compared to the balanced setting. This is a consequence of the definition of the relative treatment effect size $E(F(Y) | X_A, X_B)$ which depends on the sample sizes (Brunner and Puri, 2001).

4.2.2 Two-by-three design

To study the empirical properties of the IPS_{BG} -test (6) for the general two-way layout, we simulate data for a two-by-three design according to

$$Y = \theta_1 + \theta_2 X_A + \theta_3 X_{B1} + \theta_4 X_{B2} + \theta_5 X_A X_{B1} + \theta_6 X_A X_{B2} + \varepsilon, \quad (14)$$

where $X_A = 1$ if $A = 1$ and $X_A = -1$ if $A = 2$, $X_{B1} = 1$ if $B = 1$, $X_{B1} = 0$ if $B = 2$ and $X_{B1} = -1$ if $B = 3$, and $X_{B2} = 0$ if $B = 1$, $X_{B2} = 1$ if $B = 2$ and $X_{B2} = -1$ if $B = 3$. Table 3 gives an overview of the values of θ and the error distributions where for $E = 5$ the variance is given by $\sigma^2(X_A, X_B) = (1.5 + X_A - 1.3X_{B1} - 1.3X_{B2} + 0.4X_A X_{B1} + 0.4X_A X_{B2})^2$. For the IPS_{BG} -test, $\hat{v}$ in (6) is replaced by its upper bound.

Table 6 displays the empirical rejection rates at the 5% level of significance based on 1000 Monte-Carlo simulations. The $\mathcal{PH}$ -test is not included since it is restricted to the two-by-two design. The IPS_{BG} -test is slightly conservative for all settings, which is expected due to the conservative choice of $\hat{v}$ in (6). The test becomes less conservative with increasing sample size. The $\mathcal{RT}$ -test has an inflated type I error in the presence of main effects, which is in agreement with the conclusions of Thompson (1991). The $\mathcal{ART}$ has an inflated type I error when an outlier is present and for heteroscedastic data. The $\mathcal{RART}$ -test correctly controls the type I error rate, except for heteroscedastic data for which the test is anti-conservative. The $\mathcal{GA}$ -test has an inflated type I error for all settings, except when there are large main effects for which the test is conservative.

Overall the IPS_{BG} -test has a stable power. The $RART$ -test outperforms the IPS_{BG} -test for all settings, while the performance of both tests becomes more similar with increasing sample sizes, since the IPS_{BG} -test then becomes less conservative. The GA -test has no power in the presence of large main effects.

5 Example

To illustrate the interaction probability of superiority as an effect size measure, we consider the cross sectional part of the childhood respiratory disease study as provided by Rosner (1999). It is of interest to study the association between smoking and the lung capacity of children. The outcome is the Forced Expiratory Volume (FEV in litres), which is an index for the pulmonary function and we consider the Age (in years) and the Smoking status of the child as predictors. We consider children that are 11 or 15 years old. There are 81 children of 11 years that did not smoke and 9 that smoked. At the age of 15, there are 9 non-smokers and 10 smokers. The left panel in Figure 2 gives the boxplots and stripcharts of the FEV according to age and smoking status. The right panel of Figure 2 gives the effect outcome of smoking (i.e. $Z = FEV_{NS,11} - FEV_{S,11}$ and $Z^* = FEV_{NS,15} - FEV_{S,15}$, where S stands for smoker and NS for non-smoker) obtained by constructing all pairwise differences in FEV of non-smoker compared to smokers and according to age. Larger effect outcomes correspond to smaller FEV's for the smokers. This plot suggests that at the age of 11, there are no systematic differences between the smokers and the non-smokers in terms of the FEV. At the age of 15, however, the FEV of the non-smokers tends to exceed that of the smokers. The estimated interaction probability of superiority $P(FEV_{NS,11} - FEV_{S,11} \leq FEV_{NS,15} - FEV_{S,15})$ is 74% (95% confidence interval ranging from 50% to 89% upon using the logit link), i.e. it is more likely that the effect outcome of smoking at the age of 15 is larger than at the age of 11. Since the IPS can be interpreted in both directions, it also suggests that the age effect of smokers is less pronounced than the age effect of non-smokers.

The interaction average $(\mu_{NS,11} - \mu_{S,11}) - (\mu_{NS,15} - \mu_{S,15})$ is estimated by -0.93 (95% confidence interval ranging from -1.65 to -0.22). From a WMW-test it further follows that $\hat{P}(FEV_{NS,11} \leq FEV_{S,11}) = 58\%$ (95% confidence interval ranging from 41% to 74%) and $\hat{P}(FEV_{NS,15} \leq FEV_{S,15}) = 26\%$ (95% confidence interval ranging from 12% to 47%) confirming the larger effect outcome at the age of 15.

6 Discussion

We have proposed a measure for interaction based on the probability of superiority. We provide an unbiased and consistent estimator and work out the asymptotic distribution which can be used to construct confidence intervals and a hypothesis test. The test has a superior Pitman efficiency over the ANOVA F -test for a variety of distributions under location-shift. We further show how the test is related to the test of Bhapkar and Gore (1974) which is valid under the location-shift model. The current paper extends their work by providing the probability of superiority interpretation without assuming location-shift, by constructing confidence intervals for this population parameter, and by providing a Wald-type test. Upon using simulations we studied the small sample behavior of the estimators and tests. For a two-by-two design, the test correctly controls the type I error, even for small samples and the test has a good and stable performance in terms of power. For unbalanced designs, the type I error is inflated for small samples when the outcome variability is inversely proportional to the sample size. The empirical coverage of the confidence interval is close to its nominal level except when the probability of superiority is close the boundaries of the unit interval. However, increasing the sample size improves the empirical coverage. For a two-by-three design the test is slightly conservative which results in some loss of power. However, the test becomes less conservative with increasing sample size.

