Supporting information for "Statistical Inference and Power Analysis for Direct and Spillover Effects in Two-Stage Randomized Experiments" by Jiang, Imai, and Malani

Section S1 establishes the equivalence relationship between the regression-based inference and randomization-based inference.

Section S2 compares the two-stage randomized design with the completely randomized and cluster randomized designs.

Section S3 provides proofs of the theorems.

Section S4 provides more computation details.

Section S5 presents the simulation studies.

S1 Connections to linear regression

In this section, we establish direct connections between the proposed estimators and the least squares estimators, which is popular among applied researchers. Basse and Feller (2018) study the relationships between the ordinary least squares and randomization-based estimators for the direct and spillover effects under a particular two-stage randomized experiment. Here, we extend these previous results to a general setting with m treatment assignment mechanisms.

We consider the following linear model for the outcome,

$$Y_{ij} = \sum_{a=1}^{m} \{\beta_{1a} Z_{ij} \mathbf{1}(A_j = a) + \beta_{0a}(1 - Z_{ij})\mathbf{1}(A_j = a)\} + \epsilon_{ij}, \tag{S1}$$

where ϵ_{ij} is the error term. Unlike the two-step procedure in Basse and Feller (2018), we fit the weighted least squares regression with the following inverse probability weights,

$$w_{ij} = \frac{1}{J_{A_j}} \cdot \frac{1}{n_{jZ_{ij}}}. \tag{S2}$$

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_{11}, \widehat{\beta}_{01}, \ldots, \widehat{\beta}_{1m}, \widehat{\beta}_{0m})^\top$ be the weighted least squares estimators of the coefficients in the models of equation (S1), respectively. For the variance estimator, we need additional notation. Let $\mathbf{X}_j = (X_{1j}, \ldots, X_{n_j j})^\top$ be the design matrix of cluster j for the model given in (S1) with $X_{ij} = (Z_{ij}\mathbf{1}(A_j = 1), (1 - Z_{ij})\mathbf{1}(A_j = 1), \ldots, Z_{ij}\mathbf{1}(A_j = m), (1 - Z_{ij})\mathbf{1}(A_j = m))^\top$. Let $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_J^\top)^\top$ be the entire design matrix, $\mathbf{W}_j = \text{diag}(w_{1j}, \ldots, w_{n_j j})$ be the weight matrix for cluster j, and $\mathbf{W} = \text{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_J)$ be the entire weight matrix. We use $\hat{\boldsymbol{\epsilon}}_j = (\hat{\epsilon}_{1j}, \ldots, \hat{\epsilon}_{n_j j})$ to denote the residual vector for cluster j obtained from the weighted least squares fit of the model given in equation (S1), and $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}}_1^\top, \ldots, \hat{\boldsymbol{\epsilon}}_J^\top)^\top$ to represent the residual vector for the entire sample.

We consider the cluster-robust generalization of HC2 covariance matrix (Bell and McCaffrey, 2002),

$$\widehat{\text{var}}_{\text{hc2}}^{\text{cluster}}(\widehat{\boldsymbol{\beta}}) \;=\; (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\left\{\sum_j \mathbf{X}_j^\top\mathbf{W}_j(\mathbf{I}_{n_j} - \mathbf{P}_j)^{-1/2}\widehat{\boldsymbol{\epsilon}}_j\widehat{\boldsymbol{\epsilon}}_j^\top(\mathbf{I}_{n_j} - \mathbf{P}_j)^{-1/2}\mathbf{W}_j\mathbf{X}_j\right\}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1},$$

where $\mathbf{I}_{n_j}$ is the $n_j \times n_j$ identity matrix and $\mathbf{P}_j$ is the following cluster leverage matrix,

$$\mathbf{P}_j \;=\; \mathbf{W}_j^{1/2}\mathbf{X}_j(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_j^\top\mathbf{W}_j^{1/2}.$$

The next theorem establishes the equivalence relationship between the regression-based inference and randomization-based inference.

Theorem S1 (Equivalent Weighted Least Squares Estimators) *The weighted least squares estimators based on the model of equation* (S1) *are equivalent to the randomization-based estimators of the average potential outcomes, i.e.,* $\widehat{\boldsymbol{\beta}} = \widehat{Y}$. *The cluster-robust generalization of HC2 covariance matrix is equivalent to the randomization-based covariance matrix estimator, i.e.,* $\widehat{\text{var}}_{\text{hc2}}^{\text{cluster}}(\widehat{\boldsymbol{\beta}}) = \widehat{D}/J$.

Proof is given in Section S3.9.

S2 Theoretical Comparison of Three Randomized Experiments

Although the two-stage randomized design allows for the detection of spillover effects, this may come at the cost of statistical efficiency for detecting the average treatment effect if it turns out that

spillover effects do not exist. In this section, we conduct a theoretical comparison of the two-stage randomized design with the completely randomized design and cluster randomized design in the absence of interference between units. The latter two are the most popular experimental designs and are limiting designs of the two-stage randomized designs. That is, we compute the relative efficiency loss due to the use of the two-stage randomized design when there is no spillover effect.

Formally, when there is no interference between units, we can write $Y_{ij}(z,a) = Y_{ij}(z)$, $\overline{Y}_j(z,a) = \overline{Y}_j(z)$, and $\overline{Y}(z,a) = \overline{Y}(z)$. As a result, both the direct and marginal direct effects reduce to the standard average treatment effect. To unify the notation in the three types of experiments, we define the unit-level average treatment effect as, $\text{ATE}_{ij} = Y_{ij}(1) - Y_{ij}(0)$, the cluster-level average treatment effect as, $\text{ATE}_j = \sum_{i=1}^{n_j}\{Y_{ij}(1) - Y_{ij}(0)\}/n_j$, and the population-level average treatment effect as $\text{ATE} = \sum_{i=1}^{J}\text{ATE}_j/J$. As noted above, our comparison of three designs assumes no interference between units. The reason for this assumption is that the average treatment effect represents a different causal quantity under the three designs in the presence of interference, making the efficiency comparison across the designs less meaningful (Karwa and Airoldi, 2018).

For simplicity, consider the case when the cluster size is equal, i.e., $n_j = n$ for all j. Define the within-cluster variance of $Y_{ij}(z)$ and ATE_{ij} as,

$$\eta_w^2(z) \;=\; \frac{\sum_{j=1}^{J}\sum_{i=1}^{n}\{Y_{ij}(z) - \overline{Y}_j(z)\}^2}{nJ-1}, \quad \tau_w^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n}\{\text{ATE}_{ij} - \text{ATE}_j\}^2}{nJ-1},$$

the between-cluster variance of $Y_{ij}(z)$ and ATE_{ij} as,

$$\eta_b^2(z) = \frac{\sum_{j=1}^{J}\{\overline{Y}_j(z) - \overline{Y}(z)\}^2}{J-1}, \quad \tau_b^2 = \frac{\sum_{j=1}^{J}\{\text{ATE}_j - \text{ATE}\}^2}{J-1},$$

and the total variance of $Y_{ij}(z)$ and ATE_{ij} as,

$$\eta^2(z) \;=\; \frac{\sum_{j=1}^{J}\sum_{i=1}^{n}\{Y_{ij}(z) - \overline{Y}(z)\}^2}{nJ-1}, \quad \tau^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n}\{\text{ATE}_{ij} - \text{ATE}\}^2}{nJ-1}.$$

We can connect these variances by defining the intracluster correlation coefficient with respect to $Y_{ij}(z)$ in cluster j under treatment condition z as,

$$r_j(z) = \frac{\sum_{i\neq i'}^{n}(Y_{ij}(z) - \overline{Y}(z))(Y_{i'j}(z) - \overline{Y}(z))}{(n-1)\cdot\sum_{i=1}^{n}(Y_{ij}(z) - \overline{Y}(z))^2}.$$

and the intracluster correlation coefficient with respect to ATE_{ij} in cluster j as,

$$r'_j = \frac{\sum_{i \neq i'}^{n} (\text{ATE}_{ij} - \text{ATE})(\text{ATE}_{i'j} - \text{ATE})}{(n-1) \cdot \sum_{i=1}^{n} (\text{ATE}_{ij} - \text{ATE})^2}.$$

To further facilitate our theoretical comparison, we make additional approximation assumptions. First, the intracluster correlation coefficients are approximately the same with respect to $Y_{ij}(z)$ and ATE_{ij} across clusters and treatment conditions, i.e., $r_j(z) \approx r'_j \approx r$. Second, the cluster size is relatively small compared to the number of clusters $nJ - 1 \approx nJ \approx n(J-1)$. These approximations help simplify the expressions of the variances as

$$\begin{aligned} \eta_w^2(z) &\approx \frac{(n-1)(1-r)}{n} \cdot \eta^2(z), \quad \tau_w^2 \approx \frac{(n-1)(1-r)}{n} \cdot \tau^2, \\ \eta_b^2(z) &\approx \frac{1+(n-1)r}{n} \cdot \eta^2(z), \quad \tau_b^2 \approx \frac{1+(n-1)r}{n} \cdot \tau^2. \end{aligned} \tag{S3}$$

We consider three randomized experiments in the population with nJ units. Under the two-stage randomized design, the treatment is randomized according to Assumptions 1. Under the completely randomized design, the treatment is randomized across units,

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \frac{1}{\binom{nJ}{\sum_{a=1}^{m} J_a n p_a}},$$

for all $\boldsymbol{z}$ such that $\sum_{i,j} z_{ij} = \sum_{a=1}^{m} J_a n p_a$. Finally, under the clustered randomized design, the treatment is randomized across clusters, where all the units in each cluster is assigned to the same treatment condition, i.e.,

$$\Pr(\boldsymbol{A} = \boldsymbol{a}) = \frac{1}{\binom{J}{\sum_{a=1}^{m} J_a p_a}},$$

for all $\boldsymbol{z}$ such that $\sum_{j=1}^{J} a_j = \sum_{a=1}^{m} J_a p_a$. Note that under this setting, the number of treated units will be the same in the three types of randomized experiments.

We consider the difference in means estimator for estimating ATE,

$$\widehat{\text{ATE}} = \frac{1}{J} \sum_{j=1}^{J} \left\{ \frac{\sum_{i=1}^{n} Y_{ij} Z_{ij}}{n_{j1}} - \frac{\sum_{i=1}^{n} Y_{ij}(1 - Z_{ij})}{n_{j0}} \right\}. \tag{S4}$$

The following theorem gives the variances of this estimator under the three experimental designs.

