Practical Aspects of Design for Group Randomized Trials

Larry V. Hedges

Northwestern University

Prepared for the 2026 IES/NCER Summer Research Training Institute at Northwestern University

Topics for Today

1. Planning a research design
2. Design sensitivity
3. Choosing a particular design
4. Planning cluster randomized designs
5. Planning multisite (randomized block) designs
6. Is it ever OK to ignore levels of sampling in our analyses?

Research Questions Should Drive Design

The most important consideration driving design decisions are the research questions

Several questions are typically of interest, so the issue is ***not*** picking ***the*** research question but prioritizing among the several questions of interest

Research questions are necessarily framed in terms of outcomes, but the same basic question may pertain to one or several different outcomes

Possible framings of research questions:

-Does the treatment have an effect on the outcome?

-How large is the effect?

-Is the effect moderated by another variable?

But, too many research questions compromise design sensitivity

Multiple Significance Tests Compromise the Interpretation of Statistical Significance

Number k of tests	Probably that at least one test is statistically significant
1	0.05
2	0.10
5	0.23
10	0.40
20	0.64
50	0.92

Correcting for Multiple tests Requires a Bigger Sample Size

k	$0.05/k$	Critical value	Power at level α/k if $n = 64$	Required n for 0.80 power after adjustment
1	0.0500	1.98	0.80	64
2	0.0250	2.27	0.71	78
3	0.0167	2.43	0.66	86
4	0.0125	2.53	0.62	91
5	0.0100	2.62	0.59	96
10	0.0050	2.86	0.49	109
20	0.0025	3.09	0.40	122
50	0.0010	3.37	0.30	140

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

Remember that the design must be adequately sensitive to yield unambiguous results

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

Research Design is an Iterative Process

Research design is an iterative process:

1. First you consider what is possible and feasible
2. Then you evaluate sensitivity
3. Finding that it is too small you see if you can change the design to improve sensitivity
4. You may re-evaluate what is possible and feasible
5. …

Obvious design changes could be adding covariates or introducing imbalance, ***if that enables a larger total sample size***

Design changes can often involve implementing the design in cohorts (year 1, year 2)

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 change the design

While the effect size of a treatment may not easily be increased, use of covariates can reduce variation so that the "effective" effect size is increased

Similarly, changing the design (e.g., introducing matching or blocking) may reduce variation

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

There Are Always Potential Conflicts

Practical considerations often push us towards larger sample sizes

Larger sample sizes increase sensitivity but often introduce variation

More variation can introduce possible confounding

Example: To get more schools you may need several districts (or regions or states)—Are districts blocks and do they have random effects?

More schools increases sensitivity, but if districts have random effects that does that ***decrease*** sensitivity?

More outcomes always seem like a good idea, but taking more outcomes into account decreases design sensitivity

Multiple variations of treatment often seems like a good idea, but they require larger sample sizes for an equally sensitive design to detect treatment versus control differences

Sensitivity may differ for different outcomes

Both the main effect of treatment and analysis of moderators seems like a good idea, but sensitivity of tests for moderators may be different than that of the main effect of treatment

This means that compromises are inevitable in planning designs

We Have Focused on Three Basic Designs: The Cluster Randomized Design, the Multi-site Individually Randomized Design, the Multisite Cluster Randomized Design

Why choose one versus the other?

Is one design more sensitive than the other (for a fixed sample size)?

Are there other reasons to choose versus the other?

General Principles

Randomization at the lowest level possible gives the greatest design sensitivity

Generally:

Individually randomized designs are more sensitive than multisite individually randomized designs which are more sensitive than cluster randomized designs

Why?

(This is not obvious)

To Understand This Think About Variance

Compare the individually randomized design with sample size $N = 2mn$ to a cluster randomized design with a sample size of $N = 2mn$ (m clusters pretreatment with n observations per cluster)

The total variance of an individual observation in either design is $\sigma_T^2 = \sigma_1^2 + \sigma_2^2$

In both designs the treatment effect is a mean difference

The variance of the mean difference in the individually randomized design is

$$\frac{4\sigma_T^2}{N} = \frac{2\left(\sigma_1^2 + \sigma_2^2\right)}{mn} = \frac{2\sigma_1^2}{mn} + \frac{2\sigma_2^2}{mn}$$

The variance of the mean difference in the cluster randomized design is

$$\frac{2\left(\sigma_1^2/n + \sigma_2^2\right)}{m} = \frac{2\sigma_1^2}{mn} + \frac{2\sigma_2^2}{m}$$

