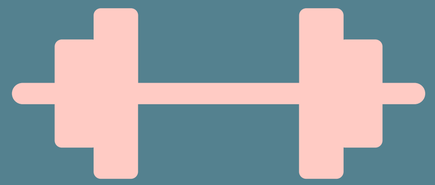
Introduction to

SENSITIVITY AND POWER



Elizabeth Tipton

TOPICS

1. Overview of design sensitivity
2. Cluster randomized trials
3. Multisite randomized trials
4. Design parameters

DESIGN SENSITIVITY

SUMMARIZING THE ATE

- In an experiment, we randomize units to T or C and then we estimate the average treatment effect (ATE). We can report this estimate $\hat{\delta}$ and its standard error $SE(\hat{\delta})$.
  - We could combine these into a 95% CI:

$$\hat{\delta} \pm t_{df} SE(\hat{\delta})$$

  - We may want to know if there is evidence that the true $\delta \neq 0$. For this, we can use a hypothesis test based upon the statistic,

$$t = \frac{\hat{\delta}}{SE(\hat{\delta})}.$$

DESIGN SENSITIVITY

- These approaches focus on the correct Type I error.
  - They keep the likelihood small (e.g., $p < .05$) that we would reject the null hypothesis and conclude that the treatment causes changes in the outcome when this is not the case.
- But they do not address Type II error.
  - That is, it is possible that we could design a study in which the 95% CI would *nearly always* include 0, or the hypothesis test would *never* be rejected, even when the intervention really did cause the outcome to change.
- **This concerns is with design sensitivity - is our study the right design, with the right sample size, etc to ensure that if there really is an effect we would be likely to find it?**

DESIGN SENSITIVITY

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
  - 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 (MDES)**: The smallest effect size for which the design has specified power
  - Typically use the 0.05 significance level, with 80% power

IMPROVING SENSITIVITY

These concepts are highly related. We will focus on both power and the MDES.

- Our approach will be:
  - Understand the relationship between different designs and parameters and sensitivity.
  - Conceive of ways that we can *improve* such sensitivity, while *also* keeping our estimand (the focus of our study) valid and useful.

CAVEAT ABOUT POWER

- Researchers often refer to the “power of the design” (e.g., “this study is underpowered”).
- **But this isn’t quite correct. Power always refers to:**
  - A specific parameter (e.g., the ATE δ)
  - A specific value of this parameter (e.g., $\delta = \delta_a$)
- It is possible for a design to have adequate power for one parameter (e.g., the ATE) and not for another (e.g., variation treatment effects). Or to have adequate power δ_a but not for another δ_b.

CAVEAT ABOUT MDES

- It is easy to get confused about what the MDES means.
- The MDES is the smallest possible value of the ATE δ that can be detected with a specified power.
  - e.g., if the MDES = 0.19 at 80% power, there is at least 80% power to detect effect sizes of 0.19, 0.20, 0.21, ... and so on.

CLUSTER RANDOMIZED DESIGNS

CLUSTER RANDOMIZED DESIGN

- We begin by focusing on a simple design in which:
  - Students are nested in schools
  - Schools (clusters) are randomly assigned to T or C
- Notice that this same design could:
  - Randomize classrooms, teachers, community centers, neighborhoods, etc
  - Key feature: groups (clusters) of units are created non-randomly before the study begins

MODEL

Let Y_{ij} be the outcome for the ith student (unit) in the jth school (cluster). Let $T_j = \pm 1/2$ indicate if school j is assigned to T. We can model this using:

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

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

Which we can combine into the model

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \eta_j + \epsilon_{ij}$$

INTRACLASS CORRELATION

The total variance is thus,

$$Var(Y_{ij}) = Var(\eta_j) + Var(\epsilon_{ij}) = \sigma_2^2 + \sigma_1^2 = \sigma_T^2$$

The proportion of total variance that is attributable to clusters is thus,

$$\rho_2 = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2}$$

Which is called the ‘intraclass correlation coefficient’ (ICC). In education, these values are typically between about 0.10 to 0.30, depending upon the outcome and population.

EFFECT SIZE AND TEST

We can turn the ATE into an effect size using $\delta = \frac{\gamma_1}{\sigma_T}$. Our NH is

$$H_0 : \delta = 0$$

And we test this using the statistic,

$$t = \frac{\hat{\delta}}{SE(\hat{\delta})}$$

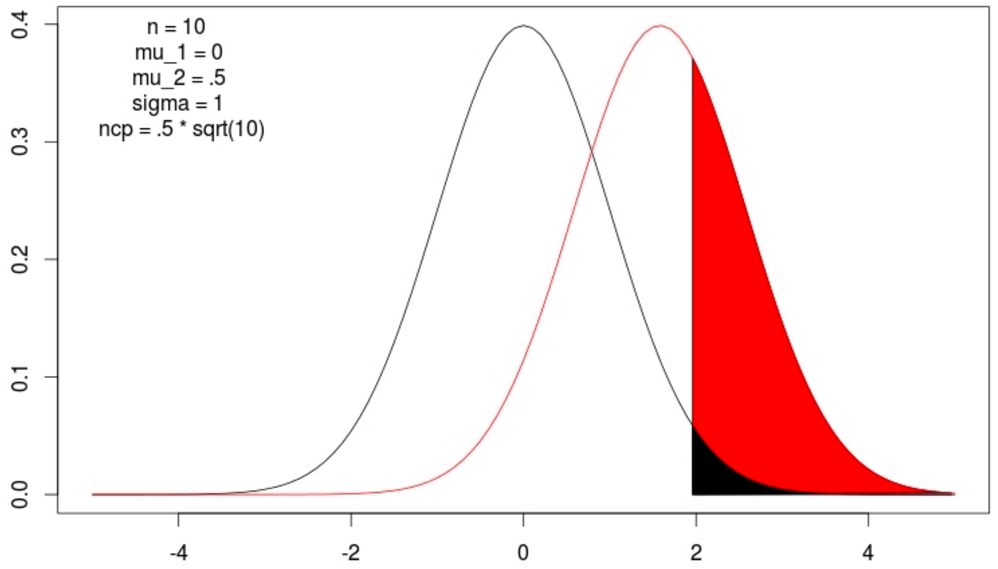
Where $SE(\hat{\delta}) = \sqrt{\frac{m_t + m_c}{m_t m_c n}} \sqrt{1 + (n-1)\rho_2}$

SAMPLING DISTRIBUTIONS

When the NH is true, i.e., when $\delta = 0$, the t-test follows a t-distribution with $df = M - 2$.