Theorem S2 (Comparison of Three Experimental Designs) *Under the approximation assumptions of equation* (S3)*, the variance of the average treatment effect estimator* $\widehat{\text{ATE}}$ *given in equation* (S4) *under the two-stage randomized design is*

$$\frac{1-r}{J^2}\sum_{a=1}^{m}\frac{J_a}{np_a}\cdot\eta^2(1)+\frac{1-r}{J^2}\sum_{a=1}^{m}\frac{J_a}{n(1-p_a)}\cdot\eta^2(0)-\frac{1-r}{nJ}\cdot\tau^2, \tag{S5}$$

the variance of $\widehat{\text{ATE}}$ *under the completely randomized design is*

$$\frac{1}{\sum_{a=1}^{m}J_a np_a}\cdot\eta^2(1)+\frac{1}{\sum_{a=1}^{m}J_a n(1-p_a)}\cdot\eta^2(0)-\frac{1}{nJ}\cdot\tau^2, \tag{S6}$$

the variance of $\widehat{\text{ATE}}$ *under the cluster randomized design is*

$$\frac{1+(n-1)r}{\sum_{a=1}^{m}J_a np_a}\cdot\eta^2(1)+\frac{1+(n-1)r}{\sum_{a=1}^{m}J_a n(1-p_a)}\cdot\eta^2(0)-\frac{1+(n-1)r}{nJ}\cdot\tau^2. \tag{S7}$$

Proof is given in Appendix S3.10. From Theorem S2, the ratio of the coefficients of $\eta^2(1)$ in equations (S5) and (S6) is

$$(1-r)\cdot\sum_{a=1}^{m}q_a p_a\cdot\sum_{a=1}^{m}\frac{q_a}{p_a}, \tag{S8}$$

whereas the ratio of the coefficients of $\eta^2(1)$ in equations (S5) and (S7) is

$$\frac{1-r}{1+(n-1)r}\cdot\sum_{a=1}^{m}q_a p_a\cdot\sum_{a=1}^{m}\frac{q_a}{p_a}. \tag{S9}$$

The ratios of the coefficients of other parameters take similar forms. Thus, our discussion focuses on equations (S8) and (S9).

Equation (S8) implies that the relative efficiency of the two-stage randomized design over the completely randomized design depends on the intracluster correlation coefficient, and the assignment probabilities at the first and the second stage of randomization. Due to the Cauchy–Schwarz inequality, equation (S8) is greater than or equal to $1-r$. The value of this quantity increases as the heterogeneity between p_a increases. Therefore, as the difference in treated proportions between clusters becomes large, the two-stage randomized design becomes less efficient for estimating the average treatment effect. On the other hand, the ability to detect spillover effects relies on the heterogeneity of p_a. This implies that there is a tradeoff between the efficiency of estimating the

average treatment effects and the ability to detect spillover effects. This finding is consistent with that of Baird *et al.* (2018).

In addition, when the treated proportion is identical across clusters, $p_a = p_{a'}$ for any a, a', the two-stage randomized design becomes stratified randomized design. In this case, equation (S8) equals $1 - r$, which is less than 1. This is consistent with the classic result that the stratified randomized design improves efficiency over the completely randomized design.

Lastly, equation (S9) implies that the relative efficiency of the two-stage randomized design with respect to the clustered randomized design depends additionally on the cluster size. As the cluster size increases, the two-stage randomized design becomes more efficient than the clustered randomized design. When cluster size is large, the two-stage randomized design may be preferable because it allows for the detection of spillover effects while maintaining efficiency in estimating the average treatment effect.

S3 Proofs of the Theorems

We can write

$$\widehat{Y}(z,a) \;=\; \frac{1}{J_a}\sum_{j=1}^{J}\widehat{Y}_j(z)\mathbf{1}(A_j = a) \;=\; \mu(z,a) + \sum_{j=1}^{J}\delta_j(z,a),$$

where

$$\begin{aligned}\mu(z,a) &= \frac{1}{J_a}\sum_{j=1}^{J}\overline{Y}_j(z,a)\mathbf{1}(A_j = a),\\ \delta_j(z,a) &= \frac{1}{J_a}\left\{\widehat{Y}_j(z) - \overline{Y}_j(z,a)\right\}\mathbf{1}(A_j = a).\end{aligned}$$

Let $\mu = (\mu(1,1), \mu(0,1), \ldots, \mu(1,m), \mu(0,m))^\top$ and $\delta_j = (\delta_j(1,1), \delta_j(0,1), \ldots, \delta_j(1,m), \delta_j(0,m))^\top$ be the vectorization of $\mu(z,a)$ and $\delta_j(z,a)$, respectively, and $\delta = \sum_{j=1}^{J}\delta_j$. We can write

$$\widehat{Y} \;=\; \mu + \delta.$$

Let $\mathcal{F}_0 = \sigma(A_1, \ldots, A_J)$ be the σ-algebra generated by $\{A_j : j = 1, \ldots, J\}$. Conditioning on $\mathcal{F}_0$, $\{\delta_j : j = 1, \ldots, J\}$ are jointly independent. Therefore, we have $\mathbb{E}(\delta \mid \mathcal{F}_0) = \mathbb{E}(\delta_j \mid \mathcal{F}_0) = 0$. From

the law of total expectation,

$$\begin{aligned}
\mathbb{E}(\delta) &= \mathbb{E}(\delta_j) = 0, \\
\mathrm{cov}(\delta) &= \sum_{j=1}^{J} \mathrm{cov}(\delta_j) = \sum_{j=1}^{J} \mathbb{E}\mathrm{cov}(\delta_j \mid \mathcal{F}_0), \\
\mathrm{cov}(\mu, \delta) &= \mathbb{E}\left\{\mathrm{cov}(\mu, \delta \mid \mathcal{F}_0)\right\} + \mathrm{cov}\{\mathbb{E}(\mu \mid \mathcal{F}_0), \mathbb{E}(\delta \mid \mathcal{F}_0)\} = 0.
\end{aligned} \tag{S10}$$

S3.1 Proof of Theorem 2

We need the following lemma for the proof.

LEMMA S1 (LI AND DING (2017), THEOREMS 3 AND 5) *In a completely randomized experiment with N units and Q treatment groups of sizes N_q $(q = 1, \ldots, Q)$, let Z_i be the treatment indicator and Y_i be the observed outcome for unit i. Let $Y_i(q)$ be the length-L vector potential outcome of i under treatment q, and $S_{qq'} = (N-1)^{-1} \sum_{i=1}^{N} \{Y_i(q) - \overline{Y}(q)\}\{Y_i(q') - \overline{Y}(q')\}$ be the finite-population covariances for $q, q' = 1, \ldots, Q$. Let $\tau = \sum_{q=1}^{Q} G_q \overline{Y}(q)$ be the population average causal effect of interest, and $\widehat{\tau} = \sum_{q=1}^{W} G_q \widehat{Y}(q)$ be the moment estimator with $\widehat{Y}(q) = N_q^{-1} \sum_{i=1}^{N} Y_i \mathbf{1}(Z_i = q)$. We have (a)*

$$\mathrm{cov}(\widehat{\tau}) = \sum_{q=1}^{Q} N_q^{-1} G_q S_{qq} G_q^{\top} - N^{-1} S_{\tau}^2,$$

where S_{τ}^2 is the finite-population covariance of $\tau_i = \sum_{q=1}^{W} G_q Y_i(q)$; (b) suppose the following conditions hold for $q, q' = 1, \ldots, Q$ as N goes to infinity:

(a) $S_{qq'}$ has a finite limit;

(b) N_q/N has a finite limit in $(0, 1)$;

(c) $\max_i |Y_i(q) - \overline{Y}(q)|^2/N = o(1)$. Then, we have

$$\sqrt{N}(\widehat{\tau} - \tau) \xrightarrow{d} N(0, V),$$

where V denotes the limiting value of $N\mathrm{cov}(\widehat{\tau})$.

We then prove Theorem 2. For simplicity, we consider the case with $m = 2$. From (S10), we have $\mathrm{cov}(\widehat{Y}) = \mathrm{cov}(\mu) + \mathrm{cov}(\delta)$. We then derive the analytic forms of $\mathrm{cov}(\mu)$ and $\mathrm{cov}(\delta)$.

For $\mathrm{cov}(\mu)$, define $B_j(a) = (\overline{Y}_j(1,a), \overline{Y}_j(0,a))^\top$ as the vector potential outcome of cluster j under $A_j = a$ with means $\overline{B}(a) = (\overline{Y}(1,a), \overline{Y}(0,a))^\top$ and covariances

$$\begin{aligned} S_b(a) &= (J-1)^{-1} \sum_{j=1}^{J} \{B_j(a) - \overline{B}_j(a)\}\{W_j(a) - \overline{B}_j(a)\}^\top \\ &= \begin{pmatrix} \sigma_b^2(1,1;a,a) & \sigma_b^2(1,0;a,a) \\ \sigma_b^2(1,0;a,a) & \sigma_b^2(0,0;a,a) \end{pmatrix}. \end{aligned}$$

We can then write μ as $(I_2, 0_{2\times 2})^\top \overline{B}(1) + (0_{2\times 2}, I_2)^\top \overline{B}(2)$. From Lemma S1, we have

$$\begin{aligned} \mathrm{cov}(\mu) &= J_1^{-1}(I_2, 0_{2\times 2})^\top S_b(1)(I_2, 0_{2\times 2}) + J_2^{-1}(0_{2\times 2}, I_2)^\top S_b(2)(0_{2\times 2}, I_2) - J^{-1} S_b \\ &= J^{-1}(H \circ S_b). \end{aligned}$$

For $\mathrm{cov}(\delta)$, theory of simple random sampling implies,

$$\begin{aligned} \mathrm{var}\left\{\widehat{Y}_j(z,a) \mid A_j = a\right\} &= \frac{1}{n_{jz}}\left(1 - \frac{n_{jz}}{n_j}\right)\sigma_j^2(z,a), \\ \mathrm{cov}\left\{\widehat{Y}_j(1,a), \widehat{Y}_j(0,a) \mid A_j = a\right\} &= -\frac{1}{n_j}\sigma_j^2(1,0;a). \end{aligned}$$

We can write

$$\mathrm{cov}\left\{\widehat{Y}_j(z,a), \widehat{Y}_j(z',a) \mid A_j = a\right\} = n_j^{-1}\{n_j/n_{jz}\mathbf{1}(z = z') - 1\}\sigma_j^2(z,z';a).$$

Therefore, we have

$$\mathrm{cov}(\delta_j \mid A_j = a) = J_a^{-2} q_a n_j^{-1}\{H_j(a) \circ S_w\} = q_a^{-1} J^{-2} n_j^{-1}\{H_j(a) \circ S_j\}, \tag{S11}$$

where

$$\begin{aligned} H_j(1) &= \mathrm{diag}(q_1^{-1}, 0) \otimes \left\{\mathrm{diag}(n_j/n_{j1}, n_j/n_{j0}) - 1_{2\times 2}\right\}, \\ H_j(2) &= \mathrm{diag}(0, q_2^{-1}) \otimes \left\{\mathrm{diag}(n_j/n_{j1}, n_j/n_{j0}) - 1_{2\times 2}\right\}. \end{aligned}$$

Because $H_j = H_j(1) + H_j(2)$, we can obtain

$$\mathrm{cov}(\delta_w) = \mathbb{E}\{\mathrm{cov}(\delta_j \mid A_j = a)\} = J^{-2} n_j^{-1}\{H_j \circ S_j\}.$$

□

S3.2 Proof of Theorem 3

Recall that $\widehat{D}$ be a $2m$ by $2m$ block diagonal matrix with the a-th matrix on the diagonal

$$\widehat{D}_a = \frac{J}{J_a}\begin{pmatrix} \widehat{\sigma}_b^2(1,a) & \widehat{\sigma}_b^2(1,0;a) \\ \widehat{\sigma}_b^2(1,0;a) & \widehat{\sigma}_b^2(0,a) \end{pmatrix}.$$