In addition to the interaction probability of superiority, the *magnitude* of the interaction effect, say τ , can be defined as the value such that $P(Z \leq Z^* + \tau) = 0.5$, where τ can be estimated by a Hodges–Lehmann type estimator $\hat{\tau} = \text{Median}_{i,j,k,l}\{(Y_{11i} - Y_{21j}) - (Y_{12k} - Y_{22l})\}$. For the childhood respiratory disease study in Section 5, for example, $\hat{\tau} = \text{Median}\{(\text{FEV}_{\text{NS},11} - \text{FEV}_{\text{S},11}) - (\text{FEV}_{\text{NS},15} - \text{FEV}_{\text{S},15})\} = -0.93$. The construction of confidence intervals for τ and the ARE comparison over $\hat{\alpha}$ is considered as future research.

The method proposed in this article makes use of ranks computed on the pairwise differences of the outcomes. This implies that the method is not invariant under monotone transformations and only applies to metric data (since we consider differences). In this respect, it is not a genuine rank-test which is typically invariant under monotone transformations and applies to ordinal outcomes as well.

Acknowledgments

The authors would like to thank the referees and Associate Editor for their constructive comments. The authors acknowledge the IAP research network P7/06 of the Belgian Government (Belgian Science Policy) and the Research Fund Flanders (FWO) research grants G020214N, V403114N and K211116N.

A Appendix

A.1 Sketch of the proof of Theorem 1

Let $\gamma = g(\beta)$ and $\hat{\gamma} = g(\hat{\beta})$. Since $\hat{\beta} \xrightarrow{p} \beta$, it holds that

$$\sqrt{N}(\hat{\gamma} - \gamma) = \dot{g}(\beta) \sqrt{N}(\hat{\beta} - \beta) + o_p(1).$$

Upon using Hájek projections, one can show that

$$\begin{aligned} \sqrt{N}(\hat{\beta} - \beta) &= \frac{\sqrt{N}}{n_{11}} \sum_{i=1}^{n_{11}} [1 - F_1(Y_{11i}) - \beta] + \frac{\sqrt{N}}{n_{21}} \sum_{j=1}^{n_{21}} [F_2(Y_{21j}) - \beta] \\ &+ \frac{\sqrt{N}}{n_{12}} \sum_{k=1}^{n_{12}} [F_3(Y_{12k}) - \beta] + \frac{\sqrt{N}}{n_{22}} \sum_{l=1}^{n_{22}} [1 - F_4(Y_{22l}) - \beta] + o_p(1), \end{aligned}$$

where F_1 is the distribution of $Y_{21} + Y_{12} - Y_{22}$, F_2 of $Y_{11} + Y_{22} - Y_{12}$, F_3 of $Y_{11} + Y_{22} - Y_{21}$ and F_4 of $Y_{21} + Y_{12} - Y_{11}$. Consequently, the asymptotic variance of $\sqrt{N}(\hat{\beta} - \beta)$ is given by

$$\tilde{\sigma}^2 = \frac{1}{\lambda_{11}} \text{Var}[F_1(Y_{11})] + \frac{1}{\lambda_{21}} \text{Var}[F_2(Y_{21})] + \frac{1}{\lambda_{12}} \text{Var}[F_3(Y_{12})] + \frac{1}{\lambda_{22}} \text{Var}[F_4(Y_{22})],$$

so that $\sigma^2 = \dot{g}(\beta)^2 \tilde{\sigma}^2$. The consistent estimator for $\tilde{\sigma}^2$ follows from noting that e.g. $\text{Var}[F_2(Y_{21})] = \text{Var}(F_2(Y_{21}) - \beta) = \text{E}[(F_2(Y_{21}) - \beta)^2]$ and where F_2 is replaced by the empirical distribution function and the expectation by the sample mean. Similar for the other terms.

A.2 Additional simulation results

Table 7 gives the empirical powers for a small interaction effect ($\theta_4 = 0.3$) and a large interaction effect ($\theta_4 = 1.1$).

References

- Acion, L., Peterson, J., Temple, S., and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25:591–602.
- Akritis, M. (1990). The rank transform method in some two-factor designs. *Journal of the American Statistical Association*, 85:73–78.
- Akritis, M. and Arnold, S. (1994). Fully nonparametric hypotheses for factorial designs I: multivariate repeated measures designs. *Journal of the American Statistical Association*, 89:336–343.
- Akritis, M. G., Arnold, S. F., and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92(437):258–265.
- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.
- Bhapkar, V. and Gore, A. (1974). A nonparametric test for interaction in two-way layouts. *Sankhya: The Indian Journal of Statistics: Series A*, 36:261–272.
- Billinton, R. and Kuruganty, P. (1980). A probabilistic index for transient stability. *IEEE Transactions on Power Apparatus and Systems*, (1):195–206.
- Brumback, L. C., Pepe, M. S., and Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25(4):575–590.
- Brunner, E. and Neumann, N. (1986). Rank tests in 2x2 designs. *Statistica Neerlandica*, 40(4):251–272.
- Brunner, E. and Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical papers*, 42(1):1–52.
- Conover, W. and Iman, R. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35:124–129.

