Sample Design for Group Randomized Trials (Sample Power I and II)

Larry V. Hedges

Northwestern University

Prepared for the 2022 IES/NCER Summer Research Training Institute at Northwestern University

Topics for Today

1. Design sensitivity and planning a design
2. Planning cluster randomized designs
3. Planning multisite (randomized block) designs
4. What effect sizes are reasonable?
5. How do we get values of other design parameters?
6. What about subgroup (moderator) effects
7. Is it ever OK to ignore levels of sampling in our analyses?

Design Sensitivity

Sound research design requires quantification of the sensitivity of research designs to detect effects

There are three related concepts of design sensitivity:

Precision of treatment effect estimates: The standard error of the treatment effect estimate

Statistical **power**: The probability of detecting and effect size of a given magnitude

Minimum detectable effect size: The smallest effect size for which the design has specified power (often 80% following Cohen's recommendation)

Power tells you the probability that a design can detect an effect of a given size (usually at the 0.05 significance level)

Minimum detectable effect size tells you what effect size a design can detect (usually at the 0.05 significance level, usually with 80% power)

Planning a Design

Planning a design is creating a data collection protocol that has adequate sensitivity to detect the effect size expected or the smallest meaningful effect size

So far that means finding a sample size that has adequate sensitivity

If resources are unlimited, this means simply obtaining an adequate sample size (look at a graph or a table of power values)

Resources (budget) are essentially ***always*** limited in research

This reality makes design a more difficult problem

If a completely randomized design is inadequately sensitive, there are two alternatives:

-Improve the existing design to increase sensitivity

-Choose a different type of design

Improving Design Sensitivity

Design sensitivity (holding significance level constant) in any design depends on effect size, sample size(s), and certain other design parameters which are different for different designs

Thus to increase design sensitivity we can: Increase sample size(s), reduce variation (which increases the effective effect size, or (sometimes) alter other design parameters

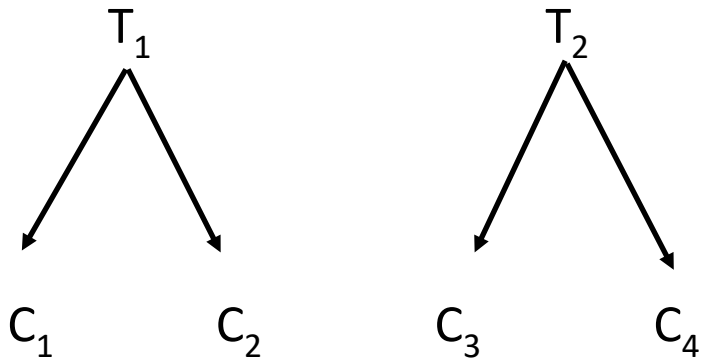
While the effect size of a treatment may not easily be increased, use of covariates and restricted sampling can reduce variation so that the “effective” effect size is increased

Depending on constraints there may be ways of increasing sample size by planned imbalance

The Cluster Randomized Design

The Cluster Randomized Design

The figure below illustrates the cluster randomized design



In the language of experimental design, clusters (C) are nested within treatments (T) (every cluster receives only one treatment)

This is also called the ***group randomized design*** or the ***hierarchical design*** in classical experimental design

Why Cluster Randomization?

Cluster randomization is less efficient (leads to less sensitive designs) than individual randomization

So why randomize clusters?

Assignment of individuals to treatments independently is sometimes ***impractical***, ***unfeasible***, or ***impossible***

For example:

It is impractical to assign students in the same classroom to different curricula, have different duty rules for interns supervised in the same clinic

It may be politically difficult to assign only some students (or teachers) in a school to a much more desirable treatment

It is theoretically impossible to assign aggregate treatments to different individuals within the same aggregate (e.g., while school behavior support, whole school trust interventions)

Why Cluster Randomization?

Contamination between treatment and control groups is sometimes a concern

This could be inadvertent or intentional

For example

Control teachers might learn of new teaching methods from their colleagues in the treatment group

Students in a tutoring intervention might bring their untutored peers to their tutoring sessions, intentionally subverting the experiment

Parents in the same school might insist that their children assigned to the control group receive the treatment

Digression: Two-Stage Cluster Sampling

The relevant sampling model for cluster randomized designs is two stage cluster sampling

Stage 1: Obtain a simple random sample of clusters

Stage 2: Obtain a simple random sample of individuals within clusters

You all know that if the population variance is σ_T^2 the variance of the mean of a simple random sample of size N from that population is σ_T^2/N

But the variance of the mean of a two-stage cluster sample from that population is not $\sigma_T^2/N,$ but

$$[1 + (n-1)\rho]\sigma_T^2/N$$

where $\rho = \sigma_2^2/\sigma_T^2 = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$ is the **intraclass correlation**

Here σ_2^2 is the between-cluster (means) variance and σ_1^2 is the within-cluster variance and the total variance is $\sigma_T^2 = \sigma_1^2 + \sigma_2^2$

The quantity $[1 + (n-1)\rho]$ is called the **design effect** and represents the penalty (in variance) for using a two-stage cluster sample instead of a simple random sample

Design Effect for Clusters of Size n

Intraclass	Individuals per Cluster (n)		
Correlation (ρ)	10	50	500
0.01	1.04	1.22	2.48
0.05	1.20	1.86	5.09
0.10	1.38	2.43	7.13
0.15	1.53	2.89	8.71
0.20	1.67	3.29	10.04
0.25	1.80	3.64	11.21

Where Does the Design Effect Come from?

Think of the i^{th} observation in the j^{th} cluster Y_{ij} as composed of a cluster mean α_i and a deviation from that cluster mean ε_{ij} then

$$Y_{ij} = \alpha_i + \varepsilon_{ij}$$

The variance of the cluster mean is

$$V\left\{\overline{Y}_{\bullet j}\right\} = V\left\{\frac{1}{n}\sum_{i=1}^{n} Y_{ij}\right\} = V\left\{\frac{1}{n}\sum_{i=1}^{n} \alpha_j + \varepsilon_{ij}\right\} = V\left\{\alpha_j + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ij}\right\} = V\left\{\alpha_j\right\} + \frac{1}{n^2} nV\left\{\varepsilon_{ij}\right\} = \sigma_2^2 + \frac{\sigma_1^2}{n}$$

It is conventional to define the intraclass correlation as $\rho = {\sigma_2}^2/({\sigma_1}^2 + {\sigma_2}^2) = {\sigma_2}^2/{\sigma_T}^2$, therefore

$$V\left\{\overline{Y}_{\bullet j}\right\} = \frac{n\sigma_2^2 + \sigma_1^2}{n} = \frac{\left(\frac{n\sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)\left(\sigma_1^2 + \sigma_2^2\right)}{n} = \frac{\left[n\rho + (1-\rho)\right]\left(\sigma_1^2 + \sigma_2^2\right)}{n} = \frac{\left[1 + (n-1)\rho\right]\left(\sigma_1^2 + \sigma_2^2\right)}{n} = \frac{\left[1 + (n-1)\rho\right]\sigma_T^2}{n}$$

The variance of the sample mean from a two-stage cluster sample on m clusters of size n (total size mn) is just the variance of the mean of m cluster means

$$V\left\{\overline{Y}_{\bullet\bullet}\right\} = V\left\{\frac{1}{m}\sum_{j=1}^{m} Y_{\bullet j}\right\} = V\left\{\frac{1}{m}\sum_{i=1}^{m} \frac{\left[1 + (n-1)\rho\right]\sigma_T^2}{n}\right\} = \frac{\left[1 + (n-1)\rho\right]\sigma_T^2}{mn}$$

Where Does the Design Effect Come from?

Think of the i^{th} observation in the j^{th} cluster Y_{ij} as composed of a cluster mean α_i and a deviation from that cluster mean ε_{ij} then

$$Y_{ij} = \alpha_i + \varepsilon_{ij}$$

The variance of the cluster mean is

$$V\left\{\bar{Y}_{\bullet j}\right\} = V\left\{\frac{1}{n}\sum_{i=1}^{n} Y_{ij}\right\} = V\left\{\frac{1}{n}\sum_{i=1}^{n} \alpha_j + \varepsilon_{ij}\right\} = V\left\{\alpha_j + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ij}\right\} = V\left\{\alpha_j\right\} + \frac{1}{n^2} nV\left\{\varepsilon_{ij}\right\} = \sigma_2^2 + \frac{\sigma_1^2}{n}$$

It is conventional to define the intraclass correlation as $\rho = \sigma_2^2/(\sigma_1^2 + \sigma_2^2) = \sigma_2^2/\sigma_T^2$, therefore

$$V\left\{\bar{Y}_{\bullet j}\right\} = \frac{n\sigma_2^2 + \sigma_1^2}{n} = \frac{\left(\frac{n\sigma_2^2}{\sigma_1^2+\sigma_2^2} + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)\left(\sigma_1^2+\sigma_2^2\right)}{n} = \frac{\left[n\rho + (1-\rho)\right]\left(\sigma_1^2+\sigma_2^2\right)}{n} = \frac{\left[1+(n-1)\rho\right]\left(\sigma_1^2+\sigma_2^2\right)}{n} = \frac{\left[1+(n-1)\rho\right]\sigma_T^2}{n}$$

The variance of the sample mean from a two-stage cluster sample on m clusters of size n (total size mn) is just the variance of the mean of m cluster means

$$V\left\{\bar{Y}_{\bullet\bullet}\right\} = V\left\{\frac{1}{m}\sum_{j=1}^{m} Y_{\bullet j}\right\} = V\left\{\frac{1}{m}\sum_{i=1}^{m} \frac{\left[1+(n-1)\rho\right]\sigma_T^2}{n}\right\} = \frac{\left[1+(n-1)\rho\right]\sigma_T^2}{mn}$$

This is $[1 + (n-1)\rho]$ times σ_T^2/mn, the variance of the mean of a simple random sample of size mn

Two Stage Cluster Sampling

This is relevant because the estimate of the treatment effect in a cluster randomized design is a comparison of the ***means*** of treatment ***clusters*** with the ***means*** of control ***clusters***

Thus the variance of the estimated treatment effect depends on the variance of means of two-stage cluster samples (one for treatment and one for control)

This has substantial effects on the sensitivity of the design and on the analyses required

Model and Notation: Cluster Randomized Designs

Let Y_{ij} be the outcome score for i^{th} level 1 unit (individual) in the j^{th} level 2 unit (cluster). The level 1 (individual level) model is

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}, \qquad \text{and } \varepsilon_{ij} \sim \mathrm{N}(0, \sigma_1^2)$$

where β_{0j} is the mean of the j^{th} cluster, and ε_{ij} is a level 1 residual. The level 2 (cluster level) model is

$$\beta_{0j} = \gamma_0 + \gamma_1 T_i + \eta_j, \qquad \text{and } \eta_j \sim \mathrm{N}(0, \sigma_2^2)$$

where $T_i = \pm \frac{1}{2}$ is a treatment indicator variable and η_j is a level 2 (cluster level) residual.
We could write the combined model as

$$Y_{ij} = \gamma_0 + \gamma_1 T_i + \eta_j + \varepsilon_{ij}.$$

Note that the combined residual term is $(\eta_j + \varepsilon_{ij})$ and the residuals from observations in the j^{th} cluster share the same random effect η_j

This violates the usual assumption of independence of the residuals

Effect Size and Intraclass Correlation: Cluster Randomized Designs

The total variance in the population is partitioned into between and within cluster variance

$$\sigma_T^2 = \sigma_I^2 + \sigma_I^2$$

The ratio of between cluster to total variance, the (level 2) intraclass correlation

$$\rho_2 = \sigma_2^2 \Big/ \left(\sigma_1^2 + \sigma_2^2\right) = \sigma_2^2 \Big/ \sigma_T^2$$

quantifies how much clustering there is in the population

The natural effect size in this design is a variation of Cohen's d, which in this notation is

$$\delta = \frac{\gamma_1}{\sigma_T} = \frac{\mu^T - \mu^C}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

Hypothesis Testing: Cluster Randomized Designs

The test of the hypothesis that the treatment effect is zero, that is

$$\mathrm{H}_0\colon \gamma_1 = 0$$

is based on the test statistic

$$t = \hat{\gamma}_1 \big/ SE\left(\hat{\gamma}_1\right)$$

which is taken to have the t-distribution with $M-2$ degrees of freedom (M is the number of clusters)

In a design with m^T treatment clusters and m^C control clusters all of size n

$$SE\left(\hat{\gamma}_1\right) = \left(\sqrt{\frac{1+\left(n-1\right)\rho_2}{m^T n} + \frac{1+\left(n-1\right)\rho_2}{m^C n}}\right)\sigma_T = \left(\sqrt{\frac{m^T + m^C}{m^T m^C n}}\sqrt{1+\left(n-1\right)\rho_2}\right)\sigma_T$$

Recall that $1 + (n-1)\rho_2$ is the design effect from sample surveys

Hypothesis Testing: Cluster Randomized Designs

When the null hypothesis is false, that is $\gamma_1 \neq 0$, then the test statistic has the noncentral t-distribution with $M-2$ degrees of freedom and noncentrality parameter

$$\lambda = \gamma_1 \Big/ SE\left(\hat{\gamma}_1\right)$$

Note: If the design is balanced (all clusters have the same size n) a two-sample t-test using cluster means as the data is ***exactly correct***, is equivalent to the multilevel model test, and is an optimal test of treatment effects,

Unbalanced Cluster Randomized Designs

When the design is not balanced, the standard error of the test statistic is much more complex

$$\left[SE\left(\hat{\gamma}_1\right)\right]^2 = \left[\sum_{j=1}^{m^T} n_j^T \Big/ \left(n_j^T \sigma_2^2 + \sigma_1^2\right)\right]^{-1} + \left[\sum_{j=1}^{m^C} n_j^C \Big/ \left(n_j^C \sigma_2^2 + \sigma_1^2\right)\right]^{-1}$$

$$= \left(\left[\sum_{j=1}^{m^T} n_j^T \Big/ \left[1 + \left(n_j^T - 1\right)\rho_2\right]\right]^{-1} + \left[\sum_{j=1}^{m^C} n_j^C \Big/ \left[1 + \left(n_j^C - 1\right)\rho_2\right]\right]^{-1}\right)\sigma_T^2$$

This can be better understood by seeing it as a combination of "averaged" design effects

$$SE\left(\hat{\gamma}_1\right) = \left(\frac{DEF^T}{N^T} + \frac{DEF^C}{N^C}\right)\sigma_T^2$$

where, dropping the T or C superscripts

$$DEF = \left[\frac{1}{N}\sum_{j=1}^{m} \frac{n_j}{1 + \left(n_j - 1\right)\rho_2}\right]^{-1}$$

How Does this Complex SE Arise?