We calculate the expectation of each term in $\widehat{D}_a$. We have

$$\begin{aligned}
& \mathbb{E}\{\widehat{\sigma}_b^2(z,a)\} \\
= & \frac{1}{J_a-1}\mathbb{E}\left\{\sum_{j=1}^{J}\widehat{Y}_j^2(z)I(A_j=a) - J_a\widehat{Y}(z)^2\right\} \\
= & \frac{1}{J_a-1}\mathbb{E}\left(\sum_{j=1}^{J}\left[\mathrm{var}\left\{\widehat{Y}_j(z,a)\mid A_j=a\right\} + \overline{Y}_j(z,a)^2\right]I(A_j=a)\right) - \frac{J_a}{J_a-1}[\mathrm{var}\{\widehat{Y}(z,a)\} + \overline{Y}(z,a)^2] \\
= & \frac{J_a}{J(J_a-1)}\sum_{j=1}^{J}\mathrm{var}\{\widehat{Y}_j(z)\mid A_j=a\} + \frac{J_a}{J(J_a-1)}\sum_{j=1}^{J}\overline{Y}_j(z,a)^2 - \frac{J_a}{J_a-1}[\mathrm{var}\{\widehat{Y}(z,a)\} + \overline{Y}(z,a)^2] \\
= & \frac{J_a}{J(J_a-1)}\sum_{j=1}^{J}\mathrm{var}\left\{\widehat{Y}_j(z)\mid A_j=a\right\} + \frac{J_a(J-1)}{(J_a-1)J}\sigma_b^2(z,a) - \frac{J_a}{J_a-1}\mathrm{var}\{\widehat{Y}(z,a)\} \\
= & \frac{J_a}{J(J_a-1)}\sum_{j=1}^{J}\mathrm{var}\left\{\widehat{Y}_j(z)\mid A_j=a\right\} + \frac{J_a(J-1)}{(J_a-1)J}\sigma_b^2(z,a) \\
& -\frac{J_a}{J_a-1}\left[\left(1-\frac{J_a}{J}\right)\frac{\sigma_b^2(z,a)}{J_a} + \frac{1}{J_aJ}\sum_{j=1}^{J}\mathrm{var}\left\{\widehat{Y}_j(z,1)\mid A_j=a\right\}\right] \\
= & \sigma_b^2(z,a) + \frac{1}{J}\sum_{j=1}^{J}\mathrm{var}\left\{\widehat{Y}_j(z,a)\mid A_j=a\right\} \\
= & \sigma_b^2(z,a) + \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_{jz}}\left(1-\frac{n_{jz}}{n_j}\right)\sigma_j^2(z,a).
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
\mathbb{E}\left\{\widehat{\sigma}_b^2(1,0;a)\right\} &= \sigma_b^2(1,0;a) + \frac{1}{J}\sum_{j=1}^{J}\mathrm{cov}\left\{\widehat{Y}_j(1,a),\widehat{Y}_j(0,a)\mid A_j=a\right\} \\
&= \sigma_b^2(1,0;a) - \frac{1}{J}\sum_{j=1}^{J}\frac{\sigma_j^2(1,0;a)}{n_j}.
\end{aligned}$$

Finally, we prove that $\widehat{D}$ is a conservative estimator for D. Denote $R = \mathbb{E}(\widehat{D}) - D$ with the (k,l)-th element r_{kl}. We have

$$r_{2a-1,2a-1} = \sigma_b^2(1,a), \quad r_{2a,2a} = \sigma_b^2(0,a), \quad r_{2a-1,2a} = \sigma_b^2(1,0;a)$$

for $a = 1, \ldots, m$. For $a \neq a'$, we have

$$r_{2a-1,2a'-1} \;=\; \sigma_b^2(1; a, a'), \quad r_{2a,2a'-1} = \sigma_b^2(1, 0; a, a').$$

Therefore, for any vector $c = (c_1, \ldots, c_{2m})$, $cRc^\top$ is the between-cluster variance of $\sum_{a=1}^m c_{2a-1} Y_{ij}(1, a) + \sum_{a=1}^m c_{2a} Y_{ij}(0, a)$. As a result, $\widehat{D}$ is a conservative estimator for D and is unbiased for D if $\overline{Y}_j(z, a)$ is constant across clusters. □

S3.3 Proof of Theorem 4

S3.3.1 Lemmas

Let $*$ denote convolution. We need the following lemmas for the proof.

LEMMA S2 (OHLSSON (1989), THEOREM A.1) *For $j = 1, \ldots, J$, let $\{\xi_{J,j} : j = 1, \ldots, J\}$ be a martingale difference sequence relative to the filtration $\{\mathcal{F}_{J,j} : j = 0, 1, \ldots, J\}$, and let X_J be an $\mathcal{F}_{J,0}$-measurable random variable. Denote $\xi_J = \sum_{j=1}^J \xi_{J,j}$. Suppose that the following conditions hold as J goes to infinity:*

(a) $\sum_{j=1}^J \mathbb{E}(\xi_{J,j}^4) = o(1)$.

(b) For some sequence of non-negative real numbers $\{\beta_J : J = 1, 2, \ldots\}$ with $sup_J \beta_J < \infty$, we have $\mathbb{E}\left[\left\{\sum_{j=1}^J \mathbb{E}(\xi_{J,j}^2 \mid \mathcal{F}_{J,j-1}) - \beta_J^2\right\}^2\right] = o(1)$.

(c) For some probability distribution $\mathcal{L}_0$, $\mathcal{L}(X_J) * N(0, \beta_J^2) \xrightarrow{d} \mathcal{L}_0$.

Then, $\mathcal{L}(X_J + \xi_J) \xrightarrow{d} \mathcal{L}_0$ as J goes to infinity.

LEMMA S3 *Suppose that Assumptions 1, 2, 3, and Condition 1 hold. Then*

$$J^2 \sum_{j=1}^J \mathbb{E}(||\delta_j||_2^4 \mid A_j = a) \;=\; o(1), \quad J^2 \sum_{j=1}^J \mathbb{E}(||\delta_j||_2^4) \;=\; o(1)$$

Proof. It suffices to verify the first equality. Note that for j with $A_j = a$ $(a = 1, 2)$

$$||\delta_j||_2^2 \;=\; \frac{1}{J_a^2}\left[\left\{\widehat{Y}_j(1) - \overline{Y}_j(1, a)\right\}^2 + \left\{\widehat{Y}_j(0) - \overline{Y}_j(0, a)\right\}^2\right].$$

From the Cauchy–Schwarz inequality, we have

$$||\delta_j||_2^4 \leq \frac{2}{J_a^4}\left[\left\{\widehat{Y}_j(1)-\overline{Y}_j(1,a)\right\}^4+\left\{\widehat{Y}_j(0)-\overline{Y}_j(0,a)\right\}^4\right].$$

Because $\left\{\widehat{Y}_j(z)-\overline{Y}_j(z,a)\right\}^4 \leq 8\widehat{Y}_j^4(z)+8\overline{Y}_j^4(z)$,

$$\begin{aligned}
&\sum_{j=1}^{J}\mathbb{E}\left[\left\{\widehat{Y}_j(z)-\overline{Y}_j(z,a)\right\}^4 \mid A_j=a\right] \\
= \;& 8\sum_{j=1}^{J}\mathbb{E}\left\{\widehat{Y}_j^4(z)\mid A_j=a\right\}+8\sum_{j=1}^{J}\overline{Y}_j^4(z,a) \\
= \;& 8\sum_{j=1}^{J}\mathbb{E}\left\{\widehat{Y}_j^4(z)\mid A_j=a\right\}+o(J^2).
\end{aligned}$$

From the power-mean inequality, for $A_j=a$,

$$\widehat{Y}_j^4(z) \;\leq\; \frac{1}{n_{jz}}\sum_{j=1}^{J}Y_{ij}^4(z,a)\cdot\mathbf{1}(Z_{ij}=z) \;\leq\; \frac{1}{n_{jz}}\sum_{j=1}^{J}Y_{ij}^4(z,a) \;\leq\; \epsilon^{-1}\overline{Y_j(z,a)} \;=\; o(J^2). \tag{S12}$$

Therefore,

$$\sum_{j=1}^{J}\mathbb{E}\left[\left\{\widehat{Y}_j(z)-\overline{Y}_j(z,a)\right\}^4 \mid A_j=a\right] \;=\; o(J^2).$$

As a result,

$$\begin{aligned}
J^2\sum_{j=1}^{J}\mathbb{E}(||\delta_j||_2^4 \mid A_j=a) \;&\leq\; \frac{2}{q_a^2J^2}\sum_{z=0,1}\sum_{j=1}^{J}\mathbb{E}\left[\left\{\widehat{Y}_j(z)-\overline{Y}_j(z,a)\right\}^4 \mid A_j=a\right] \\
&=\; o(1).
\end{aligned}$$

□

S3.3.2 Proof of the asymptotic normality

For simplicity, we focus on the case with $m=2$. We only need to show that for any unit vector η with length 4,

$$\eta^\top\sqrt{J}(\widehat{Y}-\overline{Y})=\eta^\top\sqrt{J}(\mu-\overline{Y}+\delta)\xrightarrow{d}N(0,\eta^\top D\eta).$$

Let $X_J=\eta^\top\sqrt{J}(\mu-\overline{Y})$ and $\xi_{J,j}=\eta^\top\sqrt{J}\delta_j$. It suffices to verify the conditions in Lemma S2. We will suppress J in the subscripts when no confusion arises.

First, $\mathcal{F}_{J,0}$ contains the information from the first stage randomization, and $\mathcal{F}_{J,j}$ contains the information from the first stage randomization plus the second stage randomization in the first j clusters. Therefore, $\{\mathcal{F}_{J,j} : j = 0, \ldots, J\}$ is a filtration.

For Lemma S2 condition (a), from the Cauchy–Schwarz inequality, we have

$$\xi_{J,j}^4 \;=\; J^2(\eta^\top \delta_j)^4 \le J^2||\eta||_2^4 \cdot ||\delta_j||_2^4 \;=\; J^2 \cdot ||\delta_j||_2^4.$$

From Lemma S3, we have

$$\sum_{j=1}^{J} \mathbb{E}(\xi_{J,j}^4) \;\le\; J^2 \sum_{j=1}^{J} \mathbb{E}(||\delta_j||_2^4) \;=\; o(1).$$

For Lemma S2 condition (b), we have from Theorem 2,

$$\beta_J^2 \;=\; \mathrm{var}(\xi_J) \;=\; J^{-1}\eta^\top \left\{ \sum_{j=1}^{J} n_j^{-1}\{H_j \circ S_j\} \right\} \eta.$$

Because $\mathbb{E}\{\xi_{J,j}^2 \mid \mathcal{F}_{J,j-1}\} = \mathbb{E}\{\xi_{J,j}^2 \mid \mathcal{F}_{J,0}\} = \mathrm{var}\{\xi_{J,j} \mid \mathcal{F}_{J,0}\}$ and $\mathbb{E}(\xi_{J,j} \mid \mathcal{F}_{J,0}) = 0$, we have

$$\sum_{j=1}^{J} \mathbb{E}\{\xi_{J,j}^2 \mid \mathcal{F}_{J,j-1}\} \;=\; \sum_{j=1}^{J} \mathrm{var}\{\xi_{J,j} \mid \mathcal{F}_{J,0}\} \;=\; \mathrm{var}(\xi_J \mid \mathcal{F}_{J,0})$$

and $\beta_J^2 = \mathbb{E}\{\mathrm{var}(\xi_J \mid \mathcal{F}_{J,0})\}$. Therefore,

$$\mathbb{E}\left[\left\{ \sum_{j=1}^{J} \mathbb{E}(\xi_{J,j}^2 \mid \mathcal{F}_{J,j-1}) - \beta_J^2 \right\}^2 \right] \;=\; \mathrm{var}\left\{ \mathrm{var}(\xi_J \mid \mathcal{F}_{J,0}) \right\}.$$

