UCLA School of Education & Information Studies
Department of Education

EDUC 231D
Advanced Quantitative Methods: Multilevel Analysis
Winter 2025

Multisite and Cluster Randomized Design

Lecture 9 Presentation Slides

February 4 & 6, 2025

Today’s Topics

- Overview of randomized designs
- Multisite randomized design
- Variation in treatment effects across sites
- Cluster randomized design

Overview of randomized designs

Randomized designs

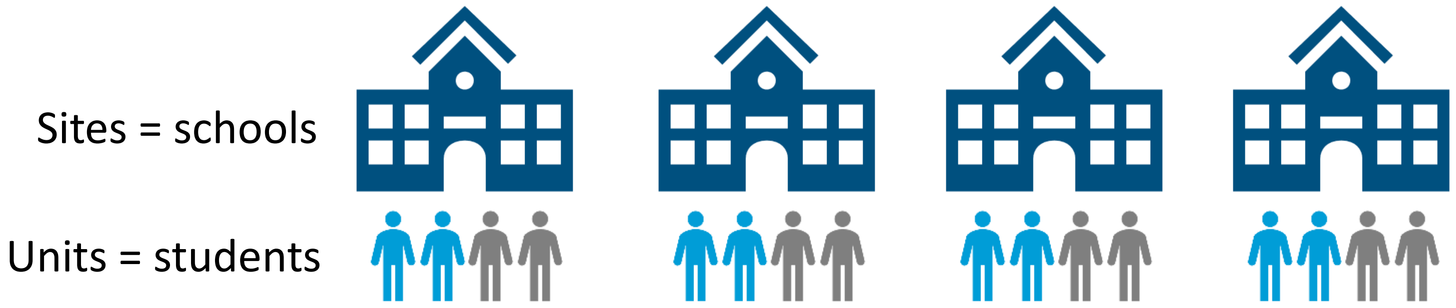
- Randomized designs, sometimes called experimental designs, are considered the “gold standard” for estimating the causal effect of a treatment/intervention
- Units are randomly assigned to different groups
  - Random assignment should result in groups (e.g., treatment and control groups) that are equivalent, on average, in terms of preexisting or pretreatment characteristics
  - Strong internal validity

Randomized designs

- In traditional experiments, the units of analysis are randomly assigned to treatment conditions (individual random assignment design)
- Studies in multilevel settings introduce additional considerations for the random assignment design and analysis
  - Units could be randomly assigned to treatments within sites (e.g., students in schools)
  - Existing groups/clusters could be randomly assigned to treatments so that all units within those groups are assigned to the same treatment (e.g., schools are assigned treatments)

Multisite individual random assignment design

- Treatment assigned to units (e.g., students)
- Units are nested within sites (e.g., schools)

