The Annals of Applied Statistics
2018, Vol. 12, No. 4, 2151–2174
https://doi.org/10.1214/18-AOAS1147

RANK TESTS IN UNMATCHED CLUSTERED RANDOMIZED TRIALS APPLIED TO A STUDY OF TEACHER TRAINING

BY PENG DING[1] AND LUKE KEELE

University of California, Berkeley and Georgetown University

In the Teacher and Leader Performance Evaluation Systems study, schools were randomly assigned to receive new measures of teacher and principal performance. One outcome in the study, measured at the teacher level, was truncated at zero, and displayed a long tail. Rank-based statistics are one natural method to apply to such outcomes, since inferences will be robust and exact, and we can avoid assumptions about the model that generated the data. We investigate four different possible rank statistics that vary in the form of weighting applied to clusters. Each test statistic has the correct level but may vary in terms of the power to detect departures from the null. We conduct simulations for power comparing to linear mixed models with Normal, t, and Cauchy errors. We obtain a point estimate and construct confidence intervals by applying the Tobit model of effects, which assumes that treatment increases the outcome by a constant amount but only if the response under control would be positive. We also develop a formal randomization-based method for testing the appropriateness of the Tobit model of effects. In the data from the study, we find no evidence against the Tobit model of effects.

1. Introduction.

1.1. *Clustered experimental designs.* In many contexts, it is useful to apply treatments to groups of individuals rather than to individual subjects. For example, a clustered randomized trial was used to evaluate Success For All, a reading intervention for elementary schools students [Borman et al. (2005)]. In this study, 21 intact schools were assigned to the Success For All reading curriculum, while 20 schools were assigned to the control condition. Alternatively, clustered experimental designs are used to examine the effects at the level of medical practices. In the POST trial, 52 general practices were randomized to study the effect of mailing letters to patients to improve follow up care for coronary heart disease [Feder et al. (1999)]. One advantage of clustered treatment assignment is that treatments can spillover to units within the same cluster and not bias the estimates of treatment effects [Imbens and Wooldridge (2008)]. However, clustered treatment assignment tends to reduce efficiency compared to assigning treatments at the individual level.

Received May 2017; revised October 2017.

[1]Supported by the Institute for Education Science (IES) Grant R305D150040 and NSF Grant DMS-1713152).

Key words and phrases. Fisher randomization test, clustered randomization, model checking, rank statistic, Tobit model.

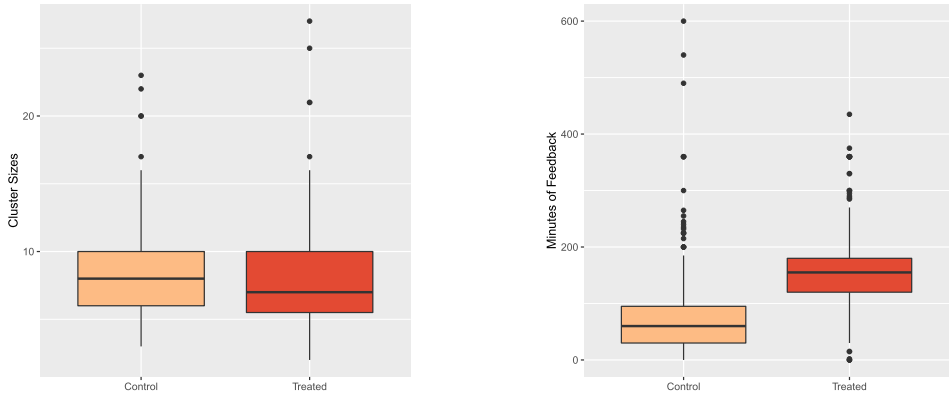
This will be especially true when units in the same cluster tend to have the same response for reasons unrelated to treatment [Cornfield (1978)].

Clustered randomized trials are commonly analyzed using the linear mixed model (LMM) assuming additive effects and cluster-level random effects [Song and Ahn (2002), Murnane and Willett (2010), Hayes and Moulton (2009)]. Critically, the LMM assumes the data are generated from Normal distributions. Tests from LMMs may have a substantial loss of power when the data are from heavy-tailed distributions. Moreover, the inferential properties of the LMM also depend on modeling assumptions. These modeling assumptions and asymptotic approximations may be particularly suspect since many clustered randomized trials have sample sizes of less than 50–75 clusters.

Alternatively, Rosner, Glynn and Lee (2003, 2006), Datta and Satten (2005) and Dutta and Datta (2016) have used rank-based statistics to test the null hypothesis of zero treatment effect in settings with clustering. In order to derive asymptotic distributions for these rank-based tests, they restricted the form of the test statistics, and they must also implicitly assume that units' outcomes have the same marginal distribution under the null hypothesis regardless of whether the units are in the same cluster or not. Rank statistics of this type are more resistant to violations of parametric modeling assumptions and heavy-tailed outcomes, but they still require assumptions about the data generating process and a large number of clusters for asymptotic approximations. We seek to develop rank-based tests under fewer assumptions. Before introducing these new methods, we describe the Teacher and Leader Performance Evaluation Systems (TLPES) study, a clustered randomized trial which motivates the methods we develop.

1.2. *The TLPES intervention*. Recent research has highlighted the importance of high quality teaching [Chetty, Friedman and Rockoff (2014)], and the TLPES was one proposed intervention to increase teacher and principal quality. The TLPES is a clustered randomized trial designed to examine the effect of implementing new measures of teacher and principal performance. In the trial, there were 63 treated schools and 64 control schools. The number of teachers in each school ranges from 2 to 27 [Wayne et al. (2016)]. Figure 1(a) contains box plots for the number of teachers (units) within schools (clusters) by treatment status. Cluster sizes are well balanced under treatment and control with p-value $= 0.60$ from a t-test and p-value $= 0.31$ from using the Wilcoxon rank sum test.

The goal in the study was to implement higher quality measures of teachers and principal performance in hopes of improving classroom practice and principal leadership. In the trial, teachers in treated schools received four rounds of teaching observation where the new measures of teacher performance were implemented. The teaching observations were then followed by feedback sessions between the observer and the teacher. The teacher evaluations were carried out by either the principal or by a district selected observer. Both principals and district selected observers received three–four days of training on teacher evaluation methods to



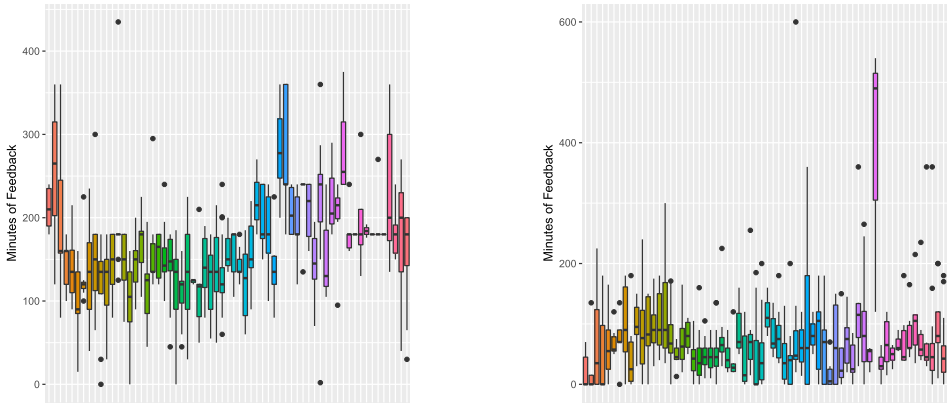
(a) Number of units by cluster (b) Feedback session lengths

FIG. 1. *Cluster sizes and outcomes for the TLPES study by treatment status.*

support the teaching observations and feedback sessions. The intention was that by clarifying expectations and providing additional feedback, teachers would spend more time on professional development and high performing teachers would be encouraged to improve classroom practice. The primary outcomes in TLPES were measures of teacher classroom practice and student achievement [Wayne et al. (2016)].

The study also measured whether the intervention changed the behavior of teachers in the treatment group. One of these measures was the length of the feedback session measured in minutes, which was collected for teachers in both the treatment and control arms of the study. Figure 1(b) contains box plots of the amount of feedback a teachers received as measured by the number of minutes in the feedback session by treatment condition collected after the first year. Figure 2 contains box plots of the outcome by school and treatment condition. Treated schools tend to have longer feedback sessions but also display greater variability compared to control schools. In the control arm, while many teachers received feedback, a large fraction of observations recorded zero minutes of feedback producing point masses at zero in the outcome measure. We observe the mass of the density at zero with other point masses at 30, 60, 90 and 180 minute intervals. This outcome measure is far from a Normally distributed outcome. In this study, we focus on the length of the feedback session as a single outcome in the analysis. If future decisions are based on additional outcomes, then corrections for multiple testing should be applied. We ignore this issue in the current paper, since we did not have access to the full set outcomes collected in the study.