When the NH is false, i.e., when $\delta \neq 0$, the t-test follows a non-central t-distribution with $df = M - 2$ and non-centrality parameter

$$\lambda = \frac{\delta}{SE(\hat{\delta})}$$



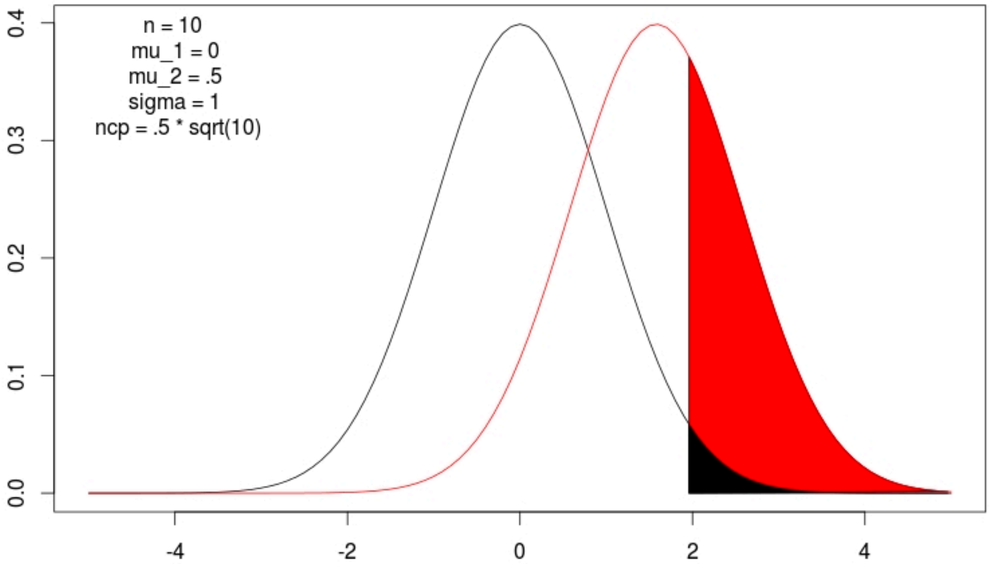
STATISTICAL POWER

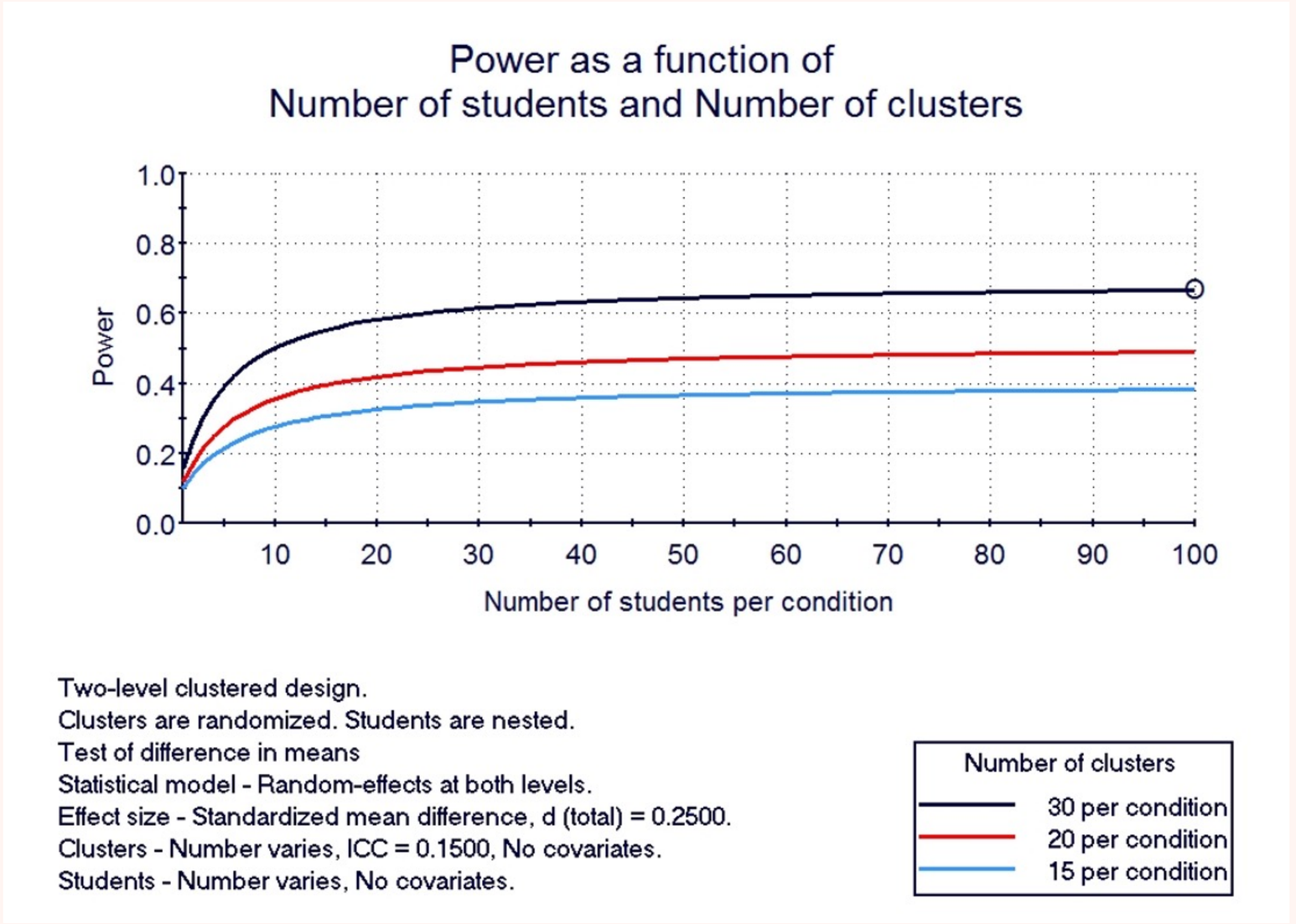
The **black** shaded area is the **Type I error** (e.g., 0.05).