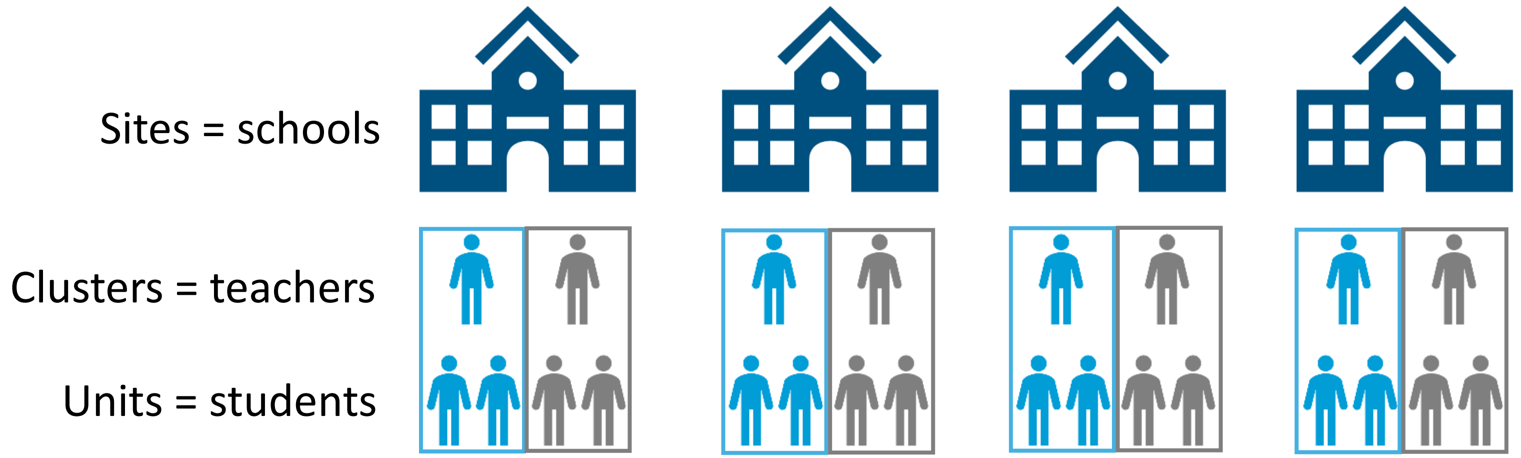


Cluster random assignment design

- Treatment assigned at the group level (e.g., schools)
- Units of analysis (e.g., students) are nested within "clusters"

Multisite cluster random assignment design

- Treatment assigned at the group level (e.g., teachers)
- Units of analysis (e.g., students) are nested within “clusters”
- Clusters are nested within “sites” (e.g., schools)



Multisite Randomized Designs

Multilevel model to analyze a multisite randomized design

- What is the average treatment effect across sites?
- Does the average treatment effect differ across sites?

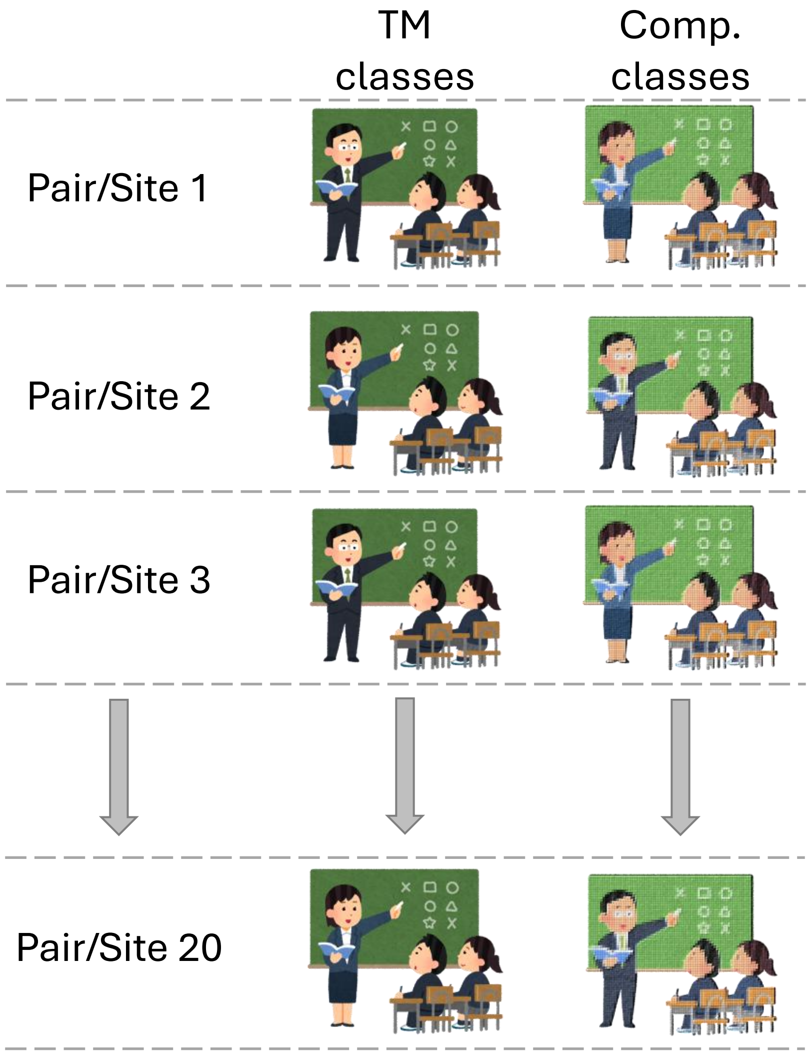
$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(Trt_{ij} - \overline{Trt}_{.j}\right) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \qquad u_{1j} \sim N(0, \tau_{11})$$

Multisite randomized design: TM example

- Transition Mathematics (TM) prealgebra curriculum
- Evaluation took place in 20 classroom pairs from schools across the U.S.
- Within each pair, one class used TM and the other class used the existing curriculum
- Random assignment was used in 10 pairs but not the other 10 pairs



TM example: What's the average effect of TM on geometry readiness?

