Basics of Experimental Design

Spyros Konstantopoulos
spyros@msu.edu

Michigan State University

Prepared for the IES Summer Research Training Institute 2022

Experimental Design

- "Experimental Design" encompasses:
  1. Strategies for organizing data collection
  2. Involves knowledge of data generating processes
  3. Data analysis procedures *matched* to those data collection strategies
- The researcher is interested in determining the effect of some treatment (e.g., school intervention) on some units-subjects outcome (e.g., student achievement)
- Typically, two groups are created: one treatment and one control group
- Typically, the designs are balanced (i.e., equal sample sizes in both groups)
- The effect is the change in the outcome of interest (e.g., achievement) by some intervention/treatment
- This change in the outcome is designed to have a beneficial effect (e.g., increase achievement)

Experimental Design: Analysis

- Analysis of Variance (ANOVA) is a traditional analysis procedure applied to experimental designs, especially for Randomized Control Trials (RCTs)
- Other appropriate analytic procedures include:
  - Regression models
  - Multilevel or hierarchical models
  - Statistical models applied to aggregates (classroom or school means)
- All these procedures estimate the mean difference in an outcome between treatment and control groups
- Analytic procedures should *match* research hypotheses, design, and a priori power analyses

Why Do We Need Experimental Design?

- Aim to identify treatment effects in the presence of ***variability*** (differences) of units and/or responses
- Variability exists because:
  - Units (students, teachers, & schools) are not identical
  - Units respond in different ways to treatments
- We need experimental design to control this variability (i.e., equate treatment and control groups on average at the beginning of the experiment) and then identify treatment effects on outcomes of interest
- It is the best way to identify what causes a change in an outcome of interest (when threats to the internal validity of the experiment are minimized)

History

- The idea of controlling variability by creating similar – equivalent groups through design has a long history

- In 1753 Sir James Lind's published the *treatise of the scurvy* describing his study where 12 scurvy patients (sailors who spent much time in the sea) were assigned to six similar groups that received different treatments (proposed remedies)

- One of the treatments involved consumption of oranges and lemons. People in that group showed dramatic improvement compared to the other groups

History

- In the late 1890's, Fibiger examined the effectiveness of diphtheria antitoxin in treating diphtheria patients and assigned patients to a treatment (received antitoxin) or a control group (standard treatment) according to the day they were admitted (i.e., every other day patients were assigned to different groups)

- In the 1930s, Amberson et al. (1931) used random assignment via a coin-toss to create equivalent groups to examine the effects of sanocrysin on pulmonary tuberculosis

History

- The first modern randomized clinical trial in medicine is considered to be the trial of streptomycin for treating tuberculosis

- It was conducted by the British Medical Research Council in 1946 and reported in 1948

- Patients were randomly assigned to a group that took streptomycin and a group that did not

History

- Another renowned RCT was the polio vaccine field trial conducted in the U.S. in 1954

- Children ages 6-9 were assigned to a treatment group that received the polio vaccine or a control group that received a placebo

History

- Studies in crop variation I – VI (1921 – 1929)
- In 1919 a statistician named Fisher was hired at Rothamsted agricultural station
- They had a lot of observational data on crop yields and hoped a statistician could analyze it to find effects of various treatments

History

- In a series of studies, within 8 years, Fisher invented the basic principles of experimental design and analysis of variance and covariance

- He also invented control of variation by random assignment

History

- In the field of education two eminent books introduced Fisher's methodological foundations of experimental design and analysis
- In 1940 Lindquist published his book about *Statistical Methods in Educational Research* that discussed random allocation of units and principles of experimental design and analyses
- In the 1960s, Campbell and Stanley (1966) outlined methodologies for designing experiments and quasi-experiments as well as analyzing appropriately data from experiments

History

- In the field of education, a noteworthy large-scale RCT was conducted in the mid-1980s in the state of Tennessee, known as the Tennessee class size experiment or Project STAR (Student Teacher Achievement Ratio)
- A four-year experiment that followed a cohort of kindergarten students in 79 schools through third grade. In the first year of the study, within each school, kindergarten students and teachers were randomly assigned to either a small class, a regular size class, or a regular size class with a full-time teacher assistant

History

- Since 2002 mainly due to the IES funding streams and the emphasis IES has placed on rigorous research designs there has been an abundance of RCTs

- IES has funded nearly 350 RCTs since its inception

Principles of Experimental Design

- Objective: Control variability and identify systematic effects of treatments on outcomes
  - ⇨ Create sample groups that are on average equivalent at the beginning of the experiment
    - Measures of traits are similar across groups
    - Groups would have the same response if given the same treatment.

- Methods to achieve this goal include:
  1. ***Random Assignment*** ← "True" experiments
  2. ***Matching*** ← Quasi-experiments, Observational studies
  3. ***Statistical Adjustment*** ← Quasi-experiments, Observational studies

Control by Random Assignment

Controls for the effects of ***all*** characteristics:

- observables or non-observables
- known or unknown

  ⇨ Makes treatment and control groups equivalent *on* ***average*** on ***all*** characteristics

- Differences in outcomes after treatment that are larger than would be expected by chance can be attributed to the *treatment effect* and not to preexisting differences between the groups (causal inference)
- Each unit (e.g., student, classroom, school) is assigned to a treatment or control condition by chance (a random mechanism)
- Treatment and control groups are equivalent on average in the beginning of the study and changes in outcomes should be due to manipulating an independent variable (the treatment) only

Control by Random Assignment

- Considered to be the gold standard in clinical research. The last 20 years arguably, it is considered to be the gold standard in education research.

- Currently used frequently in education.
  ⇨ Strongest design for causal inference

Control by Matching

- *Known* sources of variation may be eliminated by matching (i.e., matching is conducted using measured/observed variables)
- For example, eliminate district, school, or classroom effects before comparing students (e.g., compare students in similar classrooms, schools or districts)
- Matching can take place in the design phase of a study or in data analysis. For example, propensity score matching is one method that creates similar groups to estimate treatment effects using observed covariates
- Matching methods “mimic” random assignment (i.e., aim to balance baseline variables in treatment and control groups)

Control by Matching

- Matching can only be done on **known and observable** characteristics that have been measured
- Perfect matching is not always possible
- It is critical to measure the right variables that will minimize variability (e.g., prior achievement, SES)
- Limits generalizability by removing possibly informative variation (e.g., differences in teachers)
- May reduce the sample size (because the variation is reduced) needed for the study

Control by Statistical Adjustment

- A form of post-hoc pseudo-matching
- Uses statistical associations between outcomes and controls/covariates to simulate matching
- Reduces variation of outcomes in regression models
- Controlling for covariates increases precision of regression estimates (i.e., smaller standard errors)
- Statistical control is possible using **known and observable** characteristics only
- Does not necessarily address *all* preexisting differences prior to assignment to treatment or control conditions (ideally all relevant variables should be measured)

Using Principles of Experimental Design

- When using random assignment, we do not have to know a lot to use it effectively
  - ⇨ Simply conduct random assignment of sample units to treatment and control conditions

- Nonetheless it is good practice to measure important covariates at the baseline of the experiment (e.g., prior achievement, SES) and included them in the analysis for more precise estimation

- It is also imperative to monitor the experiment to ensure random assignment is not compromised

Using Principles of Experimental Design

- When using matching or statistical control, we have to think carefully, ahead of time, about which variables would be important to measure and control for in the analyses (to avoid omitted variable bias)

- Some thorough thinking when designing a quasi-experiment or an observational study is necessary in order to measure all variables that are important to use in the study (i.e., that will produce equivalent groups)

Using Principles of Experimental Design

- Random assignment may not be as efficient as matching or statistical control (i.e., may require larger sample sizes for the same power)

- However, if covariates have been measured, they could/should be used in the power and the statistical analyses

- Including these covariates in a regression model would reduce variability in the outcome and result in a more precise estimation (higher power of tests)

Basic Ideas of Design: Independent Variables

- Categorical independent variables are also called ***factors.***
- The *categories* of factors are called levels
- Some independent variables can be manipulated, others cannot:
  - ***Treatments*** are independent variables that can be manipulated
  - ***Blocks*** (e.g., grades, schools) and ***covariates*** (e.g., gender, race) are independent variables that cannot be manipulated
- Units can be randomly assigned to treatment levels, but *not* to blocks. For example, students within a school (the block) can be assigned randomly to a treatment or a control condition