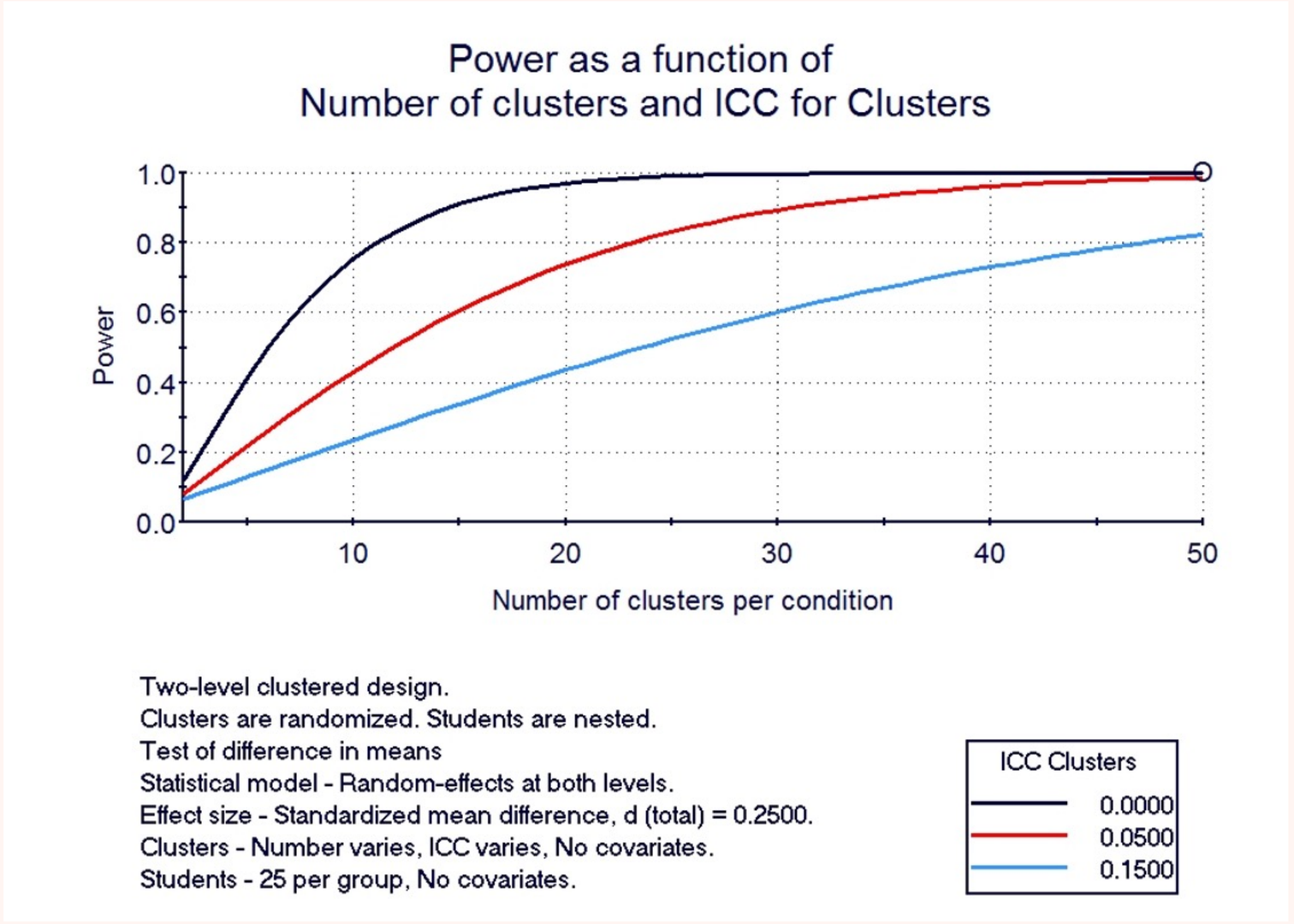
The **red** shaded area is the **Type II error** (e.g., 0.20) for a specific $\delta = \delta_a$.

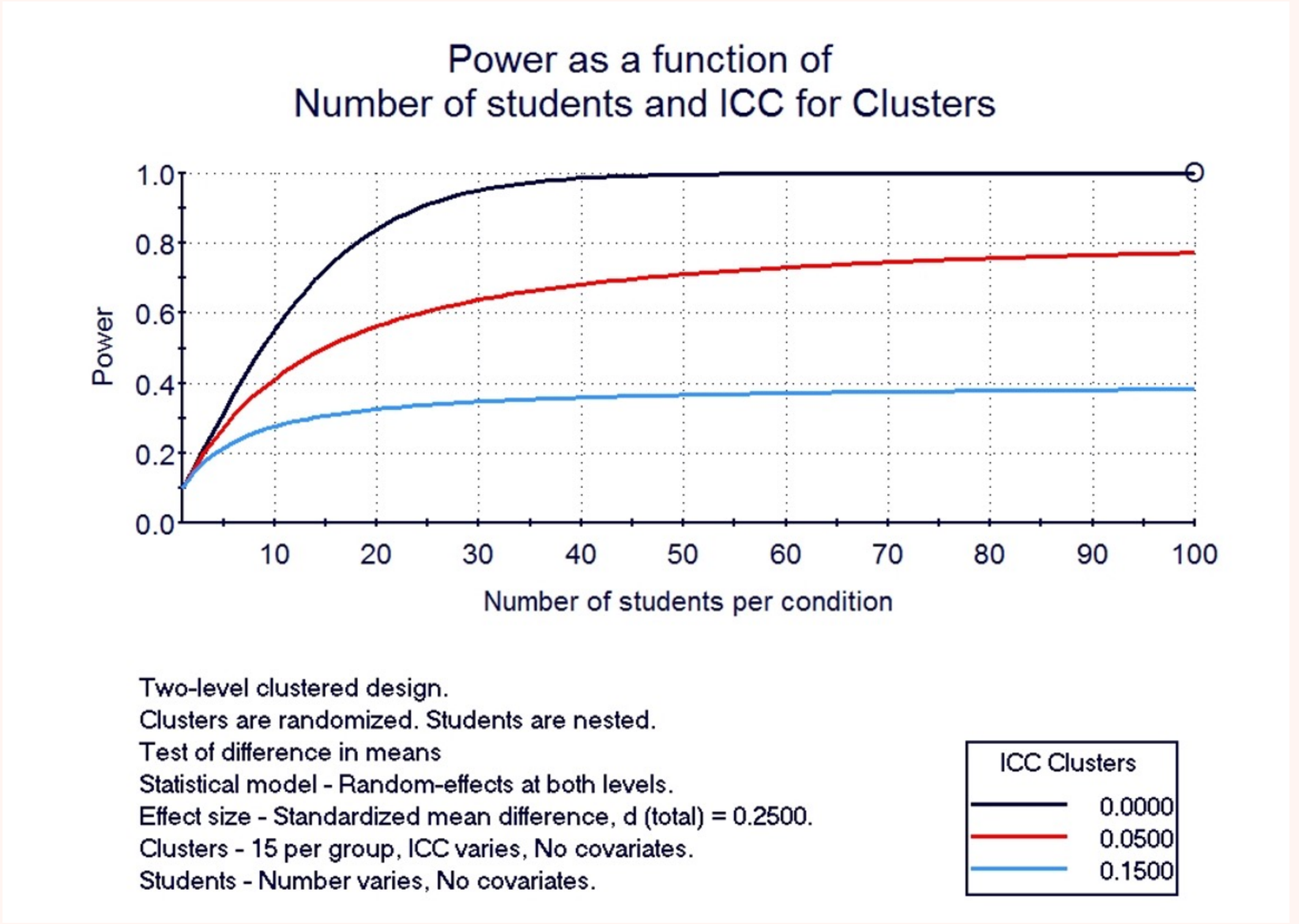
The **statistical power** is

$$Power(\delta_a) = 1 - Type\ II\ Error = 1 - F(t_{\alpha/2} \mid df, \lambda) + F(-t_{\alpha/2} \mid df, \lambda)$$