- What if we ignored the nested structure of the data?

$$Y_i = \beta_0 + \beta_1(Trt_i) + r_i$$

(Note: outcome is score on a geometry readiness test where student scores range from 1 to 19 and the standard deviation is 4.22.)

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	8.703	0.251	34.678	0.0000	***
trtmnt	1.389	0.356	3.902	0.0001	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 4.169 on 547 degrees of freedom

Multiple R-squared: 0.02708, Adjusted R-squared: 0.0253

F-statistic: 15.23 on 547 and 1 DF, p-value: 0.0001

TM example: What's the average effect of TM on geometry readiness?

- What if we estimate a separate OLS regression for every site?

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Trt_i - \overline{Trt}_{.j}) + r_{ij}$$

- Mean of $\hat{\beta}_{1j} = 1.44$
- Variance of $\hat{\beta}_{1j} = 3.65$

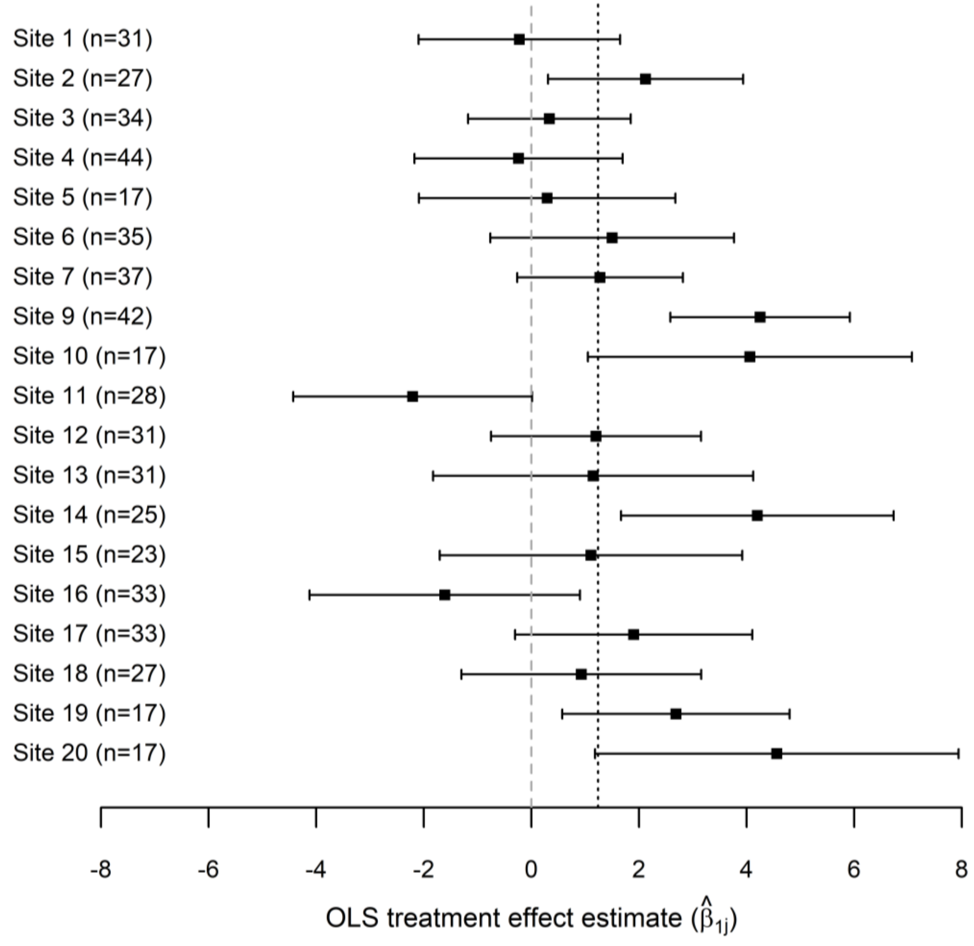
Site (j)	Size (n_j)	Site Mean ($\hat{\beta}_{0j}$)	SE($\hat{\beta}_{0j}$)	TM Effect ($\hat{\beta}_{1j}$)	SE($\hat{\beta}_{1j}$)	95% CI($\hat{\beta}_{1j}$)
1	31	13.52	0.48	-0.23	0.96	(-2.10 , 1.65)
2	27	6.59	0.46	2.12	0.93	(0.31 , 3.93)
3	34	5.15	0.38	0.33	0.77	(-1.18 , 1.84)
4	44	7.86	0.49	-0.24	0.99	(-2.18 , 1.70)
5	17	8.47	0.61	0.29	1.22	(-2.09 , 2.67)
6	35	11.54	0.57	1.50	1.15	(-0.76 , 3.76)
7	37	13.97	0.39	1.27	0.79	(-0.27 , 2.81)
9	42	6.98	0.43	4.25	0.85	(2.58 , 5.92)
10	17	8.71	0.73	4.06	1.54	(1.05 , 7.07)
11	28	6.68	0.56	-2.21	1.13	(-4.43 , 0.01)
12	31	14.42	0.50	1.20	1.00	(-0.75 , 3.15)
13	31	10.87	0.76	1.15	1.52	(-1.83 , 4.12)
14	25	10.12	0.58	4.20	1.29	(1.66 , 6.73)
15	23	10.83	0.71	1.11	1.43	(-1.70 , 3.92)
16	33	11.45	0.64	-1.61	1.28	(-4.12 , 0.90)
17	33	8.70	0.56	1.90	1.13	(-0.31 , 4.11)
18	27	6.81	0.56	0.93	1.14	(-1.30 , 3.15)
19	17	5.29	0.53	2.69	1.08	(0.57 , 4.80)
20	17	7.12	0.82	4.56	1.72	(1.18 , 7.94)

