Biometrics 68, 75–84
March 2012

DOI: 10.1111/j.1541-0420.2011.01622.x

A Powerful and Robust Test Statistic for Randomization Inference in Group-Randomized Trials with Matched Pairs of Groups

Kai Zhang,* Mikhail Traskin, and Dylan S. Small*****

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

**email:* zhangk@wharton.upenn.edu

***email:* mtraskin@wharton.upenn.edu

****email:* dsmall@wharton.upenn.edu

Summary. For group-randomized trials, randomization inference based on rank statistics provides robust, exact inference against nonnormal distributions. However, in a matched-pair design, the currently available rank-based statistics lose significant power compared to normal linear mixed model (LMM) test statistics when the LMM is true. In this article, we investigate and develop an optimal test statistic over all statistics in the form of the weighted sum of signed Mann-Whitney-Wilcoxon statistics under certain assumptions. This test is almost as powerful as the LMM even when the LMM is true, but it is much more powerful for heavy tailed distributions. A simulation study is conducted to examine the power.

Key words: Causal effect; Group-randomized trials; Randomization inference; Rank-based statistics; Robustness.

1. Introduction

Group-randomized trials are commonly used in medical research. In a group-randomized trial, the randomization is assigned to groups rather than to individuals. For example, the Prospect Study (Prevention of Suicide in Primary Care Elderly: Collaborative Trial) was a group-randomized trial of treatments for depression among adults over the age of 60 (Bruce et al., 2004). Twenty primary care practices of various sizes were grouped into 10 pairs on the basis of region (urban/other), affiliation, overall size, and population type, and randomization was used to select one practice in each pair for the intervention; the other practice served as a control. The intervention provided the practice with a depression care manager who was either a social worker, a nurse, or a psychologist. The manager was supervised on a weekly basis by a psychiatrist. The depression care manager provided guideline-based treatment recommendations to the physicians in the practice and interpersonal psychotherapy to some patients. The control practices provided "usual care" with certain enhancements involving diagnosis and the education of the physicians about the treatment guidelines. The outcome is the change in the Hamilton depression score, a 24-item measure of depression severity, from baseline to 4 months. The mean changes within the 20 groups are summarized in Table 1. Note that there is sometimes a substantial difference in the numbers in the treated and control group in a pair (e.g., in pair 4). Although practices were paired on the overall size of primary care practices, they were not paired on the number of patients eligible for the study; see Bruce et al. (2004) for the eligibility criteria.

Group-randomized trials have been discussed by Cornfield (1978); Gail et al. (1996); Brookmeyer and Chen (1998); Donner (1998); Murray (1998); Braun and Feng (2001); Frangakis, Rubin, and Zhou (2002); and Murray et al. (2006), among others. Group-randomized trials can be thought of as trials about the effectiveness of changing the infrastructure (practice management in the case of the Prospect Study) for the benefit of present and future persons. The essential role of matching in group-randomized trials is discussed by Imai, King, and Nall (2009). Matching techniques are discussed by Zhang and Small (2009).

A common approach for analyzing group-randomized trials is to assume that the treatment effect is additive and to use the linear mixed model (LMM). The LMM further assumes that the data are generated from normal distributions. This additional assumption of normality is strong and leads to a test that loses substantial power when the true distributions are heavy tailed. The level of the test might also be away from the nominal level if the assumption is violated (see Section 4 for details).

From the practical point of view, we would like to have tests that have approximately the desired nominal level under nonnormal distributions. Among these tests, we would further prefer the tests that have powers comparable to the optimal tests under a wide range of distributions. To achieve such a robust test, we consider rank-based statistics. In particular, to ensure that the level of the test is correct, we utilize randomization inference, inference that relies only on the null hypothesis under test and the randomization actually used in the experiment, rather than relying on any additional assumptions about the stochastic process that generated the data (Fisher, 1935; Pitman, 1937; Welch, 1937; Kempthorne, 1952; Scheffé, 1959, chapter 9; Lehmann, 1998, section 1; Cox and Reid, 2000, section 2.2.5; Rosenbaum, 2002).

© 2011, The International Biometric Society

Table 1
Summary of data from prospect study

Pair	Size of control group	Size of treated group	Mean change in control group	Mean change in treated group
1	44	49	−4.7	−4.6
2	31	6	−0.3	−7.3
3	5	27	−2.0	−8.2
4	22	1	−3.7	−3.0
5	29	26	−2.8	−7.7
6	5	37	−6.6	−5.9
7	29	17	−5.0	−9.9
8	22	40	−4.7	−8.7
9	23	20	−4.9	−9.1
10	24	30	−5.9	−9.1
All	234	253	−3.9	−7.5

Small, Ten Have, and Rosenbaum (2008) developed a rank-based statistic W, which is a weighted sum of signed Mann-Whitney-Wilcoxon (MWW) statistics. The weights in W are calculated based on the cluster sizes only. Small et al. (2008) showed that when the variation in an observation is mainly explained by the within-group variation rather than between-group variation, W has similar power to the LMM when the data are indeed generated from the LMM model, and has higher power than the LMM when the data are not normally distributed. The details of this test statistic are discussed in the next section.

The test based on W relaxes the normal assumption in the LMM and is robust against deviations from normality. However, when most of the variation of an observation is explained by the between-group variation, the power of the W tests can fall considerably below that of the LMM test, when the data are indeed generated from the LMM model. To find a test statistic that achieves robust inference against nonnormal distributions and has high power when the data are generated from the LMM, we propose a new method that accounts for the intracluster correlation in group-randomized trials. Our method can adjust for the covariates with no change in the logic; see Rosenbaum (2002).

The remainder of this article is organized as follows. In Section 2, we introduce the additive effect model and our notation. In Section 3, we describe the methodology of constructing the new test statistic and explain how we determine the optimal weights. Simulation results are shown in Section 4 to compare the powers of test statistics. We apply the new test statistic to the Prospect Study in Section 5. We summarize the article in Section 6. We outline how to derive the optimal weights in the Appendix.

2. Setup and Notation

2.1 *The Additive Effect Model*

In a group-randomized trial with n subjects, a subject k has two potential responses: let r_{T_k} denote the outcome the subject would have if k's group was assigned to treatment (subject k's potential response under treatment) and r_{C_k} denote the outcome the subject would have if k's group was assigned to control (subject k's potential response under control). General descriptions on potential responses are given by Neyman (1923) and Rubin (1974). The quantity $r_{T_k} - r_{C_k}$ measures the effect on subject k of assigning k's group to treatment rather than control. Also, note that for a subject k whose group was assigned to the treatment, we observe r_{T_k} but not r_{C_k}; however, because it is a randomized experiment, the subject assigned to the control group provides an approximation to the counterfactual r_{C_k}'s. General descriptions on counterfactuals are given in Lewis (1973a, 1973b) and Robins (1986, 1987).