The SE in unbalanced designs is quite complicated, but there is a way to understand it

The treatment effect (parameter) γ_1 is the difference between the means of the treatment and control groups, so is the estimate of γ_1 (the estimated treatment effect)

The cluster means in the treatment group have expected value $\gamma_0 + \gamma_1/2$ and the cluster means in the control group have expected value $\gamma_0 - \gamma_1/2$

Therefore any normalized weighted average of the cluster means in a particular treatment group estimates the mean of that treatment group (normalized means weights sum to 1)

Maximum likelihood generates efficient estimates by estimating the weighted average that is most precise

The most precisely estimated weighted mean is the inverse variance weighted mean (this is just like in meta-analysis) using weights

$$w_j = \left(1/v_j\right) \Big/ \sum 1/v_i$$

How Does this Complex SE Arise?

The variance of the j^{th} cluster mean is $v_j = \sigma_2^2 + \sigma_1^2/n_j$ and $1/v_j = n_j/(n_j\sigma_2^2 + \sigma_1^2)$

The variance of the inverse variance weighted mean is the reciprocal of the sum of the weights

Therefore the variance of the estimate of each treatment group mean is (dropping the T or C superscripts) is

$$\frac{1}{\sum_{j=1}^{m} 1/v_j} = \left[\sum_{j=1}^{m} \frac{n_j}{1+\left(n_j-1\right)\rho_2}\right]^{-1}$$

The variance (square of the SE) is just the sum of two terms like this (one for the treatment group and one for the control group)

Design Sensitivity: The Cluster Randomized Design

Precision

When all the treatment and control clusters are of size n, then

$$SE\left(\hat{\gamma}_1\right) = \left(\sqrt{\frac{m^T + m^C}{m^T m^C n}} \sqrt{1 + \left(n-1\right)\rho_2} \right) \sigma_T$$

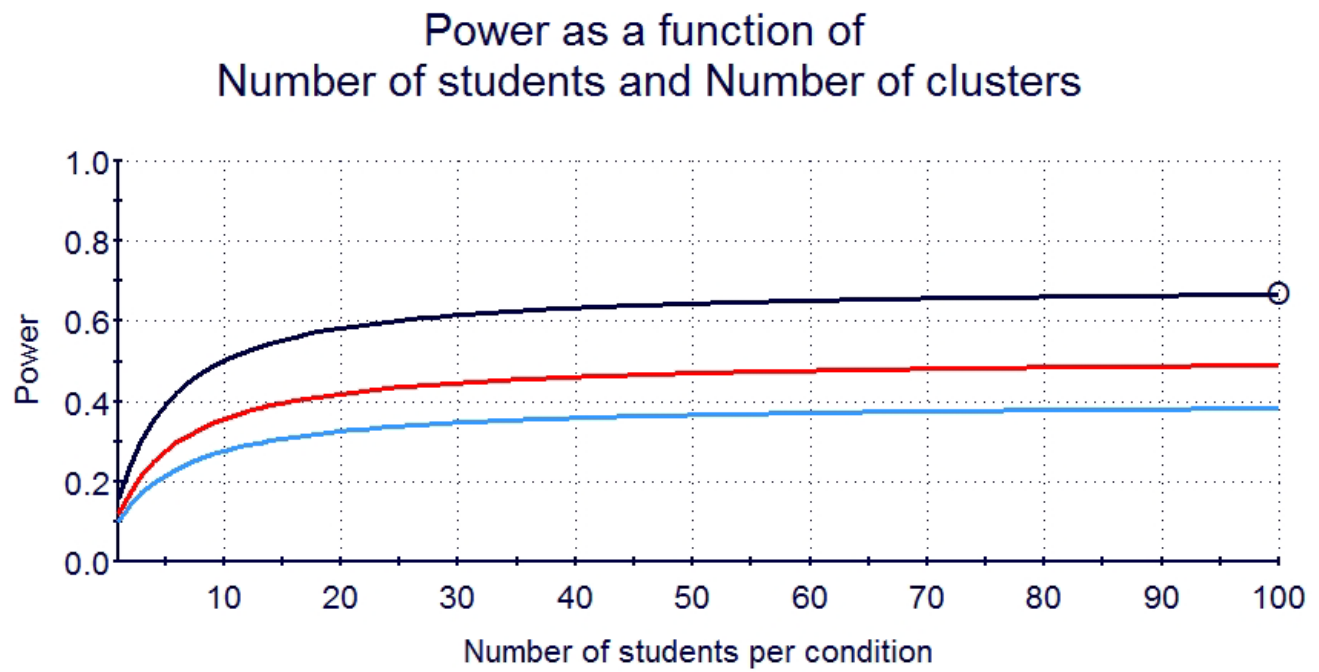
Statistical Power

$$power = 1 - \mathrm{F}(t_{\alpha/2} \mid df, \lambda) + \mathrm{F}(- t_{\alpha/2} \mid df, \lambda)$$

where $\mathrm{F}(x \mid df, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with df degrees of freedom and noncentrality parameter λ

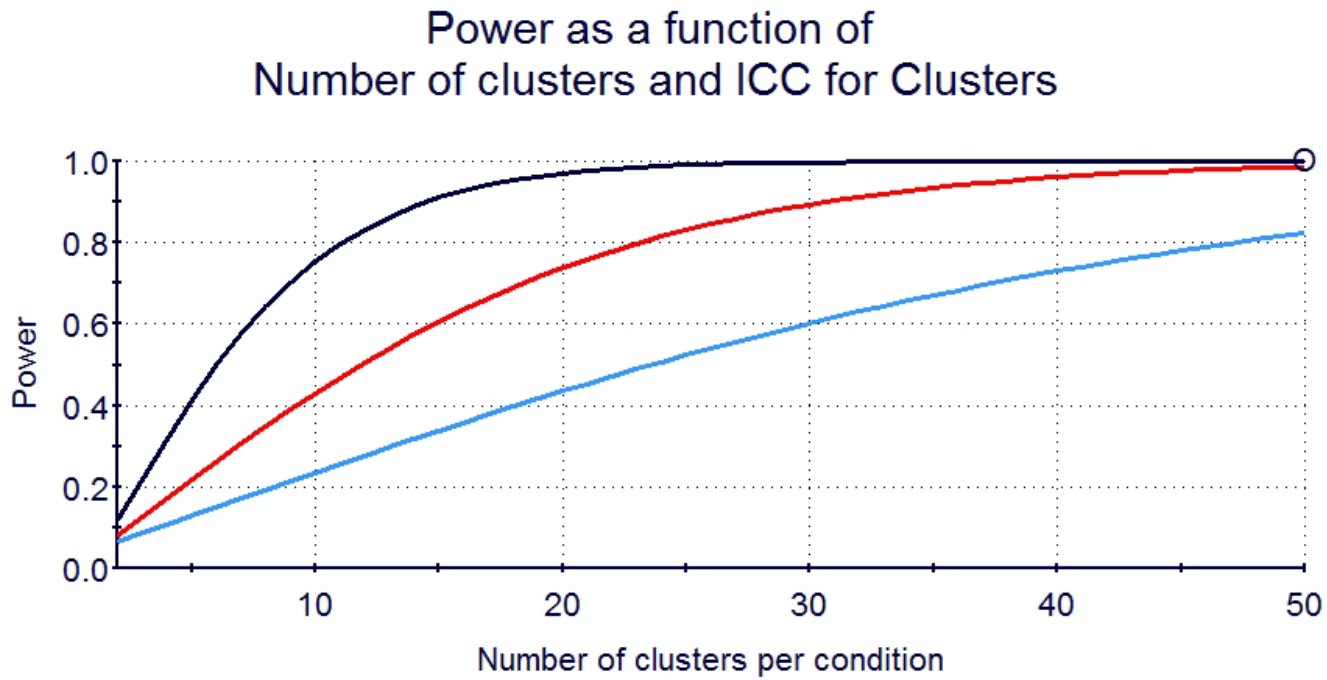
As an approximation, the noncentral t-distribution is *approximately* a translated central t-distribution (i.e., with $\lambda = 0$), so $\mathrm{F}(x \mid df, \lambda) \approx \mathrm{F}(x - \lambda \mid df, 0)$ [works well if df and x are fairly large)

This approximation is helpful because you can compute with it in spreadsheets like EXCEL



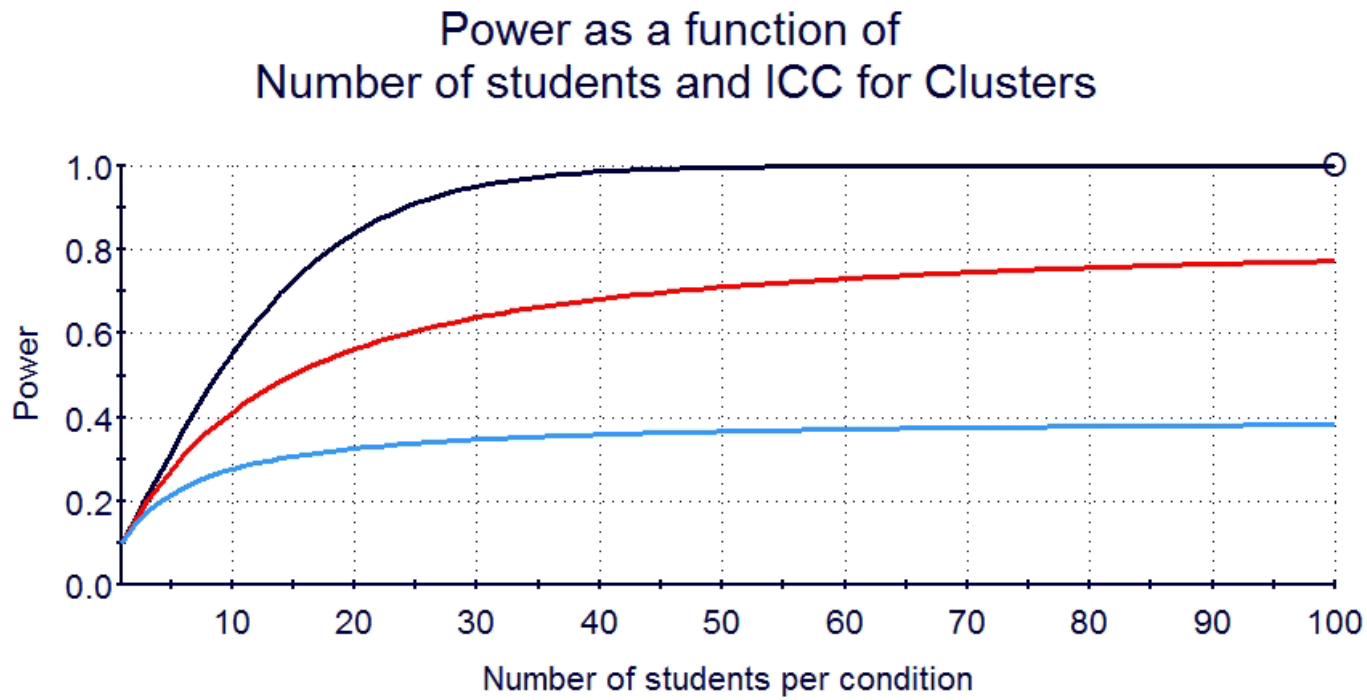
Two-level clustered design.
Clusters are randomized. Students are nested.
Test of difference in means
Statistical model - Random-effects at both levels.
Effect size - Standardized mean difference, d (total) = 0.2500.
Clusters - Number varies, ICC = 0.1500, No covariates.
Students - Number varies, No covariates.

Number of clusters
30 per condition
20 per condition
15 per condition



Two-level clustered design.
Clusters are randomized. Students are nested.
Test of difference in means
Statistical model - Random-effects at both levels.
Effect size - Standardized mean difference, d (total) = 0.2500.
Clusters - Number varies, ICC varies, No covariates.
Students - 25 per group, No covariates.