TM example: What's the average effect of TM on geometry readiness?

- What if we estimate a separate OLS regression for every site?

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Trt_i - \overline{Trt}_{.j}) + r_{ij}$$

- Mean of $\hat{\beta}_{1j} = 1.44$
- Variance of $\hat{\beta}_{1j} = 3.65$



TM example: What's the average effect of TM on geometry readiness?

- What if we use a multilevel model?

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(Trt_{ij} - \overline{Trt}_{.j}\right) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \qquad u_{1j} \sim N(0, \tau_{11})$$

Small group discussion, part 1

- In groups of 3-4, take 10 minutes to define the following parameters for the multilevel model:

- β_{0j} :
- γ_{00} :
- u_{0j} :
- τ_{00} :

- β_{1j} :
- γ_{10} :
- u_{1j} :
- τ_{11} :

Small group discussion, part 1

- In groups of 3-4, take 10 minutes to define the following parameters for the multilevel model:

- β_{0j} : mean geometry readiness score for site *j*
- γ_{00} : grand-mean geometry readiness score
- u_{0j} : deviation of site *j*'s geometry readiness mean score from the grand-mean
- τ_{00} : between-site variance in site mean geometry readiness scores

- β_{1j} : mean difference in geometry readiness scores between treatment and control students in site *j*; or average treatment effect in site *j*
- γ_{10} : grand-mean average treatment effect
- u_{1j} : deviation of the average treatment effect in site *j* from the grand-mean average treatment effect
- τ_{11} : between-site variance in site average treatment effects

TM example: What's the average effect of TM on geometry readiness?

- What if we use a multilevel model?

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(Trt_{ij} - \overline{Trt}_{.j}\right) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \qquad u_{1j} \sim N(0, \tau_{11})$$

```
m1 <- lmer(gebtot ~ trt.gpc + (1 + trt.gpc | site),
           data = tm2
summary(m1)
```

TM example: What’s the average effect of TM on geometry readiness?

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: gebtot ~ trt.gpc + (1 + trt.gpc | site)
   Data: tm2

REML criterion at convergence: 2846.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1314 -0.6874  0.0252  0.6597  3.0634

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 site     (Intercept) 7.941    2.818
          trt.gpc     2.115    1.454    -0.21
 Residual             9.100    3.017
Number of obs: 549, groups:  site, 19

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)   9.2241     0.6603 18.1048  13.970 3.87e-11 ***
trt.gpc       1.3440     0.4282 17.3130   3.139  0.00588 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