We say the treatment effect is additive if there is a constant τ, such that $\tau = r_{T_k} - r_{C_k}$ for all subjects $k = 1, \ldots, n$. In this case, the distribution of r_{T_k}'s is shifted by τ from the distribution of r_{C_k}'s. However, the shape and scale of the distribution of the r_{T_k}'s are the same as those of the r_{C_k}'s.

For example, consider the Prospect Study. Figure 1 shows the boxplots of these data.

From the table and the boxplots, we see that the shape and scale of clusters within each pair are very similar. However, the means of the clusters seem to be shifted. Therefore, the additive effect model seems to be a reasonable model for these data.

The details of the additive effect model for a matched-pair group-randomized trial are as follows. Suppose there are S pairs of clusters. Within the sth pair, $s = 1, \ldots, S$, there are two clusters of units. Within each pair, we use $j = T$ or $j = C$ to denote the treated cluster and the control cluster. The sizes of the clusters in pair s are denoted by n_{sj}.

We denote the observed response from the kth individual in the sth pair, jth cluster, as R_{sjk}. Under the additive effect model, we can write the observed response for subject sjk, R_{sjk} as follows.

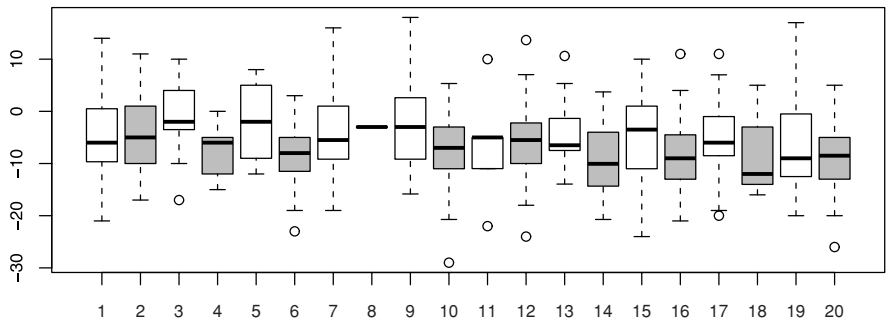

Figure 1. Boxplots of data from the Prospect Study. Every two clusters form a pair. Treated clusters are shaded.

$$R_{sjk} = \tau Z_{sj} + \phi_s + \gamma_{sj} + \zeta_{sjk}; \quad (1)$$

where τ is the treatment effect; Z_{sj} is the treatment indicator on the sth pair and j cluster; ϕ_s is the fixed pair effect on the sth pair; γ_{sj} is the cluster effect on the sth pair and jth cluster; and ζ_{sjk} are individual effects. The response from each treated individual is the sum of the treatment effect, the effect from his/her pair, the effect from his/her cluster, and an individual effect. Note that we will focus on deviations of observed responses within a pair so that the pair effects ϕ_s's do not affect the inference. More general descriptions of the model can be found in Murray (1998).

In this notation, the potential outcomes can be written as follows:

$$r_{T_{sjk}} = \tau + \phi_s + \gamma_{sj} + \zeta_{sjk} \quad \text{and} \quad r_{C_{sjk}} = \phi_s + \gamma_{sj} + \zeta_{sjk}. \quad (2)$$

The LMM assumes the distributions of the cluster effects and the individual errors are independent and identically distributed (i.i.d.) normal. As noted in Section 1, this assumption is strong and leads to problems when it is violated.

Our problem of interest is the following test:

$$H_0 : \tau = \tau_0 \quad \text{versus} \quad H_1 : \tau < \tau_0, \quad (3)$$

or alternatively, we can consider $H_0^{'} : \tau = \tau_0$ versus $H_1^{'} : \tau > \tau_0$, or $H_0^{''} : \tau = \tau_0$ versus $H_1^{''} : \tau \neq \tau_0$.

2.2 *Current Test Statistics and Heuristics of Improvements*

As stated before, to achieve robust inferences against nonnormal distributions, we consider rank-based statistics and use randomization inference.

We refer to $e_{sjk} = R_{sjk} - \tau_0 Z_{sj}$ as the adjusted response of unit sjk under $H_0 : \tau = \tau_0$ (adjusted response for short). One of the classical rank-based statistics, the signed MWW statistic for pair s, W_s is,

$$W_s = \sum_{l=1}^{n_{sT}} \sum_{m=1}^{n_{sC}} \{\mathbf{I}(e_{sTl} > e_{sCm}) - \mathbf{I}(e_{sCm} > e_{sTl})\}. \quad (4)$$

W_s is the number of pairs of individual treated-control units in the pair of clusters s for which the treated unit has bigger adjusted response minus the number of pairs for which the control unit has bigger adjusted response.

Small et al. (2008) developed a rank-based statistic W, which is defined as

$$W = \sum_{s=1}^{S} \frac{1}{n_{sT} + n_{sC} + 1} W_s. \quad (5)$$

W is essentially a weighted sum of S signed MWW statistics, with weights $\{(n_{sT} + n_{sC} + 1)^{-1}\}_{s=1}^{S}$.

Before discussing the property of this test statistic W, we first introduce the concept of the intracluster correlation in group-randomized trials: when the variance of the cluster effects, σ_γ^2, and the variance of the individual errors, σ_ζ^2, are finite, the intracluster correlation λ is defined as $\lambda = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\zeta^2}$.

Thus, because the pair effects are not random, the intracluster correlation λ can be regarded as the proportion of variance of an observation that is explained by the cluster effects.

Small et al. (2008) showed that when the intracluster correlation λ is small, W has similar power to the LMM when the data are indeed generated from the LMM model, and has higher power than the LMM when the data are not normally distributed. However, when the intracluster correlation λ approaches 1, W loses a great deal of power compared to the LMM. This is because W assigns the weights of W_s according to the pair sizes only and ignores the intracluster correlation. Actually, Braun and Feng (2001) discussed this problem and proposed that cluster sizes are not the best weights to use, and that instead the weights should incorporate the intracluster correlation.

To this end, we note that the higher the intracluster correlation λ is, the more similar the outcomes from a given cluster are. Therefore, the weights assigned to each cluster should be closer to each other because the proportion of the variation from each unit can be explained more by the cluster mean. In the extreme case, when the intracluster correlation λ is 1, the information from each cluster can be obtained by taking the cluster mean. In this case, we should equally weight the clusters.

Based on the above logic, we investigate a rank-based test statistic that is in the form of a weighted sum of the MWW statistics. The weights are determined by the data and lead to a test statistic that satisfies the following criteria: when λ is small, it has similar power to W; when λ approaches 1, it assigns nearly equal weights to each pair; when the distributions of the cluster effects and the individual errors are not normal, it maintains the nominal level and has higher power compared to the LMM test and the W test.

3. Construction of the New Test Statistic $U_{\mathbf{w}}$

As discussed in Section 2, the pair effects do not affect the inference; therefore, without loss of generality, we assume the pair effects are 0's, i.e., $\phi_s = 0$, $s = 1, \ldots, S$.