1.3. *Contribution and structure of the paper.* Given the distribution of the outcome and the relatively small number of clusters, we might wish to avoid the use



(a) Treated Schools (b) Control Schools

FIG. 2. *Within school variability of the outcome, length of feedback session, by treatment condition.*

of asymptotic approximations in the analysis of the data from the TLPES intervention. In general, it may be preferable to use tests that have power under a wider range of distributions. To ensure that the level of the test is correct, we use randomization inference. Under randomization inference, we use the actual randomization to simulate the distribution of the test statistics rather than relying on assumptions about the process which generated the outcome data [Fisher (1935), Rosenbaum (2002a), Imbens and Rubin (2015)]. Rank-based tests under randomization inference are well developed for clustered randomized trials with clusters paired prior to the assignment of treatment [Small, Ten Have and Rosenbaum (2008), Zhang, Traskin and Small (2012)], but no such work exists for clustered randomized trials that are unpaired. Unlike previous robust inference for clustered randomized trials [Rosner, Glynn and Lee (2003, 2006), Datta and Satten (2005), Dutta and Datta (2016)], which dealt with only the testing problem, we are also interested in point and interval estimation of the treatment effect without imposing the LMM assumptions. In the randomization inference framework, the test of no effect can be inverted to provide distribution-free confidence intervals, and the Hodges–Lehmann method produces point estimates. See Rosenbaum (2002a), Chapter 2, for details.

Because the control outcomes have a point mass at zero, the usual constant treatment effect model [Rosenbaum (2002a), Ding, Feller and Miratrix (2016)] does not hold. Instead, we estimate the treatment effect under the Tobit model [Tobin (1958), Rosenbaum (2002a)], where the control outcomes are truncated at zero. Moreover, we propose an exact randomization test for the appropriateness of the Tobit model itself, extending previous work for testing treatment effect heterogeneity, that is, the constant treatment effect model [Ding, Feller and Miratrix (2016)]. The formal test for the goodness of fit of the Tobit model is another contribution to the literature.

The paper proceeds as follows. Section 2 introduces the potential outcomes notation for clustered randomized experiments and the basis for randomization-based test. Section 3 proposes several possible choices of test statistics in clustered randomized experiments, and discusses the exact distribution of general test statistics and the asymptotic distributions of a class of test statistics that can be represented as summation of cluster-level characteristics. Section 4 uses simulations to study the powers of the proposed test statistics. Section 5 discusses point estimation, confidence interval and model checking under the Tobit model. Section 6 applies the proposed methodology to the TLPES study, obtaining point and interval estimates of the treatment effect under a Tobit model and testing the appropriateness of the Tobit model itself. Section 7 concludes.

2. Treatments assigned to clusters.

2.1. *Notation*: *Treatment effects for units in schools*. We have C clusters, and cluster i has total number of units n_i, $i = 1, \ldots, C$. We label the units by (i, j), with i denoting the cluster number and j denoting the number of units within cluster, $i = 1, \ldots, C$ and $j = 1, \ldots, n_i$. In the TLPES study, schools are clusters, and the teachers are the units. The total number of experimental units is $N = \sum_{i=1}^{C} n_i$. Before conducting the experiment, unit (i, j) has unit-level pretreatment covariates X_{ij}, and cluster i has cluster-level pretreatment covariates W_i. In a cluster-randomized experiment, we randomly assign C_1 clusters to receive treatment and C_0 clusters to receive control; within each cluster, all units receive the same treatment level. Let Z_i be the treatment indicator for cluster i, with 1 for treatment and 0 for control. Therefore, $Z = (Z_1, \ldots, Z_C)$ is the treatment vector for clusters, and its realized value $z = (z_1, \ldots, z_C) \in \{0, 1\}^C$ must satisfy $\sum_{i=1}^{C} z_i = C_1$. A cluster-randomized experiment is a completely randomized experiment at the cluster level, that is, for all z with C_1 under treatment and C_0 under control,

$$P(Z = z) = 1 \Big/ \binom{C}{C_1}. \tag{1}$$

We use the potential outcomes framework to define causal effects. The potential outcome $Y_{ij}(z)$ is a function of the experimental unit (i, j) and the treatment, and thus is treated as fixed [Neyman (1923), Rubin (1974)]. It is often reasonable to assume that there is no interference between units across different clusters, which simplifies the potential outcome as $Y_{ij}(z_i)$. That is, since the units available for treatment are clusters and not teachers, it is unlikely that teachers in different schools interfere with each other. As such under this notation, we allow within-cluster interference but rule out between-cluster interference. Thus we rule out the possibility of interference between units under the usual stable unit treatment value assumption [Rubin (1986)]. Because all units within the same cluster receive the same treatment, we can write the potential outcomes of unit (i, j) as $Y_{ij}(1)$ and

$Y_{ij}(0)$ if cluster i receives treatment and control, respectively. The observed outcome Y_{ij} is a deterministic function of the potential outcomes and the cluster-level treatment assignment:

$$Y_{ij} = Z_i Y_{ij}(1) + (1 - Z_i) Y_{ij}(0).$$

A primary goal of the experiment is to test whether the treatment affects the outcome, that is, the following sharp null hypothesis of zero causal effects on each experimental unit:

$$H_0 : Y_{ij}(1) = Y_{ij}(0) \qquad (i = 1, \ldots, C;\ j = 1, \ldots, n_i).$$

2.2. *Exact inference in a clustered randomized trial.* Under the sharp null hypothesis, the observed outcomes $Y_{ij} = Y_{ij}(1) = Y_{ij}(0)$ are all fixed, and the only randomness of the data comes from the treatment assignment Z. Let $Y = \{Y_{ij} : i = 1, \ldots, C;\ j = 1, \ldots, n_i\}$ be the collection of observed outcomes, $n = (n_1, \ldots, n_C)$ be the vector of cluster sizes, X the collection of unit-level covariates, and W the collection of cluster-level covariates. Ideally, we can use any function of the data, $T(Z, Y, n, X, W)$, as a test statistic against the sharp null hypothesis; its null distribution is known and can be either calculated exactly or simulated by repeatedly drawing the treatment assignment Z from its distribution (1). Of course, in practice, we must also select a test statistic to measure possible differences in outcomes caused by the treatment. Below, we review and propose possible test statistics.

3. Test statistics for clustered randomized trials.

3.1. *Test statistics based on the original outcome scale.* The first class of test statistics, based on the original scale of the outcome, is often motivated by estimation of the average causal effect for all units in the clustered randomized experiment:

$$\tau = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \{Y_{ij}(1) - Y_{ij}(0)\}}{\sum_{i=1}^{C} n_i}.$$

For example, we review some intuitive estimators of τ discussed by Middleton and Aronow (2015). To facilitate this discussion, we let $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ be the total of the observed outcomes within cluster i. The first estimator is the difference between the means of the treatment and control units:

$$\widehat{\tau}_1 = \frac{\sum_{i=1}^{C} Z_i Y_i}{\sum_{i=1}^{C} Z_i n_i} - \frac{\sum_{i=1}^{C} (1 - Z_i) Y_i}{\sum_{i=1}^{C} (1 - Z_i) n_i},$$

which is unbiased only with equal cluster size, and consistent only if the number of clusters goes to infinity. With unequal cluster sizes and finite number of clusters, $\widehat{\tau}_1$

is biased and inconsistent for τ. The second estimator modifies the denominators of $\widehat{\tau}_1$, and yields unbiased estimation for τ regardless of the cluster sizes:

$$\widehat{\tau}_2 = \frac{\sum_{i=1}^{C} Z_i Y_i}{C_1 N/C} - \frac{\sum_{i=1}^{C}(1-Z_i)Y_i}{C_0 N/C}.$$

In fact, $\widehat{\tau}_2$ can be further modified by

$$\widehat{\tau}_3 = \frac{\sum_{i=1}^{C} Z_i\{Y_i - k(n_i - N/C)\}}{C_1 N/C} - \frac{\sum_{i=1}^{C}(1-Z_i)\{Y_i - k(n_i - N/C)\}}{C_0 N/C},$$

where k is some fixed number. Middleton and Aronow (2015) motivate $\widehat{\tau}_3$ as a remedy for $\widehat{\tau}_2$, because $\widehat{\tau}_2$ is not invariant to location and scale transformations of the outcome. They essentially treat n_i as a pretreatment covariate predictive of the sum of outcomes Y_i for each cluster. Adjustment for this factor reduces variability in Y_i and consequently increase estimation precision. For a predetermined k, the unbiasedness of $\widehat{\tau}_2$ is preserved by $\widehat{\tau}_3$ after adjusted for the same term for both treatment and control clusters. In practice, k is estimated using the regression coefficient of Y_i on n_i. However, under this data-dependent value of k, the unbiasedness of $\widehat{\tau}_3$ is no longer guaranteed in general.