Blocks

- Blocks can be regions, states, cities, school districts, schools, grades, or even classrooms
- Blocks reduce variability (similar to matching)
- For example, assign randomly schools to treatment conditions within school districts (the blocks)
- Or assign randomly students or classrooms to treatment conditions within schools (the blocks)
- Block effects should be taken into account in a priori power computations and in statistical analyses. Blocks could be random of fixed effects

Basic Ideas of Design: Nesting & Crossing

- *Example:* schools are randomly assigned to treatment conditions
  - ⇨ ***schools are nested within each treatment condition***

Schools

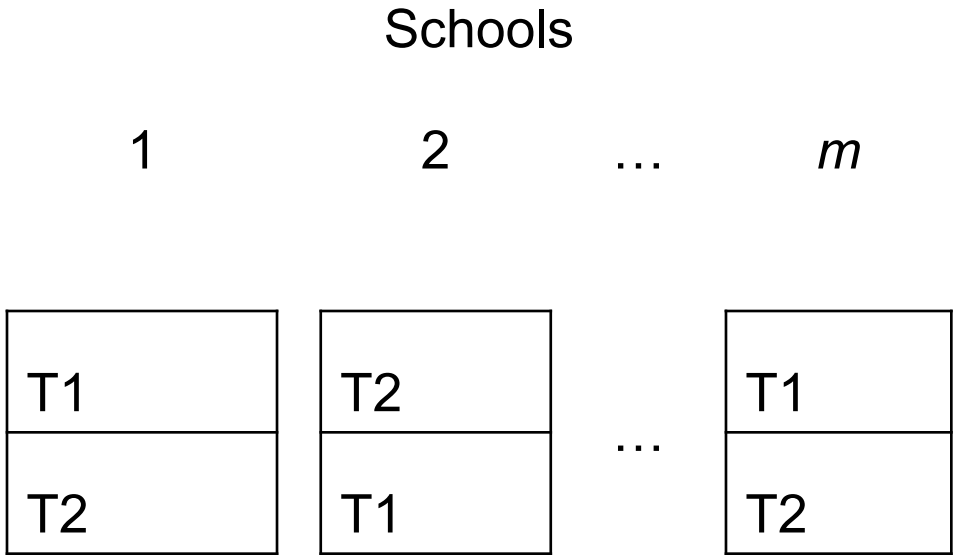
1, 2, … ,m m + 1, … , 2m

Treatments

1 2

Basic Ideas of Design: Nesting & Crossing

- *Example:* classrooms or students are randomly assigned to treatment or control conditions within schools



⇨ ***treatments are crossed with schools***

Three Basic Designs

- Completely Randomized Design
  - Treatments are randomly assigned to individuals (e.g., students). Nesting is not considered
- Cluster (or Group) Randomized Design
  - Also called a Hierarchical Design
  - For example, schools are assigned randomly to treatment or control groups and the same treatment is assigned to all units within the school (classes and students)
- Block Randomized Design
  - For example, students are assigned randomly to treatments *within* schools or grades (the blocks)
  - Larger units such as classes can also be assigned randomly to treatments *within* schools or grades (the blocks)

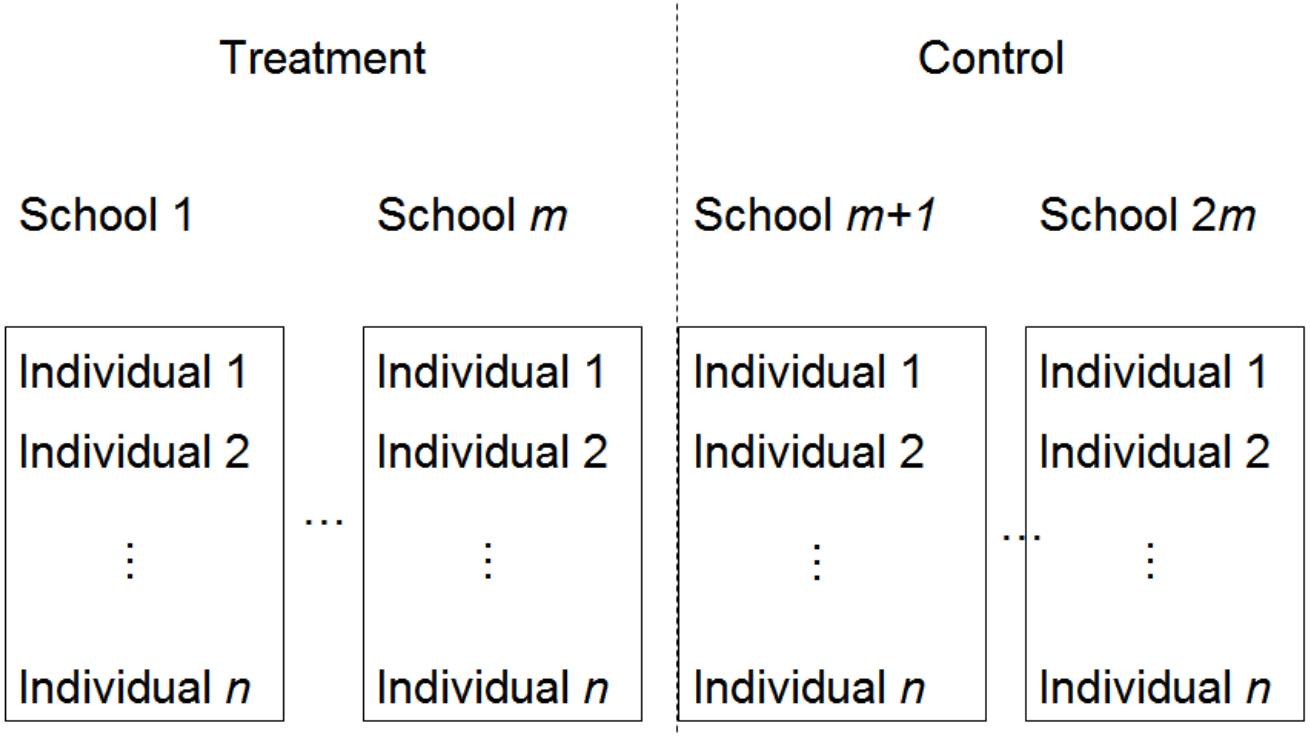
Completely Randomized Design

- Individuals are randomly assigned to one of two treatments:

Treatment	Control
Individual 1	Individual 1
Individual 2	Individual 2
⋮	⋮
Individual *n*	Individual *n*

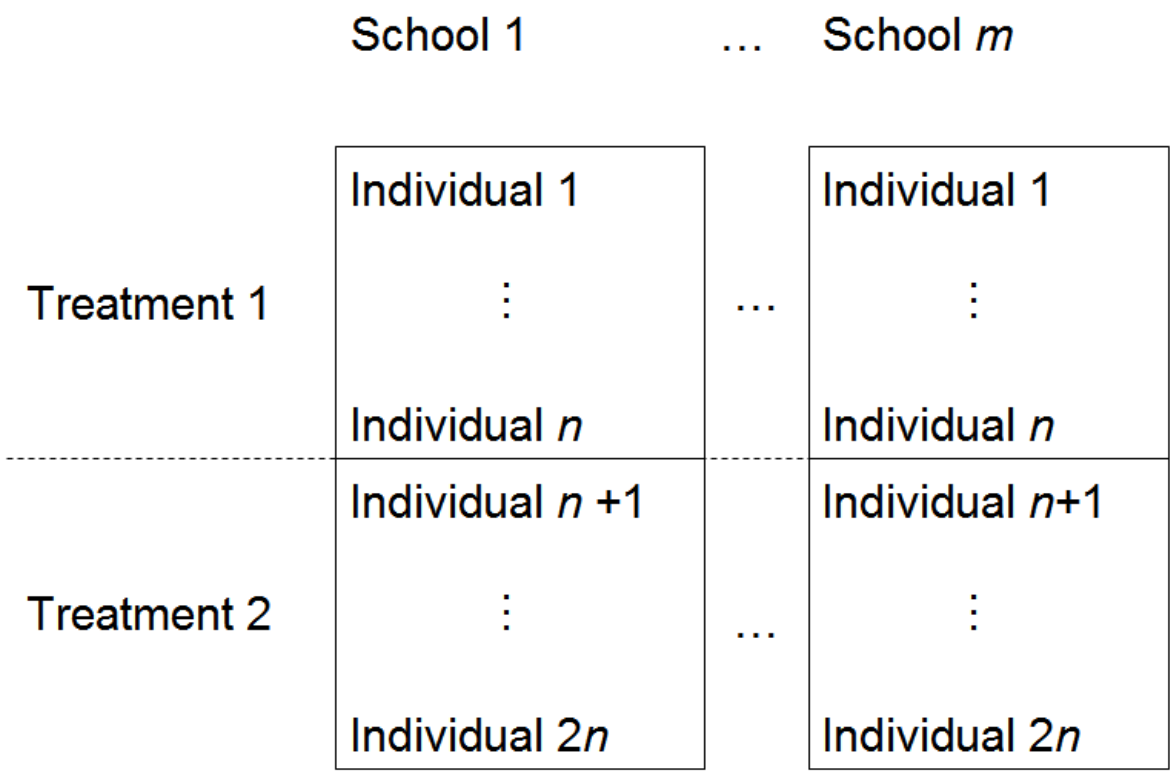
Cluster or Group Randomized Design

- Schools are randomly assigned to one of two treatments, all students within schools receive the treatment:



Block Randomized Design

- Individuals are randomly assigned to one of two treatments within their school:



Randomization Procedures

- Could use a table of random numbers, *but be sure to pick an arbitrary start point*!