3.1 *Methodology*

To find the optimal weights, we first consider the normalized signed MWW statistic for pair s, Q_s, which is described as follows.

$$Q_s = \frac{1}{n_{sT} n_{sC}} W_s. \quad (6)$$

Thus, the normalization is obtained by dividing the MWW statistic by the product of cluster sizes within the pair. After normalization, Q_s ranges from -1 to 1, $\forall s$.

With Q_s's available, the new statistic $U_{\mathbf{w}}$ is constructed in the form of

$$U_{\mathbf{w}} = \sum_{s=1}^{S} w_s Q_s, \quad (7)$$

where $\{w_s\}$ are weights assigned to pairs with $\sum_{s=1}^{S} w_s = 1$ and $w_s \geq 0$.

Therefore, $U_{\mathbf{w}}$, similarly to W, is also a weighted sum of MWW statistics W_s but its weights $w_s/(n_{sT} n_{sC})$ are different than those of W. The set of weights of Q_s's, $\{w_s\}$, is crucial for $U_{\mathbf{w}}$. In choosing these weights in $U_{\mathbf{w}}$, we will focus on maximizing the power of the test against a specific $H_1^* : \tau = \tau_1 < \tau_0$. We suggest choosing τ_1 to be an alternative of practical significance that the study is designed to have good power for.

For example, for the Prospect Study, Falissard, Lukasiewicz, and Corruble (2003) consider a difference of 2.7 or lower on the Hamilton depression score between treatment groups is clinically meaningful. To focus on having high power for this difference, we maximize the power for the testing problem: $H_0 : \tau = 0$ and $H_1 : \tau = -2.7$. After the specification of H_0 and H_1, our methodology of constructing the weights for U_{w} can be described as follows:

(1) Under the assumption that cluster and individual effects being i.i.d. normal, find the mean and variance of U_{w} under H_0 and H_1;
(2) Approximate the distribution of Q_s by a normal distribution, and then choose the set of weights $\{w_s\}$ that maximizes the power of the test;
(3) With the weights found in the second step, obtain the exact distribution of U_{w} under randomization to draw valid randomization inference.

The weights we construct are estimates of the optimal ones under the normal assumption. However, regardless of whether the normal assumption holds, our inferences are randomization inferences and therefore have the exact stated nominal level. The details of this procedure of constructing U_{w} are explained in the following subsection.

We should note that in applications, before collecting the data and constructing the test statistic, one should first specify τ_0 and τ_1 to be practically meaningful. Doing so avoids the problem of manipulating a test.

3.2 *Construction of* U_{w}

To construct U_{w} under the normal assumption, we first find its expectation and variance. We need to consider some related probabilities here. Because we are concerned about rank-based statistics, the probabilities related to comparisons between treated outcomes and control outcomes are crucial. Moreover, the values of some of these probabilities are functions of the intracluster correlation under the normal assumption. Therefore, by working with these probabilities, we incorporate the intracluster correlation in our test statistic. Our arguments are patterned after the corresponding arguments for the MWW test as in Lehmann (1998, section 2.3), with the main complication being that we have to take account of the cluster effects.

Because under H_0 the distribution of adjusted responses from two different pairs are independent of each other, we consider within-pair comparisons only. We first consider a probability needed for the expectation of U_{w}: For a treated unit and a control unit in pair s, we denote by c the probability that the adjusted response of the treated unit is larger than the control unit,

$$c = \mathbf{P}(e_{sTl} > e_{sCm}). \tag{8}$$

For example, under the null hypothesis $\tau = 0, c = \frac{1}{2}$. For c defined in (8), the expectation of Q_s as defined in (6) is given as follows.

Proposition 1.

$$\mathrm{E}[Q_s] = 2c - 1.$$

The proof of Proposition 1 can be found in Web Appendix A.

For calculating the variance of Q_s we need three further probabilities arising from the square of the sum in equation (4). These three probabilities are closely related to the intracluster correlation.

If we sample one treated unit and two control units without replacement from pair s, we denote by p_1 the probability that the adjusted response of the treated unit is larger than those of both control units,

$$p_1 = \mathbf{P}(e_{sTl} > e_{sCm_1} \text{ and } e_{sTl} > e_{sCm_2}, m_1 \neq m_2). \tag{9}$$

On the other hand, if we sample two treated units and one control unit from pair s, we denote the probability that the adjusted response of the control unit is smaller than those of both treated units by p_2,

$$p_2 = \mathbf{P}(e_{sTl_1} > e_{sCm} \text{ and } e_{sTl_2} > e_{sCm}, l_1 \neq l_2). \tag{10}$$

Finally, for two pairs of treated and control units sampled without replacement, we denote by q the probability that the adjusted responses of the treated unit are larger in their pairs,

$$q = \mathbf{P}(e_{sTl_1} > e_{sCm_1} \text{ and } e_{sTl_2} > e_{sCm_2}, l_1 \neq l_2 \text{ and } m_1 \neq m_2). \tag{11}$$

These three probabilities are functions of the intracluster correlation under the LMM. For instance, under the null hypothesis $\tau = 0$, if $\lambda = 0$, then $p_1 = p_2 = \frac{1}{3}$ and $q = \frac{1}{4}$; if $\lambda = 1$ instead, then $p_1 = p_2 = q = \frac{1}{2}$. Note also that $p_1 = p_2$ if the distributions of the cluster effects and the individual errors are symmetric. However, this equality does not necessarily hold in general.

With the above probabilities, we have the following proposition about the variance of Q_s.

Proposition 2. *For* c, p_1, p_2, *and* q *defined in* (8), (9), (10), *and* (11), *respectively, the variance of* Q_s *as defined in* (6) *is given as follows.*

$$\mathrm{Var}[Q_s] = \frac{4}{n_{sT}n_{sC}}\left\{c - p_1 - p_2 + q + (p_1 + p_2 - 2q) \times \frac{n_{sT} + n_{sC}}{2} + (q - c^2)n_{sT}n_{sC}\right\}.$$

The proof of Proposition 2 can be found in Web Appendix B. Based on Proposition 1 and Proposition 2, we have

$$\mathrm{E}[U_{\mathrm{w}}] = 2c - 1 \quad \text{and} \quad \mathrm{Var}[U_{\mathrm{w}}] = \sum_{s=1}^{S} w_s^2 \mathrm{Var}[Q_s]. \tag{12}$$

Because we are interested in the additive effect, without loss of generality, we assume that the null hypothesis is $H_0 : \tau_0 = 0$. Then if we denote under $H_0, p_{10} = \mathbf{P}(e_{sTl} > e_{sCm_1}$ and $e_{sTl} > e_{sCm_2}$, $m_1 \neq m_2 \,|\, H_0)$, $p_{20} = \mathbf{P}(e_{sTl_1} > e_{sCm}$ and $e_{sTl_2} > e_{sCm}, l_1 \neq l_2 \,|\, H_0)$, and $q_0 = \mathbf{P}(e_{sTl_1} > e_{sCm_1}$ and $e_{sTl_2} > e_{sCm_2}$, $l_1 \neq l_2$ and $m_1 \neq m_2 \,|\, H_0)$, we have

$$\mathrm{E}_{H_0}[Q_s] = 0;$$

$$\mathrm{Var}_{H_0}[Q_s] = \frac{4}{n_{sT}n_{sC}}\left\{\frac{1}{2} - p_{10} - p_{20} + q_0 + (p_{10} + p_{20} - 2q_0) \times \frac{n_{sT} + n_{sC}}{2} + \left(q_0 - \frac{1}{4}\right) n_{sT}n_{sC}\right\}.$$