Power as a function of
Number of students and Number of clusters
Power
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
60
70
80
90
100
Number of students 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 = 0.1500, No covariates.
Students - Number varies, No covariates.
Number of clusters
30 per condition
20 per condition
15 per condition


Power as a function of
Number of clusters and ICC for Clusters
Power
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
Number of clusters 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


Power as a function of
Number of students and ICC for Clusters
Power
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
60
70
80
90
100
Number of students 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 - 15 per group, ICC varies, No covariates.
Students - Number varies, No covariates.
ICC Clusters
0.0000
0.0500
0.1500

MDES

If we begin by stating the Type I and II error thresholds we can live with, we can rearrange this into the MDES,

$$\delta_M \approx M_{M-2}\sqrt{\frac{m_t + m_c}{m_t m_c n}[1 + (n-1)\rho_2]}$$

Where if $df = M - 2 > 16$ then $2.8 < M_{df} < 2.9$.

In a balanced design with $m_t = m_c = m$ this can be simplified into

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

N VS M IMPORTANCE

Which matters more: the within school sample size (n) or the number of schools (m)?

Let's look at the MDES in the balanced case:

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

Thus, if we let $n \to \infty$ then $\delta_M \to M_{2m-2}\sqrt{2\rho_2/m}$.

The point is that the number of schools (clusters) is more consequential than the number of students (units).

VALUES OF M_{df}

df	*Mdf*		*df*	*Mdf*
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

INCREASING SENSITIVITY APPROACHES

LET'S ADD COVARIATES

Suppose X_{ij}^c is a centered level 1 student (unit) covariate and W_j is a level 2 school (cluster) level covariate.

Let Y_{ij} be the outcome for the ith student (unit) in the jth school (cluster). Let $T_j = \pm 1/2$ indicate if school j is assigned to T. We can model this using:

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

$$\beta_{0j} = \gamma_0^A + \gamma_1^A T_j + \gamma_2^A W_j + \eta_j^A \qquad \text{where } \eta_j^A \sim N(0, \sigma_{A2}^2)$$

$$\beta_{1j} = \gamma_3^A$$

Which we can combine into the model

$$Y_{ij} = \gamma_0^A + \gamma_1^A T_j + \gamma_2^A W_j + \gamma_3^A X_{ij}^c + \eta_j^A + \epsilon_{ij}^A$$

WHAT ABOUT δ^A?

Does $\delta = \delta^A$?

In a randomized design, YES, these are equivalent!!

This is because $Corr(T_j, X_{ij}^c) = Corr(T_j, W_j) = 0$ as a result of randomization.

Thus in an RCT, we include covariates not to improve the estimate of the ATE, but instead to improve the sensitivity / precision of this estimate.

EFFECT OF ADJUSTMENTS

Now, we can show that the standard error can be written,

$$SE(\hat{\delta}^A) = \sqrt{\frac{m_t + m_c}{m_t m_c n}} \sqrt{\bar{R}_1^{\,2} + (\bar{R}_2^{\,2} n - \bar{R}_1^{\,2})\rho_2}$$

Where $\bar{R}^2 = 1 - R^2$.

Since $\bar{R}^2 \leq 1$ then including covariates reduces the standard error, which then:

- Increases the non-centrality parameter
- Increases statistical power
- Reduces the MDES

DESIGN PARAMETERS

We've shown that the sensitivity of the cluster randomized design is a function of:

- Numbers of clusters: m_t, m_c
- Numbers of units within clusters: n
- ICC: ρ_2
- Amount of variation explained by covariates: R_1^2, R_2^2

SOME ADVICE

- Consider optimal design information as informative but not determinative
- Small cluster sizes are dangerous: losing a few individuals can mean losing the whole cluster
- Round up to have slightly larger clusters than are necessary
- Remember that the design parameters you choose are approximations
- It’s safer to over-estimate the ICC than underestimate it
- Inclusion of cluster-level covariates has a larger effect on power than individual level covariates

MULTISITE RANDOMIZED DESIGNS

MULTISITE RANDOMIZED DESIGN

- We begin by focusing on a simple design in which:
  - Students (units) are nested in schools (sites)
  - Within each school (site) we randomize students (units) to T or C
- Notice that this same design could:
  - Randomize classrooms or teachers within schools
  - Randomize schools within districts
  - Key feature: randomization is *within* groups

AVERAGE OF AVERAGES

- **In this design, we essentially have m separate simple RCTs.**
  - Within each of m sites, we can estimate a separate ATE: $\delta_1, \delta_2, \ldots, \delta_m$.
- **Our focus is often on the average of these ATEs:** $\delta = ave(\delta_1, \ldots, \delta_m)$
  - But we can also study the variation in these ATEs, the distribution of these ATEs, and so on. *The sensitivity / precision of these estimates, however, is smaller than for the average.*

SITES AS RANDOM? FIXED?

- In the cluster-randomized design, we talked about the schools (clusters) as a random sample of schools (clusters) in the population.
- In the multisite randomized design, we can also talk about the schools (sites) as a random sample of schools (sites) in the population. In this case, we are treating the schools (sites) as **random**.
  - It is possible, however, to conceive of the schools (sites) as **fixed**. In this case, we are limiting our inferences to the schools (sites) in the study.
  - Our focus will be on sites as random.

MULTISITE DESIGN

Suppose there are m sites and the jth site randomizes n_j^t students to treatment and n_j^c students to control.

Now let $T_{ij} = \pm 1/2$ indicate if student i in site j is assigned to T. We can then write,

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \epsilon_{ij} \qquad \text{with } \epsilon_{ij} \sim N(0, \sigma_1^2)$$

$$\beta_{0j} = \gamma_0 + \eta_{0j} \qquad \text{with } \eta_{0j} \sim N(0, \sigma_2^2)$$

$$\beta_{1j} = \gamma_1 + \eta_{1j} \qquad \text{with } \eta_{1j} \sim N(0, \tau^2)$$

Which can be combined into the model,

$$Y_{ij} = \gamma_0 + \gamma_1 T_{ij} + \eta_{1j} T_{ij} + \eta_{0j} + \epsilon_{ij}$$

NOTICE

Take a look at this again: $$Y_{ij} = \gamma_0 + \gamma_1 T_{ij} + \eta_{1j} T_{ij} + \eta_{0j} + \epsilon_{ij}$$

Notice that now the treatment is showing up twice:

- γ_1 is the ATE
- $\eta_{1j} = \delta_j - \delta$ has to do with variation

The total variation is thus,

$$Var(Y_{ij}) = \sigma_1^2 + \sigma_2^2 + \tau^2$$

EFFECT SIZE OPTIONS

The Multisite Design brings with it different options for effect sizes. The differences between these have to do with the standard deviation that these are scaled against. Two options:

1. **Within-study SD** (akin to that in a simple RCT): $\delta_w = \frac{\gamma_1}{\sigma_1}$

2. **Total SD** (akin to that in a cluster RCT): $\delta_t = \frac{\gamma_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$

Importantly, notice that neither of these use the total variance - this is because the total variance includes the variation in treatment-effects. We want the effect of the treatment to only occur in the numerator.