- Could use random number generators in statistical software packages. Be sure the seed value varies each time

Post Hoc Test to Check Randomization

- It is common practice to check whether random assignment was successful using observed variables (baseline equivalence of variables)

- This is particularly important when the overall attrition and the attrition in treatment or control groups (i.e., differential attrition) is not low

- This is a post hoc method that can identify variables where random assignment did not work as expected by design (i.e., the means of baseline covariates in the treatment are different than those in the control group)

Post Hoc Test to Check Randomization

- It is unclear that this procedure can discredit random assignment altogether (e.g., a mean difference may be observed by chance) unless there is systematic evidence
- It helps us identify which observed variables to include in our regression models as statistical controls to eliminate pre-existing differences
- Differences should not be significant and the magnitude of the mean difference should not exceed 0.25 standard deviations according to WWC

Post Hoc Test to Check Randomization

- What Works Clearinghouse (WWC) offers some useful guidelines about baseline equivalence of observed variables between treatment and control groups

- WWC offers some useful suggestions about attrition as well

https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

Sampling Models

Sampling Models

- They are closely linked with the research design and the statistical analysis stages
- Example: Which sample will provide a more precise mean estimate?
  - Sample A, with N = 1,000
  - Sample B, with N = 3,000

- It is sample B because if the total population variance is σ_T^2 then the variance of the sample mean is σ_T^2/N (which indicates smaller variances of means in larger samples)

Sampling Models in Educational Research

- Simple random samples are rare in large-scale field research in education

- Educational populations have nested structures (multiple levels, units of different sizes – classes, schools, districts)
  - For example, students nested within classrooms within schools within districts within cities states and so forth

Sampling Models in Educational Research

- Survey research in education often exploits this nested structure for example by first sampling schools and then students within schools
- This sampling strategy is called *multi-stage (multilevel) cluster sampling* in survey research
- *Example:* Clusters such as schools are first sampled and then individuals such as students within clusters are sampled

  ⇨ ***Two-stage (two-level) cluster sample***
- *Example:* Schools are first sampled, then classrooms, then students

  ⇨ ***Three-stage (three-level) cluster sample***

Variance of the Mean of Clustered Sample: Two Levels

- The usual variance calculation is based on a simple random sample
- When clustering is used, the variance must reflect the dependence of individuals within a cluster
- The variance of the mean of a cluster sample:

$$\frac{\tau^2}{m} + \frac{\sigma^2}{mn} = \frac{\sigma^2 + n\tau^2}{mn}$$

where:

$\tau^2 =$ Level-2 variance, $m =$ number of Level-2 units

$\sigma^2 =$ Level-1 variance, $n =$ number of Level-1 units within a Level-2 unit

Variance of the Mean of Clustered Sample

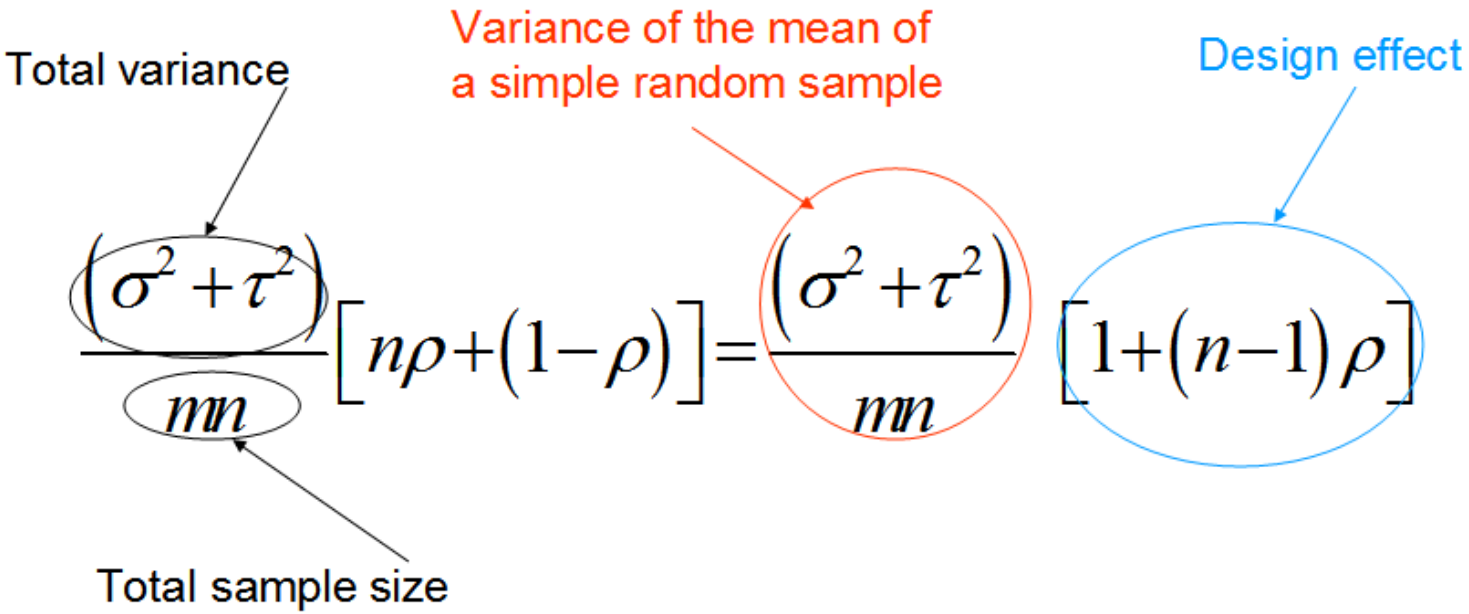
- The intraclass correlation coefficient (ICC), ρ, is the proportion of total variance at the 2nd level (and represents the clustering effect)
- If we write $\rho = \tau^2/(\sigma^2 + \tau^2)$, the variance of the mean becomes:

$$\frac{\left(\sigma^2+\tau^2\right)}{mn}\left[\left(1-\rho\right)+n\rho\right]=\frac{\left(\sigma^2+\tau^2\right)}{mn}\left[1+\left(n-1\right)\rho\right]$$

- where $\left[1+\left(n-1\right)\rho\right]$ is called the design effect (it inflates the variance by a number greater than 1 when $\rho \neq 0$) and captures the clustering effect

Variance of the Mean of Clustered Sample

- This variance can be decomposed:



- where the total population variance is

$$\sigma_T^2 = \sigma^2 + \tau^2$$

Variance of the Mean of Clustered Sample

- Now suppose we have *n* students in *p* classes in each of *m* schools
- Assume a sample size $N = mpn$, and same total population variance of σ_T^2
- If the sampling strategy had been *simple* then the variance of the mean would be:

$$\frac{\left(\sigma_T^2\right)}{mpn}$$

Variance of the Mean of Clustered Sample: Three Levels

- Let's consider a three-stage (three-level) cluster design: *n* students in *p* classes in each of *m* schools

- Assume sample size $N = mpn$, and same total population variance of σ_T^2

- We could have two levels of clustering and thus two ICCs, ρ_3 *(third or school level) and* ρ_2 *(second or classroom level)*

Variance of the Mean of Clustered Sample: Three Levels

- Suppose the variances at the first, second and third level are respectively σ^2, τ^2 and ω^2. and the total variance is the sum of the three variances