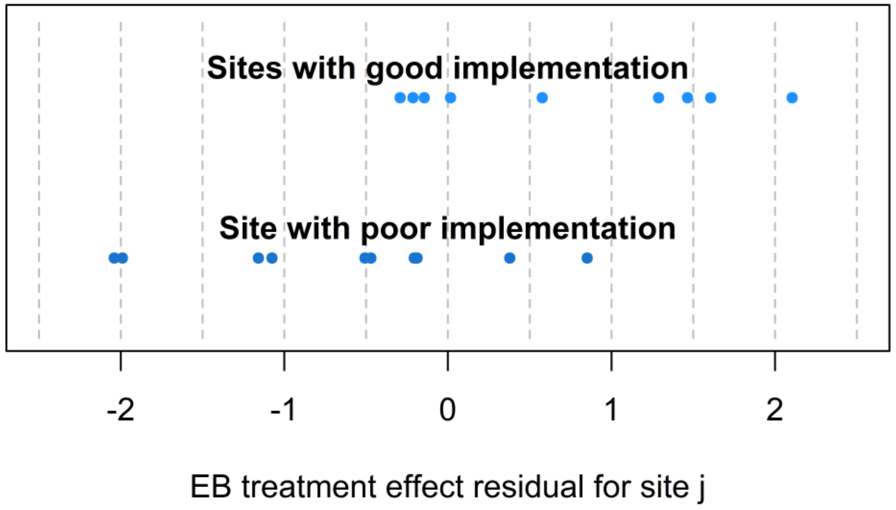
Small group discussion, part 2

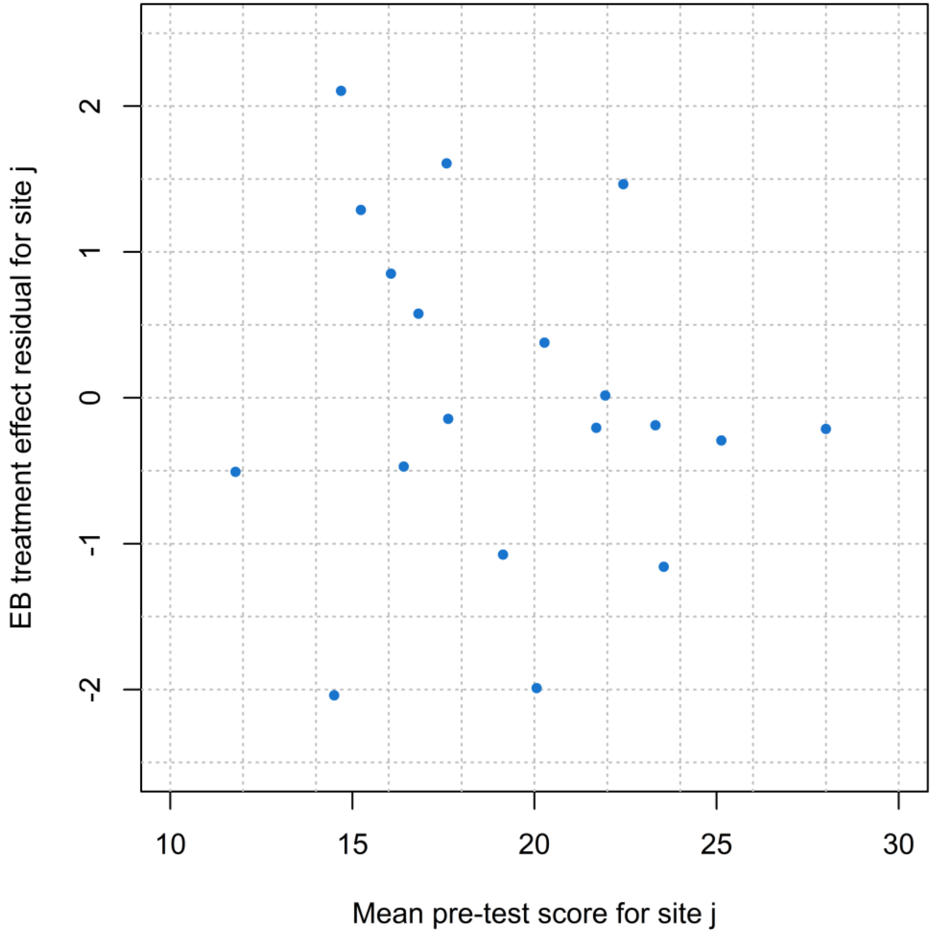
- In groups of 3-4, take 10 minutes to answer the following questions based on the model output on the previous slide:
  - What's the estimated grand-mean score on the geometry readiness test?
  - What's the estimated grand-mean effect of TM?
  - How does the standard error for the TM effect estimate from the multilevel model compare to the standard error for the TM effect estimate from the OLS model that ignored the nested data structure? Why would the standard errors from the two models differ?
  - To what extent does the average treatment effect vary between sites?
  - What's the expected average treatment effect at a site where the effect is 1 standard deviation <u>below</u> the grand-mean effect?
  - What's the expected average treatment effect at a site where the effect is 1 standard deviation <u>above</u> the grand-mean effect?
  - Do you think TM tends to be more effective in sites with higher or lower average scores on the geometry readiness test? What from the model results helped you come to that conclusion?