Therefore, we only need to verify that $\mathrm{var}\left\{\mathrm{var}(\xi_J \mid \mathcal{F}_{J,0})\right\} = o(1)$. Denote

$$\zeta_j(a) \;=\; \mathbb{E}(\xi_{J,j}^2 \mid A_j = a) \;=\; J\eta^\top \mathrm{cov}(\delta_j \mid A_j = a)\eta$$

with mean $\overline{\zeta}(a) = J^{-1}\sum_{j=1}^{J} \zeta_j(a)$, variance $S_{\zeta(a)} = (J-1)^{-1}\sum_{j=1}^{J}\{\zeta_j(a) - \overline{\zeta}(a)\}^2$, and sample mean $\widehat{\zeta}(a) = J_a^{-1}\sum_{j=1}^{J} \zeta_j(a)\mathbf{1}(A_j = a)$ for $a = 1, 2$. We have $\mathrm{var}(\xi_J \mid \mathcal{F}_{J,0}) = J_1\widehat{\zeta}(1) + J_2\widehat{\zeta}(2)$. Theory of simple random sampling implies $\mathrm{var}\{\widehat{\zeta}(a)\} = (1 - q_a)q_a^{-1}S_{\zeta(a)}$. Therefore, we have

$$\begin{aligned}
\mathrm{var}\left\{\mathrm{var}(\xi_J \mid \mathcal{F}_{J,0})\right\} &= \mathrm{var}\left\{ J_1\widehat{\zeta}(1) + J_2\widehat{\zeta}(2) \right\} \\
&\le 2\mathrm{var}\left\{ J_1\widehat{\zeta}(1) \right\} + 2\mathrm{var}\left\{ J_1\widehat{\zeta}(1) \right\} \\
&= 2Jq_1q_2(S_{\zeta(1)} + S_{\zeta(2)}).
\end{aligned}$$

From the Cauchy–Schwarz inequality,

$$\zeta_j^2(a) \;=\; \{\mathbb{E}(\xi_j^2 \mid A_j = a)\}^2 \;\le\; \mathbb{E}(\xi_j^4 \mid A_j = a) \;\le\; J^2\mathbb{E}\{||\delta_j||_2^4 \mid A_j = a).$$

Thus, from Lemma S3, we have

$$(J-1)^{-1}\sum_{j=1}^{J}\{\zeta_j(a)\}^2 \;\le\; (J-1)^{-1}J^2\sum_{j=1}^{J}\mathbb{E}\{||\delta_j||_2^4 \mid A_j = a) \;=\; o(J^{-1}). \tag{S13}$$

From (S11), we have

$$\begin{aligned}\overline{\zeta}(a) &= \eta^\top\left\{\sum_{j=1}^{J}\mathrm{cov}(\delta_j \mid A_j = a)\right\}\eta \\ &= \eta^\top\left\{q_a^{-1}J^{-2}\sum_{j=1}^{J}n_j^{-1}\{H_j(a)\circ S_w\}\right\}\eta \\ &= o(J^{-1}),\end{aligned} \tag{S14}$$

where the last equality follows from Condition 1. Combining (S13) and (S14), we have $S_{\zeta(a)} = o(J^{-1})$ for $a = 1, 2$, which leads to $\mathrm{var}\{\mathrm{var}(\xi_J \mid \mathcal{F}_{J,0})\} = o(1)$.

We then consider Lemma S2 condition (c). From Lemma S1 and Theorem 2, we have $\sqrt{J}(\mu - \overline{Y}) \xrightarrow{d} N(0, H\circ S_b)$ under Condition 1. Thus the convolution of $\mathcal{L}(X_J)$ with $N(0, \eta^\top\left\{\sum_{j=1}^{J}J^{-2}n_j^{-1}(H_j\circ S_w)\right\}\eta)$ converges in distribution to $N(0, \eta^\top D\eta)$. □

S3.4 Proof of Theorem 5

We need the following two lemmas.

Lemma S4 *Suppose that Assumptions 1, 2, 3, and Condition 1 hold. Then* $\widehat{D} - \mathbb{E}\{\widehat{D}\} = o(1)$ *a.s.*

Proof of Lemma S4. Denote

$$\widehat{T}(z, z'; a) \;=\; J_a^{-1}\sum_{j=1}^{J}\widehat{Y}_j(z)\widehat{Y}_j(z')\mathbf{1}(A_j = a).$$

We first show that $\widehat{T}(z, z'; a) - \mathbb{E}\left\{\widehat{T}(z, z'; a)\right\} = o(1)$. It suffices to verify that $\mathrm{cov}\left\{\widehat{T}(z, z'; a)\right\} = o(1)$. Denote $U_j = \widehat{Y}_j(z)\widehat{Y}_j(z')\mathbf{1}(A_j = a)$ and $\mu_j = \mathbb{E}(U_j \mid A_j = a)$. We can write

$$\mathrm{cov}\left\{\widehat{T}(z, z'; a)\right\} \;=\; J_a^{-2}\left\{\sum_{j=1}^{J}\mathrm{cov}(X_j) + \sum_{j\neq k}\mathrm{cov}(X_j, X_k)\right\}.$$

By some algebra, we have

$$\begin{aligned}
\mathbb{E}\{\mathrm{cov}(X_j \mid A_j)\} &= q_a \mathrm{cov}(X_j \mid A_j = a) \;=\; q_a \mathbb{E}(X_j^2 \mid A_j = a) - q_a \mu_j^2, \\
\mathrm{cov}\{\mathbb{E}(X_j \mid A_j)\} &= \mathbb{E}\left[\{\mathbb{E}(X_j \mid A_j)\}^2\right] - \{\mathbb{E}(X_j)\}^2 \\
&= q_a \{\mathbb{E}(X_j \mid A_j = a)\}^2 - q_a^2 \mu_j^2 \\
&= q_a \mu_j^2 - q_a^2 \mu_j^2 \\
\mathbb{E}\{\mathbb{E}(X_j \mid A_j)\mathbb{E}(X_k \mid A_k)\} &= \Pr(A_k = A_j = a)\mathbb{E}(X_j \mid A_j = a)\mathbb{E}(X_k \mid A_k = a) \\
&= q_a \frac{J_a - 1}{J - 1} \mu_j \mu_k, \\
\mathbb{E}(X_j)\mathbb{E}(X_k) &= q_a^2 \mu_j \mu_k.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{cov}(X_j) &= \mathbb{E}\{\mathrm{cov}(X_j \mid A_j)\} + \mathrm{cov}\{\mathbb{E}(X_j \mid A_j)\} \\
&= q_a \mathbb{E}(X_j^2 \mid A_j = a) - q_a^2 \mu_j^2, \\
\mathrm{cov}(X_j, X_k) &= \mathrm{cov}\{\mathbb{E}(X_j \mid A_j), \mathbb{E}(X_k \mid A_k)\} + \mathbb{E}\{\mathrm{cov}(X_j, X_k \mid A_j, A_k)\} \\
&= \mathrm{cov}\{\mathbb{E}(X_j \mid A_j), \mathbb{E}(X_k \mid A_k)\} \\
&= \mathbb{E}\{\mathbb{E}(X_j \mid A_j)\mathbb{E}(X_k \mid A_k)\} - \mathbb{E}(X_j)\mathbb{E}(X_k) \\
&= -(J-1)^{-1} q_a (1 - q_a) \mu_j \mu_k.
\end{aligned}$$

As a result, we have

$$\begin{aligned}
& J_a^2 \mathrm{cov}\left\{\widehat{T}(z, z'; a)\right\} \\
= \;& q_a \sum_{j=1}^{J} \mathbb{E}(X_j^2 \mid A_j = a) - q_a^2 \sum_{j=1}^{J} \mu_j^2 - \frac{q_a(1 - q_a)}{J - 1} \sum_{j \neq k} \mu_j \mu_k \\
= \;& q_a \sum_{j=1}^{J} \mathbb{E}(X_j^2 \mid A_j = a) - q_a^2 \sum_{j=1}^{J} \mu_j^2 + \frac{q_a(1 - q_a)}{J - 1} \sum_{j=1}^{J} \mu_j^2 - \frac{q_a(1 - q_a)}{J - 1} \sum_{j,k} \mu_j \mu_k \\
\leq \;& q_a \sum_{j=1}^{J} \mathbb{E}(X_j^2 \mid A_j = a) - \left\{q_a^2 - \frac{q_a(1 - q_a)}{J - 1}\right\} \sum_{j=1}^{J} \mu_j^2.
\end{aligned}$$

When J goes to infinity, we can obtain

$$J_a^2 \mathrm{cov}\left\{\widehat{T}(z, z'; a)\right\} \;\leq\; q_a \sum_{j=1}^{J} \mathbb{E}(X_j^2 \mid A_j = a) \;=\; q_a \sum_{j=1}^{J} \mathbb{E}\left\{\widehat{Y}_j^2(z) \widehat{Y}_j^2(z') \mid A_j = a\right\}.$$

From $\widehat{Y}_j^2(z)\widehat{Y}_j^2(z') \leq 2\widehat{Y}_j^4(z) + 2\widehat{Y}_j^4(z')$ and (S12), we then have

$$J_a^2 \text{cov}\left\{\widehat{T}(z,z';a)\right\} \;\leq\; 2q_a \sum_{j=1}^{J} \mathbb{E}\left\{\widehat{Y}_j^4(z) + \widehat{Y}_j^4(z') \mid A_j = a\right\} \;=\; o(1),$$

where the last equality follows from a similar argument in the proof of Lemma S3. Therefore, we have $\widehat{T}(z,z';a) - \mathbb{E}\left\{\widehat{T}(z,z';a)\right\} = o(1)$.

We then show that $\widehat{Y} - \overline{Y} = o(1)$. From Theorem 2, we have

$$\text{cov}(\widehat{Y}) \;=\; J^{-1}(H \circ S_b) + J^{-2}\sum_{j=1}^{J} n_j^{-1}\{H_j \circ S_j\} \;=\; o(1),$$

where the last equality follows from Condition 1.

Finally, we prove that $\widehat{D} - \mathbb{E}(\widehat{D}) = o(1)$ as J goes to infinity. The elements of $\widehat{D}$ are $J/J_a\widehat{\sigma}^2(z,z';a)$, which can be written as

$$q_a^{-1}\left\{\widehat{T}(z,z';a) - \widehat{Y}(z,a)\widehat{Y}(z,a')\right\},$$

where we ignore the difference between J_a and $J_a - 1$. Therefore, $J/J_a\widehat{\sigma}^2(z,z';a)$ converges to

$$q_a^{-1}\left[\mathbb{E}\left\{\widehat{T}(z,z';a)\right\} - \overline{Y}(z,a)\overline{Y}(z,a')\right],$$

which is equal to $J/J_a\mathbb{E}\left\{\widehat{\sigma}^2(z,z';a)\right\}$. □

LEMMA S5 *(i) If $X \sim N_k(0,A)$, Then $X^\top BX \stackrel{d}{=} \sum_{j=1}^k \lambda_j(AB)\xi_j^2$, where the $\lambda_j(AB)$'s are eigenvalues of AB, and $\xi_j \sim \chi^2(1)$ and are i.i.d. with each other.*
(ii) If $X_n \stackrel{d}{\to} N_k(0,A)$, and $B_n \stackrel{p}{\to} B$, then $X_n^\top B_n X_n \stackrel{d}{=} \sum_{j=1}^k \lambda_j(AB^{-1})\xi_j^2$. If $B - A$ is positive semidefinite, then $0 \leq \lambda_j(AB^{-1}) \leq 1$ for all j.