- Then, the second level ICC is defined as

$$\rho_2 = \tau^2/(\sigma^2 + \tau^2 + \omega^2)$$

- The third level ICC is defined as

$$\rho_3 = \omega^2/(\sigma^2 + \tau^2 + \omega^2)$$

Variance of the Mean of Clustered Sample: Three Levels

- The variance of the mean is now:

$$\frac{\left(\sigma_T^2\right)}{mpn}\left[1+\left(n-1\right)\rho_2+\left(pn-1\right)\rho_3\right]$$

  - ⇨ The three-level design effect is:

$$\left[1+\left(n-1\right)\rho_2+\left(pn-1\right)\rho_3\right]$$

and captures the clustering effect at the second and third levels

Variance of the Mean of Clustered Sample

- Treatment effects in experiments and quasi-experiments are *mean differences* between two groups

- The sampling model dictates the variance structure and estimation

- Variance impacts:
  - Precision of treatment effect estimates
  - Statistical power

Inferential Population and Inference Models

- The inferential population or the inference model has implications for analysis and therefore for the design of experiments
- Question to consider: *Do we make inferences to the schools in this sample or to a larger population of schools?*
  - Inferences to the ***sampled schools or classes*** in the sample are called **conditional inferences**
  - Inferences to a ***larger population of schools or classes*** are called **unconditional inferences**

  ⇨ *Bottom line:* Inferences are different in **conditional** or **unconditional** models

Inferential Population and Inference Models

- In a conditional inference, we are estimating the mean treatment effect in the observed schools
- In an unconditional inferences, we are estimating the mean treatment effect in the population of schools from which the observed schools were sampled
- In both cases, a mean treatment effect is estimated, but they are *different* parameters with their own respective variances.

Fixed and Random Effects

- Fixed Effects
  - The levels of a factor in a study constitute the entire inference population
  - The inference model is conditional
    - ⇨ The factor is called fixed, and its effects are called *fixed effects*
- Random Effects
  - The levels of a factor in a study are sampled
  - The inference model is unconditional
    - ⇨ The factor is called random, and its effects are called *random effects*

Specifying Analyses

Know the inference model

- Think through the levels of the design that will be included in the analysis

- Decide on the inference model for each level
  - ⇨ *Do I want to generalize to a larger universe than just the units in the sample?*

Specifying Analyses

Know the design

- Generally, ***Covariate effects*** should be fixed effects

- ***Treatment effects*** should be random effects when the design permits it (e.g., block randomized designs)

Applications to Experimental Design

- We will look in detail at the two most widely used experimental designs in education:
  - Cluster randomized designs
  - Block randomized designs

Cluster Randomized Design

The Cluster Randomized Design

- We wish to compare one treatment and one control group
- Assignment to groups is made to whole schools randomly
- Assign *2m* schools with *n* students in each school (assume balanced design)
- There are *m* schools in each treatment condition
- Assign *all* students in each school to the *same* treatment

The Cluster Randomized Design

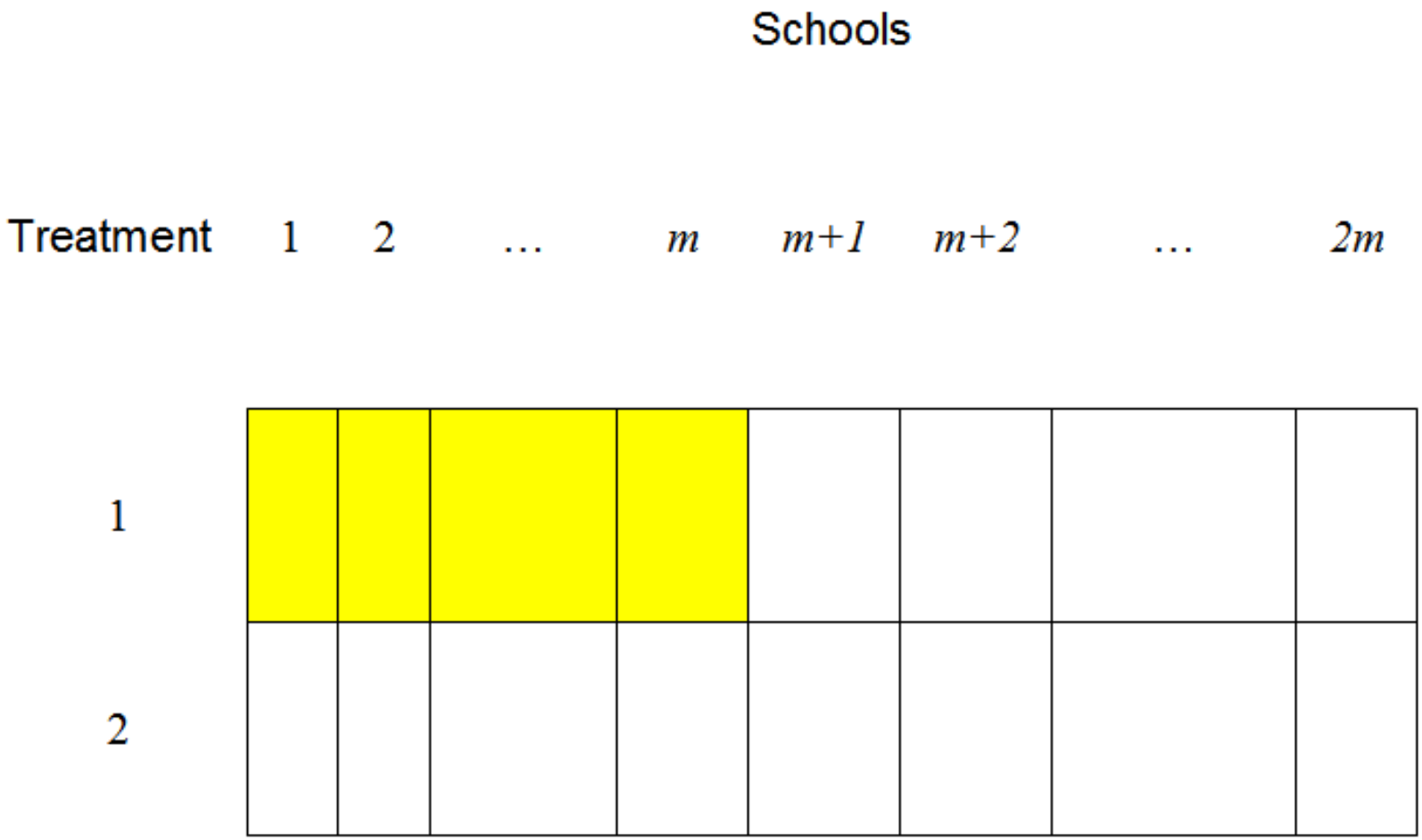
- Diagram of the Experiment:

Schools

Treatment	1	2	…	*m*	*m+1*	*m+2*	…	*2m*
1								
2								

The Cluster Randomized Design

- Treatment 1 Schools:



The Cluster Randomized Design

- Treatment 2 Schools:

Schools

Treatment	1	2	…	*m*	*m+1*	*m+2*	…	*2m*
1								
2								

Two-Level CRT Design No Covariates: Conceptual Multilevel Model

Level 1 (individual level):

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

Level 2 (school level):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \xi_{0j} \qquad \xi_{0j} \sim N\left(0, \tau^2\right)$$

The ICC is:

$$\rho = \tau^2 \Big/ \left(\sigma^2 + \tau^2\right) = \tau^2 \Big/ \sigma_T^2$$

Where σ_T^2 is the total variance

Two-Level CRT Design with Covariates: Conceptual Multilevel Model

Level 1 (individual level):

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0, \sigma_A^2\right)$$

Level 2 (school level):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \gamma_{02} W_j + \xi_{0j} \qquad \xi_{0j} \sim N\left(0, \tau_A^2\right)$$