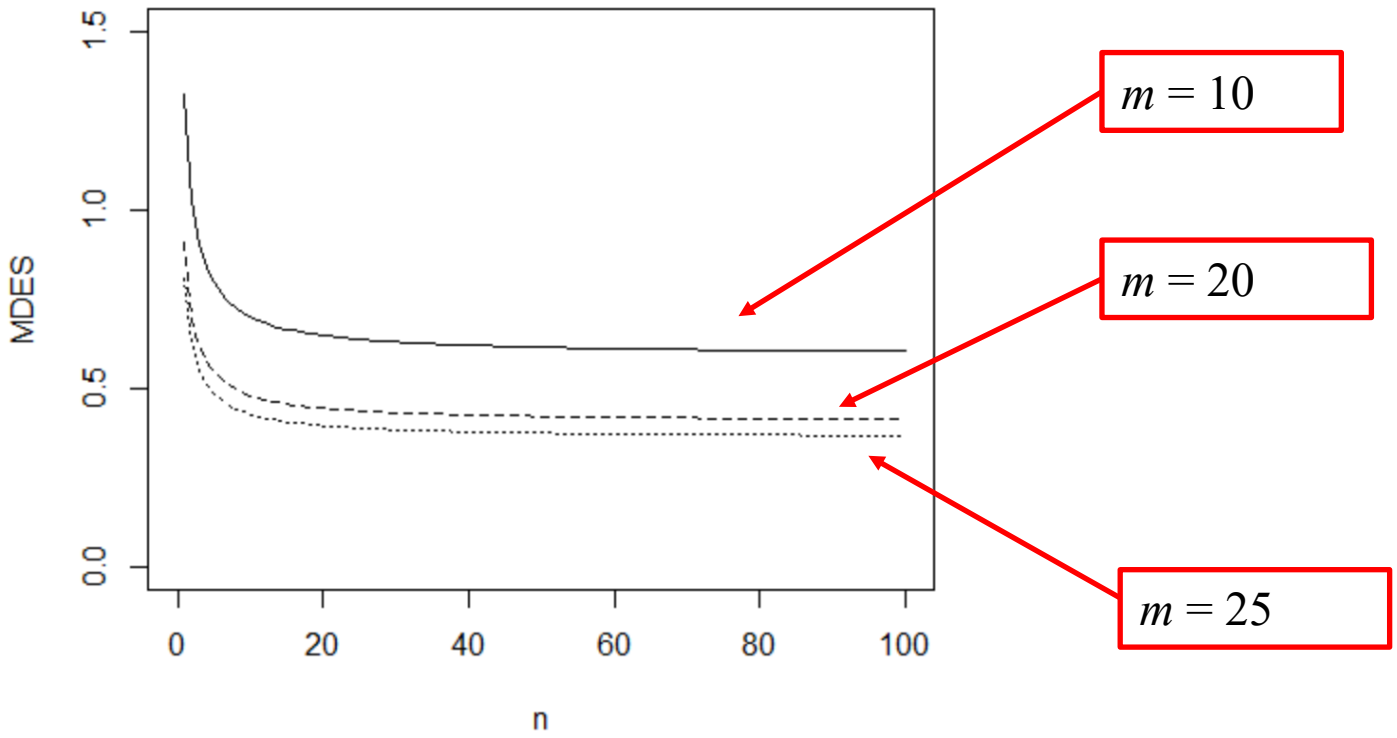
ICC Clusters
0.0000
0.0500
0.1500



Two-level clustered design.
Clusters are randomized. Students are nested.
Test of difference in means
Statistical model - Random-effects at both levels.
Effect size - Standardized mean difference, d (total) = 0.2500.
Clusters - 15 per group, ICC varies, No covariates.
Students - Number varies, No covariates.

ICC Clusters
0.0000
0.0500
0.1500

MDES in Cluster Randomized Designs as a function on n for $\rho_2 = 0.20$



Approximate Minimal Detectable Effect Size: Cluster Randomized Designs

It is useful to have an algebraic approximation to the minimum detectable effect size

$$\delta_M \approx M_{m^T+m^C-2}\sqrt{\frac{\left(m^T+m^C\right)\left[1+\left(n-1\right)\rho_2\right]}{m^T m^C n}}$$

where M_{df} is the constant depending on the degrees of freedom

M_{df} is a decreasing function of df and $M_f < 2.9$ for $df > 16$, but $M_\infty = 2.80$, so $2.8 < M_{df} < 2.9$

When the design is balanced so that $m^T = m^C = m$, then δ_M has the very simple form

$$\delta_M \approx M_{2m-2}\sqrt{\frac{2\left[1+\left(n-1\right)\rho_2\right]}{mn}}$$

Sample Size and MDES: Cluster Randomized Designs

The relation between sample size and design sensitivity is more complex in cluster randomized designs than in the completely randomized design because of two level cluster sampling

In the completely balanced design, the approximates MDES is

$$\delta_M \approx M_{2m-2}\sqrt{\frac{2\left[1+\left(n-1\right)\rho_2\right]}{mn}} = M_{2m-2}\sqrt{\frac{2\left(1-\rho_2\right)}{mn}+\frac{2\rho_2}{m}}$$

From this expression, see that as $m \to \infty$, $\delta_M \to 0$

But note that as $n \to \infty$, $\delta_M \to M_{2m-2}\sqrt{2\rho_2/m}$

Therefore total sample size is a poor indicator of design sensitivity—it depends on m and n separately (and mostly on m)

The point where increasing n has little effect occurs for relatively small n

Values of M_{df}

df	M_{df}		df	M_{df}
2	5.36		28	2.85
4	3.35		30	2.85
6	3.11		32	2.85
8	3.01		34	2.84
10	2.96		36	2.84
12	2.93		38	2.84
14	2.91		40	2.84
16	2.90		50	2.83
18	2.88		75	2.82
20	2.88		100	2.82
22	2.87		500	2.80
24	2.86			
26	2.86		∞	2.80

Calculating M_{df}

The calculation of M_{df} is based on approximating the noncentral t-distribution by a translated central t-distribution and ignoring the tail of the distribution that is opposite the effect size (usually the negative tail)

$$power = \mathrm{F}\{t < c_{\alpha/2} \mid df, \lambda\} = \mathrm{F}\{t - \lambda < c_{\alpha/2} \mid df, 0\}$$

First compute the two-sided critical value at the significance level α

Then compute the value of the t-distribution corresponding to the quantile for the desired power (e.g., for 80% = 0.8 power, the 80th percentile. Call this t_{power}

For example, if $df = 50$, $c_{\alpha/2} = 1.984$ and $t_{power} = 0.845$, so that $M_{df} = 1.984 + 0.845 = 2.83$

Increasing Design Sensitivity: Cluster Randomized Designs

Design sensitivity can be increased by using covariates

Covariates can be added at either level 1 (the individual level) or level 2 (the cluster level)

The effect of a covariate at a particular level can be understood as decreasing the (effective) variance at that level

Because the level 2 variance component has the largest effect on uncertainty of the treatment effect, level 2 covariates will generally have the largest effect on design sensitivity

Multilevel Model with Covariates: Cluster Randomized Designs with Covariates

Suppose X is a centered level 1 (individual level) covariate and W is a level 2 (cluster level) covariate

Let Y_{ij} be the outcome score for i^{th} level 1 unit (individual) in the j^{th} level 2 unit (cluster). The level 1 (individual level) model is

$$Y_{ij} = \beta_{0j}^{A} + \beta_{1j}^{A} X_{ij} + \varepsilon_{ij}^{A}, \qquad \text{and } \varepsilon_{ij}^{A} \sim \mathrm{N}(0, \sigma_{A1}^{2})$$

The level 2 (cluster level) model is

$$\beta_{0j}^{A} = \gamma_{00}^{A} + \gamma_{01}^{A} T_j + \gamma_{02}^{A} W_j + \eta_j^{A}, \text{ and } \eta_j^{A} \sim \mathrm{N}(0, \sigma_{A2}^{2})$$

$$\beta_{1j}^{A} = \gamma_{10}^{A}$$

where $T_j = \pm \frac{1}{2}$ is a treatment indicator variable, γ_{01}^{A} is the treatment effect, γ_{02}^{A} is the effect of the cluster level covariate, and η_j^{A} is a level 2 (cluster level) residual (we illustrate only one covariate at each level)

We could write the combined model as

$$Y_{ij} = \gamma_{00}^{A} + \gamma_{10}^{A} X_{ij} + \gamma_{01}^{A} T_j + \gamma_{02}^{A} W_j + \eta_j^{A} + \varepsilon_{ij}^{A}.$$

Hypothesis Testing: Cluster Randomized Designs with Covariates

The test of the hypothesis that the treatment effect is zero, that is

$$\mathrm{H}_0\colon \gamma_{01} = 0$$

is based on the test statistic

$$t = \hat{\gamma}_{01} \big/ SE\left(\hat{\gamma}_{01}\right)$$

which is taken to have the t-distribution with $M - 2 - q$ degrees of freedom (M is the number of clusters), where q is the number of level 2 (cluster level) covariates

In a balanced design with m^T treatment clusters and m^C control clusters all of size n

$$SE\left(\hat{\gamma}_{01}\right) = \left(\sqrt{\frac{m^T + m^C}{m^T m^C n}} \sqrt{\bar{R}_1^2 + \left(\bar{R}_2^2 n - \bar{R}_1^2\right)\rho_2} \right) \sigma_T$$

where

$$\bar{R}_1^2 = 1 - R_1^2 = \sigma_{A1}^2 \big/ \sigma_1^2 \quad \text{and} \quad \bar{R}_2^2 = 1 - R_2^2 = \sigma_{A2}^2 \big/ \sigma_2^2$$

Hypothesis Testing: Cluster Randomized Designs with Covariates

When the null hypothesis is false, that is when $\gamma_{01} \neq 0$, then t is taken to have the noncentral t -distribution with $M-2-q$ degrees of freedom and noncentrality parameter

$$\lambda_A = \gamma_{01} \Big/ SE\left(\hat{\gamma}_{01}\right)$$

Note that now is the covariate adjusted standard error

$$SE\left(\hat{\gamma}_{01}\right) = \left(\sqrt{\frac{m^T + m^C}{m^T m^C n}} \sqrt{\overline{R}_1^2 + \left(\overline{R}_2^2 n - \overline{R}_1^2\right)\rho_2} \right) \sigma_T$$

Comparing the covariate adjusted to the unadjusted noncentrality parameters, see that

$$\frac{\lambda_A}{\lambda} = \sqrt{\frac{1+\left(n-1\right)\rho_2}{\overline{R}_1^2 + \left(\overline{R}_2^2 n - \overline{R}_1^2\right)\rho_2}}$$

Design Sensitivity: The Cluster Randomized Design with Covariates

Precision

When all the treatment and control clusters are of size n, then

$$SE\left(\hat{\gamma}_{01}\right) = \left(\sqrt{\frac{m^T + m^C}{m^T m^C n}} \sqrt{\bar{R}_1^2 + \left(\bar{R}_2^2 n - \bar{R}_1^2\right)\rho_2} \right) \sigma_T$$

Statistical Power

$$power = 1 - \mathrm{F}(t_{\alpha/2} \mid df, \lambda) + \mathrm{F}(- t_{\alpha/2} \mid df, \lambda)$$

where $\mathrm{F}(x \mid df, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with df degrees of freedom and noncentrality parameter λ

As an approximation, the noncentral t-distribution is *approximately* a translated central t-distribution (i.e., with $\lambda = 0$), so $\mathrm{F}(x \mid df, \lambda) \approx \mathrm{F}(x - \lambda \mid df, 0)$ [works well if df and x are fairly large)

This approximation is helpful because you can compute with it in spreadsheets like EXCEL

Approximate Minimal Detectable Effect Size: Cluster Randomized Design with Covariates

It is useful to have an algebraic approximation to the minimum detectable effect size

$$\delta_M \approx M_{m^T+m^C-2-q}\sqrt{\frac{\left(m^T+m^C\right)\left[\bar{R}_1^2+\left(\bar{R}_2^2 n-\bar{R}_1^2\right)\rho_2\right]}{m^T m^C n}}$$

where M_{df} is the constant depending on the degrees of freedom discussed before

Recall that M_{df} is a decreasing function of df and $M_f < 2.9$ for $df > 16$

When the design is balanced so that $m^T = m^C = m$, then δ has the very simple form

$$\delta_M \approx M_{2m-2-q}\sqrt{\frac{2\left[\bar{R}_1^2+\left(\bar{R}_2^2 n-\bar{R}_1^2\right)\rho_2\right]}{mn}}$$

Comparing Sensitivity with and without Covariates: Cluster Randomized Designs

Compare expressions for the MDES with and without covariates

$$\frac{\delta_{AM}}{\delta_M} \approx \frac{M_{2m-2-q}}{M_{2m-2}} \sqrt{\frac{\overline{R}_1^2 + \left(\overline{R}_2^2 n - \overline{R}_1^2\right)\rho_2}{1+(n-1)\rho_2}} \approx \sqrt{\frac{\overline{R}_1^2 + \left(\overline{R}_2^2 n - \overline{R}_1^2\right)\rho_2}{1+(n-1)\rho_2}} = \sqrt{\frac{(1-\rho_2)\overline{R}_1^2 + n\overline{R}_2^2\rho_2}{(1-\rho_2)+n\rho_2}}$$

(because M_{df} changes little with df if df is moderate in size)

Because $1-\rho_2$ is typically smaller than $n\rho_2$, the ratio of MDES values is ***very*** approximately (for large n)

$$\frac{\delta_{AM}}{\delta_M} = \sqrt{\overline{R}_2^2} = \sqrt{1-R_2^2}$$

This is analogous to the result for the completely randomized design

A single level 2 covariate with $R_2 = 0.7$ ($R_2{}^2 = 0.49$) would reduce the MDES by a factor of

$$\sqrt{1-0.49} = 0.71$$

Limiting Values of MDES: Cluster Randomized Designs

Note that even if the covariates at the individual level explained all of the variance at the individual level, so that $R_1^2 = 1$ and $\overline{R}_1^2 = 0$, $\delta_{AM} > 0$ in that case

$$\delta_{AM} \approx M_{2m-2-q}\sqrt{\frac{2\overline{R}_2^2\rho_2}{m}}$$

If the covariates at the cluster level explained all of the variance at the cluster level, so that $R_2{}^2 = 1$, then $\delta_{AM} > 0$. In that case

$$\delta_{AM} \approx M_{2m-2-q}\sqrt{\frac{2\overline{R}_1^2\left(1-\rho_2\right)}{mn}}$$