Variation in Treatment Effects Across Sites

Exploration of between-site treatment effect variation

- Can get a better sense of treatment effect variation by looking at the site-level random effects (u^*_{1j})





Describing between-site treatment effect variation

- What site-level factors/characteristics are associated with the within-site treatment effect?
  - Cross-level interactions to address questions of moderation/mediation

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(Trt_{ij} - \overline{Trt}_{.j}\right) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}, \qquad u_{1j} \sim N(0, \tau_{11})$$

TM example: Does implementation matter?

- Can test whether the quality of TM implementation is related to the site-specific effect estimates
  - And let's control for site-mean pretest scores while we're at it

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(Trt_{ij} - \overline{Trt}_{.j}\right) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\left(\overline{Pretest}_{.j} - \overline{Pretest}_{..}\right) + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Imp_j + u_{1j}, \qquad u_{1j} \sim N(0, \tau_{11})$$

(Note: *imp* = 1 for sites with good implementation of TM and *imp* = 0 for sites with poor implementation of TM)

TM example: Does implementation matter?

- Can test whether the quality of TM implementation is related to the site-specific effect estimates

```
m2 <- lmer(gebtot ~ trt.gpc + sitemgm.gdm + imp:trt.gpc
          + (1 + trt.gpc | site), data = tm2
```

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 site     (Intercept) 1.2485   1.1174
          trt.gpc     0.8736   0.9346   -0.55
 Residual             9.0975   3.0162
Number of obs: 549, groups:  site, 19

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  9.21152    0.28883 16.55912  31.892 2.66e-16 ***
trt.gpc      0.22948    0.45553 16.74412   0.504  0.62099
sitemgm.gdm  0.59769    0.06648 15.80116   8.991 1.31e-07 ***
trt.gpc:imp  2.28970    0.65070 16.49168   3.519  0.00274 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cluster Randomized Designs

Multilevel model to analyze a cluster randomized design

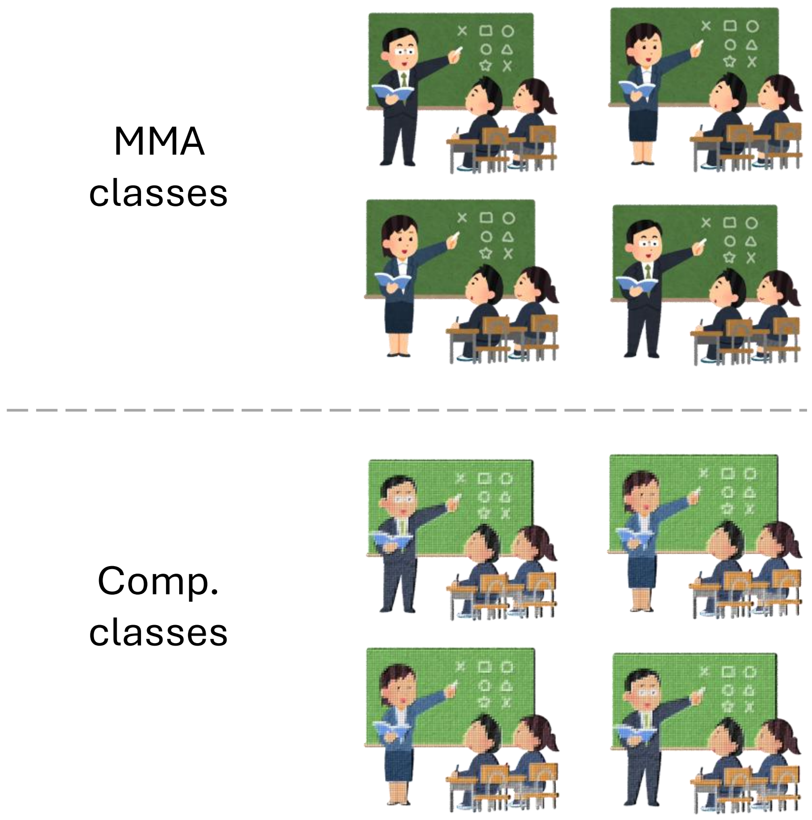
- What is the average treatment effect across sites?
- Does the average treatment effect differ across sites?