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

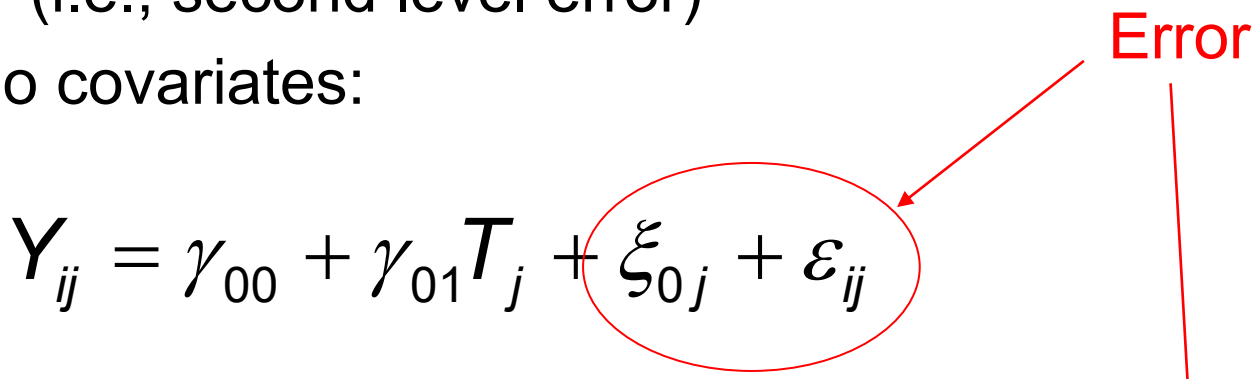
Note that σ_A^2 and τ_A^2 are adjusted

Covariate effect $\beta_{1j} = \gamma_{10}$ is fixed

Two-Level CRT Design: Single Level Model with Random Effects

The previous models can be written as regression models (or mixed models) with additional errors terms (i.e., second level error)

No covariates:



Covariates:

$$Y_{ij} = \gamma_{00} + \gamma_{01} T_j + \gamma_{02} W_j + \gamma_{10} X_{ij} + \xi_{0j} + \varepsilon_{ij}$$

Three-Level CRT Design No Covariates: Conceptual Multilevel Model

Level 1 (individual level):

$$Y_{ijk} = \pi_{0jk} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

Level 2 (class level):

$$\pi_{0jk} = \beta_{00k} + \xi_{0jk} \qquad \xi_{0jk} \sim N\left(0, \tau^2\right)$$

Level 3 (school level):

$$\beta_{00k} = \gamma_{000} + \gamma_{001} T_k + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega^2\right)$$

There are two ICCs:

$$\rho_3 = \omega^2 \Big/ \left(\sigma^2 + \tau^2 + \omega^2\right) = \omega^2 \Big/ \sigma_T^2 \quad \text{(School)}$$

$$\rho_2 = \tau^2 \Big/ \left(\sigma^2 + \tau^2 + \omega^2\right) = \tau^2 \Big/ \sigma_T^2 \quad \text{(Classroom)}$$

Three-Level CRT Design with Covariates: Conceptual Multilevel Model

Level 1 (individual level):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} X_{ijk} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N\left(0, \sigma_A^2\right)$$

Level 2 (classroom level):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} Z_{jk} + \xi_{0jk} \qquad \xi_{0jk} \sim N\left(0, \tau_A^2\right)$$

$$\pi_{1jk} = \beta_{10k}$$

Level 3 (school level):

$$\beta_{00k} = \gamma_{000} + \gamma_{001} T_k + \gamma_{002} W_k + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega_A^2\right)$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{10k} = \gamma_{100}$$

Covariate effects $\pi_{1jk} = \beta_{10k} = \gamma_{100}$ and $\beta_{01k} = \gamma_{010}$ are fixed

Three-Level CRT Design: Single Level Model with Random Effects

The previous three-level models can be written as regression models (or mixed models) with additional errors terms (i.e., second, third level errors)

No covariates:

$$Y_{ijk} = \gamma_{000} + \gamma_{001} T_k + \eta_{00k} + \xi_{0jk} + \varepsilon_{ijk}$$

Error

Covariates:

$$Y_{ijk} = \gamma_{000} + \gamma_{001} T_k + \gamma_{002} W_k + \gamma_{010} Z_{jk} + \gamma_{100} X_{ijk} +$$

$$\eta_{00k} + \xi_{0jk} + \varepsilon_{ijk}$$

Standard Errors of Regression Estimates and Clustering

- Appropriate analyses of two and three level data must take into account the multilevel structure (clustering)
- Otherwise, the standard errors of regression estimates and statistical tests are incorrect
- The standard errors of treatment effect estimates are typically smaller when clustering is ignored
- This results to higher values of *t*-tests and higher probabilities of finding a significant effect (committing a Type I error)

Standard Errors of Regression Estimates and Clustering

- There are at least three ways of adjusting standard errors for clustered data
  - Conduct the analysis using multilevel models (e.g., SAS proc MIXED, SPSS linear mixed models, HLM, MIwin, Stata mixed, R lmer)
  - Post hoc corrections:
    - Use the design effect: multiply the square root of the design effect with the standard error of the estimate.
    - Use clustered standard errors (e.g., Stata) that adjust for clustering

Power Analysis
In CRD

Statistical Power of Treatment Effect

- Power is the probability of detecting an anticipated treatment effect
- Alternatively, power is the probability of accepting the research hypothesis when it is true or rejecting the null hypothesis when it is false
- It is a critical component in the design of experiments

Statistical Power of Treatment Effect

- Power analysis translates to sample sizes needed to detect a treatment effect
- For example, for a specific significance level (e.g., 0.05) and for certain clustering effects (ICC values) how many students, classes, schools are needed to detect an anticipated treatment effect that is meaningful (e.g., 0.20 standard deviations) with a power probability greater than 0.80?
- Typically, we use two-tailed tests
- Typically, we assume one treatment and one control group and a balanced design (makes computations much easier)

Statistical Power of Treatment Effect

- Power in simple random sample designs (no clustering) depends on:
  - Significance level (larger p-value higher power)
  - Effect size (magnitude of treatment effect – larger effect size higher power)
  - Sample size (larger sample size higher power)
- Typically, we assume non-directional hypotheses and two-tailed tests with significance level fixed at 0.05
- We use treatment effect estimates that have been documented in previous work (primary studies or meta-analyses)
- We calculate the sample size necessary to achieve adequate power

Statistical Power of Treatment Effect

- Power analysis reduces to figuring out the sample size that will give us a high probability of detecting the treatment effect
- Every researcher's goal: High power to detect a treatment effect
  - The low bound of power is typically considered 0.80 as
  - Ideally, we want power to be as close to 1 as possible (and the Type II error to be close to 0)
  - Because power computations are not exact it is good practice to ensure power is much greater than 0.80
- In simple random samples designs we can look power up in a table for specific sample sizes and effect sizes (Cohen 1988)

Power of two-sample two-tailed t-test at .05 level

n	d											
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20
2	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.08	0.09	0.10	0.10	0.11
3	0.05	0.05	0.06	0.07	0.08	0.09	0.10	0.12	0.14	0.16	0.18	0.21
4	0.05	0.06	0.07	0.08	0.09	0.11	0.13	0.16	0.19	0.22	0.26	0.30
5	0.05	0.06	0.07	0.09	0.11	0.13	0.16	0.20	0.24	0.29	0.33	0.39
6	0.05	0.06	0.08	0.10	0.12	0.16	0.20	0.24	0.29	0.35	0.41	0.47
7	0.05	0.06	0.08	0.11	0.14	0.18	0.23	0.28	0.34	0.41	0.47	0.54
8	0.05	0.07	0.09	0.12	0.15	0.20	0.26	0.32	0.39	0.46	0.54	0.61
9	0.05	0.07	0.09	0.13	0.17	0.22	0.29	0.36	0.43	0.51	0.59	0.67
10	0.06	0.07	0.10	0.14	0.19	0.25	0.32	0.40	0.48	0.56	0.64	0.72
11	0.06	0.07	0.10	0.15	0.20	0.27	0.35	0.43	0.52	0.61	0.69	0.76
12	0.06	0.08	0.11	0.16	0.22	0.29	0.37	0.47	0.56	0.65	0.73	0.80
13	0.06	0.08	0.11	0.16	0.23	0.31	0.40	0.50	0.60	0.69	0.77	0.84
14	0.06	0.08	0.12	0.17	0.25	0.33	0.43	0.53	0.63	0.72	0.80	0.86
15	0.06	0.08	0.12	0.18	0.26	0.35	0.46	0.56	0.66	0.75	0.83	0.89
16	0.06	0.09	0.13	0.19	0.28	0.38	0.48	0.59	0.69	0.78	0.85	0.91
17	0.06	0.09	0.14	0.20	0.29	0.40	0.51	0.62	0.72	0.81	0.87	0.92
18	0.06	0.09	0.14	0.21	0.31	0.42	0.53	0.65	0.75	0.83	0.89	0.94
19	0.06	0.09	0.15	0.22	0.32	0.44	0.56	0.67	0.77	0.85	0.91	0.95
20	0.06	0.09	0.15	0.23	0.34	0.46	0.58	0.69	0.79	0.87	0.92	0.96
21	0.06	0.10	0.16	0.24	0.35	0.48	0.60	0.72	0.81	0.89	0.94	0.97
22	0.06	0.10	0.16	0.25	0.37	0.49	0.62	0.74	0.83	0.90	0.95	0.97
23	0.06	0.10	0.17	0.26	0.38	0.51	0.64	0.76	0.85	0.91	0.95	0.98
24	0.06	0.10	0.17	0.27	0.40	0.53	0.66	0.77	0.86	0.92	0.96	0.98
25	0.06	0.11	0.18	0.28	0.41	0.55	0.68	0.79	0.88	0.93	0.97	0.99
26	0.06	0.11	0.19	0.29	0.42	0.56	0.70	0.81	0.89	0.94	0.97	0.99
27	0.07	0.11	0.19	0.30	0.44	0.58	0.71	0.82	0.90	0.95	0.98	0.99
28	0.07	0.11	0.20	0.31	0.45	0.60	0.73	0.84	0.91	0.96	0.98	0.99
29	0.07	0.12	0.20	0.32	0.46	0.61	0.75	0.85	0.92	0.96	0.98	0.99
30	0.07	0.12	0.21	0.33	0.48	0.63	0.76	0.86	0.93	0.97	0.99	1.00