Fortunately, for generating valid randomization tests, we do not need to worry about the unbiasedness or consistency of these estimators, because we simply treat them as candidate test statistics, that is, special cases of $T(Z, Y, n, X, W)$. For other discussions on estimation the average treatment or testing the null hypothesis based on these statistics, see Gail et al. (1992, 1996), Hansen and Bowers (2009), Schochet (2013), and Aronow, Middleton et al. (2013).

3.2. *Test statistics based on ranks.* For heavy-tailed outcome distributions, the behavior of test statistics based on the means in Section 3.1 may be driven by outliers. Another popular class of test statistics is based on the ranks of the outcomes. Let R_{ij} be the rank of Y_{ij} among the outcomes for all units, $R_i = \sum_{j=1}^{n_i} R_{ij}$ be the total rank of the units within cluster i, and $R = (R_1, \ldots, R_C)$ be the vector of the total ranks for all clusters.

In general, we can use

$$\omega = \sum_{i=1}^{C} Z_i \psi(R_i, n_i)$$

as a test statistic where ψ is any general function of both the total ranks and the sizes of the clusters. Next, we propose intuitive rank-based test statistics as special cases of ω. Each test statistic weights cluster level information differently. The first one mimics the classical Wilcoxon rank sum statistic:

$$\omega_s = \sum_{i=1}^{C} Z_i R_i,$$

with $\psi(R_i, n_i) = R_i$ depending only on the summed total of the treated ranks. The second test statistic is based on average rank of each cluster:

$$\omega_a = \sum_{i=1}^{C} Z_i R_i / n_i,$$

with $\psi(R_i, n_i) = R_i/n_i$. The third one uses n_i as a weight for the total rank R_i:

$$\omega_m = \sum_{i=1}^{C} Z_i R_i n_i,$$

with $\psi(R_i, n_i) = R_i n_i$. The fourth one adjusts the total sum of the cluster ranks by cluster sizes:

$$\omega_n = \sum_{i=1}^{C} Z_i \{R_i - k(n_i - N/C)\},$$

where k can be the regression coefficient of R_i on n_i. Similar to the construction of $\widehat{\tau}_3$, we treat n_i as a cluster-level pretreatment covariate and use it to reduce the the variability of the cluster total rank.

Although all of the above test statistics are useful candidates for randomization tests, the power of each test statistic will largely depend on the generating model of the outcomes under alternative hypotheses. As such, the power of each test statistic may differ depending on how treatment effects possible vary with cluster size. For example, if the treatment effects for different clusters increase with cluster size, then over-weighting R_i by n_i, that is, using ω_m, will be more likely to yield increased power than if ω_s is used. Alternatively, if the treatment effects for different clusters decrease with cluster sizes, then down-weighting R_i by $1/n_i$, that is, using ω_a, will tend to yield larger power than using ω_s. We explore the these possibilities using simulations in Section 4.

3.3. *Model-assisted test statistics*. The test statistics above may ignore important covariate information about the sub-units and clusters. Such covariate information is easily incorporated in using regression models. Ignoring the clustering of units, one commonly used model is the linear model

$$Y_{ij} = a + \tau Z_i + \beta' W_i + \gamma' X_{ij} + \varepsilon_{ij},$$

where the ε_{ij}'s are independent and identically distributed (IID) with mean zero and variance σ^2. Alternatively, a LMM may be employed. With a random effect for the clustering of units, the LMM has the following form:

$$Y_{ij} = a + \tau Z_i + \beta' W_i + \gamma' X_{ij} + c_i + \varepsilon_{ij},$$

where the cluster-level random effects c_i's are IID with mean zero and variance δ^2, the ε_{ij}'s are IID with mean zero and variance σ^2, and the c_i's and ε_{ij}'s are independent and follows joint Normality. See Middleton (2008) and Schochet (2013)

for a discussion of the properties and especially the disadvantages of regression analysis for estimating the average causal effects in clustered randomized trials.

The inferential advantages of randomization tests and linear models or LMMs are not incompatible [Rosenbaum (2002b)]. In fact, we can use these models to adjust for covariates, but still preserve the exactness of randomization tests. There are at least two ways of combining regression models with randomization tests. The first strategy is straightforward: we simply choose $\widehat{\tau}$, the fitted parameter of the linear model or LMM, as the test statistic. Usually, practitioners interpret the model parameter τ as the average causal effect for the finite population of interest, which may not be justified by randomization. However, we need not assume that such models hold exactly.

The second strategy exploits the fact that (Y, X, W) are all fixed under the sharp null hypothesis. We calculate the residuals, Y_{ij}^r, from

$$Y_{ij} = a + \beta' W_i + \gamma' X_{ij} + \varepsilon_{ij}$$

or

$$Y_{ij} = a + \beta' W_i + \gamma' X_{ij} + c_i + \varepsilon_{ij},$$

and treat the Y_{ij}^r's as transformed outcomes. Under the sharp null, the (Y_{ij}, W_i, X_{ij})'s are all fixed, and so are the transformed outcomes. We can use Y_{ij}^r, or its rank R_{ij}^r, to construct test statistics as in Sections 3.1 and 3.2.

Neither of the above strategies rely on modeling assumptions, and in fact we can fit more flexible models possibly with more complicated functional forms and random effects structures. As such, the rank based test statistics we proposed in the previous section can easily incorporate information from baseline covariates.

3.4. *Exact and asymptotic null distributions.* All the test statistics discussed thus far have the form $T(Z, Y, n, X, W)$, with Z being the only random component under the sharp null hypothesis. Theoretically, we can compute the randomization distribution of $T(Z, Y, n, X, W)$ by enumerating all possible values of Z according to (1); practically, we can approximate the randomization distribution by taking a simple random sample from all possible values of Z if the total number is excessively large.

Generally, we can represent the test statistics above as the sum of cluster characteristics, which allows for an asymptotic approximation when the number of clusters is large. We consider a special case with

$$T(Z, Y, n, X, W) = \sum_{i=1}^{C} Z_i \psi_i \equiv \sum_{i=1}^{C} Z_i \psi(O_i, n_i, X_i, W_i), \tag{2}$$

where ψ is a function not depending on Z, and O_i may be Y_i, R_i, Y_i^r or R_i^r. All the rank-based test statistics in Section 3.2 are within this class of test statistics with different choices of the function $\psi(\cdot)$. Define $\bar{\psi} = \sum_{i=1}^{C} \psi_i / C$ as the mean and

$S_\psi^2 = \sum_{i=1}^{C}(\psi_i - \bar{\psi})^2/(C-1)$ as the variance of the ψ_i's for all clusters. Clustered randomization trials are completely randomized trials on the cluster level, and therefore the clusters under treatment are a simple random sample of the finite clusters, and the test statistic (2) is the sample total of the ψ's. According to the standard calculations in survey sampling [Cochran (1977)], the mean and variance of (2), over all possible randomizations, are

$$\mathbb{E}\{T(Z,Y,n,X,W)\} = C_1\bar{\psi}, \qquad \mathrm{Var}\{T(Z,Y,n,X,W)\} = C_1C_0S_\psi^2/C.$$

If the number of clusters is sufficiently large, the null distribution can be well approximated by a Normal distribution with the above mean and variance due to the finite population central limit theorem [Li and Ding (2017)].

4. Simulations for type I error and power. The rank statistics we outline in Section 3.2 are used as randomization tests. The test will reject when the significance level is less than or equal to α, and randomization ensures that the test has type I error α. However, each test statistic uses a set of weights. The choice of weights does not affect whether the test has the correct type I error, but the choice of weights does affect the power of the test. Next, we conduct a simulation study to understand whether the power of the test varies across the different test statistics under different data generating processes.