Similarly, if under H_1, $c_1 = \mathbf{P}(e_{sTl} > e_{sCm} \,|\, H_1)$, $p_{11} = \mathbf{P}(e_{sTl} > e_{sCm_1}$ and $e_{sTl} > e_{sCm_2}, \quad m_1 \neq m_2 \,|\, H_1)$, $\quad p_{21} = \mathbf{P}(e_{sTl_1} > e_{sCm}$ and $e_{sTl_2} > e_{sCm}, l_1 \neq l_2 \,|\, H_1)$, $\quad$ and $\quad q_1 = \mathbf{P}(e_{sTl_1} > e_{sCm_1}$ and $e_{sTl_2} > e_{sCm_2}, l_1 \neq l_2$ and $m_1 \neq m_2 \,|\, H_1)$, then we have

$$\mathrm{E}_{H_1}[Q_s] = 2c_1 - 1;$$

$$\mathrm{Var}_{H_1}[Q_s] = \frac{4}{n_{sT} n_{sC}} \Bigg\{ c_1 - p_{11} - p_{21} + q_1 + (p_{11} + p_{21} - 2q_1) \times \frac{n_{sT} + n_{sC}}{2} + (q_1 - c_1^2) n_{sT} n_{sC} \Bigg\}.$$

Denote the $\mathrm{E}_{H_1}[Q_s]$, $\mathrm{Var}_{H_0}[Q_s]$ and $\mathrm{Var}_{H_1}[Q_s]$ by E (note that $\mathrm{E}_{H_1}[Q_s]$ is the same for all s), V_{s0}, and V_{s1}, respectively. Then

$$\mathrm{E}_{H_0}[U_{\mathbf{w}}] = 0 \quad \text{and} \quad \mathrm{Var}_{H_0}[U_{\mathbf{w}}] = \sum_{s=1}^{S} w_s^2 V_{s0};$$

$$\mathrm{E}_{H_1}[U_{\mathbf{w}}] = E \quad \text{and} \quad \mathrm{Var}_{H_1}[U_{\mathbf{w}}] = \sum_{s=1}^{S} w_s^2 V_{s1}.$$

We notice that when the numbers of observations in the clusters are moderately large, we may approximate the distributions of Q_s's by normal distributions. In this case, the power of a level α test (3) based on $U_{\mathbf{w}}$ with $\tau_0 = 0$ and $\tau_1 < 0$ will be

$$\mathbf{\Phi}\left(\frac{z_\alpha \sqrt{\sum_{s=1}^{S} w_s^2 V_{s0}} - E}{\sqrt{\sum_{s=1}^{S} w_s^2 V_{s1}}} \right)$$

where z_α is the α-quantile of the standard normal distribution.

Therefore, to maximize the power of the test statistic based on $U_{\mathbf{w}}$, we choose the weights $\mathbf{w}^* = \{w_i^*\}$ such that

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{z_\alpha \sqrt{\sum_{s=1}^{S} w_s^2 V_{s0}} - E}{\sqrt{\sum_{s=1}^{S} w_s^2 V_{s1}}} \tag{13}$$

is maximized. The technical details of the maximization are discussed in the Appendix.

3.3 *Finding the Probabilities c, p_1, p_2, and q*

To find the probabilities in the formulas of E, V_{s0}'s and V_{s1}'s, we first estimate σ_γ^2 and σ_ζ^2 by $\widehat{\sigma}_\gamma^2$ and $\widehat{\sigma}_\zeta^2$ through the analysis of variance (ANOVA) method, i.e., $\widehat{\sigma}_\zeta^2 = \sum_{sjk}(e_{sjk} - \overline{e}_{sj\cdot})^2 / (\sum_{sj} n_{sj} - 2S)$ and $\widehat{\sigma}_\gamma^2 = (\sum_{sjk}(e_{sjk} - \overline{e}_{s\cdots})^2 - \sum_{sjk}(e_{sjk} - \overline{e}_{sj\cdot})^2 - (2S-1)\sigma_\zeta^2)_+ / (\sum_{sj} n_{sj} - \sum_{sj} n_{sj}^2 / \sum_{sj} n_{sj})$ where $\overline{e}_{sj\cdot} = \sum_k e_{sjk}/n_{sj}$ and $\overline{e}_{s\cdots} = \sum_{sjk} e_{sjk} / \sum_{sj} n_{sj}$. Note that under $H_0, \widehat{\sigma}_\gamma^2$ and $\widehat{\sigma}_\zeta^2$ are the same for all randomizations so that the weights are the same for all randomizations. Therefore, in computing the randomization inference by looking at the test statistic under each randomization under the null hypothesis, we only need to compute the weights once.

We then use normal distribution probabilities as a guideline to calculate those probabilities. Suppose $\gamma_{sj} \sim N(0, \sigma_\gamma^2)$ and $\zeta_{sjk} \sim N(0, \sigma_\zeta^2)$. Then if $l_1 \neq l_2$ and $m_1 \neq m_2$, the joint distribution of $(e_{sTl_1}, e_{sTl_2}, e_{sCm_1}, e_{sCm_2})$ is

$$\begin{pmatrix} e_{sTl_1} \\ e_{sTl_2} \\ e_{sCm_1} \\ e_{sCm_2} \end{pmatrix} \sim N\left(\begin{pmatrix} \tau \\ \tau \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\gamma^2 + \sigma_\zeta^2 & \sigma_\gamma^2 & 0 & 0 \\ \sigma_\gamma^2 & \sigma_\gamma^2 + \sigma_\zeta^2 & 0 & 0 \\ 0 & 0 & \sigma_\gamma^2 + \sigma_\zeta^2 & \sigma_\gamma^2 \\ 0 & 0 & \sigma_\gamma^2 & \sigma_\gamma^2 + \sigma_\zeta^2 \end{pmatrix} \right). \tag{14}$$

Therefore, we can easily find the following distributions related to those probabilities.