MDES in Cluster Randomized Designs with Covariates as a Function of R_2 ($n = 20$)

	R_2									
m	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	0.99	1.01	1.00	0.97	0.94	0.90	0.85	0.78	0.69	0.58
6	0.88	0.89	0.88	0.86	0.83	0.79	0.75	0.69	0.61	0.51
7	0.80	0.80	0.79	0.78	0.75	0.72	0.68	0.62	0.55	0.46
8	0.74	0.74	0.73	0.71	0.69	0.66	0.62	0.57	0.51	0.42
9	0.69	0.69	0.68	0.67	0.64	0.62	0.58	0.53	0.47	0.40
10	0.65	0.65	0.64	0.63	0.61	0.58	0.54	0.50	0.44	0.37
12	0.59	0.59	0.58	0.57	0.55	0.52	0.49	0.45	0.40	0.34
15	0.52	0.52	0.51	0.50	0.48	0.46	0.43	0.40	0.36	0.30
18	0.47	0.47	0.46	0.45	0.44	0.42	0.40	0.36	0.32	0.27
20	0.45	0.44	0.44	0.43	0.42	0.40	0.37	0.34	0.30	0.26
25	0.40	0.40	0.39	0.38	0.37	0.35	0.33	0.30	0.27	0.23
30	0.36	0.36	0.35	0.35	0.34	0.32	0.30	0.28	0.25	0.21
35	0.33	0.33	0.33	0.32	0.31	0.30	0.28	0.26	0.23	0.19
40	0.31	0.31	0.31	0.30	0.29	0.28	0.26	0.24	0.21	0.18
50	0.28	0.28	0.27	0.27	0.26	0.25	0.23	0.21	0.19	0.16

Improving Sensitivity by Planned Imbalance: Cluster Randomized Designs

For a fixed total sample size, balanced designs have the greatest sensitivity (smallest MDES)

Sometimes a greater total sample size can be obtained by assigning a larger number of clusters to one of the two treatment groups (usually the control group)

This can result in greater power, greater sensitivity, and smaller MDES that the balanced design (with smaller number of clusters)

For example, the number treated can be limited by resources, research personnel, equipment, etc.

Alternatively, increasing the probability of being randomized to treatment can increase acceptability of random assignment

Improvements in power and design sensitivity are not large, but can be meaningful

Unbalanced Cluster Randomized Designs

Suppose that the number of clusters in the intervention group is fixed at m^T, but the number in the comparison group is not, i.e., $m^C = cm^T$, for some $c \geq 1$ and the size of each cluster is n

The variance of the treatment effect estimate is therefore

$$SE\left(\hat{\gamma}_{01}\right) = \left(\sqrt{\frac{m^T + m^C}{m^T m^C n}} \sqrt{1 + \left(n-1\right)\rho_2} \right) \sigma_T$$

The minimum detectable effect size in the unbalanced design is approximately

$$\delta_{MU} = M_{N-2} \sqrt{\frac{\left(c+1\right)\left[1 + \left(n-1\right)\rho_2\right]}{cm^T n}}$$

The ratio of MDES in unbalanced to balanced designs is approximately

$$\frac{\delta_{MU}}{\delta_M} = \sqrt{\frac{c+1}{2c}}$$

Note that the ratio under the radical cannot be smaller than 1/2 if $c \geq 1$

Unbalanced Cluster Randomized Designs

Note that the maximum possible decrease in MDES is a factor of $\sqrt{1/2} \approx 0.71$

This is achievable only with an infinite sample size

More realistic might be doubling or tripling the size of the control group resulting in a decrease of MDES by a factor of $\sqrt{3/4} = 0.87$ or $\sqrt{4/6} = 0.82$

This might not seem to be a meaningful reduction in MDES, but some funding agencies require that designs achieve 80% power and this can sometimes be essential to do so

Planned imbalance can be combined with covariates to increase sensitivity

Combining Imbalance and Covariates: Cluster Randomized Designs

The test of the hypothesis that the treatment effect is zero, that is

$$\mathrm{H}_0\colon \gamma_{01} = 0$$

is based on the test statistic

$$t = \hat{\gamma}_{01} \Big/ SE\left(\hat{\gamma}_{01}\right)$$

which is taken to have the *t*-distribution with $M - 2 - q$ degrees of freedom (M is the number of clusters), where q is the number of cluster level covariates

In a balanced design with m^T treatment clusters and m^C control clusters all of size *n*

$$SE\left(\hat{\gamma}_{01}\right) = \left(\sqrt{\frac{c+1}{cm^T n}} \sqrt{\overline{R}_1^2 + \left(\overline{R}_2^2 n - \overline{R}_1^2\right)\rho_2} \right) \sigma_T$$

where $m^C = cm^T$,

$$\overline{R}_1^2 = 1 - R_1^2 = \sigma_{A1}^2 \Big/ \sigma_1^2 \quad \text{and} \quad \overline{R}_2^2 = 1 - R_2^2 = \sigma_{A2}^2 \Big/ \sigma_2^2$$

Combining Imbalance and Covariates: Cluster Randomized Designs

When the null hypothesis is false, that is when $\gamma_{01} \neq 0$, then t is taken to have the noncentral t -distribution with $M-2-q$ degrees of freedom and noncentrality parameter

$$\lambda_A = \gamma_{01} \Big/ SE\left(\hat{\gamma}_{01}\right)$$

Comparing the noncentrality parameter of the design with imbalance and covariates to the balanced design (with the same number of treated clusters and no covariates), see that

$$\frac{\lambda_A}{\lambda} = \sqrt{\left(\frac{2c}{c+1}\right)\frac{1+\left(n-1\right)\rho_2}{\overline{R}_1^2+\left(\overline{R}_2^2 n-\overline{R}_1^2\right)\rho_2}}$$

Thus the proportional effects of covariates and imbalance on design sensitivity are multiplicative

The effect of doubling or tripling sample size and covariates can double the noncentrality parameter or cut the MDES and standard error of the estimated treatment effect by half

Combining Imbalance and Covariates: Cluster Randomized Designs

Suppose that the number of clusters in the intervention group is fixed at m^T, but the number in the comparison group is $m^C = cm^T$, there are q level 2 covariates, and the size of each cluster is n

The variance of the treatment effect estimate is therefore

$$SE\left(\hat{\gamma}_{01}\right) = \left(\sqrt{\frac{m^T + m^C}{m^T m^C n}} \sqrt{\bar{R}_1^2 + \left(\bar{R}_2^2 n - \bar{R}_1^2\right)\rho_2} \right) \sigma_T$$

The minimum detectable effect size in the unbalanced design is approximately

$$\delta_{MAU} = M_{N-2} \sqrt{\left(\frac{c+1}{cm^T n}\right)\left(\bar{R}_1^2 + \left(\bar{R}_2^2 n - \bar{R}_1^2\right)\rho_2\right)}$$

The ratio of MDES in unbalanced to balanced designs with covariates to balanced designs without covariates is approximately

$$\frac{\delta_{MAU}}{\delta_M} = \sqrt{\left(\frac{c+1}{2c}\right)\left(\frac{\bar{R}_1^2 + \left(\bar{R}_2^2 n - \bar{R}_1^2\right)\rho_2}{1 + \left(n-1\right)\rho_2}\right)}$$

Cost Efficiency and Optimal Designs: Cluster Randomized Designs

In cluster randomized designs, sensitivity depends on both number of cluster m and cluster size n

Designs with different configurations of m and n can therefore have the same sensitivity

For example, a design with $m = 25$ and $n = 5$ or $m = 15$ and $n = 50$ both have a MDES of 0.49

Similarly, a design with $m = 30$ and $n = 10$ or $m = 40$ and $n = 50$ both have a MDES of 0.34

Which design should be chosen?

One principle for making the choice is **cost efficiency**

Choose the design gives the greatest sensitivity for a fixed cost

Personal view: This principle is helpful in informing design choices, but should never be followed blindly for two reasons:

-It can lead to obviously unwise choices in some cases

-The cost models are used are simplistic and costs can only be crudely approximated

Linear Cost Model

Linear cost model

Assume costs of three types:

Fixed costs of doing the experiment that do not depend (or depend weakly} on size (cost of principle investigator, administration, staff that supervise field operations, statistical analysis, etc.)

Variable costs that depend strongly on sample size (either m or n) are primarily he costs associated with field operations (e.g., recruitment, incentives, materials, and data collection)

The variable costs may be different for different levels of the design

Variable costs can be difficult to know exactly, but can often be estimated approximately based on experience and extrapolation

Variable Costs at the Cluster Level

Recruitment: Costs associated with obtaining agreement to participate in the experiment

-Travel to sites for research team members to explain the study (one or more trips)

-Expendable materials for use in recruiting

Incentives: Pure financial incentive costs are easy to calculate

-Replacement staff (e.g., if teachers need to be removed from classes to be trained)

-Costs of professional development (this can be substantial particularly if a treatment involves all the teachers in a schools)

-Materials for deferred adoption of treatment in control clusters (if offered)

Materials: Expendable material or equipment used in treatment

Variable Costs at the Cluster Level

Data Collection: All costs of obtaining covariate, implementation, and outcome data

-Shipping assessment instruments to and from sites

-Obtaining covariate data at the site level

-Coordinating staff on site to facilitate data collection

-Travel costs for data collection personnel (collection of implementation data via observations is particularly costly)

-Costs for personnel doing qualitative studies of clusters

Feedback to clusters about progress and results (e.g., reports of each cluster's performance)

Variable Costs at the Individual Level

Incentives: Any incentives provided to individuals e.g., students) who participate

Treatment itself: Books, hardware, software, materials needed for the treatment

Data collection: All costs of data collection that can be associated with the individual

-Consumable tests and scoring

-Staff time for interviews, individually administered tests, etc.

Linear Cost Model

Goal: To obtain greatest sensitivity for a fixed cost

Fixed costs do not matter

Compute the cost for each additional **cluster**: Call this c_2

Compute the cost for each additional **individual** in an existing cluster: Call this c_1

The total (variable) cost of and experiment with m clusters per treatment of size n is

$$C = 2mc_2 + 2mnc_1$$

Solving this equation for m yields $m = C/(2nc_1+2c_2)$

Inserting this expression for m into the expression for the variance of the treatment effect and maximizing for n (here C, c_1, and c_2 are fixed) yields n_O the optimal n

Optimal Cluster Size: Cluster Randomized Designs

With no covariates the optimal n has a surprisingly simple form

$$n_O = \sqrt{\left(\frac{c_2}{c_1}\right)\left(\frac{1-\rho_2}{\rho_2}\right)}$$

The qualitative implications are what you would expect

-The larger the (relative) cost of each cluster, (c_2/c_1) the larger n_O becomes

-The larger the intraclass correlation, the smaller n_O becomes

It is also useful to understand this in terms of the level 1 and level 2 variance components

Because ρ_2 is proportional to σ_2^2 and $1-\rho_2$ is proportional to σ_1^2 then

$$n_O = \sqrt{\left(\frac{c_2}{c_1}\right)\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}$$

The larger the (relative) individual variance (σ_1^2/σ_2^2), the larger n_O becomes

Optimal Cluster Sizes for Cluster Randomized Designs as a Function of c_2/c_1 and ρ_2

	ρ_2					
c_2/c_1	0.01	0.05	0.10	0.15	0.20	0.25
1	9.9	4.4	3.0	2.4	2.0	1.7
2	14.1	6.2	4.2	3.4	2.8	2.4
5	22.2	9.7	6.7	5.3	4.5	3.9
10	31.5	13.8	9.5	7.5	6.3	5.5
20	44.5	19.5	13.4	10.6	8.9	7.7
30	54.5	23.9	16.4	13.0	11.0	9.5
40	62.9	27.6	19.0	15.1	12.6	11.0
50	70.4	30.8	21.2	16.8	14.1	12.2
75	86.2	37.7	26.0	20.6	17.3	15.0
100	99.5	43.6	30.0	23.8	20.0	17.3

Optimal Design:
Cluster Randomized Designs

We obtain the m for the experiment by first picking n_O and then selecting the m required to achieve the required design sensitivity

Note that optimal cluster sizes are not integers (rounding is obviously needed)

What surprises most researchers is how small the optimal cluster size often is

For example, if the relative cost of clusters is 10 times that of individuals and the intraclass correlation is 0.20, the optimal cluster size is 6

Few researchers would plan an experiment using only 6 students per school, many might think that 20 – 50 students per school would be needed

The reason these results are possible is that design sensitivity depends so weakly on n

Optimal Cluster Size with Covariates: Cluster Randomized Designs

With covariates the form of the optimal n is only slightly more complex

$$n_O = \sqrt{\left(\frac{c_2}{c_1}\right)\left(\frac{\left(1-R_1^2\right)(1-\rho_2)}{\left(1-R_2^2\right)\rho_2}\right)}$$

The qualitative implications are what you would expect

-The larger the (relative) cost of each cluster, (c_2/c_1) the larger n_O becomes

-The larger the intraclass correlation, the smaller n_O becomes

-The larger $\left(1-R_1^2\right)\Big/\left(1-R_2^2\right)$ becomes, the larger n_O becomes

As an empirical generalization, $R_2{}^2$ is often bigger than $R_1{}^2$, so $1-R_1{}^2 > 1-R_2{}^2$

Therefore the use of covariates often increases n_O

Optimal Cluster Size with Covariates: Cluster Randomized Designs

This can also be better understood in terms of adjusted variance components

$$n_O = \sqrt{\left(\frac{c_2}{c_1}\right)\left(\frac{\sigma^2_{A1}}{\sigma^2_{A2}}\right)}$$

The qualitative relationship with covariates is the same as that without covariates