In the simulation study, we compare the type I error and the power of each test statistic based on ranks from data that was generated under a LMM. Specifically, we generated the data from a LMM of the form

$$Y_{ij} = a + \tau Z_i + \beta n_i + \gamma Z_i n_i + c_i + \varepsilon_{ij}.$$

Under this model, $Z_i = 1$ if cluster i was assigned to treatment and $Z_i = 0$ if the cluster was assigned to control, so τ is a measure of the individual level treatment effect. In the model, c_i is a cluster-level random effect, β is a specific cluster-level effect that varies with the size of the cluster, and γ allows the treatment effect to vary by cluster size. When $\tau = 0$, the proportion of rejects of $H_0 : \tau = 0$ estimates the type I error of a test that should have level α. For $\tau \neq 0$, the proportion of rejections estimates power against this alternative. In the simulations, we varied the distributions of the error terms c_i and ε_{ij}. We allowed the error distributions to be one of three distributions: a Normal distribution with expectation zero; a Cauchy distribution centered at zero; or a t-distribution with five degrees of freedom.

The intraclass correlation λ is $\mathrm{Var}(c_i)/\{\mathrm{Var}(c_i) + \mathrm{Var}(\varepsilon_{ij})\}$ when c_i and ε_{ij} have finite variance. We adjusted the scale of the cluster distribution errors c_i, so that we can vary the value of λ in the simulations. With Cauchy errors, λ does not exist. Although we cannot measure the intraclass correlation by the variance ratio, we can measure it by the ratio of the scale parameters of the error terms. For example, if $c_i \sim V_c \times$ Cauchy and $\varepsilon_{ij} \sim V_\varepsilon \times$ Cauchy, then the generalized intraclass correlation is $\lambda = V_c^2/(V_c^2 + V_\varepsilon^2)$. This definition applies to the errors with or without finite variances.

Within this framework, we conduct two sets of simulations. In the first simulation, we study how type I rate changes as the number of clusters increases. In this simulation, we fix the ICC and do not allow treatment effects to vary with cluster sizes. In the second simulation, we fix the number of clusters and study type I error and power, while varying other simulation parameters.

4.1. *Simulations of type I error with varying cluster sizes.* In the first set of simulations, we set $\gamma = 0$ and $\beta = 0$, so that treatment effects do not vary with the cluster sizes. Here, we set $\lambda = 0.15$. We varied the number of clusters using the following sequence: 8, 10, 20, 30, 50, 80, 100, and 200. We repeated each simulation 1000 times. We allowed the number of units per cluster to vary stochastically in the simulations. For each of the C clusters, we drew from a uniform distribution on the interval of 10 to 75. This implies that the number of units per cluster could be between 10 and 75 with equal probability. This mimics clustered randomized experiments like TLPES where the clusters are schools and there can be considerable variability in cluster sizes. We also included the LMM in all the simulations to compare the randomization tests against a more commonly-used model-based approach. For the LMM, we use the t-distribution as the reference distribution for calculating p-values. While the LMM might perform well, especially when the errors are Normal, its inferential performance does not depend on random assignment and may be wrong if the model is misspecified. We should also note that the LMM tests the hypothesis that the treatment is zero on average, unlike the rank–based tests which test the sharp null. This difference in hypotheses may increase the power of tests from the LMM slightly [Ding (2017)]. In this simulation, we used the Normal approximations for the randomization-based tests to decrease the computing time required for the simulations. Thus, the results for these tests are not exact for small sample sizes. For this set of simulations, we estimated the typical simulation to simulation variability to be 0.014.

Figure 3 contains the results as the number of clusters increases when the errors are Normal. When the number of clusters is at the minimum of eight, all the randomization-based tests reach the nominal 0.05 level, at least within simulation variability. Note that this does not represent a failure of our test statistics, but instead is a result of a dearth of information in a clustered randomized experiment with such a small number of units. For example, if 4 of the 8 clusters are assigned to treatment, then the minimum p-value that we can obtain from exact randomization test is $1/\binom{8}{4} = 0.014$. This minimum p-value is close to the 0.05 threshold.

When we use the LMM, the rejection rates do not fall within expected levels of simulation variation until the number of clusters is 30. This is not surprising. The LMM depends on an asymptotic approximation and requires a larger number of clusters for the rejection rates to converge to the correct rate. When the errors are t-distributed, results for all methods mirror those under Normal errors. Results based on the LMM converge to the nominal 0.05 level as the number of clusters increase. Due to this similarity, we omit the figure for t-distributed errors.

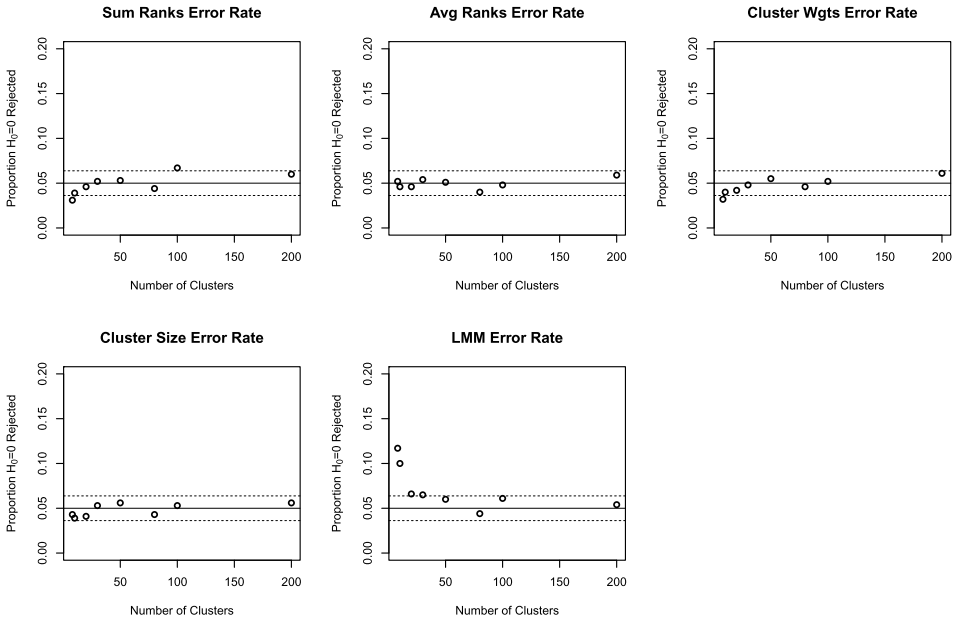


FIG. 3. *Type I error for rank based randomization tests versus linear mixed model (LMM) as the number of clusters increases with Normal errors. Dotted lines represent within simulation variability.*

Figure 4 contains the results when the errors are Cauchy. All the rank-based test statistics have nearly identical performance. Rejection rates are lower than 0.05 when there are less than 10 clusters in the clustered randomized experiments, but for larger sample sizes the level is close to the nominal 0.05 level for all randomization based tests. The LMM, however, does not reach the 0.05 level for any cluster size, even when the cluster sizes are greater than 100. In additional simulations, we increased the number of clusters to 500, 750 and 1000. Even for these larger sample sizes, the LMM did not reach the 0.05 level.