To Understand This Think About Variance

Compare the individually randomized design with sample size $N = 2mn$ to a cluster randomized design with a sample size of $N = 2mn$ (m clusters pretreatment with n observations per cluster)

The total variance of an individual observation in either design is $\sigma_T^2 = \sigma_1^2 + \sigma_2^2$

In both designs the treatment effect is a mean difference

The variance of the mean difference in the individually randomized design is

$$\frac{4\sigma_T^2}{N} = \frac{2\left(\sigma_1^2 + \sigma_2^2\right)}{mn} = \frac{2\sigma_1^2}{mn} + \frac{2\sigma_2^2}{mn}$$

The variance of the mean difference in the cluster randomized design is

$$\frac{2\left(\sigma_1^2/n + \sigma_2^2\right)}{m} = \frac{2\sigma_1^2}{mn} + \frac{2\sigma_2^2}{m}$$

Note the difference

Interpretation

The same components of variance exist in the observations

We reduce variance in the treatment group means by averaging

Randomization of individuals allows us to average ***every*** component of variance over ***all*** observations in each group

Randomization over clusters means that the cluster-specific component of variance is only ***averaged over*** m the number of ***clusters***, not mn, the number of observations in each group

Different research designs partition the variance so that different components of variance contributed to the variance (the uncertainty)of the treatment

Compare Multisite Individual Randomization to Cluster Randomization

Compare the individually multisite individually randomized design with sample size $N = 2mn$ (m sites with n individuals per treatment per site) to a cluster randomized design with a sample size of $N = 2mn$ (m clusters per treatment with n observations per cluster)

The total variance of an individual observation is $\sigma_T^2 = \sigma_1^2 + \sigma_{2'}^2 + \tau_2^2$

In both designs the treatment effect is a mean difference

The variance of the mean difference in the individually randomized design is (as before)

$$\frac{2\left(\sigma_1^2 + \sigma_{2'}^2 + \tau_2^2\right)}{mn} = \frac{2\sigma_1^2}{mn} + \frac{2\sigma_{2'}^2}{mn} + \frac{2\tau_2^2}{mn}$$

The variance of the mean difference in the multisite individually randomized design is

$$\frac{2\sigma_1^2/n + \tau_2^2}{m} = \frac{2\sigma_1^2}{mn} + \frac{\tau_2^2}{m}$$

Compare Multisite Individual Randomization to Cluster Randomization

Compare the individually multisite individually randomized design with sample size $N = 2mn$ (m sites with n individuals per treatment per site) to a cluster randomized design with a sample size of $N = 2mn$ (m clusters per treatment with n observations per cluster)

The total variance of an individual observation is $\sigma_T^2 = \sigma_1^2 + \sigma_{2'}^2 + \tau_2^2$

In both designs the treatment effect is a mean difference

The variance of the mean difference in the individually randomized design is (as before)

$$\frac{2\left(\sigma_1^2 + \sigma_{2'}^2 + \tau_2^2\right)}{mn} = \frac{2\sigma_1^2}{mn} + \frac{2\sigma_{2'}^2}{mn} + \frac{2\tau_2^2}{mn}$$

Note that what was σ_2^2 previously is now $\sigma_{2'}^2 + \tau_2^2$

The variance of the mean difference in the multisite individually randomized design is

$$\frac{2\sigma_1^2/n + \tau_2^2}{m} = \frac{2\sigma_1^2}{mn} + \frac{\tau_2^2}{m}$$

Compare Multisite Individual Randomization to Cluster Randomization

Compare the individually multisite individually randomized design with sample size $N = 2mn$ (m sites with n individuals per treatment per site) to a cluster randomized design with a sample size of $N = 2mn$ (m clusters per treatment with n observations per cluster)

The total variance of an individual observation is $\sigma_T^2 = \sigma_1^2 + \sigma_{2'}^2 + \tau_2^2$

In both designs the treatment effect is a mean difference

The variance of the mean difference in the individually randomized design is (as before)

$$\frac{2\left(\sigma_1^2 + \sigma_2^2 + \tau_2^2\right)}{mn} = \frac{2\sigma_1^2}{mn} + \frac{2\sigma_{2'}^2}{mn} + \frac{2\tau_2^2}{mn}$$

Note the difference

The variance of the mean difference in the multisite individually randomized design is

$$\frac{2\sigma_1^2/n + \tau_2^2}{m} = \frac{2\sigma_1^2}{mn} + \frac{\tau_2^2}{m}$$

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

What is the Alternative to Cluster Randomization?