-The larger the (relative) cost of each cluster, (c_2/c_1) the larger n_O becomes

-The larger the (relative) covariate adjusted individual variance ($\sigma_{A1}/\ \sigma_{A2}{}^2$), the larger n_O becomes

Comments on Optimal Cluster Randomized Designs

Consider optimal design information as informative but not determinative

Small cluster sizes are dangerous: Loss of a few individuals can lead to loss of an entire cluster

Round up to have slightly larger clusters than are necessary

Design parameters (costs, intraclass correlation, and R^2 values) used are often approximate

Work on robustness suggests that underestimation of intraclass correlations impairs efficiency more than overestimation, so assume slightly larger intraclass correlations than expected

The optimal design computed if the intraclass correlation is overestimated by 75%, is 90% as efficient as the truly optimal design

Design Parameters: The Cluster Randomized Design

Design sensitivity of the cluster randomized design depends (for a given significance level and effect size) on three things:

-Sample sizes (m^T, m^C, and n)

-Covariate outcome correlations (if covariates are used) (R_1^2 and R_2^2)

-Intraclass correlation ρ_2

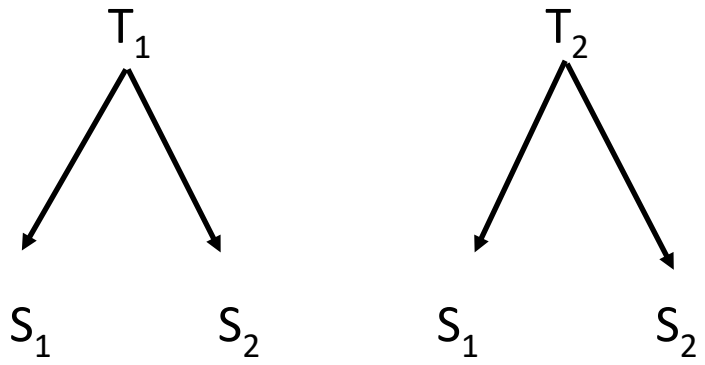
These are called **design parameters**

Information about the design parameters R_1^2 , R_2^2 , and ρ_2 is essential to plan cluster randomized experiments

The Multisite Design

The Multisite Design

The figure below illustrates the multisite design (also called the multisite individually randomized design)



In the language of experimental design, treatments (T) are crossed with sites (S) (every treatment appears in every site)

Why Use a Multisite Design?

Multisite designs are potentially more efficient than individually randomized designs

Multisite designs distribute the benefits of treatment more widely than cluster randomized designs (every site receives some treatment)

Multisite designs require a smaller commitment by sites to treatment than cluster randomized designs (not everyone gets randomized to treatment)

But

Multisite designs are administratively more complex

Contamination between treatment groups in the same site is a possibility

There may be practical, political, or theoretical difficulties in assigning individuals in the same site to different treatments

Sites, Clusters, and Blocks

The term “site “ in this design can be misleading

In experimental design, this design is called the (generalized) **randomized block design** to emphasize that sites are a kind of block—a preexisting aggregate of individuals (you cannot, or do not, randomly assign individuals to blocks)

Blocks may be sites like schools, clinics, or districts

Blocks may also be cohorts of individuals, randomization groups (when there are waves of randomization), grade levels, or treatment providers (therapists, specialists, etc.), or other matched groups of individuals

An extreme example is a design in which pairs of individuals are matched on covariates, then one of each pair is assigned to each treatment group—in this case the pairs are “sites”

Note that this design uses the design principles of matching ***and*** randomization

Fixed and Random Effects and Models for Generalization

Multisite designs introduce a conceptual complexity that does not arise in simpler designs (or it is obscured, as in cluster randomized designs)

What role should statistical inference play in the generalizations drawn from the study?

Alternatively, what, specific parameter are we estimating or testing hypotheses about?

(Statisticians would say, “What is the ***estimand***?”)

In the multisite design there are at least two options:

1) Inferences are about the average treatment effect **in the sites included in the experiment**
2) Inference are about the average treatment effect **in the (super)population of sites** from which those in the experiment are a random sample

Fixed and Random Effects and Models for Generalization

Option #1 (infer to sites included in the experiment) is called the fixed effects estimand

Option #2 (infer to the superpopulation of sites) is called the random effects estimand

Because the statistical inference is about different parameters, the analyses required are different and so are the factors that determine design sensitivity

Both can be technically correct, the choice must be based on extra-statistical considerations

Choosing requires addressing a deep issue of scientific methodology of the limits of statistical inference and its place in scientific inference

Inference and Generalization (Fixed Effects)

The formal **statistical inference** is about the average treatment effect in the sites included in the experiment

But we believe the inferences also apply (generalize to) a broader range of sites

Which sites?

Sites that are “sufficiently similar” to the sites in the experiment that they have the same effects

Sufficiently similar is either a tautology or an **extra-statistical** claim that must be justified based on subject matter expertise

Statistical inference carries relatively little of the burden of inference to scientific conclusions

As we will see, the statistical inference is stronger (higher design sensitivity)

Inference and Generalization (Random Effects)

The formal **statistical inference** is about the average treatment effect in the (super)population of sites (sites in the experiment are a random sample from the superpopulation)

The statistical inferences also apply (generalize) to a broader range of sites

Which sites?

Sites that are included in the superpopulation from which observed sites are a random sample

Without probability sampling of sites, the definition of the superpopulation is a post hoc **extra-statistical** construction that must be justified based on subject matter expertise

Statistical inference carries more of the burden of inference to scientific conclusions, but ***not all of it***

As we will see, the statistical inference is also weaker (lower design sensitivity)

The Place of Experiments in Scientific Inference

Scientific inference involves drawing conclusions about the scope of applicability of conclusions drawn from a study

Scientific inference is a form of argument

Using language from Stephen Toulman's theory of argumentation:

The study design and statistical inference provide part of the warrant for conclusions, but rarely if ever, is that warrant so strong that there are not potential objections that statistics cannot resolve

Those objections can only be resolved through subject matter considerations that are extra-statistical

The Parable of the Two Bridges of Inference

In an important paper on experimental design, Cornfield and Tukey (1956) described two spans of the bridge of inference. They noted that:

> *Inference from the observations to the real conclusions has two parts, only the first of which is statistical*
>
> *Take the simile of the bridge crossing a river by way of an island, there is a statistical span near the bank to the island, and a subject-matter span from the island to the far bank*
>
> *By modifying* [the experimental design and analysis] *we can move the island nearer to or farther from the distant bank, and the statistical span can be made stronger or weaker*
>
> *It is easy to forget the second span … yet a balanced understanding of, and choice among the statistical possibilities requires constant attention to the second span*
>
> *It may often be worthwhile to move the island nearer to the distant bank, at the cost of weakening the statistical span—particularly when the subject-matter span is weak* (p. 913)

Multisite Designs with Random Site Effects

Preview:
Multisite Design (Random Site Effects)

Recall the idea of **simple main effects** of treatments (site-specific treatment effects)

Let $\mu^T{}_a$ and $\mu^C{}_a$ be the treatment and control mean parameters in site a and let $Y^T{}_a$ and $Y^C{}_a$ be their estimates

Then the simple main effect parameter and estimate at site a are $\theta_a = \mu_a{}^T - \mu_a{}^C$ and $T_a = Y_a{}^T - Y_a{}^C$

When sites have random effects, sites are treated as a sample from a population of sites

Thus the simple main effect parameters (the θ_a) are a random sample from a population of effects

The estimand is not the mean of the θ_a's that are observed, but the mean of the entire population of θ_a's (including those that belong to sites that are ***not*** included in the experiment)

If the θ_a's in the experiment were observed, we would know that the mean of the θ_a's would be an estimate of the population mean of the θ_a's and the uncertainty of the sample mean would depend on the variance of the θ_a's

Preview:
Multisite Design (Random Site Effects)

In the multisite design, we do not observe any of the θ_a's directly (they are unknown parameters)

But we do observe estimates of the θ_a's (the T_a's)

It follows that the uncertainty of any estimate of the mean of the θ_a's using the T_a's must depend on the uncertainty (variance) of the θ_a's

The fact that the variance of the treatment effect estimate depends on the variance of the θ_a's makes the analysis of multisite designs with random site effects more complex

This fact also makes multisite designs with random site effects less sensitive than if site effects are fixed

Model and Notation: Multisite Design (Random Site Effects)

Suppose that there are m sites and the j^{th} site assigns n_j^T individuals to treatment and n_j^C individuals to the control condition

This is a true multilevel model

If Y_{ij} is the i^{th} level 1 unit in the j^{th} level 2 unit , the level 1 (individual level) model is

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \varepsilon_{ij}, \text{ and } \varepsilon_{ij} \sim \mathrm{N}(0, \sigma_1^2)$$

where $T_i = \pm \frac{1}{2}$ is a treatment indicator variable. The level 2 (site level) model is

$$\beta_{0j} = \gamma_{00} + \eta_{0j} \quad \text{and} \quad \eta_{0j} \sim \mathrm{N}(0, \sigma_2^2)$$
$$\beta_{1j} = \gamma_{10} + \eta_{1j} \quad \text{and} \quad \eta_{1j} \sim \mathrm{N}(0, \tau_2^2)$$

Note that τ_2^2 is the variance of the θ_a's (the simple main effect parameters)

As a one level model

$$Y_{ij} = \gamma_{00} + \gamma_{10} T_{ij} + \eta_{1j} T_{ij} + \eta_{0j} + \varepsilon_{ij},$$

Note that the residual terms ($\eta_{1j} T_{ij} + \eta_{0j} + \varepsilon_{ij},$) for observations in the same site are not independent

Effect Size:
Multisite Design (Random Site Effects)

The mathematically natural effect size in this design is

$$\delta = \frac{\gamma_{10}}{\sigma_1}$$

This is (almost) the same as in the completely randomized design but different than the cluster randomized design

This choice of effect size for this design is not universal (in the past I advocated using $\delta = \gamma_{10} \Big/ \sqrt{\sigma_1^2 + \sigma_2^2}$)

The two effect sizes are related via the intraclass correlation $\rho_2 = \sigma_2{}^2/(\sigma_1{}^2 + \sigma_2{}^2) = \sigma_2{}^2/\sigma_T{}^2$

$$\frac{\gamma_{10}}{\sigma_1} = \frac{\gamma_{10}}{\sqrt{\sigma_1^2 + \sigma_2^2}} \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2}} = \frac{\gamma_{10}}{\sqrt{\sigma_1^2 + \sigma_2^2}} \sqrt{\frac{1}{1-\rho_2}}$$

However, given that intraclass correlations are typically small ($\rho < 0.25$), the two effect sizes are not very different

For example , if $\rho_2 = 0.2$, then $\sqrt{1/(1-0.2)} = 1.13$

Hypothesis Testing: Multisite Design (Random Site Effects)

The test of the hypothesis that the treatment effect is zero, that is

$$\mathrm{H}_0\colon \gamma_{10} = 0$$

is based on the test statistic

$$t = \hat{\gamma}_{10} \big/ SE\left(\hat{\gamma}_{10}\right)$$

which is taken to have the t-distribution with $m - 1$ degrees of freedom

(The degrees of freedom with fixed site effects were larger: $N - 2m = 2mn - 2m$)

When the null hypothesis is false, that is $\gamma_{10} \neq 0$, then the test statistic has the noncentral t-distribution with $m - 1$ degrees of freedom and noncentrality parameter

$$\lambda = \gamma_{10} \big/ SE\left(\hat{\gamma}_{10}\right)$$

Hypothesis Testing: Multisite Design (Random Site Effects)

In a balanced design where all the $n_j^T = n_j^C = n$ the standard error is

$$\left[SE\left(\hat{\gamma}_{10}\right)\right]^2 = \frac{\tau_2^2 + \sigma_1^2/n}{m} = \frac{\tau_2^2}{m} + \frac{2\sigma_1^2}{mn} = \frac{2\sigma_1^2}{mn}\left[n\tau_2^2/2 + 1\right] = \frac{\sigma_1^2}{mn}\left(n\omega_2^2 + 2\right)$$

where $\omega_2{}^2 = \tau_2{}^2/\sigma_1{}^2$ is the variance of the θ_a divided by σ_1 : The effect size variance

Note that this definition of is also not universal (in the past I advocated using $\omega_2{}^2 = \tau_2{}^2/\sigma_2{}^2$)

In an unbalanced design the standard error is more complex

$$\left[SE\left(\hat{\gamma}_{10}\right)\right]^2 = \left[\sum_{j=1}^{m} \frac{\tilde{n}_j}{\tilde{n}_j\tau_2^2 + \sigma_1^2}\right]^{-1} \text{ where } \tilde{n}_j = \frac{n_j^T n_j^C}{n_j^T + n_j^C}$$

In both cases the variance of the average treatment effect is the inverse of the sum of the inverse variances of the simple main effects

Design Sensitivity: The Multisite Design (Random Site Effects)

Precision

When all the $n_j^T = n_j^C = n$, then

$$SE\left(\hat{\gamma}_{10}\right) = \sqrt{\frac{\left(n\omega_2^2 + 2\right)}{mn}}\sigma_1$$

Statistical Power

$$power = 1 - \mathrm{F}(t_{\alpha/2} \mid df,\, \lambda) + \mathrm{F}(-\, t_{\alpha/2} \mid df,\, \lambda)$$

where $\mathrm{F}(x \mid df,\, \lambda)$ is the cumulative distribution function of the noncentral *t*-distribution with *df* degrees of freedom and noncentrality parameter λ

As an approximation, the noncentral *t*-distribution is *approximately* a translated central *t*-distribution (i.e., with $\lambda = 0$), so $\mathrm{F}(x \mid df,\, \lambda) \approx \mathrm{F}(x - \lambda \mid df,\, 0)$ [works well if *df* and x are fairly large)