4.2. *Simulations of type I error and power with treatment effect heterogeneity.* In the second set of simulations, we fixed the number of clusters at 30, but varied a much wider set of simulation parameters. Like in the first simulation, we varied the error distributions, again using Normal, Cauchy and t-distributed errors. We also varied the intraclass correlation and whether the treatment effect varies with cluster size. In this simulation, we set $\gamma = -0.01$, 0, or 0.01 with $\beta = 0.01$. This allows the treatment effect to have a negative, positive or zero association with the size of each cluster. We also varied λ to be consistent with values from clustered randomized experiments in education. Hedges and Hedberg (2007) report the range of estimated intraclass correlation from 41 clustered randomized experiments in education. They find that the λ's range from 0.07 to 0.31, with an average value of 0.17. Note that these values of λ are large relative to those typical in public health

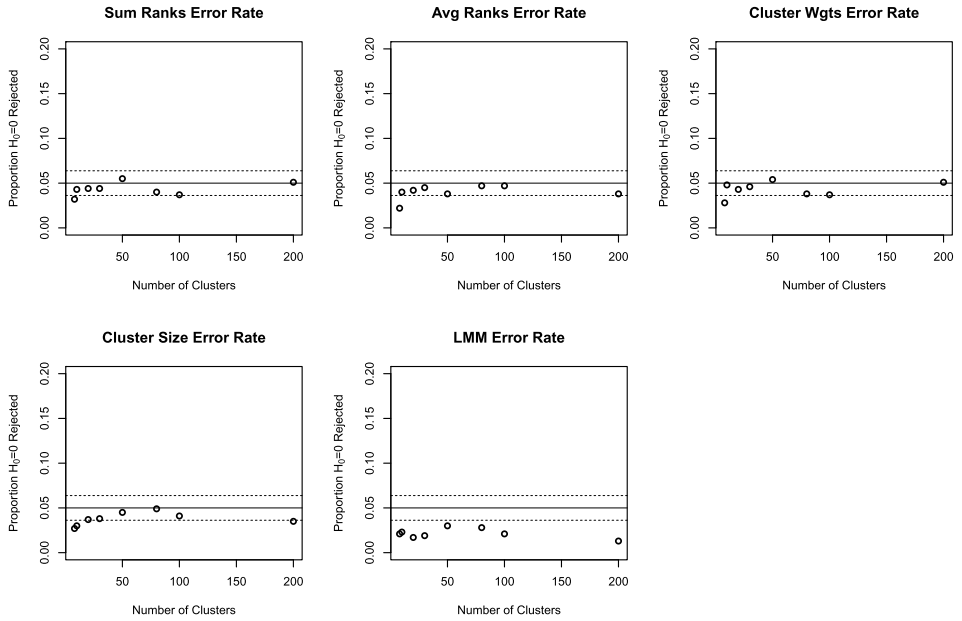


FIG. 4. *Type I error for rank based randomization tests versus linear mixed model (LMM) as the number of clusters increases with Cauchy errors. Dotted lines represent within simulation variability.*

clustered randomized experiments. Small, Ten Have and Rosenbaum (2008) conduct a similar simulation using values of λ in the range of 0.002 to 0.03 which are more typical in public health interventions that target clusters such as hospitals, clinics or villages. As a result, in this set of simulations, we use three different values for λ: 0.05, 0.15 and 0.25.

Table 1 contains the results when $\tau = 0$. The results in this simulation mirror those when we varied the number of clusters. When errors are Normal or t-distributed, all the methods reach the nominal 0.05 level. Variation in the intraclass correlation or an association between the treatment effect and cluster size do not affect the results. When errors are Cauchy, the LMM fails to reach the nominal 0.05 level. We do note that when $\lambda = 0.05$, the LMM rejection rates are at their lowest level.

Table 2 contains the results when for when $\tau = 1$. First, a few broad patterns. The test statistics ω_s and ω_m, the sum of the ranks and the weighted average ranks respectively, perform poorly across most of these scenarios. More specifically, when there is a negative association between the treatment effect and the size of the clusters, these tests have low power. The LMM has reasonable power unless the errors are Cauchy. Here, the LMM has little power. Overall, ω_n, which adjusts directly for cluster sizes, and ω_a, which is based on average ranks, both perform well across all the simulation scenarios. In fact, their performance is nearly identical in the simulations. Overall, we find that adjusting for the sizes of the clusters

TABLE 1

Type I error rates for rank-based randomization tests and LMM based on simulation when $\tau = 0$

λ	γ	Errors	Summed Ranks	Average Ranks	Weighted Ranks	Cluster Size Adj.	LMM
0.050	−0.010	t, DF = 5	0.055	0.038	0.048	0.030	0.029
0.150	−0.010	t, DF = 5	0.055	0.066	0.053	0.060	0.064
0.250	−0.010	t, DF = 5	0.065	0.051	0.050	0.045	0.054
0.050	0.000	t, DF = 5	0.058	0.067	0.054	0.059	0.072
0.150	0.000	t, DF = 5	0.060	0.047	0.051	0.045	0.051
0.250	0.000	t, DF = 5	0.048	0.046	0.057	0.048	0.048
0.050	0.010	t, DF = 5	0.062	0.051	0.057	0.046	0.053
0.150	0.010	t, DF = 5	0.062	0.050	0.060	0.055	0.052
0.250	0.010	t, DF = 5	0.052	0.071	0.055	0.064	0.068
0.050	−0.010	Normal	0.050	0.043	0.051	0.047	0.046
0.150	−0.010	Normal	0.042	0.047	0.033	0.044	0.051
0.250	−0.010	Normal	0.046	0.043	0.055	0.046	0.049
0.050	0.000	Normal	0.045	0.055	0.044	0.055	0.058
0.150	0.000	Normal	0.060	0.065	0.057	0.048	0.068
0.250	0.000	Normal	0.041	0.051	0.041	0.049	0.054
0.050	0.010	Normal	0.055	0.063	0.063	0.051	0.065
0.150	0.010	Normal	0.059	0.049	0.050	0.047	0.051
0.250	0.010	Normal	0.047	0.040	0.051	0.032	0.046
0.050	−0.010	Cauchy	0.050	0.054	0.044	0.043	0.018
0.150	−0.010	Cauchy	0.049	0.047	0.050	0.046	0.018
0.250	−0.010	Cauchy	0.055	0.047	0.054	0.048	0.017
0.050	0.000	Cauchy	0.061	0.056	0.066	0.043	0.015
0.150	0.000	Cauchy	0.035	0.058	0.038	0.057	0.024
0.250	0.000	Cauchy	0.045	0.040	0.044	0.049	0.024
0.050	0.010	Cauchy	0.054	0.046	0.051	0.044	0.014
0.150	0.010	Cauchy	0.052	0.041	0.049	0.053	0.013
0.250	0.010	Cauchy	0.059	0.053	0.055	0.048	0.017

is important for analyzing clustered randomized trials, which conforms to previous discussions for clustered data analysis under various settings [Cochran (1977), Williamson, Datta and Satten (2003), Rosner, Glynn and Lee (2003), Datta and Satten (2005), Dutta and Datta (2016)]. However, the form of those adjustments can be quite simple in form of averaging ranks within each cluster.

In sum, in many scenarios, ranked-based methods and the LMM performed similarly. The LMM had clear deficiencies when the number of clusters was small or the error distributions were heavy tailed. Across all scenarios, a test based on ω_n or ω_a, which adjusts for the number of units within each cluster, had good performance. In general, these test statistics appeared to be unaffected by the level of λ and performed well in small sample sizes.

TABLE 2
Power for rank-based randomization tests versus LMM based on simulations when $\tau = 1$

λ	γ	Errors	Summed Ranks	Average Ranks	Weighted Ranks	Cluster Size Adj.	LMM
0.050	−0.010	t, DF = 5	0.225	0.979	0.098	0.913	0.955
0.150	−0.010	t, DF = 5	0.180	0.725	0.085	0.545	0.667
0.250	−0.010	t, DF = 5	0.145	0.474	0.080	0.318	0.440
0.050	0.000	t, DF = 5	0.615	1.000	0.276	1.000	1.000
0.150	0.000	t, DF = 5	0.502	0.993	0.247	0.983	0.985
0.250	0.000	t, DF = 5	0.407	0.908	0.196	0.863	0.884
0.050	0.010	t, DF = 5	0.910	1.000	0.526	1.000	1.000
0.150	0.010	t, DF = 5	0.797	1.000	0.451	1.000	1.000
0.250	0.010	t, DF = 5	0.716	0.999	0.419	0.998	0.994
0.050	−0.010	Normal	0.279	1.000	0.113	0.980	0.999
0.150	−0.010	Normal	0.219	0.877	0.100	0.718	0.875
0.250	−0.010	Normal	0.178	0.649	0.091	0.472	0.637
0.050	0.000	Normal	0.736	1.000	0.339	1.000	1.000
0.150	0.000	Normal	0.640	1.000	0.321	1.000	1.000
0.250	0.000	Normal	0.536	0.985	0.260	0.967	0.985
0.050	0.010	Normal	0.950	1.000	0.646	1.000	1.000
0.150	0.010	Normal	0.924	1.000	0.609	1.000	1.000
0.250	0.010	Normal	0.837	0.998	0.535	0.998	0.998
0.050	−0.010	Cauchy	0.098	0.331	0.067	0.244	0.025
0.150	−0.010	Cauchy	0.079	0.184	0.062	0.141	0.018
0.250	−0.010	Cauchy	0.068	0.142	0.054	0.116	0.019
0.050	0.000	Cauchy	0.235	0.700	0.118	0.644	0.032
0.150	0.000	Cauchy	0.164	0.411	0.108	0.387	0.026
0.250	0.000	Cauchy	0.133	0.320	0.088	0.299	0.019
0.050	0.010	Cauchy	0.398	0.940	0.184	0.914	0.044
0.150	0.010	Cauchy	0.315	0.708	0.176	0.702	0.034
0.250	0.010	Cauchy	0.280	0.557	0.170	0.541	0.034