$$Y_{ij} = \beta_{0j} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\left(Trt_j - \overline{Trt}_{..}\right) + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00})$$

Multisite randomized design: MMA example

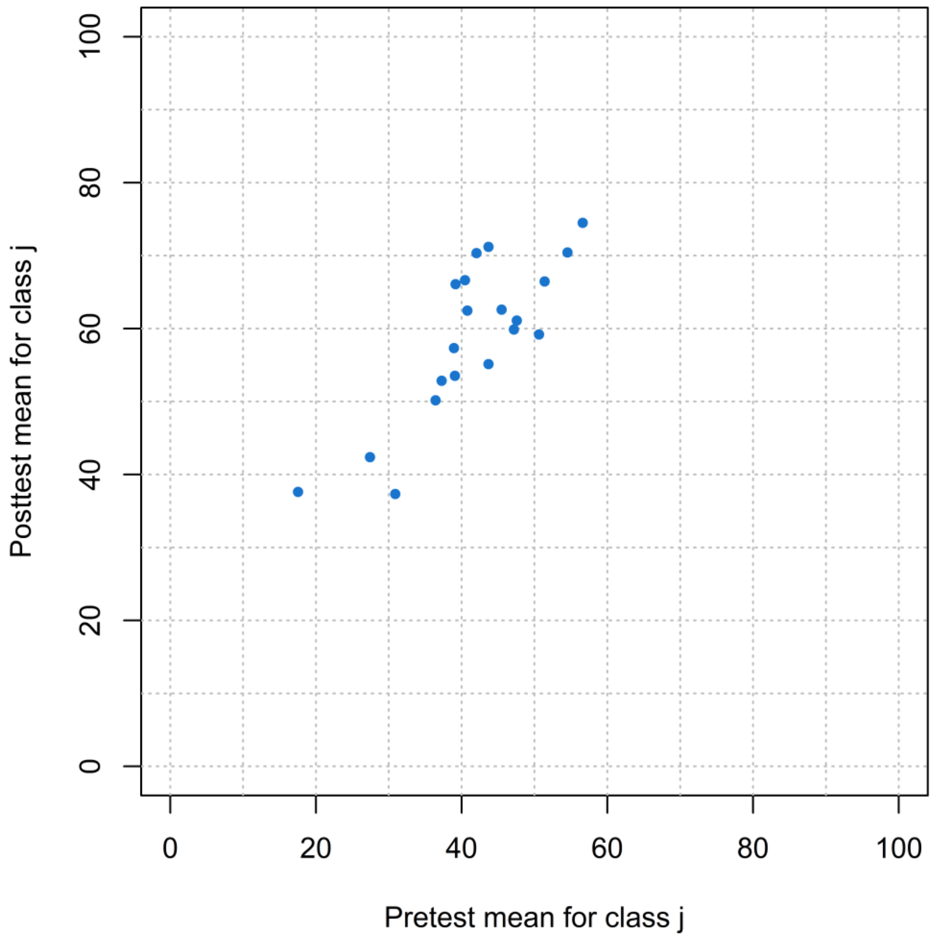
- My Math Academy (MMA) digital game-based learning supplement
- Evaluation took place in 20 kindergarten classrooms
- 10 classes randomly assigned to use MMA and the other classes used the existing curriculum



MMA example: What's the average effect of MMA on math performance?

- Percent correct on the pre- and post-tests

	Control Classes (N=195)	Treatment Classes (N=233)	Overall (N=428)
pretest			
Mean (SD)	40.7 (24.8)	43.4 (25.1)	42.2 (25.0)
Median [Min, Max]	39.0 [0, 100]	44.0 [0, 100]	44.0 [0, 100]
posttest			
Mean (SD)	55.2 (25.0)	63.2 (24.5)	59.5 (25.1)
Median [Min, Max]	56.0 [0, 100]	67.0 [0, 100]	61.0 [0, 100]



MMA example: What's the average effect of MMA on math performance?