Proof of Lemma S5. Lemma S5(i) follows form linear algebra and Lemma S5(ii) follows from Slutsky's Theorem. □

We then prove Theorem 5. From Theorem 4 and Lemma S4, we know that $\sqrt{J}(C\widehat{Y} - x) \stackrel{d}{\to} \text{N}_{2m}(0, CDC^\top)$, $C\widehat{D}C^\top \stackrel{p}{\to} C\mathbb{E}(\widehat{D})C^\top$. Because $C\mathbb{E}(\widehat{D})C^\top - CDC^\top$ is positive semi-definite, from Lemma S5(ii), we have $T \stackrel{d}{=} \sum_{j=1}^k \lambda_j\xi_j^2$, where k is the rank of C and $0 \leq \lambda_j \leq 1$ for all j. □

S3.5 Proof of Theorem 6

To prove Theorem 6, we need the following lemma.

LEMMA S6 *Suppose* $(X_1, \ldots, X_k)$ *follows a standard multivariate normal distribution. If* $0 < a_j \leq a'_j$ *for* $j = 1, \ldots, k$*, then as* J *goes to infinity,*

$$\Pr\left\{\sum_{j=1}^{k}\left(a'_j X_j + \sqrt{J}x_j\right)^2 \geq t\right\} \geq p$$

implies

$$\Pr\left\{\sum_{j=1}^{k}\left(a_j X_j + \sqrt{J}x_j\right)^2 \geq t\right\} \geq p$$

where x_j*'s,* t*, and* p *are arbitrary non-zero constants.*

Proof of Lemma S6. Without loss of generality, we can assume $x_j > 0$ for all j. Since X_j's are independent of each other, it suffices to show that

$$\Pr\left\{\left(a'_j X_j + \sqrt{J}x_j\right)^2 \geq t\right\} \geq p \tag{S15}$$

implies

$$\Pr\left\{\left(a_j X_j + \sqrt{J}x_j\right)^2 \geq t\right\} \geq p \tag{S16}$$

for all j. By some algebra, (S15) is equivalent to

$$\Phi\left(\frac{\sqrt{J}x_j - \sqrt{t}}{a'_j}\right) + \Phi\left(\frac{-\sqrt{J}x_j - \sqrt{t}}{a'_j}\right) \geq p. \tag{S17}$$

As J goes to infinity, the second term on the left-hand side of (S17) goes to 0. Therefore, we can write (S17) as

$$\Phi\left(\frac{\sqrt{J}x_j - \sqrt{t}}{a'_j}\right) \geq p. \tag{S18}$$

Similarly, we can show that (S16) is equivalent to

$$\Phi\left(\frac{\sqrt{J}x_j - \sqrt{t}}{a_j}\right) \geq p. \tag{S19}$$

Because $a'_j \geq a_j$, (S18) implies (S19). This completes the proof. $\square$

We now prove Theorem 6. The number of clusters requires for the test to have power $1-\beta$ should satisfy

$$\Pr\{J(C\widehat{Y})^\top (C\widehat{D}C^\top)^{-1}(C\widehat{Y}) \geq \chi^2_{1-\alpha}(k) \mid C\overline{Y} = x\} \geq 1-\beta.$$

Theorem 4 implies

$$\sqrt{J}(C\widehat{Y} - x) \xrightarrow{d} \mathrm{N}_k(0, CDC^\top)$$

Therefore, we can write $C\widehat{Y} = 1/\sqrt{J} \cdot (CDC^\top)^{1/2} \cdot W_k + x$, where W_k is a k-length vector following a standard multivariate normal distribution. As a result, we can write the test statistic as

$$\{(CDC^\top)^{1/2}W_k + x\}^\top (C\widehat{D}C^\top)^{-1}\{(CDC^\top)^{1/2}W_k + x\}$$

By Slutsky's theorem, it has the same asymptotic distribution as

$$\begin{aligned} T' &= \{(CDC^\top)^{1/2}W_k + \sqrt{J}x\}^\top \{C\mathbb{E}(\widehat{D})C^\top\}^{-1}\{(CDC^\top)^{1/2}W_k + \sqrt{J}x\} \\ &= [\{C\mathbb{E}(\widehat{D})C^\top\}^{-1/2}(CDC^\top)^{1/2}W_k + \sqrt{J}\{C\mathbb{E}(\widehat{D})C^\top\}^{-1/2}x]^\top \\ &\quad \cdot [\{C\mathbb{E}(\widehat{D})C^\top\}^{-1/2}(CDC^\top)^{1/2}W_k + \sqrt{J}\{C\mathbb{E}(\widehat{D})C^\top\}^{-1/2}]. \end{aligned}$$

From the matrix theory, we can write $(CDC^\top)^{1/2}\{C\mathbb{E}(\widehat{D})C^\top\}^{-1}(CDC^\top)^{1/2} = P^\top \Lambda P$, where P is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ is a diagonal matrix. Because $D_0 - D$ is positive semidefinite, $0 \leq \lambda_j \leq 1$ for all j. Denote $U = (U_1, \ldots, U_m) = PW$, which also follows a standard multivariate normal distribution. Then, we can write

$$\begin{aligned} T' &= \left[\Lambda^{1/2}U + \sqrt{J}\{C\mathbb{E}(\widehat{D})C^\top\}^{-1/2}x\right]^\top \left[\Lambda^{1/2}U + \sqrt{J}\{C\mathbb{E}(\widehat{D})C^\top\}^{-1/2}x\right] \\ &= \sum_{j=1}^{k}(\sqrt{\lambda_j}U_j + \sqrt{J}x'_j)^2, \end{aligned}$$

where x'_j is the j-th element of $\{C\mathbb{E}(\widehat{D})C^\top\}^{-1/2}x$. From Lemma S6, $\Pr(T' \geq t) \geq 1-\beta$ is implied by

$$\Pr\left\{\sum_{j=1}^{k}(U_j + \sqrt{J}x'_j)^2 \geq t\right\} \geq 1-\beta. \tag{S20}$$

Based on the definition of $s^2(q, 1-\beta, k)$, (S20) is equivalent to

$$J \sum_{j=1}^{k} x_j'^2 \geq s^2(\chi^2_{1-\alpha}(k), 1-\beta, k).$$

Because $\sum_{j=1}^{k} x_j'^2 = x^\top \{C\mathbb{E}(\widehat{D})C^\top\}x$, we obtain the sample size formula,

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m)}{x^\top \{C\mathbb{E}(\widehat{D})C^\top\}^{-1}x}.$$

□

S3.6 Proof of Theorem 7

We first derive the expression of $\mathbb{E}(\widehat{D})$ under Assumption 4. From Appendix S3.2, we have

$$\begin{aligned}
\mathbb{E}\{\widehat{\sigma}_b^2(1,a)\} &= \sigma_b^2(1,a) + \frac{1}{J}\sum_{j=1}^{J} \frac{1}{n_{j1}}\left(1 - \frac{n_{j1}}{n_j}\right)\sigma_j^2(1,a) \\
&= \sigma_b^2 + \frac{1-p_a}{np_a}\sigma_w^2 \\
&= \left\{r + \frac{(1-p_a)(1-r)}{np_a}\right\}\sigma^2,
\end{aligned}$$

where the second equality follows from conditions (a), (b), and (c) of Assumption 4. Similarly, we obtain

$$\begin{aligned}
\mathbb{E}\{\widehat{\sigma}_b^2(0,a)\} &= \sigma_b^2(0,a) + \frac{1}{J}\sum_{j=1}^{J} \frac{1}{n_{j0}}\left(1 - \frac{n_{j0}}{n_j}\right)\sigma_j^2(0,a) \\
&= \sigma_b^2 + \frac{p_a}{n(1-p_a)}\sigma_w^2 \\
&= \left\{r + \frac{p_a(1-r)}{n(1-p_a)}\right\}\sigma^2
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\{\widehat{\sigma}_b^2(1,0;a)\} &= \sigma_b^2(1,0;a) - \frac{1}{J}\sum_{j=1}^{J} \frac{\sigma_j^2(1,0;a)}{n_j} \\
&= \rho\sigma_b^2 - \frac{\rho\sigma_w^2}{n} \\
&= \rho\left(r - \frac{1-r}{n}\right)\cdot\sigma^2.
\end{aligned}$$

Therefore, under Assumption 4, $\mathbb{E}(\widehat{D}) = D_0^* = \sigma^2 \cdot \mathrm{diag}(D_{01}^*, D_{02}^*, \ldots, D_{0m}^*)$, where

$$D_{0a}^* = \frac{1}{q_a}\begin{pmatrix} r + \frac{(1-p_a)(1-r)}{np_a} & \rho\left(r - \frac{1-r}{n}\right) \\ \rho\left(r - \frac{1-r}{n}\right) & r + \frac{p_a(1-r)}{n(1-p_a)} \end{pmatrix}$$

for $a = 1, \ldots, m$.

We next prove the sample size formula. From Theorem 6, the number of clusters required for detecting the alternative hypothesis $H_1^{\mathrm{de}} : \mathrm{ADE} = x$ with power $1 - \beta$ based on T_{de} is given as,

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m)}{x^\top \{C_1 \mathbb{E}(\widehat{D}) C_1^\top\}^{-1} x},$$

which, under Assumption 4, is equivalent to

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m) \cdot \sigma^2}{x^\top \{C_1 D_0^* C_1^\top\}^{-1} x}.$$

Therefore, under the alternative hypothesis $H_1 : |\mathrm{ADE}| = \mu$ for all a, the sample size formula is

$$\begin{aligned} J &\geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m) \cdot \sigma^2}{\mu^2 \cdot \mathbf{1}_m^\top \{C_1 D_0^* C_1^\top\}^{-1} \mathbf{1}_m} \\ &= \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m) \cdot \sigma^2}{\mu^2} \cdot \frac{1}{\sum_{a=1}^m \{(1,-1) D_{0a}^* (1,-1)^\top\}^{-1}}. \end{aligned}$$

Under $r \geq 1/(n+1)$, we have $(1,-1) D_{0a} (1,-1)^\top \geq (1,-1) D_{0a}^* (1,-1)^\top$. Thus, a more conservative sample size formula is given as,

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m) \cdot \sigma^2}{\mu^2} \cdot \frac{1}{\sum_{a=1}^m \{(1,-1) D_{0a} (1,-1)^\top\}^{-1}}.$$

□

S3.7 Proof of Theorem 8

From Theorem 6, the number of clusters required for detecting the alternative hypothesis $H_1^{\mathrm{mde}} : \mathrm{MDE} = \mu$ with power $1 - \beta$ based on T_{mde} is given as,

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m)}{\mu^2 \{C_2 \mathbb{E}(\widehat{D}) C_2^\top\}^{-1}},$$

which, under Assumption 4, is equivalent to

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(1), 1-\beta, 1) \cdot \sigma^2}{\mu^2} \sum_{a=1}^m q_a^2 \left\{(1,-1) D_{0a}^* (1,-1)^\top\right\}.$$

Under $r \geq 1/(n+1)$, we have $(1,-1) D_{0a} (1,-1)^\top \geq (1,-1) D_{0a}^* (1,-1)^\top$. Thus, a more conservative sample size formula is given as,

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m) \cdot \sigma^2}{\mu^2} \cdot \sum_{a=1}^m q_a^2 \left\{(1,-1) D_{0a} (1,-1)^\top\right\}.$$