Computing Statistical Power

- Power in clustered sample designs depends on:
- Significance level (0.05 level two-tails)
- Effect size δ (standardized mean difference)
- Sample sizes at each level of sampling
  (e.g., m clusters, n individuals per cluster)
- ICC structure (the variances at higher levels such as classroom or school)

Computing Statistical Power

- One could use the power tables provided by Cohen to compute power in cluster designs

- Two things need to be addressed:
  - The number of units will now be the number of clusters (e.g., schools)
  - The effect size needs to be modified to incorporate the effect of clustering (i.e., the design effect)

- Once the new effect size and the new sample size are computed then one can compute power using methods provided by Hedges and Hedberg (2007), Hedges and Rhoads (2010), Konstantopoulos (2009)

Statistical Power in Two-Level CRD

- For example, clusters such as schools at the second or top level are randomly assigned to treatment or control groups
- Cluster sampling is assumed at the top level (i.e., clusters are random effects)
- We assume one treatment and one control group
- Individuals such as students are nested within clusters

Statistical Power: Two-Level CRD

- The power of the t-test of the treatment effect is a function of the degrees of freedom (*df*) of the test and the non-centrality parameter λ of the t-test
- The *df* are a function of the number of level-2 units (e.g., schools): $df = 2(m-1)-q$ (m = number of schools within each condition, q = number of level-2 covariates)
- The non-centrality parameter is a function of the population treatment effect, the variance at the second level τ^2 (i.e., the ICC), and the number of level-1 (e.g., students) and level-2 units (e.g., schools)
- Covariates that reduce variances increase power

Statistical Power: Two-Level CRD without Covariates

- Power increases as the ICC decreases
- Power increases as the effect size increases
- Power increases as the number of schools increases
- Power increases as the number of students increases
- The number of schools affects power much more than the number of students
- Larger proportions of variance explained at each level leads to higher power. The effect of covariates at different levels depends on the number of clusters, the units within clusters, and the centering of level-1 covariates (Konstantopoulos 2012)

Statistical Power in Two-Level CRD

- Thus, a researcher would want to

  - Sample more schools than students within schools (but within the budget)

  - Include covariates that explain much variability at each level (e.g., prior achievement, SES). However, level-2 covariates reduce the *df* of the t-test, which can reduce power to some degree (i.e., there is a trade off). Use top level covariates that explain much variance at that level

  - It is important to have an educated guess about the magnitude of the anticipated treatment effect

Statistical Power in Three-Level CRD

- Level-3 units or clusters (e.g., schools) are randomly assigned to treatment and control groups
- Level-2 units or sub-clusters (e.g., classrooms)
- Level-1 units are individuals (e.g., students)
- Cluster sampling is assumed at the middle and top levels (i.e., schools and classrooms are random effects)

Statistical Power of Treatment Effect: Three-Level CRD

- Suppose students (level-1 units) are nested within classes (level-2 units) and classes are nested within schools (level-3 units)
- The power of the t-test of the treatment effect is a function of the degrees of freedom (*df*) and the non-centrality parameter λ of the test
- The *df* are a function of the number of level-3 units (e.g., schools): $df = 2(m-1)-q$ (m = number of schools within each condition, q = number of level-3 covariates)
- The non-centrality parameter is a function of the population treatment effect, the variances (or ICCs) at the second the third levels, and the number of level-1, level-2, and level-3 units

Statistical Power of Treatment Effect: Three-Level CRD

- The *df* are a function of the number of level-3 units in the sample and the number of covariates at the third level
- The non-centrality parameter λ is a function of
  - The number of level-3 units
  - The number of level-2 units within level-3 units
  - The number of level-1 units within level-2 and level-3 units
  - The clustering at the second and third levels
  - The effect size (standardized treatment effect)

Power in Three-Level CRD: No Covariates

- Power increases as the variances (or ICCs) at the second and third levels decrease
- Power increases as the effect size increases
- Power increases as the number of schools increases
- Power increases as the number of classes increases
- Power increases as the number of students increases
- The number of schools affects power more than the number of classes or students

Statistical Power in Two-Level CRD

- Thus, a researcher would want to
  - sample more schools than classes or students within schools (but within the budget)
  - Include covariates that explain much variability at each level (e.g., prior achievement, SES). However, level-3 covariates reduce the *df* of the t-test and may reduce power to some degree
  - Have an idea of the anticipated treatment effect estimate
- Larger proportions of variance explained at each level leads to higher power. The effect of covariates at different levels depends on the number of clusters, the units within clusters, and the centering of lower level covariates (Konstantopoulos 2012)
  - Use top-level covariates that explain much variance at the that level

Cost Considerations in CRD

- One can incorporate cost functions to maximize the non-centrality parameter and in turn the power estimates
- The idea is to conduct optimal sampling of units at each level given a budget
- Two-level designs (Raudenbush, 1997)
- Three-level designs (Konstantopoulos, 2009, 2011)
- Four-level designs (Hedges & Borenstein, 2014)

Block Randomized Design

The Block Randomized Design (BRD)

- We wish to compare a treatment and a control group
- Assign randomly *n* units (e.g., students) to treatment or control conditions within blocks (e.g., grades, schools)
- Within each block there are 2*n* level-1 units (assume a balanced design)
- The block is treated as a random effect (i.e., the between-block variability is taken into account). The block is a cluster or sub-cluster and thus cluster sampling is assumed at the top level