- Cordoba, E. B., Martinez, A. C., and Ferrer, E. V. (2010). Water quality indicators: Comparison of a probabilistic index and a general quality index. The case of the Confederación Hidrográfica del Júcar (Spain). *Ecological Indicators*, 10(5):1049–1054.
- D’Agostino, R. B., Campbell, M., and Greenhouse, J. (2006). The Mann–Whitney statistic: continuous use and discovery. *Statistics in Medicine*, 25:541–542.
- De Neve, J. and Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511):1276–1283.
- Fan, C. and Zhang, D. (2014). Wald-type rank tests: A GEE approach. *Computational Statistics & Data Analysis*, 74:1–16.
- Gao, X. and Alvo, M. (2005). A nonparametric test for interaction in two-way layouts. *The Canadian Journal of Statistics*, 33:529–543.
- Grissom, R. J. and Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.
- Hollander, M. and Wolfe, D. (1999). *Nonparametric Statistical Methods*. Wiley New York.
- Kieser, M., Friede, T., and Gondan, M. (2013). Assessment of statistical significance and clinical relevance. *Statistics in Medicine*, 32:1707–1719.
- Kloke, J. and McKean, J. (2013). *Rfit: Rank Estimation for Linear Models*. R package version 0.17.
- Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, pages 525–540.
- Laine, C. and Davidoff, F. (1996). Patient-centered medicine: a professional evolution. *Journal of the American Medical Association*, 275:152–156.
- Li, Q. (2015). *StatMethRank: Statistical Methods for Ranking Data*. R package version 1.3.
- Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

- Mansouri, H. and Chang, C. (1995). A comparative study of some rank tests for interaction. *Computational Statistics and Data Analysis*, 19:85–96.
- Marden, J. and Muyot, E. (1995). Rank tests for main and interaction effects in analysis of variance. *Journal of the American Statistical Association*, 90:1388–1398.
- McKean, J. and Hettmansperger, T. (1976). Tests of hypotheses based on ranks in the general linear model. *Communications in Statistics - Theory and Methods*, 5:693–709.
- Newcombe, R. (2006). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 1: general issues and tail-area-based methods. *Statistics in Medicine*, 25:543–557.
- Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis: choosing the best statistical technique*. Routledge.
- Patel, K. and Hoel, D. (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association*, 68:615–620.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rosner, B. (1999). *Fundamentals of Biostatistics*. Pacific Grove: Duxbury.
- Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, 60(1):91–126.
- Sen, P. (1967). A note on the asymptotically distribution-free confidence bounds for $Pr\{X < Y\}$ based on two independent samples. *Sankhya: The Indian Journal of Statistics: Series A*, 29:95–102.
- Senn, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects by L. Acion, J. Peterson, S. Temple and S. Arndt. *Statistics in Medicine*, 25:3944–3948.
- Senn, S. (2011). U is for unease: reasons for mistrusting overlap measures for reporting clinical trials. *Statistics in Biopharmaceutical Research*, 3:302–309.

- Shah, D. and Madden, L. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, 94(1):33–43.
- Spurrer, J. D. (2005). Improved upper bounds for Hollander’s μ and Lehmann’s λ . *Communications in Statistics-Theory and Methods*, 34:1851–1855.
- Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012). Probabilistic index models (with discussion). *Journal of the Royal Statistical Society - Series B*, 74:623–671.
- Thompson, G. (1991). A note on the rank transform for interactions. *Biometrika*, 78:697–701.
- Tian, L. (2008). Confidence intervals for $P(Y_1 > Y_2)$ with normal outcomes in linear models. *Statistics in Medicine*, 27:4221–4237.
- Wilcox, R. (1999). Rank-based test for interactions in a two-way design. *Computational Statistics and Data Analysis*, 29:275–284.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. Elsevier.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.
- Zhou, W. (2008). Statistical inference for $P(X < Y)$. *Statistics in Medicine*, 27:257–279.

Table 1: Asymptotic relative efficiency of the IPS and $IPS_{\mathcal{BG}}$ -tests versus the ANOVA F -test.

F_ε	Uniform	Normal	Logistic	t_5	Laplace	Exponential	t_3	Lognormal
ARE	0.91	0.99	1.07	1.18	1.24	1.59	1.67	3.81

Table 2: Empirical evaluation of the asymptotic properties of Theorem 1 for a full factorial design with n replicates, and according to several standardized (i.e. mean-zero and variance one) distribution functions F_ε and several choices of θ_4 in (13). $\hat{E}(\hat{\beta})$ denotes the empirical mean of $\hat{\beta}$ based on 10000 Monte-Carlo simulations, $\hat{Var}(\hat{\beta})$ is the empirical variance, $\hat{E}(\hat{\sigma}_\beta^2)$ is the empirical mean of the estimated variances as given in Theorem 1, and 95% CI is the empirical coverage of a 95% confidence interval for β with $g(\cdot)$ the logit link.