HYPOTHESIS TESTING

We might be interested in testing if the intervention causes any sort of change in outcomes. To do so, we'd like to test the NH that $H_0 : \delta_w = 0$.

We use the test statistic: $t = \dfrac{\hat{\delta}_w}{SE(\hat{\delta}_w)}$

If indeed there is no effect (H0 true), this t-test follows a t-distribution with $df = N - 2m = 2m(n-1)$.

If the null hypothesis is false and the true treatment effect is $\delta_w = \delta_a$ then the t-test follows a non-central t-distribution with non-centrality parameter $\lambda = \delta_a / SE(\hat{\delta}_w)$.

STANDARD ERROR

For these analyses, we need the standard error. Let $\omega^2 = \tau^2/\sigma_1^2$ be the standardized treatment effect variation.

- **In balanced designs** ($n_j^t = n_j^c = n$):

$$SE(\hat{\delta}_w) = \sigma_1^{-1}\sqrt{\frac{n\tau^2 + 2\sigma_1^2}{mn}} = \sqrt{\frac{n\omega^2 + 2}{mn}}$$

- **In an unbalanced design:**

$$SE(\hat{\delta}_w) = \sigma_1^{-1}\left[\sum_{j=1}^{m}\frac{\tilde{n}_j}{\tilde{n}_j\tau^2 + \sigma_1^2}\right]^{-1/2} = \left[\sum_{j=1}^{m}\frac{\tilde{n}_j}{\tilde{n}_j\omega^2 + 1}\right]^{-1/2}$$

where $\tilde{n}_j = \dfrac{n_j^t n_j^c}{n_j^t + n_j^c}$.

POWER AND MDES

For a given effect size δ_a we can calculate $Power(\delta_a) = 1 - F(t_{\alpha/2} \,|\, df, \lambda) + F(-t_{\alpha/2} \,|\, df, \lambda)$.

Alternatively, for a given Type I and II error (and thus Power), we can calculate the MDES:

$$\delta_M \approx M_{df} \sqrt{\frac{n\omega^2 + 2}{mn}}$$

Where again, if $df > 16$ then $M_{df} < 2.9$.

TAKING THIS APART

Examining this more carefully we have: $\delta_M \approx M_{df}\sqrt{\frac{n\omega^2+2}{mn}} = M_{df}\sqrt{\frac{\omega^2}{m}+\frac{2}{mn}}.$

This gives us two insights:

- Again, increasing the number of sites (m) matters more than the number of individuals within sites (n).
- When there is a lot of variation in treatment effects (across sites), the MDES is larger. This means that it is harder to detect a non-zero average effect when there is a lot of variation in effects.

MDES

	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

INCREASING SENSITIVITY APPROACHES

COVARIATES MAY HELP?

We might think we can improve the sensitivity by including covariates.

But the only variance parameter that can be reduced here is the variation in treatment effects, ω^2.

This means two things:

- We would need to include covariates that explain variation in treatment effects across sites (not simply that explain variation in outcomes). We know far less about this.
- We would need to include treatment x covariate interactions in the model to do this.

WITH COVARIATES

Let X_{ij}^c be a centered unit level covariate and W_j be a site-level covariate. Then our model is:

$$Y_{ij} = \beta_{0j}^A + \beta_{1j}^A T_{ij} + \beta_{2j}^A X_{ij}^c + \epsilon_{ij}^A \qquad \text{with } \epsilon_{ij}^A \sim N(0, \sigma_{1A}^2)$$

$$\beta_{0j}^A = \gamma_0^A + \gamma_2^A W_j + \eta_{0j}^A \qquad \text{with } \eta_{0j}^A \sim N(0, \sigma_{A2}^2)$$

$$\beta_{1j}^A = \gamma_1^A + \gamma_3^A W_j + \eta_{1j}^A \qquad \text{with } \eta_{1j}^A \sim N(0, \tau_A^2)$$

$$\beta_{2j}^A = \gamma_4^A$$

Then our combined model can be written,

$$Y_{ij} = \gamma_0^A + \gamma_1^A T_{ij} + \gamma_2^A W_j + \gamma_3^A W_j T_{ij} + \gamma_4 X_{ij} + \eta_{1j}^A T_{ij} + \eta_{0j}^A + \epsilon_{ij}^A$$