$$e_{sTl} - e_{sCm} \sim N\big(\tau, 2\big(\sigma_\zeta^2 + \sigma_\gamma^2\big)\big); \tag{15}$$

$$\big(e_{sTl} - e_{sCm_1}, e_{sTl} - e_{sCm_2}\big)^T \sim N((\tau,\tau)^T, \mathbf{\Sigma}_p); \tag{16}$$

$$\big(e_{sTl_1} - e_{sCm}, e_{sTl_2} - e_{sCm}\big)^T \sim N((\tau,\tau)^T, \mathbf{\Sigma}_p); \tag{17}$$

$$\big(e_{sTl_1} - e_{sCm_1}, e_{sTl_2} - e_{sCm_2}\big)^T \sim N((\tau,\tau)^T, 2\mathbf{\Sigma}); \tag{18}$$

where

$$\mathbf{\Sigma}_p = \begin{pmatrix} 2\sigma_\gamma^2 + 2\sigma_\zeta^2 & 2\sigma_\gamma^2 + \sigma_\zeta^2 \\ 2\sigma_\gamma^2 + \sigma_\zeta^2 & 2\sigma_\gamma^2 + 2\sigma_\zeta^2 \end{pmatrix} = 2\big(\sigma_\gamma^2 + \sigma_\zeta^2\big) \begin{pmatrix} 2 & 1+\lambda \\ 1+\lambda & 2 \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\gamma^2 + \sigma_\zeta^2 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_\gamma^2 + \sigma_\zeta^2 \end{pmatrix} = \big(\sigma_\gamma^2 + \sigma_\zeta^2\big) \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix}.$$

Hence, we see p_1 and p_2 are indeed equal under the normal assumption by (16) and (17). Moreover, we see clearly that p_1, p_2, and q are functions of the intracluster correlation.

We replace σ_γ^2 and σ_ζ^2 by their estimates to find those probabilities.

3.4 *Local Alternative Setting*

In the local alternative case where the difference between τ and τ_0 is small, the calculation of the optimal weights can be simplified by using the approximation $V_{s1} \doteq V_{s0}$. Under this approximation, the power is given by $\Phi(z_\alpha - R)$, where

$$R = \frac{E}{\sqrt{\sum_{s=1}^{S} w_s^2 V_{s0}}}.$$

This power is optimized by $w_s^* = \frac{1/V_{s0}}{\sum_{i=1}^{S}(1/V_{i0})}$. This set of weights has the advantage of being in a closed form and being easy to calculate. As a partial adjustment for the fact that V_{s1} and V_{s0} are not exactly equal, we could instead use $w_s = \frac{1/\tilde{V}_s}{\sum_{i=1}^{S}(1/\tilde{V}_i)}$, where $\tilde{V}_s = (1/2)(V_{s0} + V_{s1})$ to approximate the optimal power. However, we shall focus on our current approach in this article, because it provides the optimal power

but does not require determining whether the hypothesis is local or not.

3.5 *Covariance Adjustment*

In group-randomized trials, randomization balances measured and unmeasured confounders only "in expectation." The number of groups is typically small so that randomization may leave substantial imbalances between the treated and control groups. It is thus desirable to adjust for measured covariates $\mathbf{X}$ that are thought to affect the outcome.

We can incorporate adjustments for covariates $\mathbf{X}$ into our method following the approach of Rosenbaum (2002) and Small et al. (2008) and regress the adjusted responses under H_0 on $\mathbf{X}$ (using a common regression across both treatment groups and control groups) and carry out our test using the residuals. Under H_0, these residuals will always be the same no matter what the randomization is, so that this is a valid randomization test. The residuals may be much more stable and less dispersed because much of the variation may be captured by the covariates $\mathbf{X}$. The imbalance between the treated and control groups may also be less after adjusting for covariates $\mathbf{X}$. Therefore, our analysis based on the residuals may give more powerful inference, assuming the regression model used is correct. Note that the analysis provides valid inferences regardless of whether the regression model used is correct.

4. Simulation Results

In the simulation, we consider the following testing problem with level 0.05.

$$H_0 : \tau = 0 \quad \text{versus} \quad H_1 : \tau = -2. \tag{19}$$

We compare the powers of tests based on $U_{\mathbf{w}^*}$ to those of the LMM tests and the W tests. We consider cases when the LMM is true with low intracluster correlations, cases when the LMM is true with high intracluster correlations, and cases when the linear model is true but the effects do not have normal distributions.

We also consider various cases of cluster sizes. In the "Actual" case, the cluster sizes are those from the actual Prospect Study, which are given in Table 1; in the "Double" case, the cluster sizes are twice as large as the cluster sizes from the Prospect Study; in the "Half" case, the cluster sizes are obtained by dividing the cluster sizes from the Prospect Study by two and then rounding up to the nearest integer; in the "Equal in pair" case, the sizes of the two clusters in each pair are the rounded mean of the two cluster sizes in that pair from the Prospect Study; in the "Extreme" case, the sizes of the clusters in the first five pairs are 10's; and the sizes of the clusters in the last five pairs are 100's.

For each case, we simulate 10,000 studies. For the LMM tests with $S = 10$ paired clusters, we consider two approaches. The first and usual approach is to compare the Wald-test statistic to the t-distribution with 9 degrees of freedom, as suggested by Feng et al. (2001, p. 175) and used in the Prospect Study. We also consider the LMM test with the correction proposed by Kenward and Roger (1997). Details and further discussion of the Kenward–Roger (KR) correction can be found in McCulloch and Searle (2001), Alnosaier (2007, unpublished PhD thesis, Oregon State University), and Rencher and Schaalje (2008). For the randomization tests based on $U_{\mathbf{w}^*}$, the weights are estimated for each simulated dataset. We compute the p-value for each study and then compute the proportion of total rejections to find the power. The results are listed in Tables 2 and 3. In these tables, the usual LMM test and the LMM test with KR correction are denoted by "LMM" and "LMM_{KR}," respectively.

In Table 2, we check the level of the tests to make sure that they are around the nominal 0.05. In the first three rows, we see that $U_{\mathbf{w}^*}$ and W tests are always valid, even when the underlying distributions are not normal. This is because the tests based on $U_{\mathbf{w}^*}$ and W are exact, randomization inference tests so that they always have the correct level. However, the level of the usual LMM Wald tests can be too high under certain circumstances. This can happen when the normality assumption of the LMM is violated. The fourth row provides such an example when the individual errors are from a scaled double-exponential (Laplace) distribution. The type I error level of the usual LMM test is 0.055 in this case. Note that the standard error of the mean of 10,000 Bernoulli observations with the probability of success 0.05 is 0.002. Therefore, the level of the usual LMM test is statistically significantly higher than the nominal 0.05. The level of the usual LMM tests can be higher than the nominal even if the LMM model is true but the cluster sizes vary substantially, because the usual LMM Wald test is not an exact test, as described in Kenward and Roger (1997). The fifth row is an example. When the cluster sizes are extreme, the usual LMM test's level is 0.059 and is statistically significantly higher than 0.05. Thus, the usual LMM test provides less accurate inference in these settings. However, the LMM test with KR correction is more conservative and has the correct level when the LMM is true and the cluster sizes are unbalanced. The rank-based tests W and $U_{\mathbf{w}^*}$, on the other hand, always have the correct level.

Table 2

Levels of the randomization test $U_{\mathbf{w}^*}$ versus those of the usual LMM test, those of the LMM test with KR correction, and those of the randomization test W for several intracluster correlations. The true treatment effect is $\tau = 0$.