F_ε	θ_4	β	$\hat{E}(\hat{\beta})$	$\hat{Var}(\hat{\beta})$	$\hat{E}(\hat{\sigma}_\beta^2)$	95% CI
$n = 5$						
Normal	-0.55	0.865	0.865	1.528	0.958	89.9
t_3	-0.55	0.903	0.902	2.065	1.094	81.3
Exponential	-0.55	0.883	0.883	2.066	1.035	81.1
Logistic	-0.55	0.869	0.868	1.663	0.981	87.8
Normal	0	0.500	0.501	0.779	0.593	94.5
t_3	0	0.500	0.501	0.760	0.591	94.3
Exponential	0	0.500	0.500	0.739	0.593	95.2
Logistic	0	0.500	0.500	0.744	0.591	94.7
Normal	0.3	0.274	0.274	1.009	0.702	93.5
t_3	0.3	0.226	0.228	1.332	0.800	91.1
Exponential	0.3	0.241	0.242	1.335	0.792	90.8
Logistic	0.3	0.268	0.267	1.058	0.721	93.4
$n = 10$						
Normal	-0.55	0.865	0.864	0.578	0.450	93.3

t_3	-0.55	0.903	0.903	1.181	0.650	88.6
Exponential	-0.55	0.884	0.883	1.075	0.598	89.1
Logistic	-0.55	0.869	0.869	0.672	0.482	92.7
Normal	0	0.500	0.499	0.297	0.273	95.4
t_3	0	0.500	0.499	0.293	0.270	95.2
Exponential	0	0.500	0.500	0.296	0.270	95.2
Logistic	0	0.500	0.498	0.291	0.271	95.6
Normal	0.3	0.274	0.275	0.354	0.318	94.8
t_3	0.3	0.227	0.226	0.479	0.373	93.6
Exponential	0.3	0.241	0.242	0.435	0.363	94.0
Logistic	0.3	0.268	0.267	0.382	0.327	94.5
$n = 20$						
Normal	-0.55	0.864	0.865	0.230	0.211	95.0
t_3	-0.55	0.903	0.903	0.429	0.318	91.8
Exponential	-0.55	0.884	0.884	0.366	0.288	92.6
Logistic	-0.55	0.870	0.869	0.255	0.226	94.4
Normal	0	0.500	0.502	0.142	0.132	94.7
t_3	0	0.500	0.500	0.135	0.130	95.0
Exponential	0	0.500	0.500	0.135	0.130	95.2
Logistic	0	0.500	0.501	0.137	0.131	95.0
Normal	0.3	0.274	0.275	0.164	0.153	95.0
t_3	0.3	0.226	0.227	0.188	0.176	94.5
Exponential	0.3	0.241	0.240	0.184	0.173	95.0
Logistic	0.3	0.268	0.267	0.169	0.156	94.5

Table 3: Several settings of the parameters associated with models (13) and (14). $T = 1$ corresponds to no main effects, $T = 2$ to moderate main effects, and $T = 3$ to large main effects. Under the E heading, the five error distributions are listed.

T				
1	2	3		
two-by-two design				
$\theta^T = (0, 0, 0, \theta_4)$	$\theta^T = (1, 1, -2, \theta_4)$	$\theta^T = (1, 10, -20, \theta_4)$		
two-by-three design				
$\theta^T = (0, 0, 0, 0, \theta_5, \theta_6)$	$\theta^T = (1, 1, -2, -2, \theta_5, \theta_6)$	$\theta^T = (1, 10, -20, -20, \theta_5, \theta_6)$		
E				
1	2	3	4	5
$\varepsilon \stackrel{d}{=} N(0, 4)$	$\varepsilon \stackrel{d}{=} t_3$	$\varepsilon \stackrel{d}{=} LN(0, 4.67)$	$\varepsilon \stackrel{d}{=} N(0, 4) + \text{outlier}$	$\varepsilon \stackrel{d}{=} N[0, \sigma^2(X_A, X_B)]$

Table 4: Empirical type I error rates and empirical powers (%) at the 5% level of significance and based on 1000 Monte-Carlo simulations for the two-by-two design. Table 3 gives the coding for T and E ; n denotes the number of observations in each cell.