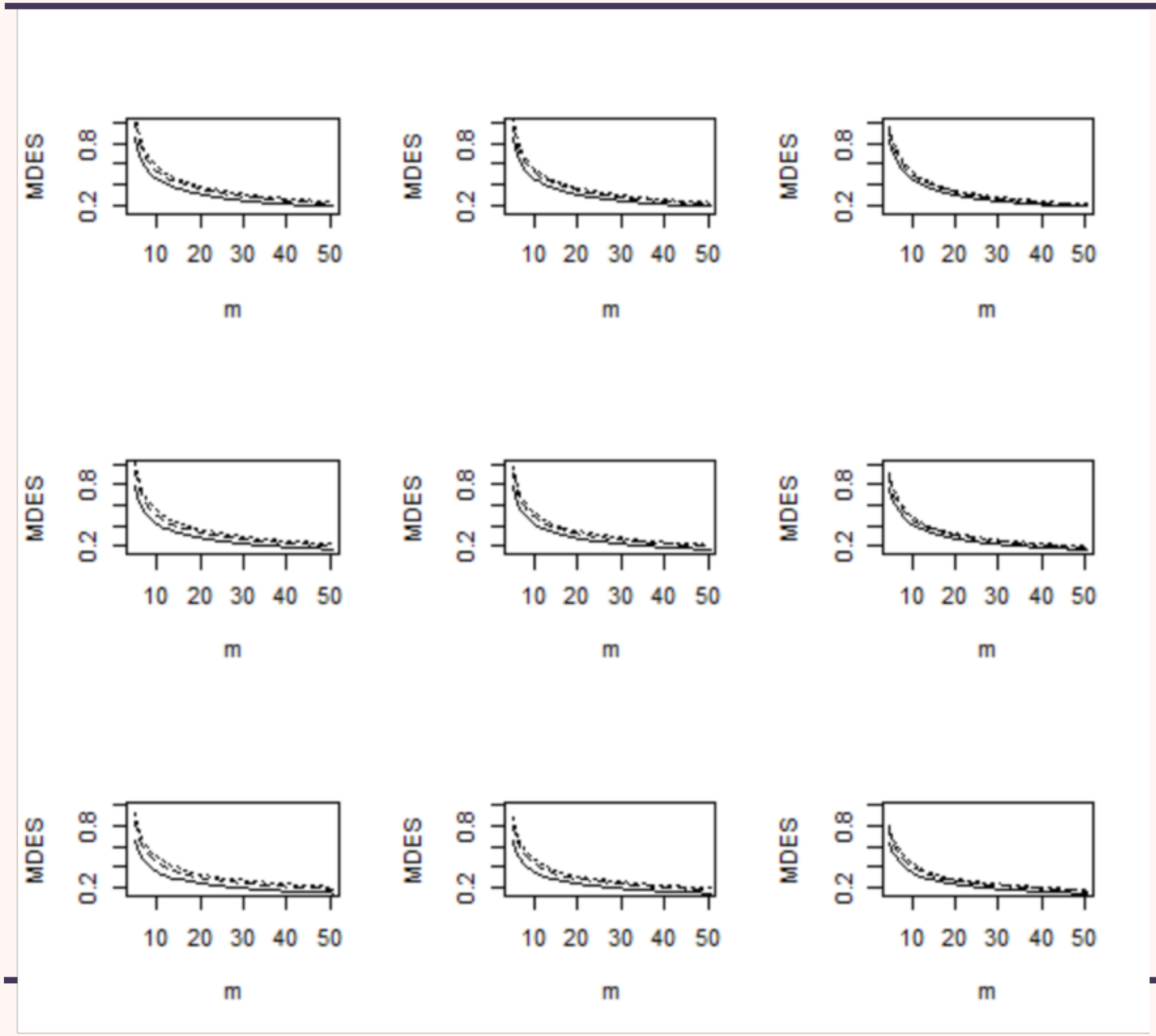
SE WITH COVARIATES

- The inclusion of covariates affects power, the MDES, and sensitivity through the standard error.
- In a balanced design (the simplest form), we have:

$$SE(\hat{\delta}_w) = \sigma_1^{-1}\sqrt{\frac{n\tau_A^2 + 2\sigma_{A1}^2}{mn}} = \sqrt{\frac{n\bar{Q}^2\omega^2 + 2\bar{R}_1^{\,2}}{mn}}$$

- Here $\bar{R}_1^{\,2} = 1 - R_1^2$ is the proportion in **outcomes** *not* explained by the covariates.
- Here $\bar{Q}^2 = 1 - Q^2$ is the proportion in **treatment effect variation *not*** explained by the covariates.

MDES

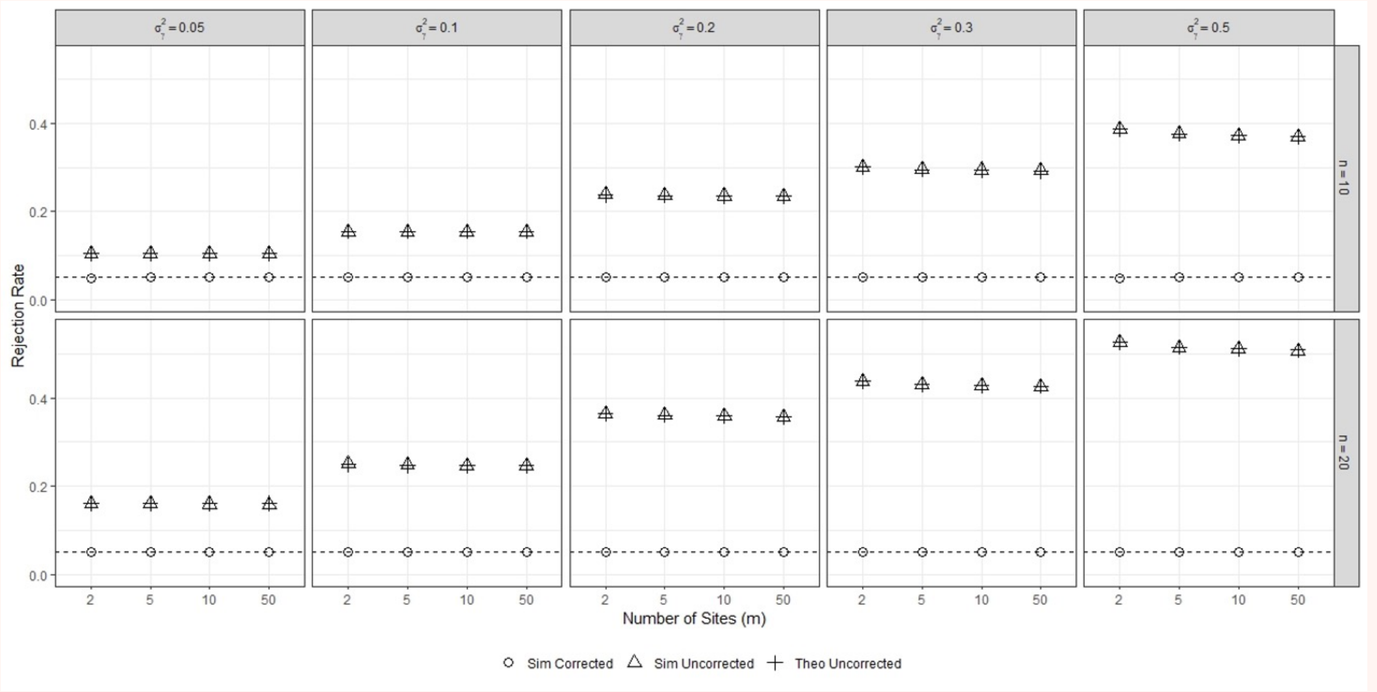


WHAT ABOUT FIXED EFFECTS?

A common question with the MSRT: Do we need to use a multi-level model or can we just use fixed effects (FE) for sites?

The short answer to this is: No.

Why? Because the treatment effect variance ends up in the residuals. The Type I error stated is thus wrong.



Actual Significance Levels of Nominal 0.05 Level Tests

From Chan and Hedges (in press)

DESIGN PARAMETERS

The Multisite Design has the following design parameters:

- Number of sites: m
- Number of units within sites: n_t, n_c
- Variation in treatment effects across sites: ω^2
- Variation explained in unit outcomes by covariates: R_1^2
- Variation explained in across site treatment effects by covariates: Q^2

A CAVEAT

Different software for power and MDES use different definitions of ω^2. For example, some use:

- By within-site variance: $\omega_w^2 = \frac{\tau^2}{\sigma_1^2}$
- By between-site variance: $\omega_b^2 = \frac{\tau^2}{\sigma_2^2}$

Clearly the scale of these differ. Be careful to read documentation to understand which definition to use when selecting credible values.