This approximation is helpful because you can compute with it in spreadsheets like EXCEL

Approximate Minimal Detectable Effect Size:
Multisite Designs (Random Site Effects)

It is useful to have an algebraic approximation to the minimum detectable effect size

$$\delta_M \approx M_{m-1}\sqrt{\frac{n\omega_2^2 + 2}{mn}}$$

where M_{df} is the constant depending on the degrees of freedom discussed before

Recall that M_{df} is a decreasing function of df and $M_f < 2.9$ for $df > 16$

It might be surprising that the MDES does ***not*** involve the intraclass correlation, but only τ_2^2 and σ_1^2 (via ω_2^2)

Recall that the treatment effect is a mean of simple main effects and the uncertainty of the simple main effect parameters depends on their variance (τ_2^2) and the estimation error in T_a as an estimate of θ_a which depends on σ_1^2

Another way to think about it is that the simple main effects are differences between site-specific means, both means contain the site effect, so the site effect disappears in the difference

Design Sensitivity and Design Parameters:
Multisite Designs (Random Site Effects)

The approximate MDES shows us how sensitivity of the multisite design depends on design parameters

$$\delta_M \approx M_{m-1}\sqrt{\frac{n\omega_2^2+2}{mn}} = M_{m-1}\sqrt{\frac{\omega_2^2+2/n}{m}}$$

We see that δ_M is decreasing function of m and n and an increasing function of $\omega_2{}^2$

We also see that as m becomes large, δ_M tends to zero

Similarly, as $\omega_2{}^2$ becomes large, δ_M becomes large

But, like in cluster randomized designs, as n becomes large δ_M tends to a positive limit

$$\delta_L = M_{m-1}\sqrt{\frac{\omega_2^2}{m}}$$

The Effect Size Variance ω_2^2

Note that we have introduced another **design parameter** ω_2^2

It may not be a parameter about which researchers have much experience or insight

The parameter ω_2^2 is best understood as the ***effect size variance*** across sites

Recall that the simple main effect is $\theta_a = \mu_a^T - \mu_a^C$ thus $\theta_a / \sigma_1 = (\mu_a^T - \mu_a^C)/\sigma_1$ is an effect size

The variance of θ_a / σ_1 is $\omega_2^2 = \tau_2^2/\sigma_2^2$ so ω_2^2 is truly the variance of the simple main effect sizes

Values of ω_2^2 depend on the treatment and the setting and they cannot be known or even estimated until the experiment is conducted

Empirical values of ω_2^2 from experiments in education and social science range from 0 to about 0.20, with the mean and median being about 0.12

Values of 0 are reported in about 30%-40% of experiments, but values of exactly 0 are somewhat suspect

MDES: Multisite Designs (Random Site Effects) as a Function of m, n, and ω_2^2

	ω_2^2						ω_2^2				
m	0	0.05	0.1	0.15	0.25		0	0.05	0.1	0.15	0.25
	$n = 10$						$n = 20$				
5	0.76	0.85	0.93	1.00	1.13		0.54	0.66	0.76	0.85	1.00
6	0.65	0.72	0.79	0.85	0.97		0.46	0.56	0.65	0.72	0.85
7	0.57	0.64	0.70	0.76	0.86		0.41	0.50	0.57	0.64	0.76
8	0.52	0.58	0.64	0.69	0.78		0.37	0.45	0.52	0.58	0.69
9	0.48	0.54	0.59	0.64	0.72		0.34	0.42	0.48	0.54	0.64
10	0.45	0.50	0.55	0.59	0.67		0.32	0.39	0.45	0.50	0.59
15	0.35	0.39	0.43	0.47	0.53		0.25	0.31	0.35	0.39	0.47
20	0.30	0.34	0.37	0.40	0.45		0.21	0.26	0.30	0.34	0.40
25	0.27	0.30	0.32	0.35	0.40		0.19	0.23	0.27	0.30	0.35
30	0.24	0.27	0.29	0.32	0.36		0.17	0.21	0.24	0.27	0.32
40	0.21	0.23	0.25	0.27	0.31		0.15	0.18	0.21	0.23	0.27
50	0.19	0.21	0.23	0.24	0.28		0.13	0.16	0.19	0.21	0.24

Comparing MDES of the Multisite Design (Random Site Effects) to that of the Cluster Randomized Design

Recall that the approximate MDES of a balanced **cluster randomized** design with mn individuals in each treatment group is

$$\delta_M \approx M_{2m-2}\sqrt{\frac{2(n-1)\rho_2 + 2}{mn}}$$

while the MDES of the **multisite** design with mn individuals in each treatment group is

$$\delta_M \approx M_{m-1}\sqrt{\frac{n\omega_2^2 + 2}{mn}}$$

Recalling that M_{df} depends weakly on df, see that the primary difference is the term $2(n-1)\rho_2$ versus the term $n\omega^2$

As an empirical finding (at least in the USA) ω^2 (mean around 0.1) tends to be smaller than $2\rho_2$ (mean around 0.4), thus the multisite design tends to be considerably more sensitive

Moreover n tends to be smaller in multisite designs than in cluster randomized design (further increasing the sensitivity of the multisite design)

Increasing Design Sensitivity:
Multisite Designs (Random Site Effects)

Design sensitivity can be increased by using covariates

Covariates can be added at either level 1 (the individual level) or level 2 (the site level)

The effect of a covariate at a particular level can be understood as decreasing the (effective) variance at that level

Because the relevant level 2 variance component (ω_2^2) may have the largest effect on uncertainty of the treatment effect, level 2 covariates will generally have the largest effect on design sensitivity

But

The relevant level 2 variance component ($\omega_2^2 = \tau_2^2/\sigma_1^2$) is the effect size variance—the function of the covariate is to explain this effect size variance (***not outcome variance***)

There is much less scientific knowledge about covariates that explain treatment effect variance than about covariates that explain outcomes

Model and Notation:
Multisite Design (Random Site Effects)

Suppose there are $q1$ level 1 covariates and $q2$ level 1 covariates

If Y_{ij} is the i^{th} level 1 unit in the j^{th} level 2 unit , the level 1 (individual level) model (one covariate) is

$$Y_{ij} = \beta_{0j}{}^{A} + \beta_{1j}{}^{A} T_{ij} + \beta_{2j}{}^{A} X_{ij} + \varepsilon_{ij}{}^{A}, \qquad \varepsilon_{ij}{}^{A} \sim \mathrm{N}(0, \sigma_{A1}{}^{2})$$

where $T_i = \pm \frac{1}{2}$ is a treatment indicator variable. The level 2 (site level) model (one covariate) is

$$\beta_{0j}{}^{A} = \gamma_{00}{}^{A} + \gamma_{01}{}^{A} W_j + \eta_{0j}{}^{A} \quad \text{and} \quad \eta_{0j}{}^{A} \sim \mathrm{N}(0, \sigma_{A2}{}^{2})$$

$$\beta_{1j}{}^{A} = \gamma_{10}{}^{A} + \gamma_{11}{}^{A} W_j + \eta_{1j}{}^{A} \quad \text{and} \quad \eta_{1j}{}^{A} \sim \mathrm{N}(0, \tau_{A2}{}^{2})$$

$$\beta_{2j}{}^{A} = \gamma_{20}{}^{A},$$

Average treatment effect

As a one level model

$$Y_{ij} = \gamma_{00}{}^{A} + \gamma_{10}{}^{A} T_{\text{ij}} + \gamma_{01}{}^{A} W_1 + \gamma_{11}{}^{A} W_j T_{ij} + \gamma_{20}{}^{A} X_{ij} + \eta_{1j}{}^{A} T_{ij} + \eta_{0j}{}^{A} + \varepsilon_{ij}{}^{A}$$

Hypothesis Testing:
Multisite Design (Random Site Effects)

The test of the hypothesis that the treatment effect is zero, that is

$$\mathrm{H}_0\colon \gamma_{10}{}^{A} = 0$$

is based on the test statistic

$$t = \hat{\gamma}_{10}^{A} \Big/ SE\left(\hat{\gamma}_{10}^{A}\right)$$

which is taken to have the t-distribution with $m - 1 - q2$ degrees of freedom, where $q2$ is the number of level 2 (site level) covariates

When the null hypothesis is false, that is when $\gamma_{10}{}^{A} \neq 0$, then t is taken to have the noncentral t -distribution with $m - 1 - q2$ degrees of freedom and noncentrality parameter

$$\lambda_A = \gamma_{10}^{A} \Big/ SE\left(\hat{\gamma}_{10}^{A}\right)$$

In a balanced design with all of the $n_j{}^{C}$ and $n_j{}^{C}$ equal to $n,$ the covariate adjusted standard error is

$$\left[SE\left(\hat{\gamma}_{10}^{A}\right)\right]^2 = \frac{n\tau_{A2}^2 + 2\sigma_{A1}^2}{mn} = \frac{\left[n\bar{Q}_2^2\omega_2^2 + 2\bar{R}_1^2\right]\sigma_1^2}{mn}$$

Hypothesis Testing:
Multisite Design (Random Site Effects)

The test of the hypothesis that the treatment effect is zero, that is

$$\mathrm{H}_0\colon \gamma_{10}{}^{A} = 0$$

is based on the test statistic

$$t = \hat{\gamma}_{10}^{A} \Big/ SE\left(\hat{\gamma}_{10}^{A}\right)$$

which is taken to have the *t*-distribution with $m - 1 - q2$ degrees of freedom, where $q2$ is the number of level 2 (site level) covariates

When the null hypothesis is false, that is when $\gamma_{10}{}^{A} \neq 0$, then t is taken to have the noncentral t -distribution with $m - 1 - q2$ degrees of freedom and noncentrality parameter

$$\lambda_A = \gamma_{10}^{A} \Big/ SE\left(\hat{\gamma}_{10}^{A}\right)$$

In a balanced design with all of the $n_j{}^{C}$ and $n_j{}^{C}$ equal to n, the covariate adjusted standard error is

$$\left[SE\left(\hat{\gamma}_{10}^{A}\right)\right]^2 = \frac{n\tau_{A2}^2 + 2\sigma_{A1}^2}{mn} = \frac{\left[n\left(1-Q_2^2\right)\omega_2^2 + 2\left(1-R_1^2\right)\right]\sigma_1^2}{mn}$$

We use $Q_2{}^2$ rather than $R_2{}^2$ to emphasize that this is explained variation in ***treatment effects***, not outcome

Comparing Design Sensitivity with and without Covariates:

Multisite Design (Random Site Effects)

Comparing the covariate adjusted to the unadjusted noncentrality parameters in a balanced design where all of the n_j^C and n_j^C are of size n, see that

$$\frac{\lambda_A}{\lambda} = \sqrt{\frac{n\omega_2^2 + 2}{n\left(1 - Q_2^2\right)\omega_2^2 + 2\left(1 - R_1^2\right)}}$$

where $R_I^2 = 1 - \sigma_{AI}^2/\sigma_A^2$ and $Q_2^2 = 1 - \tau_{A2}^2/\tau_2^2$ is the amount of variance across sites ***in treatment effects*** that is explained by covariates

This expression shows that using covariates increases and therefore statistical power, but qualitative generalizations are not obvious because the two terms in numerator and denominator are likely to be of about the same magnitude (neither can be neglected)

Design Sensitivity: Multisite Design with Covariates (Random Site Effects)

Precision

When all the treatment and control clusters are of size n, then

$$SE\left(\hat{\gamma}_{10}^{A}\right) = \sqrt{\frac{\left[n\bar{Q}_2^2\omega_2^2 + 2\bar{R}_1^2\right]\sigma_1^2}{mn}}$$

Statistical Power

$$power = 1 - \mathrm{F}(t_{\alpha/2} \mid df, \lambda) + \mathrm{F}(-t_{\alpha/2} \mid df, \lambda)$$

where $\mathrm{F}(x \mid df, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with df degrees of freedom and noncentrality parameter λ

As an approximation, the noncentral t-distribution is *approximately* a translated central t-distribution (i.e., with $\lambda = 0$), so $\mathrm{F}(x \mid df, \lambda) \approx \mathrm{F}(x - \lambda \mid df, 0)$ [works well if df and x are fairly large)

This approximation is helpful because you can compute with it in spreadsheets like EXCEL

Approximate Minimal Detectable Effect Size: Multisite Design with Covariates (Random Site Effects)

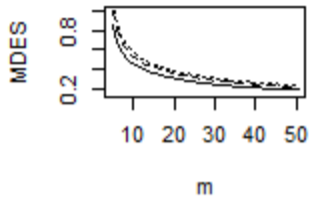
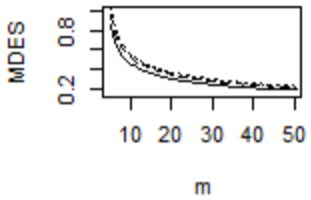
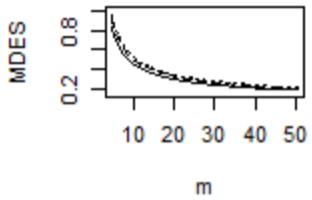
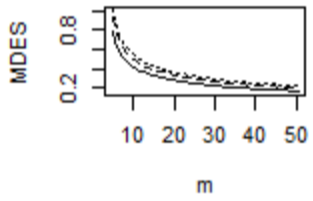
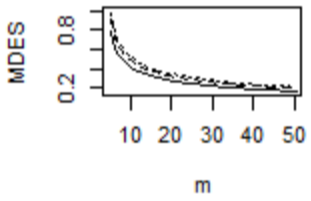
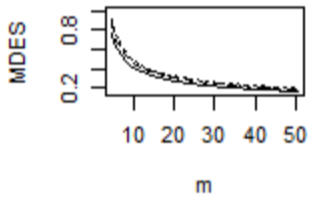
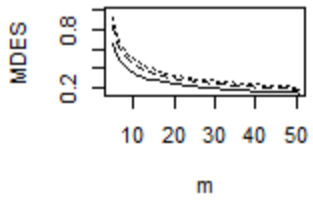
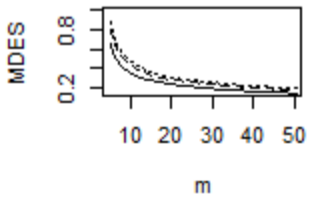
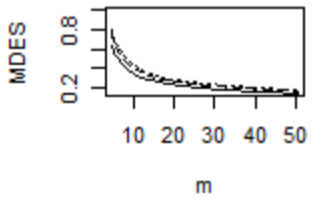
The algebraic approximation to the minimum detectable effect size in a balanced design where all of the n_j^C and n_j^C are of size n *is*

$$\delta_M \approx M_{m-1-q2}\sqrt{\frac{n\left(1-Q_2^2\right)\omega_2^2+2\left(1-R_1^2\right)}{mn}}$$

where M_{df} is the constant depending on the degrees of freedom discussed before