The multisite individually randomized (randomized block) design is an alternative you know

Compare the two designs with total sample size $N = 2mn$:

-For the CRT m clusters of size n assigned to each treatment group

-For the MSIRT m sites with n individuals assigned to each treatment at each site

Compute the (approximate) minimum detectable effect size for each design

Comparing Minimum Detectable Effect Sizes

The approximate minimum detectable effect size for the cluster randomized design is

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

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 $2.8 < M_{df} < 2.9$

The approximate minimum detectable effect size for the multisite individually randomized design is

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

Thus, the ratio of the minimum detectable effect sizes for these two designs is

$$\frac{\delta_{MCR}}{\delta_{MMS}} \approx \sqrt{\frac{2\left[1+\left(n-1\right)\rho\right]}{2+n\omega^2}} = \sqrt{\frac{1+\left(n-1\right)\rho}{1+\left(n/2\right)\omega^2}} \approx \sqrt{\frac{1+n\rho}{1+n\omega^2/2}}$$

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 $M_{df} = 1.984 + 0.845 = 2.83$

A Potential Problem with Multisite Designs: Contamination

In multisite designs, both treatments are given to individuals in the same sites

If the individuals assigned to the control group experience some of the treatment, this may weaken the treatment contrast and the treatment effect

This phenomenon is called ***contamination***

It is a ***potential*** problem with multisite designs, but how big a problem is it?

Contamination

Contamination is often less of a problem than it might seem

Some treatments require special materials or materials

Some treatments require substantial specialized training

Even if there is ***some*** contamination that weakens the treatment contrast, the multisite design may still be more powerful (despite a weakened treatment effect) (see Rhoads, 2011)

However: Contamination is a potential problem—if you use a multisite design, you are obligated to explain why contamination is not likely to be an important problem in your study

What if I Chose a Cluster Randomized Design, But I Don't Have the Required Design Sensitivity?

Consider adding covariates

The approximate minimum detectable effect size with covariates is

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

where M_{df} is the constant depending on the degrees of freedom discussed before and $\bar{R}_i^2 = 1-R_i^2$

R_1^2 and R_2^2 are the (squared) level 1 and level 2 covariate-outcome correlations

Note that R_2^2 matters ***much*** more than R_1^2

Comparing Sensitivity with and without Covariates: Cluster Randomized Designs

Compare expressions for the MDES with and without covariates

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

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

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

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

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 $\bar{R}_1^2 = 0$, $\delta_{AM} > 0$ in that case

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

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

$$\delta_{AM} \approx M_{2m-2-q}\sqrt{\frac{2\bar{R}_1^2(1-\rho)}{mn}}$$

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

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 minimum detectable effect size in the unbalanced design is approximately

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

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

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

The Magic Number 0.7

Notice that a pretty good cluster level covariate can decrease the minimum detectable effect size by a factor of about 0.7

The maximum decrease in minimum detectable effect size attainable via imbalance is about 0.7

Neither of these strategies (covariates or imbalance) can completely solve the sample size problem, but they can help

How Should I Pick the Cluster Sample Size?

Cost Efficiency and Optimal Designs: Cluster Randomized Designs

In cluster randomized designs, sensitivity depends on both number of clusters 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

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

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 $1-\rho$ 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

	ρ					
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

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

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 $\left(1-R_1^2\right)/\left(1-R_2^2\right)$ 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

The Multisite Design

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

Multisite Design (Random Site Effects)

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

Let μ^T_a and μ^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

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

The design parameter that describes the heterogeneity of treatment effects across sites is the effect size variance: $\text{Var}\{\theta_a/\sigma_1^2\} = \omega_2^2 = \tau_2^2/\sigma_1^2$

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 (**not** ω_2^2) from experiments in education and social science range from 0 to about 0.30, 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

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

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

ω_2^2 is the effect size variance of simple main 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