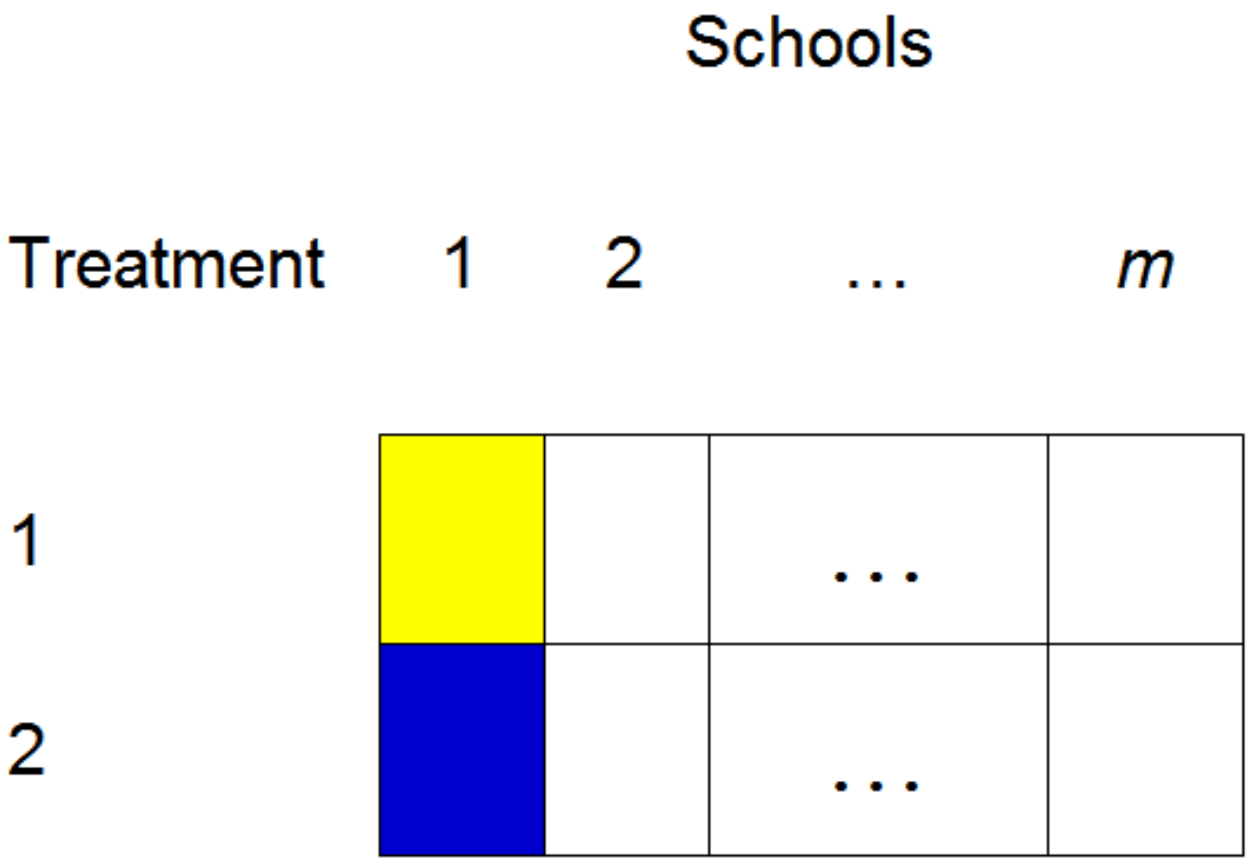
Two-Level BRD

- Diagram of the Experiment:

Schools

Treatment	1	2	…	*m*
1			…	
2			…	

Two-Level BRD



Two-Level BRD No Covariates: ANOVA Framework

- The statistical model for the observation on the i^{th} student in the j^{th} treatment in the k^{th} school:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

where:

$\mu =$ grand mean

$\alpha_j =$ average effect of being in treatment j

$\beta_k =$ average effect of being in school k

$\alpha\beta_{jk} =$ treatment by school interaction

$\varepsilon_{ijk} =$ residual

School Effect

School random effect

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

Treatment by school interaction
(random effect)

Two-Level BRD: Conceptual Multilevel Framework (MLF)

- Without covariates (student *i* in school *j*):

  Level 1 (student level):

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

  Level 2 (school level):

School random effect

$$\beta_{0j} = \gamma_{00} + \eta_{0j} \qquad \eta_{0j} \sim N\left(0, \tau^2\right)$$

$$\beta_{1j} = \gamma_{10} + \eta_{1j} \qquad \eta_{1j} \sim N\left(0, \tau_T^2\right)$$

Treatment by School interaction (random effect)

  Subscript *T* indicates treatment

Two-Level BRD

- Without covariates the mixed model is:

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

School random effect

Treatment by School interaction
(random effect)

Two-Level BRD with Covariates: Conceptual MLF

Level 1 (individual level):

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \beta_{2j} X_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0, \sigma_A^2\right)$$

Level 2 (school level):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + \eta_{0j} \qquad \eta_{0j} \sim N\left(0, \tau_A^2\right)$$

$$\beta_{1j} = \gamma_{10} + \eta_{1j} \qquad \eta_{1j} \sim N\left(0, \tau_T^2\right)$$

$$\beta_{2j} = \gamma_{20}$$

Mixed model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} T_{ij} + \gamma_{20} X_{ij} + \gamma_{01} W_j +$$

$$\eta_{0j} + T_{ij} \eta_{1j} + \varepsilon_{ij}$$ ← Error

Two-Level BRD with Covariate X as Random: MLF

Level 1 (individual level):

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \beta_{2j} X_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N\left(0, \sigma_A^2\right)$$

Level 2 (school level):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + \eta_{0j} \qquad \eta_{0j} \sim N\left(0, \tau_A^2\right)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + \eta_{1j} \qquad \eta_{1j} \sim N\left(0, \tau_{T,A}^2\right)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} W_j + \eta_{2j} \qquad \eta_{2j} \sim N\left(0, \tau_{X,A}^2\right)$$

Mixed model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} T_{ij} + \gamma_{20} X_{ij} + \gamma_{01} W_j + \gamma_{11} T_{ij} W_j + \gamma_{21} X_{ij} W_j +$$

$$\eta_{0j} + T_{ij} \eta_{1j} + X_{ij} \eta_{2j} + \varepsilon_{ij}$$ ← Error

Fixed and Random Effects

- Should blocks be fixed or random?
- Fixed Effects
  - If the inference targets the blocks (e.g., schools) in the sample, then the schools can be treated as fixed effects (i.e., block of school dummies in the regression model)
  - Then a model with one fixed student level covariate becomes

$$Y_{ij} = \gamma_0 + \gamma_1 T_{ij} + \gamma_2 X_{ij} + \boldsymbol{SC}_j \boldsymbol{\Gamma}_3 + T_{ij} \boldsymbol{SC}_j \boldsymbol{\Gamma}_4 + \varepsilon_{ij}$$

  where ***SC*** are school fixed effects (block of school dummies)
- Random Effects
  - If the inference targets a larger population of blocks (e.g., schools), then the schools can be treated as random effects. The variances of the random effects are taken into account in a weighted estimation procedure. Cluster sampling at the top level is assumed

Three-Level BRD No Covariates: ANOVA Framework

- Treatment assignment at first level (e.g., random assignment of students within a classroom). Level-2 and level-3 units are random (cluster sampling is assumed)
- The statistical model for the observation on the i^{th} student in the j^{th} treatment in the k^{th} classroom in the l^{th} school: $Y_{ijkl} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \gamma_l + \alpha\gamma_{jl} + \varepsilon_{ijkl}$

$\mu =$ grand mean

$\alpha_j =$ average effect of being in treatment j

$\beta_k =$ average effect of being in class k

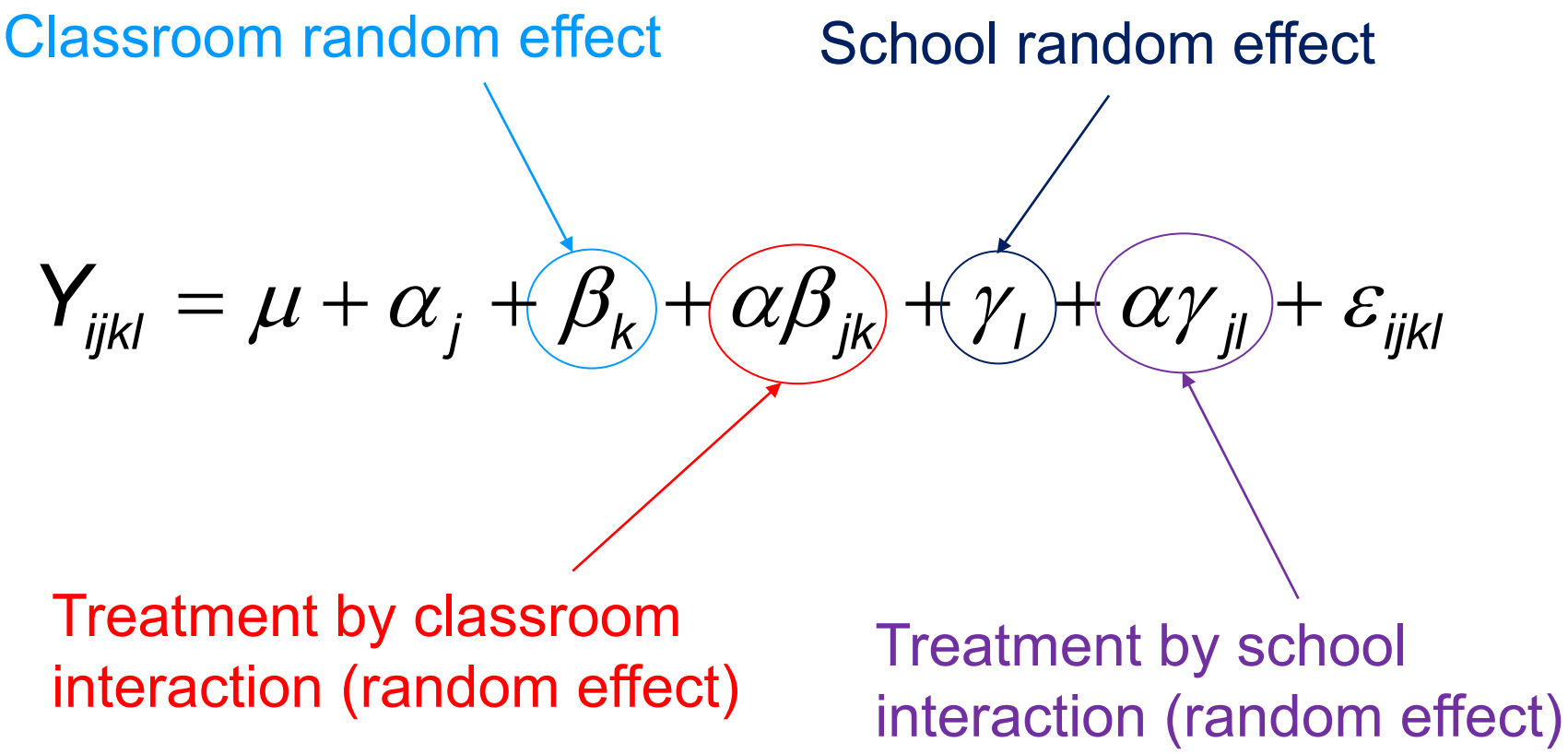
$\alpha\beta_{jk} =$ treatment by class interaction