T	E	Type I error (%) ($\theta_4 = 0$)							Power (%) ($\theta_4 = 0.7$)						
		IPS	F	RT	ART	$RAART$	PH	GA	IPS	F	RT	ART	$RAART$	PH	GA
$n = 5$															
1	1	5.4	5.6	6.6	6.7	5.7	5.1	9.8	32.1	31.3	31.3	31.4	28.1	25.6	37.4
1	2	4.8	3.7	5.3	5.3	4.3	4.1	8.6	51.6	52.1	59.9	56.7	55.5	51.4	64.9
1	3	6.1	4.5	6.1	6.0	6.3	4.2	9.3	47.4	47.9	73.5	63.3	67.0	66.9	75.8
1	4	6.1	0.0	6.1	0.0	5.3	4.6	8.7	44.6	0.0	40.3	0.0	37.8	33.3	48.7
1	5	6.2	6.6	6.4	8.1	6.9	4.2	9.3	24.2	26.6	29.9	32.4	28.4	27.9	41.0
2	1	5.7	5.5	2.4	5.9	5.1	2.1	5.6	33.2	33.1	17.2	32.7	28.4	15.4	37.9
2	2	5.4	4.9	1.6	6.0	5.6	0.9	3.8	51.8	53.1	16.6	57.9	56.0	15.1	31.3
2	3	5.2	3.7	1.9	6.0	5.4	1.5	2.5	47.6	47.7	11.6	62.2	66.8	9.8	24.1
2	4	5.9	0.0	4.1	0.0	4.7	3.9	11.1	43.0	0.0	31.8	0.0	36.9	31.0	50.2
2	5	6.0	5.9	5.7	9.3	7.5	5.0	16.2	24.4	26.9	32.9	32.3	28.1	30.8	58.8
3	1	4.1	4.4	0.0	5.1	4.3	0.0	0.0	30.6	31.5	0.0	33.0	29.4	0.0	0.0
3	2	4.6	4.1	0.0	4.7	4.1	0.0	0.0	51.3	51.7	0.0	57.0	55.3	0.0	0.0
3	3	5.6	3.8	0.0	5.2	6.3	0.0	0.0	45.0	48.2	0.0	64.2	66.0	0.0	0.0
3	4	4.8	0.0	0.0	0.0	4.2	0.0	0.0	41.7	0.0	0.0	0.0	35.7	0.0	0.0
3	5	5.5	5.5	0.0	6.8	6.0	0.0	0.0	21.7	24.5	0.0	28.2	25.6	0.0	0.0
$n = 10$															
1	1	5.0	4.9	5.0	4.7	4.9	4.7	6.5	56.4	57.0	55.5	55.7	56.5	51.0	58.3
1	2	4.4	3.7	4.4	4.4	4.9	3.7	5.5	82.9	77.7	92.0	90.4	91.4	88.3	91.3
1	3	3.8	3.1	4.2	3.6	4.7	2.9	5.0	76.3	64.5	95.9	93.0	95.1	94.3	96.1
1	4	5.0	0.0	5.8	0.0	5.7	4.4	6.7	68.7	0.0	64.3	0.0	65.4	58.0	67.4
1	5	5.4	6.2	5.7	9.1	9.2	4.6	8.3	45.8	48.7	55.7	55.6	53.3	55.3	61.0
2	1	5.0	5.3	2.5	5.3	5.4	2.2	5.4	55.7	56.7	42.5	55.9	56.2	40.1	58.1
2	2	5.8	5.3	1.3	5.8	6.0	1.4	6.0	81.6	77.1	45.7	88.2	89.7	50.0	65.7
2	3	5.8	4.6	2.8	6.3	7.4	2.8	5.3	77.7	66.2	34.9	92.0	94.8	35.3	54.5
2	4	4.5	0.0	3.2	0.0	5.1	3.7	8.0	64.8	0.0	51.6	0.0	61.6	50.7	66.7
2	5	4.9	5.2	13.0	8.1	8.3	12.2	26.0	46.4	49.6	71.6	55.0	52.2	72.5	86.8

3	1	5.9	5.7	0.0	5.4	5.4	0.0	0.0	57.1	57.5	0.0	57.0	57.5	0.0	0.0
3	2	4.8	4.2	0.0	4.7	5.1	0.0	0.0	81.7	77.7	0.0	88.9	89.8	0.0	0.0
3	3	4.6	4.3	0.0	4.9	5.5	0.0	0.0	76.7	65.8	0.0	92.1	95.2	0.0	0.0
3	4	4.6	0.0	0.0	0.0	4.8	0.0	0.0	67.4	0.0	0.0	0.0	64.8	0.0	0.0
3	5	5.6	6.5	0.0	9.2	8.7	0.0	0.0	43.5	47.3	0.0	55.1	52.9	0.0	0.0

Table 5: Empirical type I error rates and empirical powers (%) at the 5% level of significance and based on 1000 Monte-Carlo simulations for the unbalanced two-by-two design. Table 3 gives the coding for T and E where 5a denotes the setting where the sample size is inversely proportional to the variance, while for 5b the sample size is proportional to the variance. Here $n_{[i,j]}$ denotes the number of observations when $X_A = i$ and $X_B = j$.