Cluster effects	Cluster sizes	Individual errors	λ	LMM level	LMM_{KR} level	W level	$U_{\mathbf{w}^*}$ level
$N(0,0.51)$	Actual	$N(0,12.25)$	0.04	0.050	0.034	0.049	0.050
$N(0,1.67)$	Actual	Cauchy	–	0.015	0.007	0.051	0.051
Cauchy	Actual	$N(0,12.25)$	–	0.030	0.029	0.049	0.050
$N(0,1.67)$	Extreme	Laplace(0, 2.47)	0.12	0.055	0.049	0.051	0.049
$N(0,0.51)$	Extreme	$N(0,12.25)$	0.04	0.059	0.042	0.048	0.047

Table 3

Powers of the randomization test $U_{\mathbf{w}^*}$ versus those of the usual LMM test, those of the LMM test with KR correction, and those of the randomization test W for several intracluster correlations. The true treatment effect is $\tau = -2$.

Cluster effects	Cluster sizes	Individual errors	λ	LMM power	LMM_{KR} power	W power	$U_{\mathbf{w}^*}$ power
N(0,0.51)	Actual	N(0,12.25)	0.04	0.978	0.973	0.966	0.970
N(0,0.51)	Double	N(0,12.25)	0.04	0.995	0.994	0.990	0.994
N(0,0.51)	Half	N(0,12.25)	0.04	0.912	0.889	0.888	0.888
N(0,0.51)	Equal in pair	N(0,12.25)	0.04	0.991	0.991	0.987	0.988
N(0,0.51)	Extreme	N(0,12.25)	0.04	0.994	0.993	0.979	0.989
N(0,1.67)	Actual	N(0,12.25)	0.12	0.831	0.820	0.779	0.811
N(0,1.67)	Double	N(0,12.25)	0.12	0.875	0.871	0.818	0.868
N(0,1.67)	Half	N(0,12.25)	0.12	0.752	0.732	0.706	0.724
N(0,1.67)	Equal in pair	N(0,12.25)	0.12	0.872	0.871	0.847	0.860
N(0,1.67)	Extreme	N(0,12.25)	0.12	0.865	0.853	0.730	0.849
N(0,4.08)	Actual	N(0,12.25)	0.25	0.584	0.574	0.510	0.565
N(0,4.08)	Double	N(0,12.25)	0.25	0.596	0.591	0.522	0.588
N(0,4.08)	Half	N(0,12.25)	0.25	0.539	0.526	0.480	0.518
N(0,4.08)	Equal in pair	N(0,12.25)	0.25	0.610	0.609	0.577	0.599
N(0,4.08)	Extreme	N(0,12.25)	0.25	0.601	0.594	0.444	0.589
N(0,1.67)	Actual	$2.47t_4$	0.12	0.832	0.821	0.799	0.844
N(0,1.67)	Half	$2.47t_4$	0.12	0.758	0.736	0.741	0.773
N(0,1.67)	Equal in pair	$2.47t_4$	0.12	0.871	0.869	0.867	0.883
N(0,1.67)	Actual	Laplace(0, 2.47)	0.12	0.824	0.814	0.793	0.835
N(0,1.67)	Half	Laplace(0, 2.47)	0.12	0.756	0.734	0.753	0.778
N(0,1.67)	Equal in pair	Laplace(0, 2.47)	0.12	0.866	0.865	0.858	0.878
N(0,1.67)	Actual	t_2	–	0.846	0.837	0.824	0.888
N(0,1.67)	Half	t_2	–	0.812	0.799	0.816	0.867
N(0,1.67)	Equal in pair	t_2	–	0.880	0.878	0.886	0.912
N(0,1.67)	Actual	Cauchy	–	0.162	0.116	0.812	0.820
N(0,1.67)	Half	Cauchy	–	0.164	0.118	0.783	0.792
N(0,1.67)	Equal in pair	Cauchy	–	0.196	0.173	0.875	0.882
Laplace(0, 0.91)	Actual	N(0,12.25)	0.12	0.830	0.820	0.789	0.816
$0.91t_4$	Actual	N(0,12.25)	0.12	0.843	0.832	0.802	0.829
t_2	Actual	N(0,12.25)	–	0.587	0.580	0.586	0.611
Cauchy	Actual	N(0,12.25)	–	0.252	0.249	0.349	0.368

In Table 3, we compare the powers of the tests. The first five rows show that when the LMM is true and the intracluster correlation λ is relatively small, $\lambda = 0.04$, the powers of the $U_{\mathbf{w}^*}$ tests, both LMM tests and the W tests are high and close for all types of cluster sizes. We should note that because the LMM is the correct model to use in this situation, the LMM tests should have the maximum power. We should also note that because the LMM tests with KR correction are more conservative than the usual LMM tests in these cases to provide correct levels, their powers are less than those of the usual LMM tests.

As λ grows to 0.12, we see from the next five rows that the powers of the W tests become much lower than those of both LMM tests. However, the powers from the tests based on $U_{\mathbf{w}^*}$ remain close to the LMM tests for all different types of cluster sizes.

This phenomenon is more apparent in the next five rows when $\lambda = 0.25$. The differences between the powers of the LMM tests and the powers of the W tests become even larger in all cases. In particular, in the "Extreme" case, the power of the W test falls below those of the LMM tests by about 0.15, which is a substantial loss of power. Nevertheless, the powers of the $U_{\mathbf{w}^*}$ tests remain very close to the powers of both LMM tests.

We then study the powers when the LMM is not true, i.e., we checked the situations when the individual errors or the cluster effects are not normally distributed. The alternative distributions we consider are several heavy-tailed distributions—the Cauchy distribution, the t_2 distribution, and the scaled t_4 distribution—and a light-tailed distribution, the scaled Laplace distribution. For distributions with a finite variance like the t_4 and Laplace, we scaled them so that the intracluster correlations of these situations are 0.12, making the powers in these cases comparable to the LMM situations with the same intracluster correlation.

When the distributions of individual errors are scaled t_4 or scaled Laplace which have a finite variance, the LMM tests perform similarly as in the case when the LMM is true and the intracluster correlations are the same. However, we see that the powers of the $U_{\mathbf{w}^*}$ tests become higher than those of the corresponding LMM tests. The power gain for $U_{\mathbf{w}^*}$ compared to the LMM is larger when the distribution is t_2 which does not have a finite variance. Note that the standard error of the mean of 10,000 Bernoulli observations is at most 0.005

when the probability of success is 0.5. Therefore, for the above three distributions and cases of cluster sizes, the powers of $U_{\mathbf{w}^*}$ tests are statistically significantly higher than those of the LMM tests. In the extreme, when the individual effects are Cauchy distributed and do not have a finite expectation, the powers of both LMM tests drop dramatically to below 0.2. The drop is more severe for the LMM test with KR correction because this test is based on many approximations which rely on the existence of the first several moments of the underlying distribution. However, the tests based on rank statistics $U_{\mathbf{w}^*}$ and W remain very powerful.