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

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

	ω_2						ω_2				
m	0	0.1	0.2	0.3	0.5		0	0.1	0.2	0.3	0.5
	$n = 10$						$n = 20$				
5	0.76	0.78	0.83	0.91	1.13		0.54	0.56	0.63	0.74	1.00
6	0.65	0.66	0.71	0.78	0.97		0.46	0.48	0.54	0.63	0.85
7	0.57	0.59	0.63	0.69	0.86		0.41	0.43	0.48	0.56	0.76
8	0.52	0.53	0.57	0.63	0.78		0.37	0.39	0.44	0.51	0.69
9	0.48	0.49	0.53	0.58	0.72		0.34	0.36	0.40	0.47	0.64
10	0.45	0.46	0.49	0.54	0.67		0.32	0.34	0.38	0.44	0.59
15	0.35	0.36	0.39	0.42	0.42		0.25	0.26	0.30	0.34	0.47
20	0.30	0.31	0.33	0.36	0.36		0.21	0.22	0.25	0.29	0.40
25	0.27	0.27	0.29	0.32	0.32		0.19	0.20	0.22	0.26	0.35
30	0.24	0.25	0.26	0.29	0.29		0.17	0.18	0.20	0.24	0.32
40	0.22	0.23	0.24	0.27	0.27		0.16	0.17	0.19	0.22	0.29
50	0.21	0.21	0.23	0.25	0.25		0.15	0.16	0.17	0.20	0.27

What if I Chose a Multisite Individually Randomized Design, But I Don't Have the Required Design Sensitivity?

Consider adding covariates

The approximate minimum detectable effect size with covariates 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 and R_1 is the level 1 covariate-outcome correlation and Q_2 is the ***treatment effect-outcome correlation***

Note that it may be difficult to obtain covariates that explain a substantial amount of treatment effect variation, so you may not be able to rely on a large value of Q_2^2

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

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 approximate minimum detectable effect size is

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

As c becomes large δ tends to

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

Optimal Allocation within Sites with Covariates:
Multisite Individually Randomized Design

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

Optimal Allocations as a Function of c_2/c_1 and ω_2: Multisite Individually Randomized Design

	ω_2					
c_2/c_1	0.01	0.05	0.10	0.15	0.20	0.25
1	70.7	14.1	7.1	4.7	3.5	2.8
2	100.0	20.0	10.0	6.7	5.0	4.0
5	158.1	31.6	15.8	10.5	7.9	6.3
10	223.6	44.7	22.4	14.9	11.2	8.9
20	316.2	63.2	31.6	21.1	15.8	12.6
30	387.3	77.5	38.7	25.8	19.4	15.5
40	447.2	89.4	44.7	29.8	22.4	17.9
50	500.0	100.0	50.0	33.3	25.0	20.0
75	612.4	122.5	61.2	40.8	30.6	24.5
100	707.1	141.4	70.7	47.1	35.4	28.3

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

Multisite Cluster Randomized Design

Sometimes we have multiple sites (or blocks) but we can only assign intact groups within sites to treatments

The intuition is that we want to assign to treatments within sites, but it is unfeasible or impossible to assign individuals to treatments

For example,

we have multiple schools, but we must assign classrooms to treatments

we have multiple districts, but we must assign schools to treatments

This leads to a multisite cluster randomized design

Do Sites Have Fixed or Random Effects?

This decision has important consequences for design sensitivity

Suppose that there are m sites, with p clusters of size n assigned to each treatment within each site

If sites have fixed effects, the minimum detectable effect size is

$$\delta_{MF} \approx M_{2m(p-1)}\sqrt{\frac{2\left[1+\left(n-1\right)\rho_2\right]}{mpn}}$$

But if sites have random effects the minimum detectable effect size is

$$\delta_{MR} \approx M_{m-1}\sqrt{\frac{pn\omega_3^2+2\left[1+\left(n-1\right)\rho_2\right]}{mpn}}$$

If the number of sites m is small (or is big), δ^{MR} is much bigger than δMFn

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

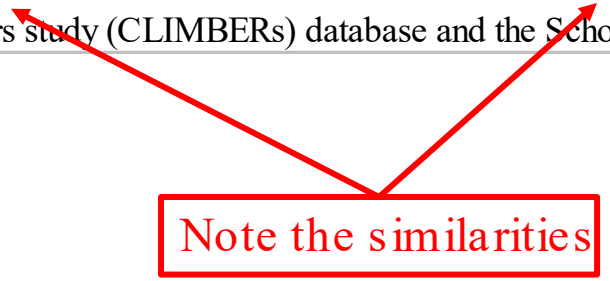
$$(\sigma_T^2/pn)\,[1 + (pn - 1)\rho_3 + (n - 1)\rho_2]$$

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!