Recall that M_{df} is a decreasing function of df and $M_f < 2.9$ for $df > 16$

MDES as a Function of m in a Multisite Design (Random Site Effects)

	$Q_1 = 0.3$	$Q_1 = 0.5$	$Q_1 = 0.7$
$R_1 = 0.3$			
$R_1 = 0.5$			
$R_1 = 0.7$			

Unbalanced Allocation in Multisite Designs (Random Site Effects)

Suppose that the number of individuals allocated to treatment in the intervention group is fixed at n^T, but the number in the comparison group is not, i.e., $n^C = cn^T$, for some $c \geq 1$

The variance of the treatment effect estimate is therefore

$$SE\left(\hat{\gamma}_{01}\right) = \left(\sqrt{\frac{cn^T\omega_2^2 + c + 1}{cmn^T}}\right)\sigma_T$$

The degrees of freedom of the test statistic are unchanged ($m - 1$) but the noncentrality parameter becomes

$$\lambda_U = \delta\sqrt{\frac{cmn^T}{cn^T\omega_2^2 + c + 1}}$$

Note that as c tends to infinity, the limiting power is not 1 but is determined by the limiting λ value

$$\lambda_L = \delta\sqrt{\frac{mn^T}{n^T\omega_2^2 + 1}}$$

Unbalanced Allocation in Multisite Designs (Random Site Effects)

The minimum detectable effect size in the unbalanced design is approximately

$$\delta_{MU} = M_{m-1}\sqrt{\frac{cn^T\omega_2^2 + c + 1}{cmn^T}}$$

Note that as c tends to infinity, the limiting value of δ_{MU} is not zero but

$$\delta_L = M_{m-1}\sqrt{\frac{n^T\omega_2^2 + 1}{mn^T}} \approx M_{m-1}\sqrt{\frac{\omega_2^2}{m}}$$

the last approximation because $n^T{\omega_2}^2$ is typically considerably larger than 1

Combining Unbalanced Allocation and Covariates: Multisite Design (Random Site Effects)

Combining unbalanced allocation and covariates, the standard error of the treatment effect is

$$\left[SE\left(\hat{\gamma}_{AU}\right)\right]^2 = \frac{\left(1-Q_2^2\right)\tau_2^2}{m} + \frac{(c+1)\left(1-R_1^2\right)\sigma_1^2}{cmn^T} = \frac{\left[cn^T\left(1-Q_2^2\right)\omega_2^2 + (c+1)\left(1-R_1^2\right)\right]\sigma_1^2}{cmn^T}$$

The test statistic has $m-1-q2$ degrees of freedom and noncentrality parameter

$$\lambda_{AU} = \delta\sqrt{\frac{cmn^T}{cn^T\left(1-Q_2^2\right)\omega_2^2 + (c+1)\left(1-R_1^2\right)}}$$

The approximate minimum detectable effect size is

$$\delta_{AUM} = M_{m-1-q2}\sqrt{\frac{cn^T\left(1-Q_2^2\right)\omega_2^2 + (c+1)\left(1-R_2^2\right)}{cmn^T}}$$

Modeling Site-Treatment Interactions Multisite Design (Random Site Effects)

In our specification of the model, we included both site effects and site treatment interactions via random effects

This is unequivocally correct if the effect size variance is zero

An alternative analysis that is sometimes used is to include dummy variables for sites, so-called “site fixed effects” and then analyze the data using ordinary least squares regression

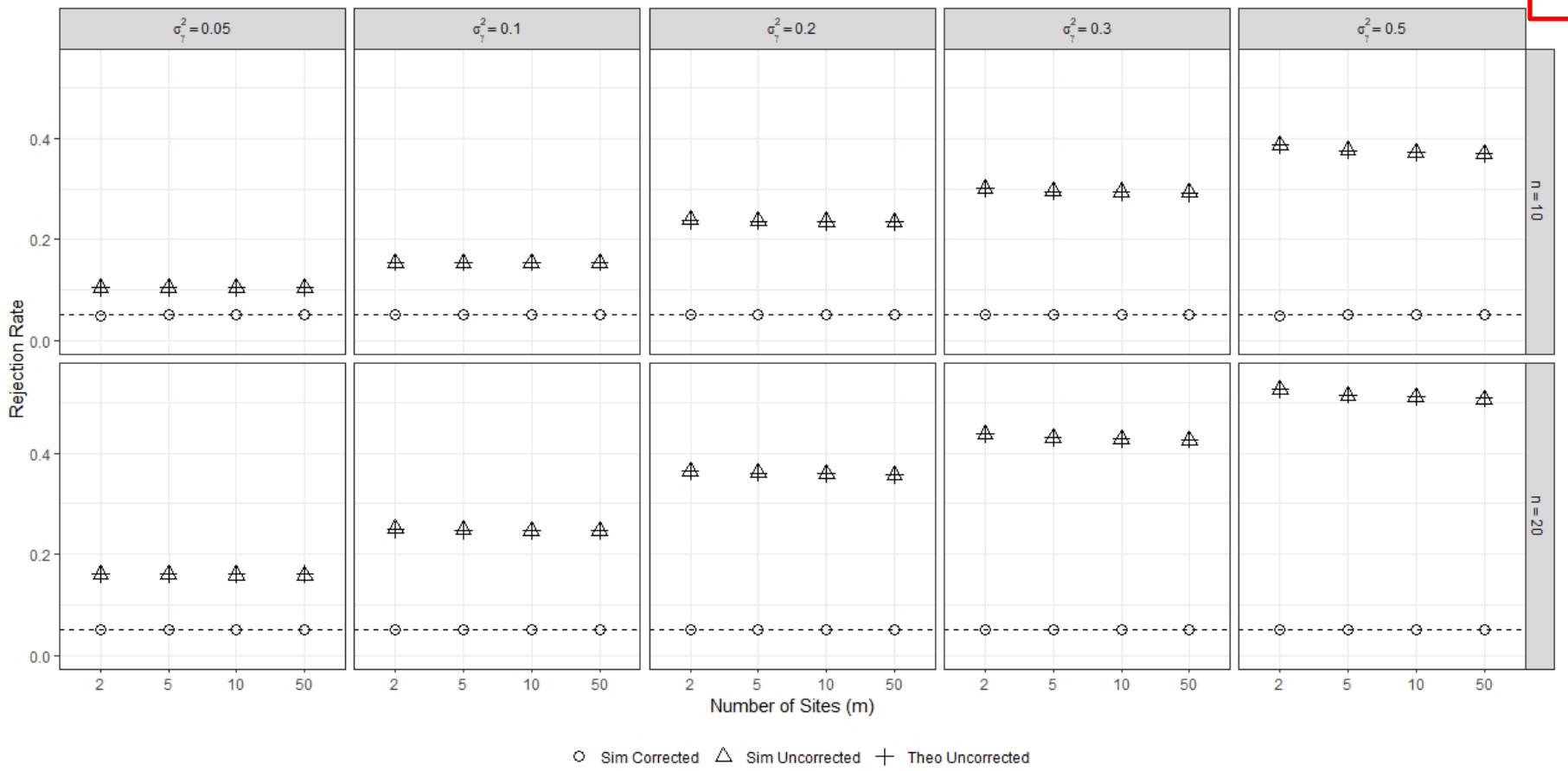
When sites have random effects, this badly inflates the significance levels of the test for treatment effects

When the effect size variance is modest ($\omega_2^2 = 0.10$) the actual rejection rate of the test with a nominal 0.05 (5%) significance level can be 25%

The next figure illustrates the magnitude of the problem

Actual Significance Levels of Nominal 0.05 Level Tests (Multisite Design with Random Site Effects)

Here $\sigma_\gamma^2 = \omega_2^2$



From Chan and Hedges (in press)

Modeling Site-Treatment Interactions in Multisite Designs (Random Site Effects)

Because the site fixed effects analysis is unequivocally correct if effect size variance is exactly zero, one might think that we could simply test for this variance

In other words, first test to see if the effect size variance is nonzero—if we find no effect size variance then use the site fixed effects model

Unfortunately, even the optimal test for effect size variance is not very powerful

Effect size variance that is large enough to seriously compromise the test for the treatment effect are essentially undetectable

The next slide shows the power of the test for effect size variance for plausible multisite designs

Minimum Detectable Effect Size Variance as a Function of m and n: Multisite Design (Random Site Effects)

	Sites Random			Sites Fixed		
m	$n = 10$	$n = 15$	$n = 20$	$n = 10$	$n = 15$	$n = 20$
10	.23	.15	.11	.19	.12	.09
15	.16	.11	.08	.14	.09	.07
20	.13	.09	.07	.12	.08	.06
25	.11	.07	.06	.10	.07	.05
50	.07	.05	.04	.07	.05	.04

Cost Efficiency and Optimal Designs

In multisite designs with random site effects, just as in cluster randomized designs, design sensitivity depends on both m and n and designs with different configurations of m and n can yield the same sensitivity

Yet costs depend differently on m and n so it is sensible to ask what designs achieve the greatest sensitivity for a fixed cost

Compute the cost for each additional **site**: Call this c_2

Compute the cost for each additional **individual** in an existing site: Call this c_1

The total (variable) cost of and experiment with m clusters per treatment of size n is

$$C = mc_2 + 2mnc_1$$

Solving this equation for m yields $m = C/(2nc_1+c_2)$

Inserting this expression for m into the expression for the variance of the treatment effect and maximizing for n (here C, c_1, and c_2 are fixed) yields n_O the optimal n

Optimal Allocation within Sites:
Multisite Design (Random Site Effects)

With no covariates the optimal n has a surprisingly simple form

$$n_O = \sqrt{\left(\frac{c_2}{2c_1}\right)\left(\frac{1}{\omega_2^2}\right)}$$

The qualitative implications are what you would expect

-The larger the (relative) cost of each cluster, (c_2/c_1) the larger n_O becomes

-The larger the effect size variance, the smaller n_O becomes

It is also useful to understand this in terms of the level 1 and level 2 variance components

Because $\omega_2 = \tau_2^2 / \sigma_1^2$ then

$$n_O = \sqrt{\left(\frac{c_2}{2c_1}\right)\left(\frac{\sigma_1^2}{\tau_2^2}\right)}$$

The larger the (relative) individual variance (σ_1^2/τ_2^2), the larger n_O becomes

Optimal Allocations as a Function of c_2/c_1 and ω_2^2: **Multisite Design (Random Site Effects)**

	ω_2^2					
c_2/c_1	0.01	0.05	0.10	0.15	0.20	0.25
1	7.1	3.2	2.2	1.8	1.6	1.4
2	10.0	4.5	3.2	2.6	2.2	2.0
5	15.8	7.1	5.0	4.1	3.5	3.2
10	22.4	10.0	7.1	5.8	5.0	4.5
20	31.6	14.1	10.0	8.2	7.1	6.3
30	38.7	17.3	12.2	10.0	8.7	7.7
40	44.7	20.0	14.1	11.5	10.0	8.9
50	50.0	22.4	15.8	12.9	11.2	10.0
75	61.2	27.4	19.4	15.8	13.7	12.2
100	70.7	31.6	22.4	18.3	15.8	14.1

Optimal Allocation within Sites with Covariates:
Multisite Design (Random Site Effects)

With covariates the form of the optimal n is only slightly more complex

$$n_O = \sqrt{\left(\frac{c_2}{2c_1}\right)\left(\frac{1-R_1^2}{\left(1-Q_2^2\right)\omega_2^2}\right)}$$

The qualitative implications are what you would expect

-The larger the (relative) cost of each cluster, (c_2/c_1) the larger n_O becomes

-The larger the effect size variance, the smaller n_O becomes

-The larger $\left(1-R_1^2\right)/\left(1-Q_2^2\right)$ becomes, the larger n_O becomes

As an empirical generalization, R_1^2 is often bigger than Q_2^2, so $1-R_1^2 < 1-Q_2^2$

Therefore the use of covariates often decreases n_O

Obtaining the Optimal Design:
Multisite Design (Random Site Effects)

We obtain the m for the experiment by first picking n_O and then selecting the m required to achieve the required design sensitivity

Note that optimal allocations are not integers (rounding is obviously needed)

What surprises most researchers is how small the optimal allocation often is

For example, if the relative cost of sites is 10 times that of individuals and the effect size variance is 0.10, the optimal cluster size is 7

Few researchers would plan an experiment using only 7 students per school to each treatment group, many might think that 25 – 30 students per school for each treatment would be needed

The reason these results are possible is that design sensitivity depends weakly on n (but the dependence is stronger than in cluster randomized designs)

Using Optimal Design Information

Optimal design calculations should inform but not completely determine design choices

Optimal designs often have very small allocations that are practically difficult to achieve because they involve singling out small groups of individuals for treatment and assessment of outcomes

Very small optimal allocations may be unwise to use because loss of those few individuals can result in loss of an entire site from the analysis (which can lead to serious reductions in design sensitivity)

Even if attrition of individuals does not lead to loss of sites, sites with very small numbers of individuals can also lead to severe imbalance that can cause serious reductions in design sensitivity

Design Parameters: Multisite Design (Random Site Effects)

Design sensitivity of the cluster randomized design depends (for a given significance level and effect size) on three things:

-Sample sizes (m, n^T, and n^C)

-Covariate outcome correlation at level 1 (if covariates are used) (R_1^2)

-Covariate treatment effect correlation (Q_2^2)

These are called **design parameters**

Information about the design parameters R_1^2 and Q_2^2 is essential to plan multisite experiments with random site effects

What Effect Sizes are Reasonable?