We study the cases when the cluster effects are not normal too. We see that when the distribution of cluster effects is a scaled t_4 distribution or a scaled Laplace distribution that has a finite variance, the LMM tests remain powerful. However, when the distribution is t_2, the power of the LMM tests are statistically significantly lower than that of the $U_{\mathbf{w}^*}$ test. The $U_{\mathbf{w}^*}$ tests and the W tests are also more powerful in the case when the cluster effects are Cauchy distributed, as shown in the last row.

In short, results from simulations show that when the LMM is true, $U_{\mathbf{w}^*}$ performs similarly to the LMM tests and has higher power than W when λ is high; on the other hand, when the LMM is not true, $U_{\mathbf{w}^*}$ has higher power than the LMM tests. We include more simulation results in Web Appendix C.

We also consider robust procedures in estimating $\widehat{\sigma}^2_\gamma$ and $\widehat{\sigma}^2_\zeta$. We utilize the robust estimators in Stahel and Welsh (1997), and the simulated powers are very close to the ones utilizing the method of moments estimates. Therefore, we will work with the ANOVA estimates.

We should also note here that no matter what distribution the effects and errors follow, the powers in the "Equal in pair" case is always higher than those in the "Actual" case. This phenomenon of the randomization test being adversely affected by imbalance is addressed in Gail et al. (1996).

5. Application to the Prospect Study

For the Prospect Study, we consider the one-sided test (3) of whether the intervention has the beneficial effect of reducing depression. Recall that Falissard et al. (2003) consider a difference of 2.7 or lower on the Hamilton depression score between treatment groups to be clinically meaningful. Therefore, we maximize the power for the testing problem: H_0 : $\tau = 0$ versus $H_1 : \tau = -2.7$.

The power of this testing problem with the test statistic $U_{\mathbf{w}^*}$ is maximized at

$$\mathbf{w}^* = \{0.2326, 0.0496, 0.0415, 0.0093, 0.1365, \\ 0.0436, 0.1064, 0.1417, 0.1060, 0.1328\}.$$

The $U_{\mathbf{w}^*}$ with this set of weights is found to be -0.227. Out of the 2^{10} possible random assignments, under $H_0 : \tau = 0$, 8 assignments yield a value of $U_{\mathbf{w}^*}$ that are less than or equal to -0.227. Hence, the p-value of this test is $8/1024 = 0.0078$. By inverting the test, we find a 95% one-sided confidence interval to be $(-\infty, -0.67]$. There is strong evidence that the intervention reduces depression, but not that the reduction is greater than -2.7. We should note that the ANOVA estimate of the intracluster correlation is 0 for these data.

We note that the set of weights $\mathbf{w}_W = \{w_{Q_s}\}$ of Q_s's in the W test (normalized so that $\sum_s w_{Q_s} = 1$) is numerically the same as $\mathbf{w}^*$ up to 4 decimal accuracy. This coincidence is because the set of weights $\mathbf{w}_W$ are found under the assumption that the intracluster correlation is zero or small. The W test is the most powerful test when this assumption is true. On the other hand, the set of weights $\mathbf{w}^*$ are found without this assumption and yields the most powerful test for any intracluster correlation. In the case of the Prospect Study data, the ANOVA estimate of the intracluster correlation happens to be numerically 0. Thus the coincidence of $\mathbf{w}^*$ and $\mathbf{w}_W$ is not surprising.

As another comparison, the usual LMM test gives a p-value of 0.002. By inverting the test, a 95% one-sided confidence interval is found to be $(-\infty, -1.80]$. Also, the LMM test with KR correction gives a p-value of 0.003 and a 95% one-sided confidence interval $(-\infty, -1.71]$. These analyses may yield less accurate inference when the underlying distribution is not normal.

We also test the hypothesis with adjustment for age and suicidal ideation. This is done by regressing the decline in Hamilton scores on age and suicidal ideation using M estimation to get residuals and applying the LMM tests, the W test and the $U_{\mathbf{w}^*}$ test to the residuals. The results are almost identical to the analysis without covariate adjustment.

6. Conclusion

In making randomization inference in group-randomized trials, both the LMM tests and the W tests have their own advantageous and disadvantageous situations. In this article, we look for a statistic that can be regarded as a compromise between the LMM tests and the W tests so that it possesses the positive features of both. Therefore, we construct a rank-based statistic $U_{\mathbf{w}^*}$ that is optimally weighted under the normal assumption, yet has the exact stated nominal level regardless of whether the normal assumption is true. This new statistic $U_{\mathbf{w}^*}$ incorporates the intracluster correlation by working with several important probabilities related to comparisons between treated outcomes and control outcomes. The test based on $U_{\mathbf{w}^*}$ is shown to be almost as powerful as the LMM tests when the LMM is true and the intracluster correlation is high, yet it is robust against nonnormal underlying distributions. It would be of interest for future study to consider the properties of $U_{\mathbf{w}^*}, W$, and the LMM tests in the local alternative setting.

7. Supplementary Materials

Web Appendices referenced in Sections 3 and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`. Software of the $U_{\mathbf{w}^*}$ test can be downloaded at `stat.wharton.upenn.edu/~zhangk/GRI.zip`.

Acknowledgements

The authors thank Professor Lawrence Brown and Professor Paul Rosenbaum for their warm and insightful suggestions. The authors also thank the editors and referees for careful and helpful comments and suggestions.

References

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. New York: Cambridge University Press.

Braun, T. M. and Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *Jornal of the American Statistical Association* **96,** 1424–1432.

Brookmeyer, R. and Chen, Y. Q. (1998). Person-time analysis of paired community intervention trials when the number of communities is small. *Statistics in Medicine* **17,** 2121–2132.

Bruce, M. L., Ten Have, T. R., Reynolds, C. F., III, Katz, I. I., Schulberg, H. C., Mulsant, B. H., Brown, G. K., Pearson, J. L., and Alexopoilos, G. S. (2004). Reducing suicidal ideation and depressive symptoms in depressed older preimary care patients: A randomized trial. *Journal of the American Medical Association* **291,** 1081–1091.

Cornfield, J. (1978). Randomization by group. *American Journal of Epidemiology* **108,** 100–102.

Cox, D. R. and Reid, N. (2000). *The Theory of the Design of Experiments*. New York: CRC Press.

Donner, A. (1998). Some aspects of the design and analysis of cluster randomized trials. *Applied Statistics* **47,** 95–113.

Falissard, B., Lukasiewicz, M., and Corruble, E. (2003). The MDP75: A new approach in the determination of the minimal clinically meaningful difference in a scale or a questionnaire. *Journal of Clinical Epidemiology* **56,** 618–621.

Feng, Z., Diehr, P., Peterson, D., and McLerran, D. (2001). Selected statistical issues in group randomized trials. *Annual Review of Public Health* **22,** 167–187.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh, U.K.: Oliver & Boyd.

Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2002). Clustered encouragement designs with individual level noncompliance. *Biostatistics* **3,** 147–177.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15,** 1069–1092.

Imai, K., King, G., and Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science* **24,** 29–53.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.

Kenward, M. and Roger, J. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53,** 983–997.