T	E	Type I error (%) ($\theta_4 = 0$)							Power (%) ($\theta_4 = 0.7$)						
		IPS	$\mathcal{F}$	RT	ART	$RART$	PH	$\mathcal{GA}$	IPS	$\mathcal{F}$	RT	ART	$RART$	PH	$\mathcal{GA}$
		$n_{[-1,-1]} = 4, \quad n_{[1,-1]} = 3,$							$n_{[-1,1]} = 7, \quad n_{[1,1]} = 6$						
1	1	6.9	4.8	4.9	4.5	5.4	4.3	9.8	35.2	30.7	28.6	30.3	30.1	24.9	8.3
1	2	6.8	5.5	5.3	5.1	5.7	3.3	8.6	53.8	50.2	60.0	56.2	56.0	47.0	6.2
1	3	6.3	3.1	5.0	5.3	5.9	3.4	9.3	49.4	44.5	67.6	58.0	57.3	55.0	7.7
1	4	8.0	0.0	5.6	0.0	6.8	4.1	8.7	43.1	0.0	37.9	0.0	38.6	31.3	8.4
1	5a	11.9	15.6	9.4	13.5	12.6	4.2	8.0	27.3	35.3	32.1	32.2	27.9	22.3	8.4
1	5b	4.4	1.7	2.8	3.1	3.6	3.8	9.5	27.9	16.5	20.7	24.1	25.2	22.2	8.4
2	1	6.6	5.6	3.0	5.4	6.1	1.4	8.4	33.3	30.2	29.7	29.5	29.6	9.4	8.2
2	2	5.9	4.1	5.5	4.7	5.0	0.4	6.8	50.0	47.3	39.3	53.3	54.5	8.6	7.7
2	3	4.9	3.8	6.0	5.8	5.5	0.4	6.2	51.7	46.1	34.7	59.3	61.6	6.3	6.2
2	4	8.5	0.0	6.9	0.0	6.3	2.3	11.9	45.0	0.0	41.3	0.0	40.9	18.4	11.5
2	5a	10.1	13.3	14.6	12.4	10.6	3.9	9.0	29.3	38.1	51.1	34.1	29.9	17.0	8.2
2	5b	4.2	1.9	1.5	3.7	4.0	2.9	5.8	27.5	15.2	10.1	23.2	24.0	23.4	4.8
3	1	6.0	4.6	0.0	4.8	4.7	0.0	0.9	31.7	27.2	0.0	27.1	27.0	0.0	1.1
3	2	5.8	4.5	0.2	5.2	5.2	0.0	0.7	53.1	47.9	0.2	53.0	52.5	0.0	1.7
3	3	6.4	4.9	0.5	6.7	6.2	0.0	1.2	45.6	41.0	0.3	55.7	57.0	0.0	1.4
3	4	8.6	0.0	100.0	0.0	5.9	0.0	4.1	48.2	0.0	100.0	0.0	42.0	0.0	2.7
3	5a	10.9	14.3	0.0	11.8	10.9	0.0	0.7	26.0	33.2	0.0	30.8	26.8	0.0	0.5
3	5b	5.4	2.0	0.0	4.3	4.8	0.0	0.4	26.0	13.8	0.0	21.6	23.4	0.0	0.4
		$n_{[-1,-1]} = 8, \quad n_{[1,-1]} = 6,$							$n_{[-1,1]} = 14, \quad n_{[1,1]} = 12$						
1	1	5.5	5.0	4.7	4.6	4.9	3.3	6.5	54.9	55.1	54.3	52.8	54.1	44.7	4.0
1	2	6.3	5.2	5.4	5.9	5.8	3.6	5.5	80.6	76.3	88.7	86.1	86.8	81.6	3.5
1	3	4.3	4.0	5.2	6.1	6.4	3.3	5.0	70.2	61.9	93.8	88.2	92.0	89.6	2.7
1	4	6.3	0.0	5.4	100.0	7.1	4.7	6.7	63.2	0.0	58.3	100.0	60.2	50.0	4.9
1	5a	7.3	14.5	9.4	13.2	12.6	5.2	6.6	31.8	52.2	51.9	47.2	45.0	41.2	3.8
1	5b	5.2	1.9	2.7	4.0	4.2	2.8	7.3	54.4	37.0	46.5	49.1	49.4	46.6	6.1
2	1	6.8	5.9	4.5	6.2	6.2	2.8	0.9	53.7	53.6	55.6	52.0	52.4	22.6	0.3
2	2	4.2	4.1	9.4	3.9	4.4	0.8	0.0	76.6	73.9	74.2	84.7	86.1	26.9	0.1
2	3	4.7	4.1	11.9	5.7	5.9	0.8	0.0	69.8	62.5	60.7	89.1	91.8	13.3	0.2
2	4	5.6	0.0	7.5	100.0	6.5	1.8	0.5	66.8	0.0	69.1	100.0	61.9	35.2	0.0
2	5a	7.0	13.1	25.1	12.9	12.3	6.1	0.3	35.8	55.3	84.4	52.8	49.2	45.5	0.2
2	5b	4.7	1.1	1.7	4.2	4.3	12.6	1.2	51.4	33.1	33.4	47.5	47.6	73.7	1.7
3	1	7.4	7.2	100.0	6.5	7.0	0.0	0.0	52.8	52.8	100.0	50.5	51.2	0.0	0.0
3	2	6.4	5.8	99.6	6.0	5.9	0.0	0.0	77.9	73.2	99.6	85.0	86.1	0.0	0.0

3	3	5.6	3.7	98.1	5.4	6.0	0.0	0.0	66.9	61.9	98.8	87.9	90.8	0.0	0.0
3	4	6.5	0.0	100.0	100.0	6.8	0.0	0.0	66.3	0.0	100.0	100.0	61.1	0.0	0.0
3	5a	4.9	13.3	100.0	11.8	10.9	0.0	0.0	34.6	52.7	100.0	50.1	47.7	0.0	0.0
3	5b	5.0	1.3	100.0	4.6	4.9	0.0	0.0	50.3	32.4	100.0	46.8	46.2	0.0	0.0

Table 6: Empirical type I error rates and empirical powers (%) at the 5% level of significance and based on 1000 Monte-Carlo simulations for the two-by-three design. Table 3 gives the coding for T and E ; n denotes the number of observations in each cell.