5. Point and interval estimation under the Tobit model. Testing the hypothesis of no effect is often one part of an analysis of a clustered randomized experiment. Typically, we also want to draw inferences about the magnitude of the treatment effect. Braun and Feng (2001) use a generalized linear mixed model with permutations of model test statistics as one method for obtaining confidence intervals and point estimates for data from a clustered randomized trial. This is a valid approach, but it is no longer strictly within the randomization inference framework, since it depends on the validity of the model. Instead, we invert each of our randomization tests to obtain a confidence interval for τ. We form a Hodges–Lehmann point estimate by equating each test statistic to its null expectation and solving for the value of $\hat{\tau}$. Such methods require an assumption about how units respond to treatment. Rosenbaum (2002a) refers to this assumption as a model of effects. One common model of effects assumes the response to treatment is constant and addi-

tive: $Y_{ij}(1) = Y_{ij}(0) + \tau_0$, where τ_0 is the individual treatment effect. Since, the length of a feedback session cannot be less than zero, a more plausible model of effects is the Tobit model [Tobin (1958), Rosenbaum (2002a)]. The hypothesis of a Tobit effect asserts

$$H_0(\tau_0) : Y_{ij}(0) = \max\{Y_{ij}(1) - \tau_0, 0\} \qquad \text{for all } (i, j). \tag{3}$$

If this hypothesis is true and τ_0 is known, then the adjusted outcome has the form $\max\{Y_{ij} - \tau_0 Z_i, 0\} = Y_{ij}(0)$ and is a fixed quantity unaffected by the treatment. The Tobit model of effects is more appropriate for a response that can equal zero but cannot be negative. Under a Tobit model, the TLPES intervention raises the length of the feedback session by the same constant τ_0, but the teacher has a positive length of time in the feedback session only if the length of time in the feedback session is positive. The Tobit hypothesis may be tested by adjusting the observed responses using the Tobit model of effects, and then applying one of the test statistics outlined in Section 3.2. See Rosenbaum (2010), Chapter 2, for more details on Tobit effects. In particular, for rank-based statistics we use the following steps to test $H_0(\tau_0)$:

1. Compute adjusted outcomes under the tobit model of effects: $\max\{Y_{ij} - \tau_0 Z_i, 0\}$ for all (i, j) using the data and value for τ_0 under the null;
2. Obtain the ranks of the adjusted outcomes, denoted by R_{ij};
3. Calculate the total ranks of the units within clusters $R_i = \sum_{j=1}^{n_i} R_i$; note that the R_i's are all fixed numbers unaffected by the treatment assignment;
4. Apply one of the rank statistics defined in Section 3.2 ω;
5. Simulate the treatment assignment $(Z_1, \ldots, Z_C)$, obtain the null distribution of ω and calculate the p-value.

Under the equivalence of confidence intervals and hypothesis tests, a $1 - \alpha$ confidence interval for a Tobit effect is formed by testing a series of hypothesis for τ_0 and retaining the values not rejected at level α as the confidence interval. A point estimate of τ_0 is obtained from the tests using the method of Hodges and Lehmann (1963). The point estimate is the value of τ_0 such that the test statistic is equal to its null expectation.

5.1. *Model checking*. Models are approximations, and they can sometimes be tested by the observed data. Under the Tobit model (3), we know that the adjusted outcome $\max\{Y_{ij} - \tau_0 Z_i, 0\}$ is unaffected by the treatment assignment. Therefore, over all randomizations, the distribution of $\max\{Y_{ij} - \tau_0 Z_i, 0\}$ in the treated and control clusters are the same on average. Thus one form of model checking for the Tobit model of effects is to plot $\max\{0, Y_i - (1 - Z_i)\hat{\tau}\}$ under treatment and control arms for each test statistic. If the Tobit model of effects were correct, in an infinite sample without bias, these residuals should be identical across treated and control groups. However, this heuristic exercise ignores the uncertainty in the

Hodges–Lehmann point estimators and provides no formal test of the model. Next, we propose a formal statistical test.

We develop a formal test by extending the methods in Ding, Feller and Miratrix (2016) to test whether the treatment effect model is Tobit, that is, to test the following null hypothesis:

$$H_0(\text{Tobit}) : Y_{ij}(0) = \max\{Y_{ij}(1) - \tau, 0\} \qquad \text{for some } \tau, \text{ for all units } (i, j).$$

Note that $H_0(\text{Tobit})$ is different from the null hypothesis in (3) where τ_0 is a known value. If τ is known in $H_0(\text{Tobit})$, then under the null hypothesis $H_0(\tau)$ defined in (3), the control potential outcomes $\max\{Y_{ij} - \tau Z_i, 0\} = Y_{ij}(0)$ are all known. To assess the difference between the empirical distributions of the control potential outcomes across the two treatment groups, we use a Kolmogorov–Smirnov-type statistic of the form

$$t_{KS}(\tau) = \max_y |\widehat{F}_1(y; \tau) - \widehat{F}_0(y)|,$$

where

$$\begin{aligned}\widehat{F}_1(y; \tau) &= \frac{\sum_{i=1}^C Z_i \sum_{j=1}^{n_i} I\{Y_{ij}(0) \le y\}}{\sum_{i=1}^C Z_i n_i} \\ &= \frac{\sum_{i=1}^C Z_i \sum_{j=1}^{n_i} I\{\max(Y_{ij} - \tau, 0) \le y\}}{\sum_{i=1}^C Z_i n_i}\end{aligned}$$

and

$$\begin{aligned}\widehat{F}_0(y) &= \frac{\sum_{i=1}^C (1 - Z_i) \sum_{j=1}^{n_i} I\{Y_{ij}(0) \le y\}}{\sum_{i=1}^C (1 - Z_i) n_i} \\ &= \frac{\sum_{i=1}^C (1 - Z_i) \sum_{j=1}^{n_i} I\{Y_{ij} \le y\}.}{\sum_{i=1}^C (1 - Z_i) n_i}\end{aligned}$$

Under $H_0(\tau)$, because the adjusted outcomes are all fixed, the only random component in $\widehat{F}_1(y; \tau)$ and $\widehat{F}_0(y)$ is the treatment assignment. Therefore, we can simulate the distribution of $t_{KS}(\tau)$, and then compute the p-value $p(\tau)$.

Even if τ is unknown, we can obtain a valid p-value by maximizing $p(\tau)$ over a confidence region of τ with some adjustment. To be more specific, we use the following steps:

1. Construct a $1 - \delta$ confidence region for τ_0, denoted by CR_δ, with $\delta \ll \alpha$.
2. Compute the p-values over this confidence region $\{p(\tau) : \tau \in \text{CR}_\delta\}$.
3. Obtain the final p-value for $H_0(\text{Tobit})$ using

$$p_\delta = \max_{\tau \in \text{CR}_\delta} p(\tau) + \delta.$$

This p-value, p_δ, is valid according to Berger and Boos (1994), and was used by Nolen and Hudgens (2011) and Ding, Feller and Miratrix (2016) in randomization inference. Importantly, the randomization test for the validity of the constant treatment effect model requires that all missing potential outcomes can be determined by the observed data and the value of the treatment effect τ_0 [Ding, Feller and Miratrix (2016)]. Under the Tobit model, however, even with a known value of τ_0, we cannot impute all the missing potential outcomes under treatment based on the observed data. However, if we follow the strategy of Rosenbaum (2002a) using the adjusted outcomes $\max\{0, Y_{ij} - (1 - Z_i)\tau\}$, we do not need to know all the values of the $Y_{ij}(1)$'s, but only the values of the $Y_{ij}(0)$'s.

6. Application to the TLPES.

6.1. *Tests of the sharp null.* Next, we apply our methods to the data from the TLPES clustered randomized experiment. We first test the sharp null hypothesis of no effect. The sharp null asserts that the outcome for each teacher is unchanged by treatment such that $Y_{ij}(1) = Y_{ij}(0)$ for all i and j. If the sharp null is true, randomization would label a teachers' school as either treated or control, but the length of their feedback sessions would be unchanged. We apply all four test statistics to the TLPES data. With all four test statistics, we are able to reject the sharp null hypothesis. Moreover, all four test statistics easily reject the sharp null. That is, with all four tests, the p-values are all smaller than 0.0001. Thus we have strong evidence that the TLPES intervention increased the length of feedback session for teachers in treated schools. That is, with statistical significance level 0.001, the additional teacher observations required by the TLPES intervention caused teachers to engage in longer feedback sessions with their principals.