□

S3.8 Proof of Theorem 9

From Theorem 6, the number of clusters required for detecting the alternative hypothesis $\text{ASE} = x$ with power $1-\beta$ based on T_{se} is given as,

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(2m-2), 1-\beta, 2m-2)}{x^\top \{C_3 \mathbb{E}(\widehat{D}) C_3^\top\}^{-1} x},$$

which, under Assumption 4, is equivalent to

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m) \cdot \sigma^2}{x^\top \{C_3 D_0^* C_3^\top\}^{-1} x}.$$

Therefore, under the alternative hypothesis $H_1^{\text{se}} : \max_{a \neq a'} |\text{ASE}(z; a, a')| = \mu$ for $z = 0, 1$, the sample size formula is

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(2m-2), 1-\beta, 2m-2) \cdot \sigma^2}{\mu^2 \cdot \min_{s \in \mathcal{S}} s^\top \{C_3 D_0^* C_3^\top\}^{-1} s},$$

where $\mathcal{S}$ is the set of $s = (\text{ASE}(0; 1, 2), \text{ASE}(0; 2, 3), \ldots, \text{ASE}(0; m-1, m), \text{ASE}(1; 1, 2), \text{ASE}(1; 2, 3), \ldots,$ $\text{ASE}(1; m-1, m))$ satisfying $\max_{z, a \neq a'} |\text{ASE}(a, a')| = 1$ for $z = 0, 1$. □

S3.9 Proof of Theorem S1

We first prove the equivalence between the point estimators. The OLS estimate can be written as,

$$\widehat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \boldsymbol{Y}.$$

Because the columns of $\mathbf{X}$ are orthogonal to each other, we can consider each element of $\widehat{\boldsymbol{\beta}}$ separately. Therefore, we have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{za} &= \left\{ \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} = z, A_j = a) w_{ij} \right\}^{-1} \left\{ \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} = z, A_j = a) w_{ij} Y_{ij} \right\} \\
&= \sum_{j=1}^{J} \sum_{i=1}^{n_j} \frac{1}{J_a n_{jz}} \cdot \mathbf{1}(Z_{ij} = z, A_j = a) Y_{ij} \\
&= \widehat{Y}(z, a).
\end{aligned}$$

We then prove the equivalence between the variance estimators. Recall the variance estimator,

$$\widehat{\text{var}}_{\text{hc2}}^{\text{cluster}}(\widehat{\boldsymbol{\beta}}) \;=\; (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \left\{ \sum_j \mathbf{X}_j^\top \mathbf{W}_j (\mathbf{I}_{n_j} - \mathbf{P}_j)^{-1/2} \widehat{\boldsymbol{\epsilon}}_j \widehat{\boldsymbol{\epsilon}}_j^\top (\mathbf{I}_{n_j} - \mathbf{P}_j)^{-1/2} \mathbf{W}_j \mathbf{X}_j \right\} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1},$$

where $\mathbf{I}_{n_j}$ is the $n_j \times n_j$ identity matrix and $\mathbf{P}_j$ is the following cluster leverage matrix,

$$\mathbf{P}_j \;=\; \mathbf{W}_j^{1/2}\mathbf{X}_j(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_j^\top\mathbf{W}_j^{1/2}.$$

Without loss of generality, suppose $A_j = 1$. We have

$$\begin{aligned}
(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} &= \boldsymbol{I}_{n\times n},\\
\mathbf{P}_j &= \mathbf{W}_j^{1/2}\mathbf{X}_j(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_j^\top\mathbf{W}_j^{1/2}\\
&= \begin{pmatrix} \frac{1}{\sqrt{J_1 n_{j1}}}\mathbf{1}_{n_{j1}} & \mathbf{0}_{n_{j1}} \\ \mathbf{0}_{n_{j0}} & \frac{1}{\sqrt{J_1 n_{j0}}}\mathbf{1}_{n_{j0}} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{J_1 n_{j1}}}\mathbf{1}_{n_{j1}} & \mathbf{0}_{n_{j1}} \\ \mathbf{0}_{n_{j0}} & \frac{1}{\sqrt{J_1 n_{j0}}}\mathbf{1}_{n_{j0}} \end{pmatrix}^\top\\
&= \begin{pmatrix} \frac{1}{J_a n_{j1}}\mathbf{1}_{n_{j1}\times n_{j1}} & \mathbf{0}_{n_{j1}\times n_{j0}} \\ \mathbf{0}_{n_{j0}\times n_{j1}} & \frac{1}{J_a n_{j0}}\mathbf{1}_{n_{j0}\times n_{j0}} \end{pmatrix},
\end{aligned}$$

where $\boldsymbol{I}_k$ is an k-dimensional identity matrix, $\mathbf{1}_k$ ($\mathbf{0}_k$) is an k-dimensional vector of ones (zeros) and $\mathbf{1}_{k_1\times k_2}$ ($\mathbf{0}_{k_1\times k_2}$) is an $k_1 \times k_2$ dimensional matrix of ones (zeros).

Since $(\mathbf{1}_{n_{j1}}^\top, \mathbf{0}_{n_{j0}}^\top)^\top$ and $(\mathbf{0}_{n_{j1}}^\top, \mathbf{1}_{n_{j0}}^\top)^\top$ are two eigenvectors of $\boldsymbol{I}_{n_j} - \mathbf{P}_j$ whose eigenvalue is $(J_1 - 1)/J_1$, we have,

$$\begin{aligned}
(\boldsymbol{I}_{n_j} - \mathbf{P}_j)^{-1/2}(\mathbf{1}_{n_{j1}}^\top, \mathbf{0}_{n_{j0}}^\top)^\top &= \sqrt{\frac{J_1}{J_1-1}}(\mathbf{1}_{n_{j1}}^\top, \mathbf{0}_{n_{j0}}^\top)^\top,\\
(\boldsymbol{I}_{n_j} - \mathbf{P}_j)^{-1/2}(\mathbf{0}_{n_{j1}}^\top, \mathbf{1}_{n_{j0}}^\top)^\top &= \sqrt{\frac{J_1}{J_1-1}}(\mathbf{0}_{n_{j1}}^\top, \mathbf{1}_{n_{j0}}^\top)^\top.
\end{aligned}$$

Thus,

$$(\boldsymbol{I}_{n_j} - \mathbf{P}_j)^{-1/2}\mathbf{W}_j\mathbf{X}_j \;=\; \sqrt{\frac{J_1}{J_1-1}}\begin{pmatrix} \frac{1}{J_1 n_{j1}}\mathbf{1}_{n_{j1}} & \mathbf{0}_{n_{j1}} & \mathbf{0}_{n_{j1}\times(2m-2)} \\ \mathbf{0}_{n_{j0}} & \frac{1}{J_1 n_{j0}}\mathbf{1}_{n_{j0}} & \mathbf{0}_{n_{j0}\times(2m-2)} \end{pmatrix}.$$

For a unit with $(A_j = 1, Z_{ij} = 1)$, we have $\widehat{\epsilon}_{ij} = Y_{ij} - \widehat{\beta}_{11} = Y_{ij} - \widehat{Y}(1,1)$, and for a unit with $(A_j = 1, Z_{ij} = 0)$, we have $\widehat{\epsilon}_{ij} = Y_{ij} - \widehat{\alpha}_{01} = Y_{ij} - \widehat{Y}(0,1)$. As a result,

$$\begin{aligned}
&\widehat{\boldsymbol{\epsilon}}_j^\top(\boldsymbol{I}_{n_j} - \mathbf{P}_j)^{-1/2}\mathbf{W}_j\mathbf{X}_j\\
=\;& \sqrt{\frac{J_1}{J_1-1}}(Y_{1j} - \widehat{Y}(1,1), \ldots, Y_{n_j j} - \widehat{Y}(0,1))\begin{pmatrix} \frac{1}{J_1 n_{j1}}\mathbf{1}_{n_{j1}} & \mathbf{0}_{n_{j1}} & \mathbf{0}_{n_{j1}\times(2m-2)} \\ \mathbf{0}_{n_{j0}} & \frac{1}{J_1 n_{j0}}\mathbf{1}_{n_{j0}} & \mathbf{0}_{n_{j0}\times(2m-2)} \end{pmatrix},
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\frac{J_1}{J_1-1}} \begin{pmatrix} \frac{1}{J_1 n_{j1}} \left\{ \sum_{i=1}^{n_j} Y_{ij} Z_{ij} - n_{j1} \widehat{Y}(1,1) \right\} \\ \frac{1}{J_1 n_{j0}} \left\{ \sum_{i=1}^{n_j} Y_{ij}(1-Z_{ij}) - n_{j0} \widehat{Y}(0,1) \right\} \end{pmatrix}^\top \\
&= \sqrt{\frac{1}{J_1(J_1-1)}} \left(\widehat{Y}_j(1) - \widehat{Y}(1,1), \widehat{Y}_j(0) - \widehat{Y}(0,1), \mathbf{0}_{2m-2}^\top \right).
\end{aligned}
$$

Similar result applies for $A_j = a$, where $a = 1 \ldots, J$. Therefore, $\widehat{\text{var}}_{\text{hc2}}^{\text{cluster}}(\widehat{\boldsymbol{\beta}})$ is a block diagonal matrix with the a-th block

$$
\begin{aligned}
&\frac{1}{J_a(J_a-1)} \sum_{j=1}^{J} \mathbf{1}(A_j = a) \left(\widehat{Y}_j(1) - \widehat{Y}(1,a), \widehat{Y}_j(0) - \widehat{Y}(0,a) \right) \left(\widehat{Y}_j(1) - \widehat{Y}(1,a), \widehat{Y}_j(0) - \widehat{Y}(0,a) \right)^\top \\
= &\begin{pmatrix} \frac{\sum_{i=1}^{J} \left\{ \widehat{Y}_j(1) - \widehat{Y}(1,a) \right\}^2 \mathbf{1}(A_j = a)}{J_a(J_a-1)} & \frac{\sum_{i=1}^{J} \left\{ \widehat{Y}_j(1,a) - \widehat{Y}(1,a) \right\} \left\{ \widehat{Y}_j(0,a) - \widehat{Y}(0,a) \right\} \mathbf{1}(A_j = a)}{J_a(J_a-1)} \\ \frac{\sum_{i=1}^{J} \left\{ \widehat{Y}_j(1,a) - \widehat{Y}(1,a) \right\} \left\{ \widehat{Y}_j(0,a) - \widehat{Y}(0,a) \right\} \mathbf{1}(A_j = a)}{J_a(J_a-1)} & \frac{\sum_{i=1}^{J} \left\{ \widehat{Y}_j(0) - \widehat{Y}(0,a) \right\}^2 \mathbf{1}(A_j = a)}{J_a(J_a-1)} \end{pmatrix} \\
= &\frac{\widehat{D}}{J}.
\end{aligned}
$$

□

S3.10 Proof of Theorem S2

First, we calculate the variance of $\widehat{\text{ATE}}$ under the two-stage randomized design. In this case, $\widehat{\text{ATE}}$ is the same as $\widehat{\text{ADE}}$. From Theorem 2, we have

$$
\text{var}\left(\widehat{\text{ADE}} \right) = \sum_{a=1}^{m} \frac{J_a^2}{J^2} \cdot \text{var}\left\{ \widehat{\text{ADE}}(a) \right\} + \sum_{a \neq a'} \frac{J_a J_{a'}}{J^2} \cdot \text{cov}\left\{ \widehat{\text{ADE}}(a), \widehat{\text{ADE}}(a') \right\}.
$$