T	E	Type I error (%) ($\theta_5 = \theta_6 = 0$)						Power (%) ($\theta_5 = \theta_6 = 0.5$)							
		$IPSBG$	$\mathcal{F}$	RT	ART	$RART$	GA	$IPSBG$	$\mathcal{F}$	RT	ART	$RART$	GA		
		$n = 5$													
1	1	3.1	4.6	5.3	4.3	4.7	14.0	30.8	36.9	36.0	35.7	35.2	51.1		
1	2	4.2	5.4	7.2	7.1	6.7	13.5	56.3	58.5	69.0	67.9	68.3	79.3		
1	3	3.2	4.5	5.4	5.5	5.5	12.8	35.9	37.1	56.4	49.2	54.1	68.9		
1	4	3.5	0.0	5.1	22.1	5.2	13.6	34.6	0.0	39.7	63.1	39.5	55.1		
1	5	4.3	7.8	6.0	9.3	8.6	12.2	15.2	23.3	22.0	25.8	23.4	27.2		
2	1	3.4	4.9	5.9	5.1	5.5	8.8	30.4	35.0	41.7	34.7	34.5	32.8		
2	2	3.4	5.4	12.5	5.5	5.3	6.4	53.9	58.3	68.1	67.4	68.7	44.2		
2	3	2.9	4.9	10.6	5.7	6.0	10.5	38.5	40.1	51.3	52.1	55.0	41.9		
2	4	3.0	0.0	8.1	24.5	4.9	9.0	34.9	0.0	42.7	62.9	41.2	42.5		
2	5	4.1	6.8	16.6	9.8	8.0	14.8	15.0	24.8	54.9	28.0	22.9	39.4		
3	1	2.9	4.5	43.3	4.6	4.4	3.3	27.0	32.3	43.3	31.8	31.7	3.3		
3	2	2.1	2.8	45.1	3.8	4.2	3.5	57.9	60.0	45.2	67.6	69.4	3.5		
3	3	3.9	4.2	42.9	6.3	7.1	3.7	39.7	41.2	42.9	53.4	56.1	3.7		
3	4	3.7	0.0	24.2	19.5	6.1	8.6	32.2	0.0	24.2	64.8	36.6	8.6		
3	5	5.1	8.3	43.1	11.0	9.8	4.4	15.7	24.5	43.4	26.6	23.3	4.4		
		$n = 10$													
1	1	3.6	4.3	4.4	4.5	4.4	7.3	62.1	65.8	63.7	64.0	62.5	69.6		
1	2	4.2	4.4	6.0	5.4	5.3	8.2	90.5	82.1	95.6	94.7	94.9	96.3		
1	3	4.6	5.9	5.4	6.2	5.9	8.2	75.7	69.1	90.6	88.2	88.5	91.0		
1	4	2.7	0.0	3.7	100.0	3.7	6.9	67.9	0.0	68.0	100.0	68.4	74.6		
1	5	4.0	6.5	4.6	10.0	7.8	6.9	28.6	37.6	39.9	42.6	37.6	36.4		
2	1	4.0	4.4	9.4	5.8	5.6	6.4	60.3	63.6	75.1	61.7	61.8	55.0		
2	2	4.3	4.6	24.3	5.5	5.8	7.5	88.6	81.8	93.7	94.1	94.3	74.2		
2	3	3.5	4.0	16.1	5.4	5.2	9.1	76.2	70.5	85.6	88.6	89.4	68.5		
2	4	3.3	0.0	12.6	100.0	4.7	5.9	70.3	0.0	80.4	100.0	71.7	66.8		
2	5	4.3	7.5	30.5	9.9	7.9	16.4	28.2	38.1	84.6	43.0	38.4	64.1		
3	1	4.9	5.3	100.0	5.1	5.0	2.5	65.4	68.1	100.0	67.4	67.1	2.5		
3	2	4.1	5.4	99.9	5.3	6.0	2.3	88.3	82.9	99.9	94.3	95.3	2.3		
3	3	4.9	6.6	100.0	6.0	6.3	2.7	72.6	66.6	100.0	86.9	88.4	2.7		
3	4	3.9	0.0	100.0	100.0	4.9	5.2	65.7	0.0	100.0	100.0	66.1	5.2		
3	5	5.0	6.7	100.0	9.4	8.0	1.8	26.7	39.6	100.0	43.9	36.3	1.8		

Table 7: Empirical powers (%) at the 5% level of significance and based on 1000 Monte-Carlo simulations for $\theta_4 = 0.3$ and $\theta_4 = 1.1$ for the two-by-two design. Table 3 gives the coding for T and E ; n denotes the number of observations in each cell.