6.2. *Tobit model*: *Estimates of the treatment effect.* We estimated point estimates and confidence intervals using all four test statistics. For the Tobit effect, if we use the summed rank test statistic, the 95% confidence interval for τ is $[65, 106]$, and the Hodges–Lehmann point estimate is $\hat{\tau} = 86$ minutes. These results suggest that the training instituted under the TLPES intervention increased the time teachers spent in feedback sessions by as little as an hour or as much as over an hour and a half. Using the test statistic based on average ranks within treated clusters, the 95% confidence interval is $[90, 114]$ and the Hodges–Lehmann point estimate is $\hat{\tau} = 100$ minutes. If we adjust for cluster size via weights, using ω_m, the 95% confidence interval is $[21, 124]$, and the Hodges–Lehmann point estimate is $\hat{\tau} = 79$ minutes. Finally, if we adjust directly for cluster size, the 95% confidence interval is $[80, 104]$, and the Hodges–Lehmann point estimate is $\hat{\tau} = 90$ minutes.[2]

[2]Wayne et al. (2016) report the estimated increase in feedback time due to treatment was 86 minutes using a constant-additive model of effects.

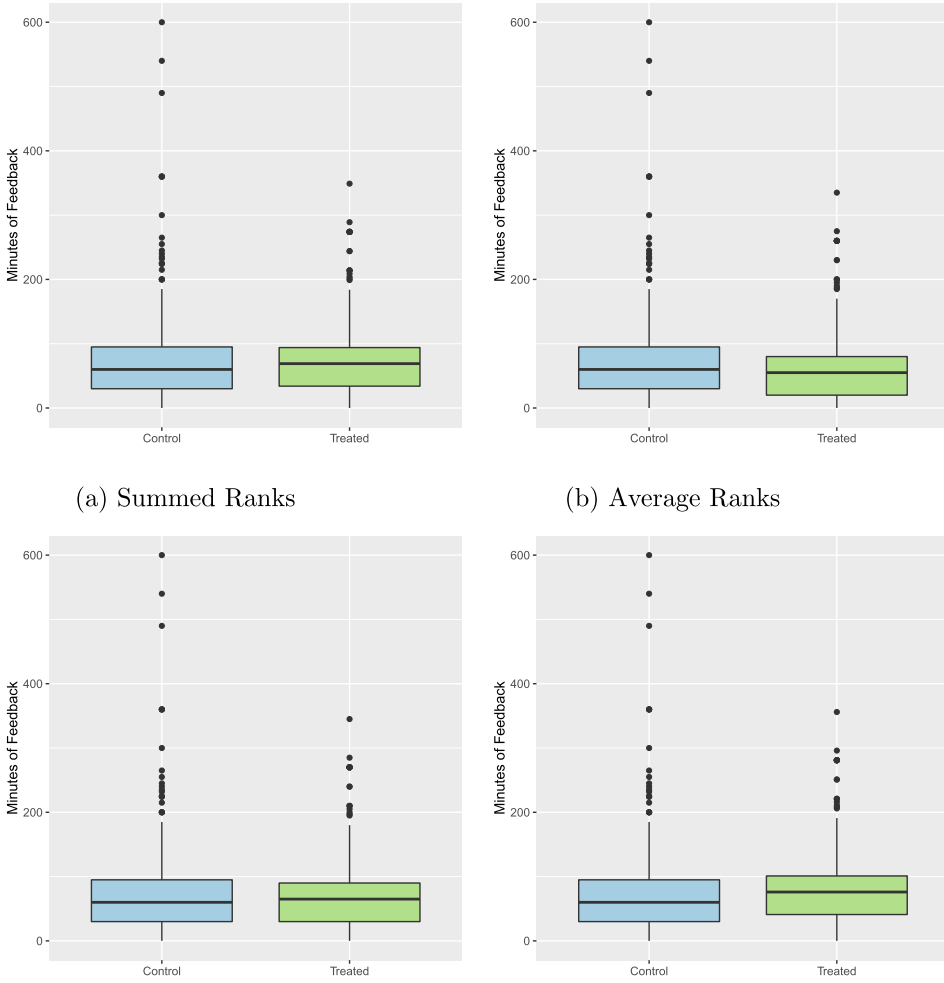
For the sake of comparison, we also used a random-effects linear mixed model to estimate the TLPES treatment effect and 95% confidence intervals. Using the LMM, the 95% confidence interval is [78, 106] and the point estimate is 92 minutes. Thus all the methods are largely in agreement. While the point estimates are all of similar magnitude, the length of the confidence intervals vary considerably. Can we draw any conclusions from this variation?

First, it appears that weighting the ranks by cluster sizes gives a wide interval estimate, which is coherent with the low powers of test statistic ω_m in a wide range of data generating processes in simulation. Second, using cluster size information in ω_n returns a confidence interval very similar to that from a LMM which suggests that both methods utilize information about cluster sizes in a similar manner.

6.3. *Tobit model*: *Model checking*. We conduct a goodness-of-fit test of the Tobit model using the exact test in Section 5.1. First, we use the more informal graphic method. Figure 5 checks the fit of the Tobit model of effects using residuals, where we plot $\max\{0, Y_i - (1 - Z_i)\hat{\tau}\}$ under treatment and control arms for each test statistic. If the Tobit model of effects were correct, in an infinite sample without bias, the residuals should be identical across treated and control groups. With the exception of several large outliers in the control groups, the boxplots are similar for all the test statistics. The data do not seem to clearly reject the Tobit model of effects as an inappropriate.

Next, we apply the formal test proposed above. With $\delta = 0.001$, the final p-value for testing the Tobit model itself is 0.111. This p-value does not depend on the test statistic we use in constructing the confidence region. The dotted line of Figure 6 shows the curve of $p(\tau)$. In sum, there is little statistical evidence against the Tobit model for fitting the TLPES data, which implies that the point and interval estimation under the Tobit model of effects is reasonable. It is possible, of course, that our test is under-powered. One alternative would be methods that explicitly relax the assumption of constant effects [Rosenbaum (2001, 2007)], though that is beyond the scope of the present analysis.

6.4. *Practical implications*. Finally, we conclude with a discussion of the practical implications of our study. One might conclude that methods we propose are of little relevance for applied data analysis given that the LMM and the rank based methods led to similar inferences in the TLPES study. That is, despite the fact that the outcomes were skewed and had point masses at zero, the LMM results are similar to the more robust rank based methods. We would argue that there are two feature of the data that favor the LMM. First, the sample sizes in TPLES are relatively large for a clustered randomized trial. In the TPLES study, there are more than 60 schools per arm. In the simulations, except in the case of the Cauchy distribution, the LMM performed well with sample sizes that large. Second, the treatment effect is quite large. Larger treatment effects will tend to be detectable irrespective of analytic method.



(a) Summed Ranks (b) Average Ranks

(c) Weighted Ranks (d) Cluster Adjusted Ranks

FIG. 5. *Residual plots from the Tobit model of effects. Each boxplot displays* $max\{0, Y_i - (1 - Z_i)\tau_0\}$ *for* $\tau_0 = \hat{\tau}$ *for each test statistic. In an infinite sample, the boxplots should be identical if the Tobit model of effects fits at the respective* $\hat{\tau}$.