When there is no interference, we have

$$
\begin{aligned}
&\text{var}\left\{ \widehat{\mathsf{DEY}}(a) \right\} \\
= &\left(1 - \frac{J_a}{J}\right) \frac{\tau_b^2}{J_a} + \frac{1}{J_a J} \sum_{j=1}^{J} \left\{ \frac{\sum_{i=1}^{n} (Y_{ij}(1) - \overline{Y}_j(1))^2}{(n-1) n p_a} + \frac{\sum_{i=1}^{n} (Y_{ij}(0) - \overline{Y}_j(0))^2}{(n-1) n (1-p_a)} - \frac{\sum_{i=1}^{n} (\text{ATE}_{ij} - \text{ATE}_j)^2}{(n-1) n} \right\} \\
= &\left(1 - \frac{J_a}{J}\right) \frac{\tau_b^2}{J_a} + \frac{nJ-1}{(n-1) n J_a J} \left\{ \frac{\eta_w^2(1)}{p_a} + \frac{\eta_w^2(0)}{1-p_a} - \tau_w^2 \right\}
\end{aligned}
$$

and $\text{cov}\left\{ \widehat{\text{ADE}}(a), \widehat{\text{ADE}}(a') \right\} = -\tau_b^2 / J$. Therefore, we can obtain

$$
\begin{aligned}
&\text{var}\left(\widehat{\text{ADE}} \right) \\
= &\sum_{a=1}^{m} J_a \left(1 - \frac{J_a}{J}\right) \frac{\tau_b^2}{J^2} - \sum_{a \neq a'}^{m} \frac{J_a J_{a'}}{J^2} \cdot \frac{\tau_b^2}{J} + \sum_{a=1}^{m} \frac{J_a^2}{J^2} \cdot \frac{nJ-1}{(n-1) n J_a J} \left\{ \frac{\eta_w^2(1)}{p_a} + \frac{\eta_w^2(0)}{1-p_a} - \tau_w^2 \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{nJ-1}{J^3(n-1)}\sum_{a=1}^{m}\frac{J_a}{np_a}\cdot\eta_w^2(1)+\frac{nJ-1}{J^3(n-1)}\sum_{a=1}^{m}\frac{J_a}{n(1-p_a)}\cdot\eta_w^2(0)-\frac{nJ-1}{J^3(n-1)}\sum_{a=1}^{m}\frac{J_a}{n}\cdot\tau_w^2 \\
&= \frac{(nJ-1)(1-r)}{nJ^3}\left\{\sum_{a=1}^{m}\frac{J_a}{np_a}\cdot\eta^2(1)+\sum_{a=1}^{m}\frac{J_a}{n(1-p_a)}\cdot\eta^2(0)-\sum_{a=1}^{m}\frac{J_a}{n}\cdot\tau^2\right\} \\
&\approx \frac{1-r}{J^2}\sum_{a=1}^{m}\frac{J_a}{np_a}\cdot\eta^2(1)+\frac{1-r}{J^2}\sum_{a=1}^{m}\frac{J_a}{n(1-p_a)}\cdot\eta^2(0)-\frac{1-r}{nJ}\cdot\tau^2,
\end{aligned}
$$

where the last line follows from the approximation assumptions in equation (S3).

Second, the variance of $\widehat{\mathrm{ATE}}$ under the completely randomized experiment with the number of the treated units equal to $\sum_{a=1}^{m} J_a np_a$ is given as,

$$
\frac{1}{\sum_{a=1}^{m} J_a np_a}\cdot\eta^2(1)+\frac{1}{\sum_{a=1}^{m} J_a n(1-p_a)}\cdot\eta^2(0)-\frac{1}{Jn}\cdot\tau^2.
$$

Third, we calculate the variance of $\widehat{\mathrm{ATE}}$ under cluster randomized experiments with the same number of treated units. In the cluster randomized experiments, the units in each cluster get the same treatment condition. Thus, the number of the treated clusters is $\sum_{a=1}^{m} J_a p_a$. As a result, the variance of $\widehat{\mathrm{ATE}}$ is given as,

$$
\begin{aligned}
&\frac{\eta_b^2(1)}{\sum_{a=1}^{m} J_a p_a}+\frac{\eta_b^2(0)}{\sum_{a=1}^{m} J_a(1-p_a)}-\frac{\tau_b^2}{J}\cdot \\
\approx\ &\frac{1+(n-1)r}{\sum_{a=1}^{m} J_a np_a}\cdot\eta^2(1)+\frac{1+(n-1)r}{\sum_{a=1}^{m} J_a n(1-p_a)}\cdot\eta^2(0)-\frac{1+(n-1)r}{nJ}\cdot\tau^2,
\end{aligned}
$$

where the last line follows from the approximation assumptions in equation (S3). □

S4 Computational details

We provide a strategy for numerically calculating the required number of clusters in Theorem 9. We focus on the following optimization problem,

$$
\min_{s\in\mathcal{S}} s^\top\{C_3 D_0 C_3^\top\}^{-1}s,
$$

where $a=(\mathrm{ASE}(0;1,2),\mathrm{ASE}(0;2,3),\ldots,\mathrm{ASE}(0;m-1,m),\mathrm{ASE}(1;1,2),\mathrm{ASE}(1;2,3),\ldots,\mathrm{ASE}(1;m-1,m))$ satisfies the constraint $\max_{a\neq a'}|\mathrm{ASE}(z;a,a')|=1$ for $z=0,1$.

We consider all the possible cases in which $\max_{a\neq a'}|\mathrm{ASE}(z;a,a')|=1$ holds for $z=0,1$. First, using quadratic programming, we can obtain the minimum of $s^\top\{C_3 D_0 C_3^\top\}^{-1}s$ under the constraint

$\text{ASE}(1;1,2) = 1$, $\text{ASE}(0;1,2) = 1$ and $-1 \leq \text{ASE}(z;a,a') \leq 1$ for all z, a, a'. We denote it by $l(1,2;1,2)$. Similarly, we can obtain $l(a_1, a_1'; a_0, a_0')$ for all a_1, a_1', a_0, a_0' by implementing this procedure for each of the possible cases satisfying $\max_{a \neq a'} |\text{ASE}(z;a,a')| = 1$ for $z = 0, 1$. As a result, the solution to the optimization problem is $\min l(a_1, a_1'; a_0, a_0')$.

S5 Simulation Studies

We conduct simulation studies to evaluate the empirical performance of the sample size formulas for the direct, marginal direct, and spillover effects. We consider a two-stage randomized experiment with three different treatment assignment mechanisms ($m = 4$), under which the treated proportions are 20%, 40%, 60%, and 80%, respectively. We generate the treatment assignment mechanism A_j with $\Pr(A_j = a) = 1/4$ for $a = 1, 2, 3, 4$ such that $J_a = J/4$. We then completely randomize the treatment assignment Z_{ij} within each cluster according to the selected assignment mechanism.

Our data generating process is as follows. First, we generate the cluster-level average potential outcomes as,

$$\overline{Y}_j(0,a) \sim \text{N}(\theta_{0a}, \sigma_b^2), \quad \overline{Y}_j(1,a) \sim \text{N}(\theta_{1a} + \rho\{\overline{Y}_j(0,a) - \theta_{0a}\}, (1-\rho^2)\sigma_b^2)$$

for $a = 1, 2, 3, 4$. Second, we generate the individual-level average potential outcomes $Y_{ij}(z,a)$ as,

$$\begin{pmatrix} Y_{ij}(1,a) \\ Y_{ij}(0,a) \end{pmatrix} \sim \text{N}_2 \left(\begin{pmatrix} \overline{Y}_j(1,a) \\ \overline{Y}_j(0,a) \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \rho\sigma_w^2 \\ \rho\sigma_w^2 & \sigma_w^2 \end{pmatrix} \right)$$

for $a = 1, 2, 3, 4$. In this super population setting, the direct effect under treatment assignment mechanism a is given by $\theta_{1a} - \theta_{0a}$ for $a = 1, 2, 3, 4$, whereas the marginal direct effect equals $(\theta_{11} + \theta_{12} + \theta_{13} + \theta_{14})/4 - (\theta_{01} + \theta_{02} + \theta_{03} + \theta_{04})/4$. The spillover effect comparing treatment assignment mechanisms a and a' under treatment condition z is $\theta_{za} - \theta_{za'}$ for $z = 0, 1$ and $a, a' = 1, 2, 3, 4$. However, our target causal quantities of interest are finite-sample causal effects ($\text{ADE}(a)$, MDE, $\text{ASE}(z;a,a')$), which generally do not equal their super-population counterparts due to sample variation. Therefore, we center the generated potential outcomes so that the finite-sample and super-population causal effects are equal to one another, i.e., $\overline{Y}(z,a) = \theta_{za}$ for $z = 0, 1$ and $a = 1, 2, 3, 4$.

We choose different values of θ's based on the different alternative hypotheses for our three causal effects of interest. For the direct effect, we generate θ_{0a} $(a = 1, 2, 3, 4)$ from a uniform distribution on the interval $[-0.3, 0.3]$, and set $\theta_{1a} = 0.3 + \theta_{0a}$ for all a; the generated potential outcomes satisfy $|\text{ADE}(a)| = 0.3$ for all a. For the marginal direct effect, we generate θ_{0a} $(a = 1, 2, 3, 4)$ from a uniform distribution on the interval $[-0.3, 0.3]$ and set $\theta_{11} = 0.12 + \theta_{01}$, $\theta_{12} = 0.48 + \theta_{02}$, $\theta_{13} = 0.24 + \theta_{03}$, and $\theta_{14} = 0.36 + \theta_{04}$; the generated potential outcomes satisfy MDE $= 0.3$. For the spillover effect, we generate θ_{0a} and θ_{1a} from a uniform distribution on the interval $[-0.15, 0.15]$ for $(a = 1, 2, 3)$ and set $\theta_{z4} = 0.3 + \min(\theta_{z1}, \theta_{z2}, \theta_{z3})$ for $z = 0, 1$; the generated potential outcomes satisfy $\max_{a \neq a'} |\text{ASE}(Z; a, a')| = 0.3$ for $z = 0, 1$.

We first consider the scenario with equal cluster size n for all clusters, the total variance $\sigma^2 = 1$, and two levels of cluster size ($n = 20$ and $n = 100$). We choose three values of the correlation coefficient between potential outcomes, $\rho = 0, 0.3, 0.6$. Because the sample size formulas in equations (12), (14), and (17) assume $\rho = 0$, the simulation settings with $\rho = 0.3, 0.6$ evaluate their robustness to the misspecification of this design parameters. In each setting, we vary the intracluster correlation coefficient $r = \sigma_b^2/(\sigma_w^2 + \sigma_b^2)$ from 0 to 1, which also determines the values of σ_w^2 and σ_b^2. We compute the required number of clusters using the sample size formulas and then generate the data based on the resulting number of clusters. The statistical power is estimated under each setting by averaging over $1,000$ Monte Carlo simulations.