- What if we ignored the nested structure of the data and analyzed the student-level data?

$$Y_i = \beta_0 + \beta_1(Trt_i) + \beta_1(Pretest_i - \overline{Pretest}_{..}) + r_i$$

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	56.327	1.108	50.845	0.0000	***
trt	5.902	1.502	3.928	0.0001	***
pretest.gdc	0.775	0.030	25.850	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 15.46 on 425 degrees of freedom

Multiple R-squared: 0.6212, Adjusted R-squared: 0.6194

F-statistic: 348.5 on 425 and 2 DF, p-value: 0.0000

MMA example: What's the average effect of MMA on math performance?

- What if we ignored the nested structure of the data and analyzed the class-level data?

$$\bar{Y}_{.j} = \beta_0 + \beta_1(Trt_j) + \beta_1(\overline{Pretest}_{.j} - \overline{Pretest}_{..}) + r_i$$

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	56.264	1.894	29.706	0.0000	***
trt	5.179	2.713	1.909	0.0733	.
cmeanpre.gdc	0.892	0.151	5.925	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 5.911 on 17 degrees of freedom

Multiple R-squared: 0.7309, Adjusted R-squared: 0.6992

F-statistic: 23.08 on 17 and 2 DF, p-value: 0.0000

MMA example: What's the average effect of MMA on math performance?

- What if we use a multilevel model?

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Pretest_i - \overline{Pretest}_{..}) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\left(Trt_j - \overline{Trt}_{..}\right) + \gamma_{02}\left(\overline{Pretest}_{.j} - \overline{Pretest}_{..}\right) + u_{0j},$$

$$u_{0j} \sim N(0, \tau_{00})$$

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

Small group discussion, part 1

- In groups of 3-4, take 10 minutes to define the following parameters for the multilevel model:

- β_{0j} :
- γ_{00} :
- γ_{01} :
- γ_{02} :
- γ_{10} :
- u_{0j}:
- τ_{00} :

Small group discussion, part 1

- In groups of 3-4, take 10 minutes to define the following parameters for the multilevel model:

- β_{0j} : expected posttest score in class *j* for a student with an average pretest score
- γ_{00} : grand-mean posttest score for a student with an average pretest score (or for the average class in the study)
- γ_{01} : mean difference in posttest scores between treatment and control classes, controlling for student pretest score and class-mean pretest score; or average treatment effect
- γ_{02} : relationship between class-mean pretest score and class-mean posttest score, controlling for student pretest score; or contextual effect of class-mean pretest score
- γ_{10} : grand-mean relationship between student pretest score and student posttest score, controlling for class-mean pretest score; or within-class relationship between posttest and pretest scores
- u_{0j}: deviation of class j's mean posttest score, after accounting for student pretest score, class-mean pretest score, and treatment condition
- τ_{00} : between-class variation in class-mean posttest score, after accounting for student pretest score, class-mean pretest score, and treatment condition

MMA example: What's the average effect of MMA on math performance?

```
m1 <- lmer(posttest ~ trt.gdc + pretest.gdc + cmeanpre.gdc
              + (1 | tchid), data = mmax)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 tchid    (Intercept)  23.33    4.83
 Residual             219.03   14.80
Number of obs: 428, groups:  tchid, 20

Fixed effects:
              Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   59.28752    1.29968  16.58325  45.617   <2e-16 ***
trt.gdc        5.32591    2.65452  16.28173   2.006   0.0617 .
pretest.gdc    0.76205    0.03039 406.47061  25.075   <2e-16 ***
cmeanpre.gdc   0.12929    0.15496  19.71908   0.834   0.4141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Small group discussion, part 2

- In groups of 3-4, take 10 minutes to answer the following questions based on the model output on the previous slide:
  - What's the estimated grand-mean score on the post-test?
  - What's the estimated average effect of MMA?
  - How does the standard error for the MMA effect estimate from the multilevel model compare to the standard error for the MMA effect estimate from the student-level and class-level OLS models that ignored the nested data structure? Why would the standard errors from the models differ?
  - What can you say about how the average treatment effect varies between classes?