Effect Size Conventions According to Cohen and Lipsey

Cohen (speculative)	Lipsey (empirical)
Small = 0.2	Small = 0.15
Medium = 0.5	Medium = 0.45
Large = 0.8	Large = 0.90

Five-year Impacts of the Tennessee Class-size Experiment

Treatment:

13-17 versus 22-26 students per class

Effect sizes:

0.11 to 0.22 for reading and math

Findings are summarized from Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The Long-Term Effects of Small Classes: A Five-Year Follow-up of the Tennessee Class Size Experiment. *Educational Evaluation and Policy Analysis*, *21*, 127-142.

Annual Reading and Math Growth

Grade Transition	Reading Growth Effect Size	Math Growth Effect Size
K - 1	1.52	1.14
1 - 2	0.97	1.03
2 - 3	0.60	0.89
3 - 4	0.36	0.52
4 - 5	0.40	0.56
5 - 6	0.32	0.41
6 - 7	0.23	0.30
7 - 8	0.26	0.32
8 - 9	0.24	0.22
9 - 10	0.19	0.25
10 - 11	0.19	0.14
11 - 12	0.06	0.01

Based on work in progress using documentation on the national norming samples for the CAT5, SAT9, Terra Nova CTBS, Gates MacGinitie (for reading only), MAT8, Terra Nova CAT, and SAT10.
95% confidence intervals range in reading from +/- .03 to .15 and in math from +/- .03 to .22

Effect Size of Performance Gap Between 50th Percentile and 10th Percentile Schools

Subject and grade	District I	District II	District III	District IV
Reading				
Grade 3	0.31	0.18	0.16	0.43
Grade 5	0.41	0.18	0.35	0.31
Grade 7	.025	0.11	0.30	NA
Grade 10	0.07	0.11	NA	NA
Math				
Grade 3	0.29	0.25	0.19	0.41
Grade 5	0.27	0.23	0.36	0.26
Grade 7	0.20	0.15	0.23	NA
Grade 10	0.14	0.17	NA	NA

Source: Bloom et al. District I outcomes are based on ITBS scaled scores, District II on SAT 9 scaled scores, District III on MAT NCE scores, and District IV on SAT 8 NCE scores.

Demographic Performance Gap in Reading and Math: NAEP Scores

Subject and grade	Black-White	Hispanic-White	Male-Female	Eligible-Ineligible for free/reduced price lunch
Reading				
Grade 4	-0.83	-0.77	-0.18	-0.74
Grade 8	-0.80	-0.76	-0.28	-0.66
Grade 12	-0.67	-0.53	-0.44	-0.45
Math				
Grade 4	-0.99	-0.85	0.08	-0.85
Grade 8	-1.04	-0.82	0.04	-0.80
Grade 12	-0.94	-0.68	0.09	-0.72

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Reading Assessment and 2000 Mathematics Assessment.

Effect Size Results from Randomized Studies

Achievement Measure	*n*	Mean
Elementary School	389	0.33
Standardized test (Broad)	21	0.07
Standardized test (Narrow)	181	0.23
Specialized Topic/Test	180	0.44
Middle Schools	36	0.51
High Schools	43	0.27

How Do We Get Information About Design Parameters Such as ρ or R^2

Empirical Questions about the Predictive Power of Covariates

- What values of intraclass correlations ρ are reasonable?
- What values of R_1^2 and R_2^2 are reasonable?
- How useful are school-level versus student-level pretests?
- How useful are earlier vs. later follow-up years
- Do reading and math achievement behave differently?
- Do earlier and later grades behave differently?

Three Empirical Strategies

Look at national probability samples

Look at large school districts

Look at state census (state assessment) data

National Intraclass Correlations in Reading Achievement (K - 6)

	No Covariates	Demographic Covariates		Pretest Covariate	
Grade	ρ	R_2^2	R_1^2	R_2^2	R_1^2
K	0.233	0.434	0.081	0.742	0.621
1	0.239	0.608	0.084	0.790	0.640
2	0.204	0.559	0.110	0.830	0.522
3	0.271	0.741	0.079	0.759	0.478
4	0.242	0.704	0.100	0.812	0.540
5	0.263	0.798	0.101	0.830	0.565
6	0.260	0.634	0.076	0.882	0.510

National Intraclass Correlations in Reading Achievement (7 - 12)

	No Covariates	Demographic Covariates		Pretest Covariate	
Grade	ρ	R_2^2	R_1^2	R_2^2	R_1^2
7	0.174	---	---	---	---
8	0.197	---	---	---	---
9	0.250	0.424	0.111	0.349	0.459
10	0.183	0.717	0.093	0.856	0.529
12	0.174	0.748	0.091	0.892	0.617
M =	0.224	0.665	0.092	0.774	0.548
a =	0.251	0.691	0.089	0.790	0.566
b =	-0.005	0.013	0.001	-0.003	-0.004

National Intraclass Correlations in Mathematics Achievement (K - 6)

	No Covariates	Demographic Covariates		Pretest Covariate	
Grade	ρ	R_2^2	R_1^2	R_2^2	R_1^2
K	0.243	0.616	0.080	0.857	0.621
1	0.228	0.614	0.079	0.823	0.624
2	0.236	0.436	0.0.88	0.676	0.505
3	0.241	0.639	0.088	0.805	0.594
4	0.232	0.435	0.066	0.679	0.485
5	0.216	0.442	0.072	0.632	0.506
6	0.264	0.117	0.069	0.740	0.502

National Intraclass Correlations in Mathematics Achievement (7 - 12)

	No Covariates	Demographic Covariates		Pretest Covariate	
Grade	ρ	R_2^2	R_1^2	R_2^2	R_1^2
7	0.191	0.638	0.096	---	---
8	0.185	0.433	0.084	0.822	0.653
9	0.216	0.523	0.097	0.895	0.724
10	0.234	0.78	0.092	0.919	0.649
11	0.138	0.739	0.121	0.835	0.73
12	0.239	0.782	0.102	0.975	0.798
M =	0.220	0.447	0.087	0.805	0.616
a =	0.242	0.460	0.083	0.276	0.482
b =	-0.004	0.016	0.002	0.014	0.017

Empirical Analysis (District Sample)

- Estimate ρ, R_2^2 and R_1^2 from data on students from a sample of schools, during multiple years at five urban school districts
- Summarize these estimates for reading and math in grades 3, 5, 8 and 10
- Compute implications for minimum detectable effect sizes

Estimated Parameters for Reading with a School-level Pretest Lagged One Year

	School District				
	A	**B**	**C**	**D**	**E**
Grade 3					
ρ	0.20	0.15	0.19	0.22	0.16
R_2^2	0.31	0.77	0.74	0.51	0.75
Grade 5					
ρ	0.25	0.15	0.20	NA	0.12
R_2^2	0.33	0.50	0.81	NA	0.70
Grade 8					
ρ	0.18	NA	0.23	NA	NA
R_2^2	0.77	NA	0.91	NA	NA
Grade 10					
ρ	0.15	NA	0.29	NA	NA
R_2^2	0.93	NA	0.95	NA	NA

Minimum Detectable Effect Sizes for Reading with a School-Level or Student-Level Pretest Lagged One Year

	Grade 3	Grade 5	Grade 8	Grade 10
40 schools randomized				
No covariate	0.39	0.38	0.42	0.42
School Level	0.26	0.26	0.17	0.11
Student Level	0.26	0.27	0.19	0.10

Empirical Analysis (State Census)

- Estimate ρ_2, ρ_3, R_3^2, R_2^2, and R_1^2 from data on students from all schools in a state and subsets of the state using a three level model (students, schools, districts)
- Summarize these estimates for reading and math in all grades available
- Compute implications for minimum detectable effect sizes

Searchable Website

There is a website with empirical estimates of design parameters for the nation and 11 states (whose development was supported by IES and NSF)

It is searchable by subject matter, grade, state, subset of the state (e.g., low SES, low achievement, etc.)

http://stateva.ci.northwestern.edu/?_gl=1*1rzbp2t*_ga*OTYxNTA0NzAxLjE2NDkyMDA2MzU.*_ga_W4YV5ZK52D*MTY1NTgzNjg4OS4xLjEuMTY1NTgzNjkxNS4zNA..

Design parameters from more states and more different kinds of design parameters are being added as we get them

Key Findings

- Using a pretest improves precision dramatically.
- This improvement increases appreciably from elementary school to middle school to high school because R_2^2 increases.
- School-level pretests produce as much precision as do student-level pretests.
- The effect of a pretest declines somewhat as the time between it and the post-test increases.
- Adding a second pretest increases precision slightly.
- Using a pretest for a different subject increases precision substantially.
- Narrowing the sample to schools that are similar to each other does not improve precision beyond that achieved by a pretest.

References

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, *29*, 30 – 59.

Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlations for planning group-randomized experiments in education. *Educational Evaluation and Policy Analysis*, *29*, 60-87.

Hedges, L. V. & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning 2 and 3 level cluster randomized experiments in education. *Evaluation Review*, *37*, 13-57.

Subgroup (Moderator) Effects

Subgroup (Moderator) Effects

We often want to know if the treatment effects are different for some subgroup of subjects

This is often called a ***moderator analysis*** where the variable defining the subgroups is the moderator variable

The sensitivity of analyses of moderator effects depends on whether the moderator is a property of the groups (clusters or sites) or the individuals within groups

Cluster/site level moderators include characteristics of those sites or ***assigned*** variations of treatment (***not*** observed variations of treatment)

Individual level moderators include characteristics of subjects or assigned (to subjects) variations of treatment (***not*** observed variations of treatment)

Subgroup (Moderator) Effects

Analyses of the effects of cluster/site level moderators compare the effects of a sub-experiment having one value of the moderator with a sub-experiment having a different value of the moderator

Therefore analyses of the effects of cluster/site level moderators are ***less sensitive*** than analyses of the main (overall mean) effect of treatment—often ***much*** less sensitive

Analyses of the effects of cluster/site level moderators compare the effects within clusters/sites of subjects having one value of the moderator with subjects having a different value of the moderator

Therefore analyses of the effects of cluster/site level moderators are ***not*** less sensitive than analyses of the main (overall mean) effect of treatment—they can even be ***more*** sensitive

Must We Match Sampling and Analysis Models?

The Issue

General Question: What happens when you design a study with randomized groups that comprise three levels based on data which do not account explicitly for the middle level?

Specific Example: What happens when you design a study that randomizes schools (with students clustered in classrooms in schools) based on data for students clustered in schools?

Short Answer

Ignoring the top randomized level (e.g., schools) is ***never*** OK

Ignoring a ***middle level*** *in the analysis* has no impact on the accuracy of significance tests in a balanced design

Ignoring a middle level *in the analysis* has little impact on the accuracy of significance tests in most unbalanced designs

Variance component estimates for *both* levels may be biased

The power of the two level analysis is therefore tricky to calculate

Thus specifying sample sizes for the design can be tricky

How Can it be OK to Omit a Level?

The analysis of cluster randomized trials does a *t*-test on cluster means

If there is two-level sampling, the cluster means have variance

$$(\sigma_T^2/n)[1 + (n - 1)\rho_3]$$

If there is three-level sampling, (p subclusters per clusters, n individuals per subcluster) the cluster means have variance

$$(\sigma_T^2/pn)\,[1 + (pn - 1)\rho_3 + (n - 1)\rho_2]$$

The key point is that the assumptions of the *t*-test are still valid—cluster means are independent and have identical variances (in balanced designs), but the precision is different

Example: 3-level vs. 2-level MDES

	MDES			
	3-Level Model		2-Level Model	
Outcomes	**Unconditional**	**Conditional**	**Unconditional**	**Conditional**
Expressive Vocabulary (Spring)	0.482	0.386	0.495	0.311
Stanford 9 Total Math Scaled Score	0.259	0.184	0.259	0.184
Stanford 9 Total Reading Scaled Score	0.261	0.148	0.264	0.150

Sources: The Chicago Literacy Initiative: Making Better Early Readers study (CLIMBERs) database and the School Breakfast Pilot Project (SBPP) d

Variance Component Estimates: 3-level vs. 2-level Analyses

	Variance Components								
	3-Level Model					2-Level Model			
Outcomes	**School**	**Class**	**Student**	**Total**		**School**		**Student**	**Total**
Expressive Vocabulary (Spring)	19.84	32.45	306.18	358.48		38.15		321.11	359.26
Stanford 9 Total Math Scaled Score	115.14	36.40	1273.15	1424.69		131.39		1293.24	1424.63
Stanford 9 Total Reading Scaled Score	108.75	158.95	1581.86	1849.56		181.77		1666.48	1848.25

Sources: The Chicago Literacy Initiative: Making Better Early Readers study (CLIMBERs) database and the School Breakfast Pilot Project (SBPP) datab

Thank You!