Figure S1 shows the required number of clusters calculated from equations (12), (14), and (17) for the statistical power of 80%. The parameters are set to $\sigma^2 = 1$, $\mu = 0.3$, $\alpha = 0.05$, and $\beta = 0.2$ with the intracluster correlation coefficient varying from 0 to 1 (horizontal axis). The required number of clusters for the marginal direct effect (middle panel) is much less than those for the direct and spillover effects (left and right panels, respectively). Across all settings, the required cluster number increases linearly with the intracluster correlation coefficient. The difference between the settings with a small cluster size $n = 20$ and a moderate cluster size $n = 100$ is not substantial. This is because the conservative variance (covariance) matrix estimators rely solely on the estimated

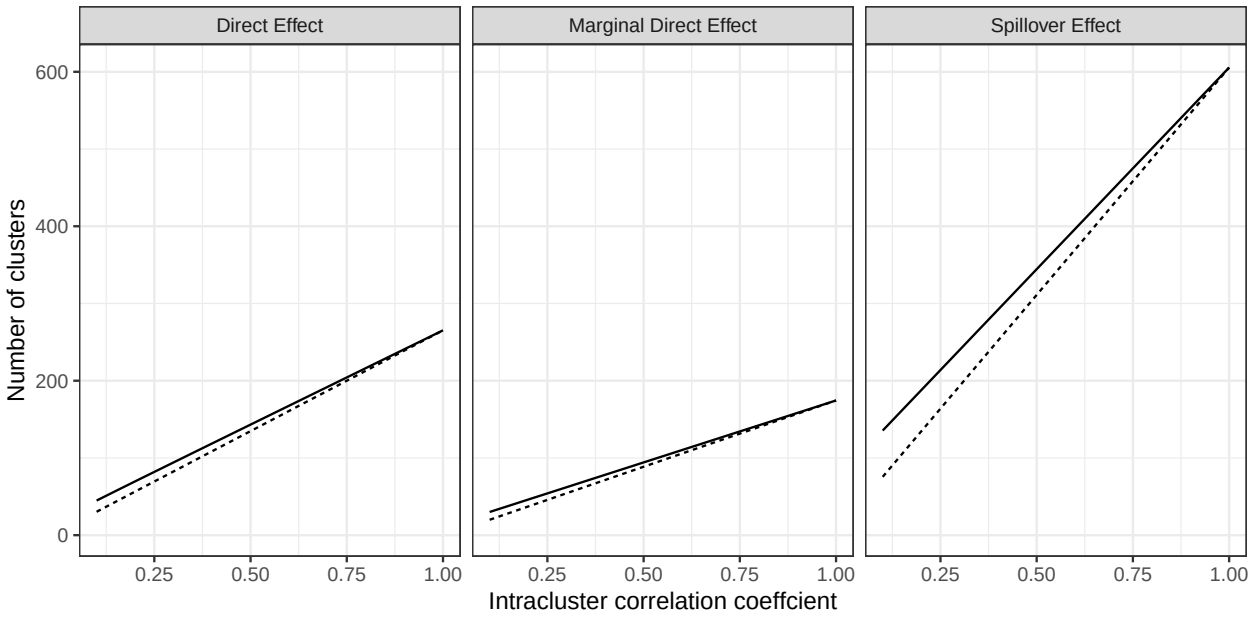
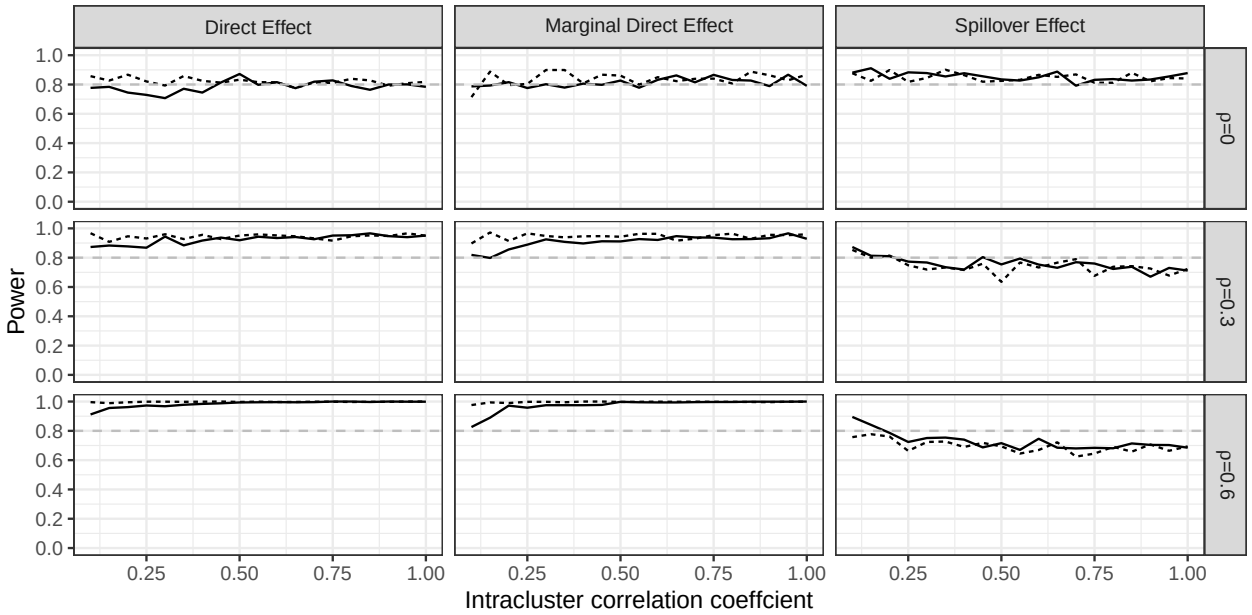


Figure S1: The required number of clusters calculated from equations (12), (14), and (17) for the statistical power of 80%. The parameters are set to $\sigma^2 = 1$, $\mu = 0.3$, $\alpha = 0.05$, $\beta = 0.2$ with the intracluster correlation coefficient varying from 0 to 1 (horizontal axis). The solid lines indicate the setting with cluster size of $n = 20$, and the dashed lines indicate the setting with $n = 100$.

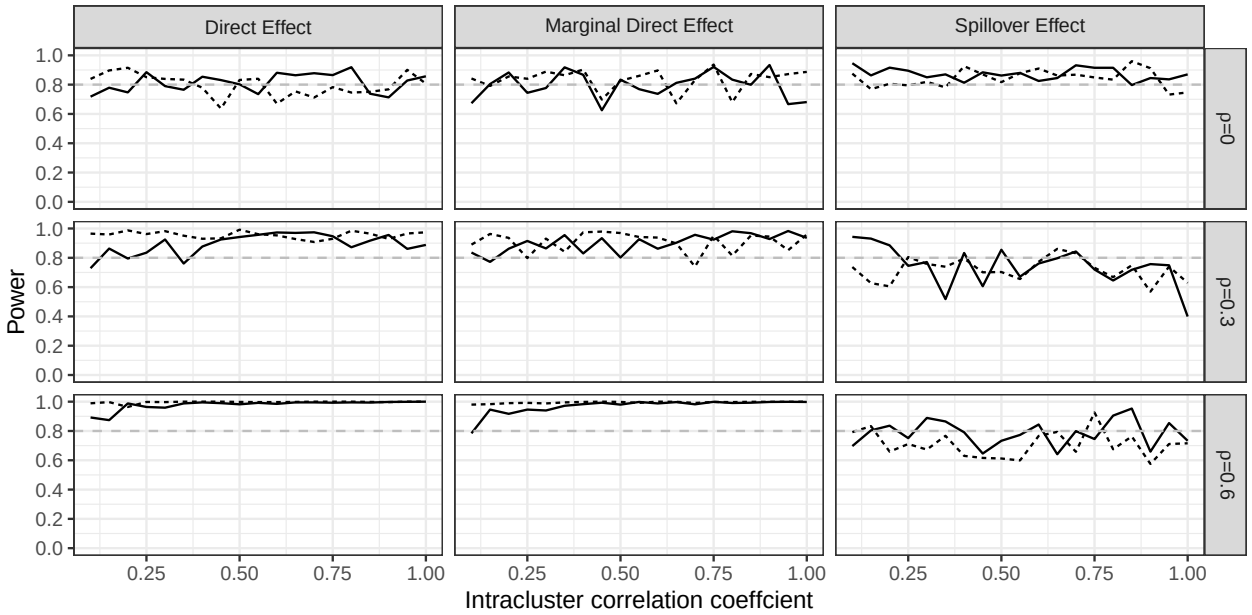
between-cluster variances, in which the cluster size plays a minimal role. As a result, having a large cluster size does not affect the required number of clusters significantly.

Figure 2(a) presents the estimated statistical power for testing the alternative hypotheses concerning the direct, marginal direct, and spillover effects in the left, middle, and right plots, respectively. With the correct specification of the correlation coefficient ρ, the achieved power is close to its expected level (0.8) under almost all settings for the direct effect, marginal direct effect, and spillover effect. When the intracluster correlation coefficient is small, the statistical power for the direct effect and marginal direct effect is sometimes below the nominal level of 0.8. This may arise because the required number of clusters is small under these settings (e.g., $J \geq 20$ for the marginal direct effect when the intracluster correlation coefficient is 0.1), reducing the accuracy of the asymptotic approximation used by the sample size formulas.

With the misspecified values of correlation coefficient $\rho = 0.3, 0.6$, the power is close to 1 for the



(a) Equal cluster size



(b) Unequal cluster size

Figure S2: Estimated statistical power for testing the alternative hypotheses the direct, marginal direct, and spillover effects. The solid lines indicate the setting with cluster size of $n = 20$, and the dashed lines indicate the setting with $n = 100$. In each plot, we vary the correlation between potential outcomes ρ as well as the intracluster correlation coefficient (horizontal axis).

direct effect and marginal direct effect when the intracluster correlation coefficient is moderate or large. This suggests that the sample size formula is conservative for these quantities. In contrast, the power is smaller than the expected level for the spillover effect, especially with a large value of the intracluster correlation coefficient. This suggests that the sample size formula for the spillover effect may not be robust to the misspecification of the correlation coefficient.

Next, we consider the scenario with unequal cluster size. We generate each cluster size from a categorical distribution distribution taking values in $\{0.5n, 0.75n, n, 1.5n, 2n, 2.5n\}$ with probabilities $\{0.25, 0.1, 0.1, 0.1, 0.2, 0.25\}$, respectively. We then generate the data using the number of clusters calculated from the sample size formulas. The parameter $\bar{n}$ in these formulas is calculated based on the distribution of the cluster sizes. Other parameters are the same as those of the case with equal cluster size.

Figure 2(b) shows the results. The results for the direct effect and marginal direct effect are largely similar to those presented in Figure 2(a). For the spillover effect, the variation of power is larger with unequal cluster size than with equal cluster size when the intracluster correlation coefficient is misspecified. These results show that the sample size formulas are robust to the unequal cluster sizes.

The simulation results also suggest that the sample size formulas are robust to the violation of the simplifying conditions used in Assumption 4. The reason is that the variances in the generated data do not satisfy these simplifying conditions due to finite sample variation.

References

Baird, S., Bohren, J. A., McIntosh, C., and Ozler, B. (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics* **100**, 5, 844–860.

Basse, G. and Feller, A. (2018). Analyzing multilevel experiments in the presence of peer effects. *Journal of the American Statistical Association* **113**, 521, 41–55.

Bell, R. M. and McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology* **28**, 2, 169–181.

Karwa, V. and Airoldi, E. M. (2018). A systematic investigation of classical causal inference strategies under mis-specification due to network interference. *arXiv preprint arXiv:1810.08259* .

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* **112**, 520, 1759–1769.

Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probability theory and related fields* **81**, 3, 341–352.