DESIGN PARAMETERS

WHAT YOU CAN CONTROL

- There are a lot of design parameters to consider. Some can be ‘chosen’ by you (e.g., n, m) whereas others are not as in your control.
- For a specified population of interest:
  - The ICC is not in your control
  - The degree of variation in treatment effects is not in your control
  - The ATE that is ‘true’ is not in your control
- **It can be tempting to change the population in order to ‘improve’ the design (e.g., reduce the ICC). But this means changing the purpose of the study!**

EFFECT SIZE

GENERAL CONSIDERATIONS

You don’t want to be too optimistic or pessimistic when considering the effect size that you will power your study to detect.

- **Optimistic:** This intervention is fabulous! We should focus on an effect size of 0.75 because I just know it will work.
- **Pessimistic:** This intervention should work but I’m just not sure about anything, so I would like to power it for an effect size of 0.02 just in case.

THE JUST RIGHT APPROACH

Thus you might consider this framing instead:

“What is an ES that would would consider ‘meaningful’ - that if the effect was smaller than this, you’d think it didn’t really ‘matter’, given the cost, the type of intervention, and so on?”

RESOURCE

- https://steppcenter.northwestern.edu/education-training/statistical-power-resources.html

2. Resources for Estimating Effect Sizes in Education

- 2.1 Magnitude of Effect Sizes
- 2.1.1 Math and/or Reading/ELA Outcomes
- 2.1.2 Science Outcomes
- 2.1.3 Social-Emotional/Cognitive/Behavioral Outcomes
- 2.2 Benchmarks for Effect Sizes

1. RELATIVE TO TYPICAL LEARNING

Annual achievement gain: Mean effect sizes across seven nationally normed tests

Grade Transition	Reading	Math	Science	Social Studies
Grade K - 1	1.52	1.14	--	--
Grade 1 - 2	0.97	1.03	0.58	0.63
Grade 2 - 3	0.60	0.89	0.48	0.51
Grade 3 - 4	0.36	0.52	0.37	0.33
Grade 4 - 5	0.40	0.56	0.40	0.35
Grade 5 - 6	0.32	0.41	0.27	0.32
Grade 6 - 7	0.23	0.30	0.28	0.27
Grade 7 - 8	0.26	0.32	0.26	0.25
Grade 8 - 9	0.24	0.22	0.22	0.18
Grade 9 - 10	0.19	0.25	0.19	0.19
Grade 10 - 11	0.19	0.14	0.15	0.15
Grade 11 - 12	0.06	0.01	0.04	0.04

NOTES: Adapted from Lipsey, et al., (2012). Spring-to-spring differences are shown. The means shown are the simple (unweighted) means of the effect sizes from all or a subset of seven tests: CAT5, SAT9, Terra Nova-CTBS, Gates-MacGinitie, MAT8, Terra Nova-CAT, and SAT10.

2. ACCOUNT FOR OUTCOME TYPES

How Effect Size Magnitude Relates to Outcomes: Kraft (2020)

Ask	Interpret	Large effect sizes when?
Is the outcome the result of short-term decisions and effort or a cumulative set of decisions and sustained effort over time?	Expect outcomes affected by short-term decisions and effort to be larger than outcomes that are the result of cumulative decisions and sustained effort over time.	Outcomes affected by short-term decisions.
How closely aligned is the intervention with the outcome?	Expect outcomes more closely aligned with the intervention to have larger effect sizes.	Outcomes closely aligned with the intervention.
How long after the intervention was the outcome assessed?	Expect outcomes measured immediately after the intervention to have larger effect sizes than outcomes measured later.	Outcomes measured immediately after the intervention.
How reliably is the outcome measured?	Expect measures with lower reliability to have smaller effect sizes than comparable measures with higher reliability.	Outcomes measured with higher reliability.

Achievement effect sizes from randomized studies broken out by type of test and grade level

Type of Test	Grade Level	N of Effect Sizes	Mean	Standard Deviation
Specialized Topic or Test, Researcher Developed	Elementary	230	0.40	0.55
	Middle	27	0.43	.048
	High	43	0.34	.038
	Total	300	0.39	0.53
Standardized Test, Narrow Scope	Elementary	374	0.25	0.42
	Middle	30	0.32	0.26
	High	22	0.03	0.07
	Total	426	0.24	0.40
Standardized Test, Broad Scope	Elementary	89	0.08	.027
	Middle	13	0.15	0.33
	High	1	--	--
	Total	103	0.08	0.28
Total	Elementary	693	0.28	0.46
	Middle	70	0.33	0.38
	High	66	0.23	0.34
	Total	829	0.28	0.45

Note: This table is reproduced from Lipsey, et al. (2012)

3. CONSIDER RESEARCH DESIGN

Relating Effect Sizes to Subjective Decisions About Research Design: Kraft (2020)

Ask	**Interpret**	**Large effect sizes when?**
Are study participants a broad sample or a subgroup most likely to benefit from the intervention?	Expect studies with more targeted samples to have larger effect sizes than studies with more diverse and representative samples.	Studies with more targeted samples (most likely to benefit).
What sample produced the standard deviation used to estimate effect sizes?	Expect effect sizes that are standardized using more homogeneous and less representative samples to have larger effect sizes.	Studies with more homogenous samples.
How similar or different was the experience of the treatment group compared to the control or comparison group?	Expect studies to have smaller effect sizes when control groups have access to resources or services similar to the treatment group.	Studies with interventions very different than the comparison, using resources not easily available.

4. RELATIVE TO OTHER INTERVENTIONS

Table 10. Achievement effect sizes from randomized studies broken out by type of intervention and target recipients

	N of Effect Sizes	Median	Mean	Standard Deviation
Type of Intervention				
Instructional format	52	.13	.21	.36
Teaching technique	117	.27	.35	.47
Instructional component or skill training	401	.27	.36	.50
Curriculum or broad instructional program	227	.08	.13	.32
Whole school program	32	.17	.11	.31
Total	829	.18	.28	.45
Target Recipients				
Individual students	252	.29	.40	.53
Small group	322	.22	.26	.40
Classroom	176	.08	.18	.41
Whole school	35	.14	.10	.30
Mixed	44	.24	.30	.33
Total	829	.18	.28	.45

NOTE: Standardized mean difference effect sizes from 181 samples. No weighting was used in the calculation of the summary statistics and no adjustments were made for multiple effect sizes from the same sample.