Lehmann, E. L. (1998). *Nonparametrics*. Upper Saddle River, New Jersey: Prentice-Hall.

Lewis, D. (1973a). Causation. *The Journal of Philosophy* **70,** 556–567.

Lewis, D. (1973b). *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Murray, D. (1998). *Design and Analysis of Group Randomized Trials*. New York: Oxford University Press.

Murray, D., Hannan, P. J., Pals, S. P., McCowen, R. G., Baker, W. L., and Blitstein, J. L. (2006). A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Statistics in Medicine* **25,** 375–388.

Neyman, J. (1923). On the application of probability theory to agricultural experiments, reprinted in *Statistical Science*, 1990, **5,** 463–480.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society, Series B* **4,** 119–130.

Rencher, A. and Schaalje, G. (2008). *Linear Models in Statistics*, 2nd edition. New York: Wiley.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7,** 1393–1512.

Robins, J. M. (1987). Addendum to a new approach to causal inference in mortality studies with sustained exposure period—application to control of the healthy worker survivor effect. *Computers and Mathematics with Applications* **14,** 923–945.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17,** 286–327.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons.

Small, D. S., Ten Have, T. R., and Rosenbaum, P. R. (2008). Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. *Journal of the American Statistical Association* **103,** 271–279.

Stahel, W. A. and Welsh, A. (1997). Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference* **57,** 295–319.

Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29,** 21–52.

Zhang, K. and Small, D. (2009). Comment: The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science* **24,** 59–64.

Received May 2010. *Revised April* 2011.
Accepted April 2011.

Appendix

Finding Optimal Weights

To find weights w_s, $s = 1, \ldots, S$, that correspond to the most powerful test, we solve the following optimization problem

$$\mathbf{w}^* = \arg\max_{\mathbf{w}\in\Delta_S} \frac{z_\alpha \sqrt{\sum_{s=1}^{S} w_s^2 V_{s0}} - E}{\sqrt{\sum_{s=1}^{S} w_s^2 V_{s1}}}, \tag{A.1}$$

where Δ_S is an S-dimensional simplex, $E < 0$ is a constant, and $z_\alpha < 0$ is an α-quantile ($\alpha < 0.5$) of a standard normal distribution. As the following Proposition 3 shows, this optimization problem may have several local maxima.

Proposition 3. *Optimization problem A.*1 *may have several local maxima.*

Proof. We shall reduce a multidimensional optimization problem to a one-dimensional problem that may have two local maxima. This will imply that the original problem may have two local maxima also.

Let $\mathbf{o}$ and $\mathbf{e}$ be any two points in Δ_S. Then a line going through these two points is given by $\mathbf{w}(t) = \mathbf{o} + t\mathbf{d}$, where $\mathbf{d} = \mathbf{e} - \mathbf{o}$ and $t \in \mathbb{R}$, and

$$f(t) = f(\mathbf{w}(t)) = \frac{z_\alpha \sqrt{\mathbf{o}^T \mathbf{V}_0 \mathbf{o} + 2t\mathbf{o}^T \mathbf{V}_0 \mathbf{d} + t^2 \mathbf{d}^T \mathbf{V}_0 \mathbf{d}} - E}{\sqrt{\mathbf{o}^T \mathbf{V}_1 \mathbf{o} + 2t\mathbf{o}^T \mathbf{V}_1 \mathbf{d} + t^2 \mathbf{d}^T \mathbf{V}_1 \mathbf{d}}},$$

where with a slight abuse of the notation $\mathbf{V}_h$, $h = 0, 1$ stands for a diagonal matrix with elements $\mathbf{V}_{sh}$, $s = 1, \ldots, S$ on the diagonal. Because $\mathbf{V}_0$ and $\mathbf{V}_1$ are positive definite, then expressions under square roots above are always positive. We can compactly rewrite $f(t)$ as

$$f(t) = \frac{z_\alpha \sqrt{at^2 + bt + c} - E}{\sqrt{dt^2 + et + f}}.$$

By taking a derivative of $f(t)$, we conclude that this function of scalar argument may have two local maxima (See Web Appendix D for an example). Therefore, expression (A.1) may also have more than one local maximum. In addition, because we are optimizing over simplex, the maximum may be achieved on the boundary. □

From an algorithmic perspective, the easiest way to deal with these issues of having more than one maximum is to use multiple starting points for a gradient-ascent projection algorithm. However, there is one special case discussed below, where we have simple guarantees of finding a maximum.

PROPOSITION 4. *If there exists a point $\tilde{w}$, where numerator of (A.1) is positive, then gradient-ascent algorithm can be modified to be guaranteed to find a global maximum.*

Proof. Here we show that over the subset of the simplex where the function we are optimizing is positive, this function is quasiconcave (see Definition 1). Because quasiconcave functions are unimodal, any gradient ascent algorithm will eventually (approximately) find the maximum. □

To proceed, we need the following definition (Boyd and Vandenberghe, 2004, Section 3.4).

DEFINITION 1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called quasiconcave if its domain and all its superlevel sets*

$$S_\alpha = \{\mathbf{x} \in \mathrm{dom} f \mid f(\mathbf{x}) \geq \alpha\}$$

for $\alpha \in \mathbb{R}$ are convex.

Quasiconcave functions are unimodal.

Notice that $\sqrt{\sum_{s=1}^S w_s^2 V_s}$ is a convex function of $\mathbf{w}$ for $\mathbf{V} = (V_s) \succ 0$ (which is exactly our case); hence, we can represent the function that we are maximizing as

$$f(\mathbf{w}) = \frac{\phi(\mathbf{w})}{\psi(\mathbf{w})},$$

where $\phi(\cdot)$ is concave and $\psi(\cdot)$ is convex and positive. By proposition's assumption $\mathcal{P} = \{\mathbf{w} \in \Delta_S \mid \phi(\mathbf{w}) > 0\} \neq \emptyset$. Then for sufficiently small $t > 0$, we have that

$$\phi(\mathbf{w}) - t\psi(\mathbf{w}) \geqslant 0$$

defines a nonempty convex subset of $\mathcal{P}$. Moreover, for any fixed $\mathbf{w}$ function $g(t) = \phi(\mathbf{w}) - t\psi(\mathbf{w})$ is nonincreasing in t; therefore, $f(\mathbf{w})$ is quasiconcave over $\mathcal{P}$ and its maximum over set $\mathcal{P}$ can be found with any gradient ascent algorithm. Notice that for $\mathbf{w} \notin \mathcal{P}$, we have $f(\mathbf{w}) \leqslant 0$; therefore, maximization over $\mathcal{P}$ yields maximum over Δ_S. □

It is very easy to check whether the positivity condition is satisfied: for a given vector $\mathbf{V}_0 = (V_{s0})$, one only has to check if $\phi(\mathbf{w}_0) > 0$, where $\mathbf{w}_0 = (w_{s0})$

$$w_{s0} = \frac{1}{\left(\sum_{s=1}^S \frac{1}{V_{s0}}\right) V_{s0}}$$

is a point of maximum of $\phi(\cdot)$ over Δ_S.