As such, the simulation evidence is the more relevant guide for applied practice. That is, applied analysts should not solely rely on evidence from an LMM when sample sizes are below 30 schools per treatment arm, even when outcomes appear normally distributed. Moreover, if outcomes have heavy tails, results from the LMM should always be compared against a rank based alternative, since the LMM may be wrong regardless of the sample size. Thus, we recommend the use of rank based methods, at the very least, as a diagnostic, whenever investigators

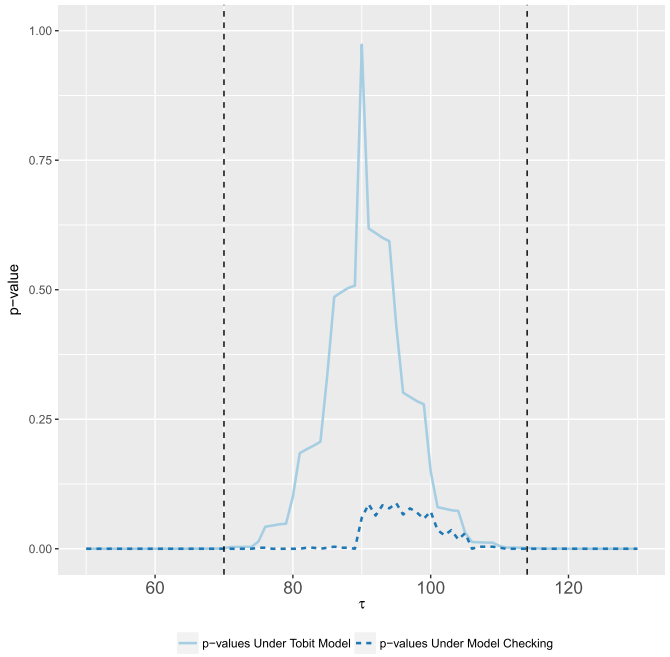


FIG. 6. *Plot for model checking test which contains p-values and confidence interval for* τ_0. *The grey line shows the p-values for inverting randomization tests using* ω_n *to construct a confidence interval* (*within two vertical dotted lines*) *for* τ_0 *under the Tobit model. The dotted lines shows the p-values for testing the Tobit model itself using* t_{KS}.

suspect the assumptions of the LMM might be in doubt. Alternatively when LMM assumptions are clearly suspect, rank based methods should be the primary mode of analysis.

7. Summary. In a clustered randomized trial for teaching training, we used the randomization of schools as the basis for inferences about treatment effects on teachers in schools. Our inferences do not depend on assumptions about the distributional properties of a model. We also tested for hypothesized Tobit effects, and extended previous work to allow for a formal test of whether the Tobit model is appropriate. While we did not perform covariance adjustments, we demonstrated how such adjustments could be performed without the need for the correct model of adjustment.

We found evidence that the additional training provided by the TLPES intervention increased the time teachers spent in feedback sessions with their supervisors. Based on a test statistic that used average ranks, the 95% confidence interval suggest that the new teacher training system increased the time spent with supervisors by at least 90 minutes. We were unable to reject the null hypothesis of a constant

treatment effect under the Tobit model, which suggests that the effect of the TLPES intervention varies little from teacher to teacher and school to school.

An advantage of randomization inference is its flexibility with respect to different types of experimental designs. For instance, we can extend the current analysis to clustered randomized experiments with blocking. As part of the study design, the 127 participating schools were randomly assigned within 37 blocks defined by baseline covariates [Wayne et al. (2016)]. We do not account for this blocking in our analysis, since we found that blocking did little to alter the results of the study for the outcome we examine. Moreover, in multi-site trials the treatment is randomly assigned to teachers within schools, which is essentially a randomized block design or stratified experiment and can be analyzed using existing methods [Imbens and Rubin (2015)]. As long as the investigator understand how treatment assignments were generated, exact tests are possible [Rosenbaum (2002a), Imbens and Rubin (2015), Ding, Feller and Miratrix (2016)].

Finally, the length of the feedback sessions were self-reported, and it appears that many of these outcome measurements were rounded to the nearest half or full hour, suggesting that a model for the measurement error process would be desirable. Although Neyman (1935)'s early analysis of experiments incorporated additive "technical errors" in the potential outcomes, they have been viewed as nuisances in randomization-based causal inference and have often been ignored [Rosenbaum (2010), Imbens and Rubin (2015)]. The data in this application suggest a more complete exploration of such measurement error, but we leave this to future research.

Acknowledgments. We thank Dr. Shu Yang for comments and suggestions. This study was conducted under contract by the American Institutes for Research for the Institute for Education Sciences.

REFERENCES

ARONOW, P. M., MIDDLETON, J. A. et al. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference* **1** 135–154.

BERGER, R. L. and BOOS, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. MR1294746

BORMAN, G. D., SLAVIN, R. E., CHEUNG, A., CHAMBERLAIN, A. M., MADDEN, N. A. and CHAMBERS, B. (2005). Success for all: First-year results from the national randomized field trial. *Educ. Eval. Policy Anal.* **27** 1–22.

BRAUN, T. M. and FENG, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *J. Amer. Statist. Assoc.* **96** 1424–1432. MR1946587

CHETTY, R., FRIEDMAN, J. N. and ROCKOFF, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* **104** 2593–2632.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. MR0474575

CORNFIELD, J. (1978). Randomization by group. *Am. J. Epidemiol.* **108** 100–102.

DATTA, S. and SATTEN, G. A. (2005). Rank-sum tests for clustered data. *J. Amer. Statist. Assoc.* **100** 908–915. MR2201018

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345. MR3695995

DING, P., FELLER, A. and MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671. MR3506797

DONNER, A. and KLAR, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Wiley, New York.

DUTTA, S. and DATTA, S. (2016). A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics* **72** 432–440. MR3515770

FEDER, G., GRIFFITHS, C., ELDRIDGE, S. and SPENCE, M. (1999). Effect of postal prompts to patients and general practitioners on the quality of primary care after a coronary event (POST): Randomised controlled trial. *BMJ* **318** 1522–1526.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London.

GAIL, M. H., BYAR, D. P., PECHACEK, T. F., CORLE, D. K., GROUP, C. S. et al. (1992). Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). *Control. Clin. Trials* **13** 6–21.

GAIL, M. H., MARK, S. D., CARROLL, R. J. and GREEN, S. B. (1996). On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* **15** 1069–1092.

HANSEN, B. B. and BOWERS, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *J. Amer. Statist. Assoc.* **104** 873–885. MR2562000

HAYES, R. and MOULTON, L. (2009). *Cluster Randomised Trials*. Chapman& Hall/CRC, London.

HEDGES, L. V. and HEDBERG, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educ. Eval. Policy Anal.* **29** 60–87.

HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.* **34** 598–611. MR0152070

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

IMBENS, G. M. and WOOLDRIDGE, J. M. (2008). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47** 5–86.

LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* **112** 1759–1769. MR3750897

MIDDLETON, J. A. (2008). Bias of the regression estimator for experiments using clustered random assignment. *Statist. Probab. Lett.* **78** 2654–2659. MR2542462

MIDDLETON, J. A. and ARONOW, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy* **6** 39–75.

MURNANE, R. J. and WILLETT, J. B. (2010). *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press, Oxford.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dąbrowska and T. P. Speed. MR1092986

NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *Suppl. J. R. Stat. Soc.* **2** 107–180.

NOLEN, T. L. and HUDGENS, M. G. (2011). Randomization-based inference within principal strata. *J. Amer. Statist. Assoc.* **106** 581–593. MR2847972

ROSENBAUM, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* **88** 219–231. MR1841270

ROSENBAUM, P. R. (2002a). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487

ROSENBAUM, P. R. (2007). Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics* **63** 1164–1171, 1313. MR2414594

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROSNER, B., GLYNN, R. J. and LEE, M.-L. T. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. *Biometrics* **59** 1089–1098. MR2025134

ROSNER, B., GLYNN, R. J. and LEE, M.-L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* **62** 185–192, 318. MR2226572

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **6** 688–701.

RUBIN, D. B. (1986). Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.

SCHOCHET, P. Z. (2013). Estimators for clustered education RCTs using the Neyman model for causal inference. *J. Educ. Behav. Stat.* **38** 219–238.

SMALL, D. S., HAVE, T. R. T. and ROSENBAUM, P. R. (2008). Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. *J. Amer. Statist. Assoc.* **103** 271–279. MR2420232

TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462

WAYNE, J. A., GARET, M. S., BROWN, S., RICKLES, J., SONG, M. and MANZESKE, D. (2016). Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback Year 1 Report. Technical report, American Institutes of Research, Washington, DC.

WILLIAMSON, J. M., DATTA, S. and SATTEN, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59** 36–42. MR1978471

ZHANG, K., TRASKIN, M. and SMALL, D. S. (2012). A powerful and robust test statistic for randomization inference in group-randomized trials with matched pairs of groups. *Biometrics* **68** 75–84. MR2909855

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
425 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: pengdingpku@berkeley.edu

MCCOURT SCHOOL OF PUBLIC POLICY
GEORGETOWN UNIVERSITY
37TH & O ST, NW
WASHINGTON, DC 20057
USA
E-MAIL: lk681@georgetown.edu