Achievement effect sizes by intervention type and target recipients (from Lipsey et al)

Table 1
Empirical Distributions of Effect Sizes From Randomized Control Trials of Education Interventions With Standardized Achievement Outcomes

		Subject		Sample Size					Scope of Test		DoE
	Overall	Math	Reading	≤100	101–250	251–500	501–2,000	>2,000	Broad	Narrow	Studies
Mean	0.16	0.11	0.17	0.30	0.16	0.16	0.10	0.05	0.14	0.25	0.03
Standard deviation	0.28	0.22	0.29	0.41	0.29	0.22	0.15	0.11	0.24	0.44	0.16
Mean (weighted)	0.04	0.03	0.05	0.29	0.15	0.16	0.10	0.02	0.04	0.08	0.02
P1	−0.38	−0.34	−0.38	−0.56	−0.42	−0.29	−0.23	−0.22	−0.38	−0.78	−0.38
P10	−0.08	−0.08	−0.08	−0.10	−0.14	−0.07	−0.05	−0.06	−0.08	−0.12	−0.14
P20	−0.01	−0.03	−0.01	0.02	−0.04	0.00	−0.01	−0.03	−0.03	0.00	−0.07
P30	0.02	0.01	0.03	0.10	0.02	0.06	0.03	0.00	0.02	0.05	−0.04
P40	0.06	0.04	0.08	0.16	0.07	0.10	0.06	0.01	0.06	0.11	−0.01
P50	0.10	0.07	0.12	0.24	0.12	0.15	0.09	0.03	0.10	0.17	0.03
P60	0.15	0.11	0.17	0.32	0.17	0.18	0.12	0.05	0.14	0.22	0.05
P70	0.21	0.16	0.23	0.43	0.25	0.22	0.15	0.08	0.20	0.34	0.09
P80	0.30	0.22	0.33	0.55	0.35	0.29	0.19	0.11	0.29	0.47	0.14
P90	0.47	0.37	0.50	0.77	0.49	0.40	0.27	0.17	0.43	0.70	0.23
P99	1.08	0.91	1.14	1.58	0.93	0.91	0.61	0.48	0.93	2.12	0.50
k (number of effect sizes)	1,942	588	1,260	408	452	328	395	327	1,352	243	139
n (number of studies)	747	314	495	202	169	173	181	124	527	91	49

Note. A majority of the standardized achievement outcomes (95%) are based on math and English language art test scores, with the remaining based on science, social studies, or general achievement. Weights are based on sample size for weighted mean estimates. For details about data sources, see Appendix A, available on the journal website. DoE = U.S. Department of Education.

Distribution of Effect Sizes from Kraft (2020)

4. CAVEAT

- Notice that these are distributions based on a lot of studies.
- Why not just use the results from a prior study of this exact intervention?
  - Because the effect size from the prior study is an estimate, not the real effect.
  - e.g., if you estimated the ES to be 0.20 in an underpowered pilot study, the true effect could be much larger or smaller than this! You might have gotten lucky!

SOURCES TO CONSIDER

Konstantopoulos, S., & Hedges, L. V. (2008). How large of an effect can we expect from school reforms? *Teachers College Record, 110(8)*, 1613-1640

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

OTHER PARAMETERS

EFFECT SIZE VARIATION

3. Resources for Estimating Effect Size Variability in Education

- 3.1 Math and/or Reading/ELA Outcomes
- 3.2 Other Outcomes (post-secondary outcomes and labor/workforce outcomes)

- https://steppcenter.northwestern.edu/education-training/statistical-power-resources.html

ICC AND R^2

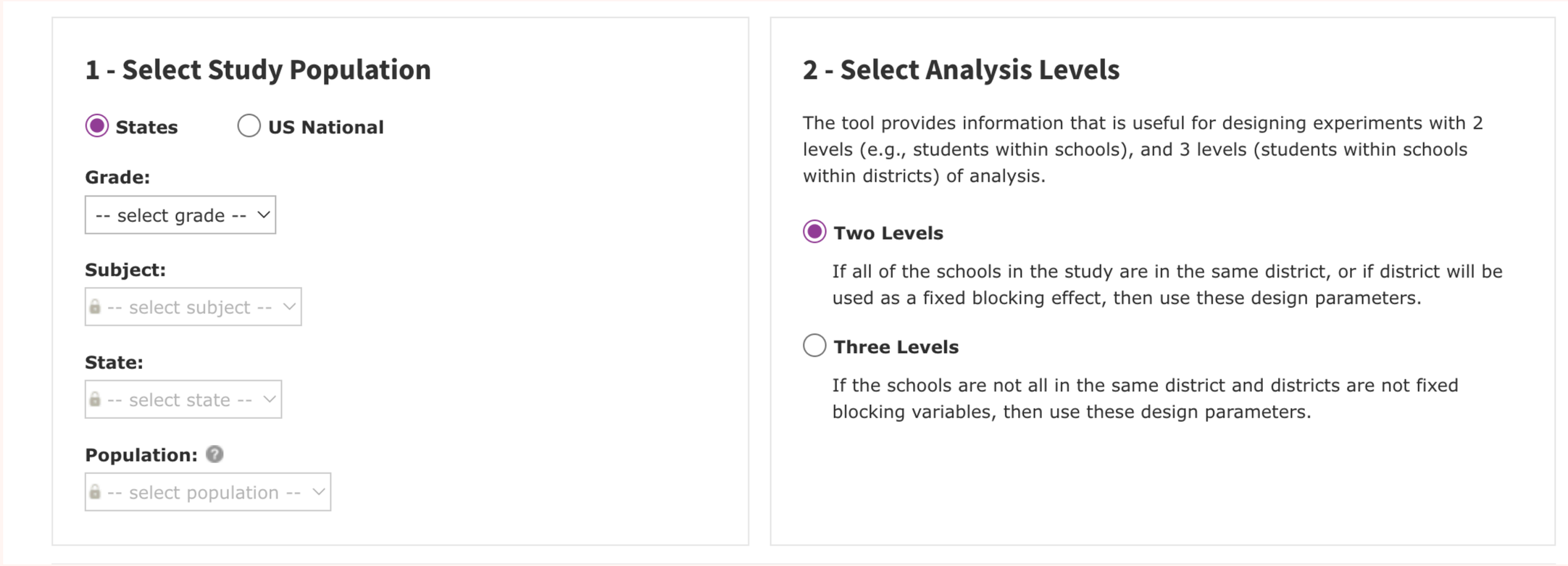
1. Resources for Estimating ICCs and/or R2s in Education

- 1.1 Math and/or Reading/ELA Outcomes
- 1.2 Science Outcomes
- 1.3 Social-Emotional/Cognitive/Behavioral Outcomes
- 1.4 Other Outcomes (nutritional outcomes)
- 1.5 Teacher Outcomes

- https://steppcenter.northwestern.edu/education-training/statistical-power-resources.html

VARIANCE ALMANAC

http://stateva.ci.northwestern.edu/



A PREVIEW (NATIONAL READING)

	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
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

A PREVIEW (NATIONAL MATH)

	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

	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

PRETEST: THE MVP

- **Covariate that matters the most: School level pre-test**
  - This matters even more at higher grades
  - This is less useful if there is too much time between pre- and post-test
  - Adding a second pre-test doesn’t add much
  - Subject specific pre-tests are best