T	E	Power (%) ($\theta_4 = 0.3$)							Power (%) ($\theta_4 = 1.1$)						
		IPS	$\mathcal{F}$	$\mathcal{RT}$	$\mathcal{ART}$	$\mathcal{RART}$	$\mathcal{PH}$	$\mathcal{GA}$	IPS	$\mathcal{F}$	$\mathcal{RT}$	$\mathcal{ART}$	$\mathcal{RART}$	$\mathcal{PH}$	$\mathcal{GA}$
		$n = 5$													
1	1	9.9	9.7	9.6	9.8	9.0	8.1	13.3	59.4	60.2	59.4	60.7	56.2	51.5	66.0
1	2	14.5	12.7	17.4	16.6	14.5	13.1	22.9	78.7	81.3	90.0	87.2	86.6	84.2	86.3
1	3	18.4	14.8	30.3	21.5	24.1	23.0	37.0	67.9	74.0	89.6	86.4	86.4	86.9	78.5
1	4	15.7	0.0	14.6	0.0	13.2	10.3	19.8	74.7	0.0	71.1	0.0	67.6	62.2	76.9
1	5	9.1	10.6	10.4	14.0	11.9	8.3	16.2	50.3	56.1	61.7	63.5	56.6	58.0	71.1
2	1	10.8	10.0	4.3	11.8	9.6	3.7	12.3	64.3	62.8	45.2	63.0	58.7	40.4	66.2
2	2	15.1	13.9	3.6	16.1	15.9	2.8	8.9	79.6	83.8	51.4	88.9	89.0	52.3	72.1
2	3	17.1	13.2	2.8	19.5	21.8	2.0	6.7	66.5	72.8	52.9	86.5	87.3	50.5	69.7
2	4	16.9	0.0	12.6	0.0	13.7	11.9	23.4	72.1	0.0	58.3	0.0	67.2	58.0	77.0
2	5	10.4	11.1	15.0	13.3	11.6	12.9	32.0	51.2	54.6	54.7	59.6	54.8	51.1	77.3
3	1	10.4	10.0	0.0	10.7	8.6	0.0	0.0	65.8	65.9	0.0	64.5	60.6	0.0	0.0
3	2	14.4	11.5	0.0	15.3	13.8	0.0	0.0	77.4	81.1	0.0	88.2	87.8	0.0	0.0
3	3	17.2	13.8	0.0	20.1	23.2	0.0	0.0	66.7	73.7	0.0	85.8	87.2	0.0	0.0
3	4	14.4	0.0	0.0	0.0	11.4	0.0	0.0	74.8	0.0	0.0	0.0	67.5	0.0	0.0
3	5	8.6	8.5	0.0	11.7	10.1	0.0	0.0	46.6	51.8	0.0	57.1	50.7	0.0	0.0
		$n = 10$													
1	1	14.1	14.0	13.4	13.8	13.9	10.9	15.3	90.1	90.2	88.9	88.8	88.9	86.0	89.5
1	2	23.1	20.6	29.1	27.5	28.4	23.1	31.7	99.4	97.2	100.0	99.8	99.9	100.0	98.3
1	3	30.4	20.5	53.0	41.0	49.8	46.6	56.0	90.8	87.0	99.9	99.5	99.9	99.8	98.2
1	4	22.1	0.0	19.4	0.0	19.3	15.3	22.6	96.4	0.0	94.3	0.0	94.9	91.5	94.4
1	5	13.1	14.6	15.6	18.4	17.4	14.2	19.2	83.3	85.9	89.5	88.4	87.3	89.7	88.5
2	1	15.2	15.4	9.0	15.4	15.7	9.9	20.4	92.9	92.6	82.3	91.9	91.9	79.0	89.9
2	2	27.0	24.4	7.9	31.7	33.2	9.6	20.8	97.9	96.4	88.1	99.9	100.0	91.2	92.9
2	3	32.9	23.8	5.3	42.3	50.7	6.3	14.8	93.1	89.2	83.7	99.3	99.8	89.9	86.7
2	4	21.3	0.0	13.8	0.0	20.3	14.1	25.1	95.5	0.0	88.3	0.0	93.9	87.7	93.2
2	5	11.9	13.2	35.3	19.7	18.5	34.6	53.9	83.7	86.6	92.2	87.7	86.3	92.0	97.9
3	1	15.7	16.1	0.0	15.6	16.1	0.0	0.0	91.1	91.1	0.0	90.2	90.4	0.0	0.0
3	2	27.9	25.2	0.0	30.1	32.1	0.0	0.0	98.3	96.1	0.0	99.8	99.8	0.0	0.0
3	3	29.8	22.3	0.0	41.8	50.4	0.0	0.0	93.1	89.2	0.0	99.3	99.5	0.0	0.0
3	4	22.7	0.0	0.0	0.0	21.4	0.0	0.0	95.8	0.0	0.0	0.0	94.5	0.0	0.0
3	5	14.1	14.4	0.0	19.0	18.1	0.0	0.0	82.4	85.8	0.0	87.8	86.6	0.0	0.0

Figure 1: Left: outcome densities for $Y_{11} \stackrel{d}{=} Y_{12} \stackrel{d}{=} Y_{21} \stackrel{d}{=} N(0, 0.5)$ and $Y_{22} \stackrel{d}{=} N(-3, 0.5)$ (top) and associated effect outcome densities for $Z = Y_{11} - Y_{21} \stackrel{d}{=} N(0, 1)$ and $Z^* = Y_{12} - Y_{22} \stackrel{d}{=} N(3, 1)$ (bottom). Right: outcome densities for $Y_{11} \stackrel{d}{=} Y_{21} \stackrel{d}{=} N(0, 0.5)$, $Y_{12} \stackrel{d}{=} N(0, 16)$ and $Y_{22} \stackrel{d}{=} N(-3, 16)$ (top), and associated effect outcome densities for $Z = Y_{11} - Y_{21} \stackrel{d}{=} N(0, 1)$ and $Z^* = Y_{12} - Y_{22} \stackrel{d}{=} N(3, 32)$ (bottom). $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 .

Figure 2: Left: FEV according to smoking status of the child (NS: non-smoker and S: smoker) and the age (11 or 15 years). Right: effect outcome of smoking (i.e. $FEV_{NS} - FEV_S$) according to age.