$\gamma_l =$ average effect of being in school l

$\alpha\gamma_{jl} =$ treatment by school interaction

$\varepsilon_{ijkl} =$ residual

Classroom and School Effects



Three-Level BRD: Conceptual MLF

Without covariates (student *i* in classroom *j* in school *k*)

Level 1 (individual level):

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} T_{ijk} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

Level 2 (class level):

$$\beta_{0jk} = \gamma_{00k} + \xi_{0jk} \qquad \xi_{0jk} \sim N\left(0, \tau^2\right)$$

$$\beta_{1jk} = \gamma_{10k} + \xi_{1jk} \qquad \xi_{1jk} \sim N\left(0, \tau_T^2\right)$$

Level 3 (school level):

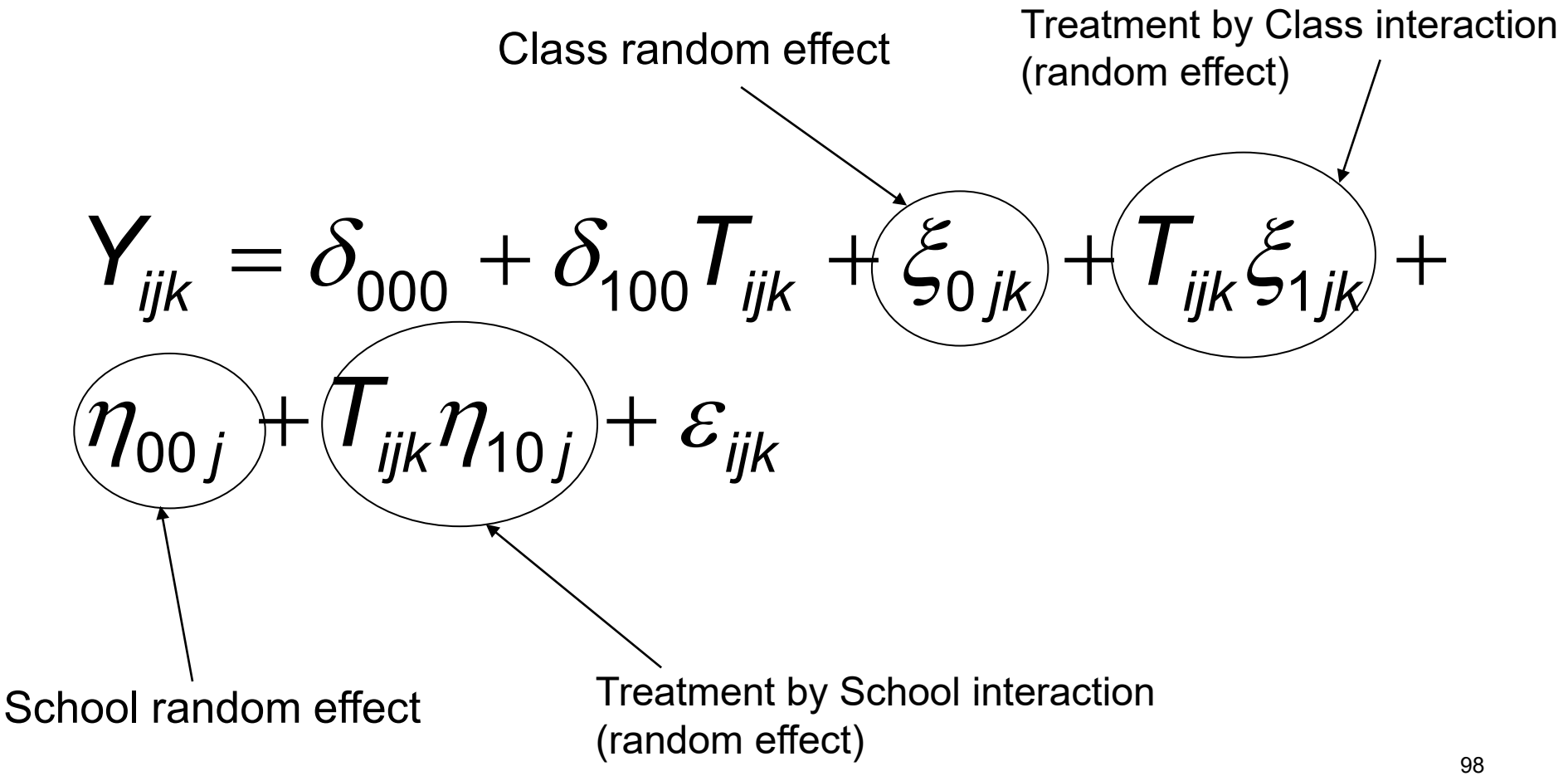
$$\gamma_{00k} = \delta_{000} + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega^2\right)$$

$$\gamma_{10k} = \delta_{100} + \eta_{10k} \qquad \eta_{10k} \sim N\left(0, \omega_T^2\right)$$

Subscript *T* indicates treatment.

Three-Level BRD

- Without covariates the mixed model is:



Three-Level BRD: Conceptual MLF

With covariates (student *i* in classroom *j* in school *k*)

Level 1 (individual level):

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} T_{ijk} + \beta_{2jk} X_{ijk} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N\left(0, \sigma_A^2\right)$$

Level 2 (class level):

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k} Z_{jk} + \xi_{0jk} \qquad \xi_{0jk} \sim N\left(0, \tau_A^2\right)$$

$$\beta_{1jk} = \gamma_{10k} + \xi_{1jk} \qquad \xi_{1jk} \sim N\left(0, \tau_T^2\right)$$

$$\beta_{2jk} = \gamma_{20k}$$

Level 3 (school level):

$$\gamma_{00k} = \delta_{000} + \delta_{001} W_k + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega_A^2\right)$$

$$\gamma_{10k} = \delta_{100} + \eta_{10k} \qquad \eta_{10k} \sim N\left(0, \omega_T^2\right)$$

$$\gamma_{01k} = \delta_{010}$$

$$\gamma_{20k} = \delta_{200}$$

Fixed and Random Effects

- Random Effects
  - If the inference targets a larger population of schools, then schools are treated as random effects
  - The mixed model is:

$$Y_{ijk} = \delta_{000} + \delta_{100}T_{ijk} + \delta_{200}X_{ijk} + \delta_{010}Z_{jk} + \delta_{001}W_k + \xi_{0jk} + T_{ijk}\xi_{1jk} + \eta_{00k} + T_{ijk}\eta_{10k} + \varepsilon_{ijk}$$

- Fixed Effects
  - If the inference targets the schools in the sample, then the schools are treated as fixed effects (i.e., block of school dummies)
  - Then the model becomes:

$$Y_{ijk} = \delta_0 + \delta_1 T_{ijk} + \delta_2 X_{ijk} + \delta_3 Z_{jk} + \delta_4 W_{jk} + \xi_{0jk} + T_{ijk}\xi_{1jk} + \boldsymbol{SC_k}\boldsymbol{\Delta}_5 + T_{ijk}\boldsymbol{SC_k}\boldsymbol{\Delta}_6 + \varepsilon_{ijk}$$

where ***SC*** are school fixed effects (dummy variables)

Three-Level BRD No Covariates: ANOVA Framework

- Treatment assignment at second level (e.g., random assignment of classrooms within a grade/school). Level-2 and level-3 units are random (cluster sampling is assumed)
- The statistical model for the observation on the i^{th} student in the j^{th} classroom in the k^{th} treatment in the l^{th} school:

$$Y_{ijkl} = \mu + \alpha_k + \beta_j + \gamma_l + \alpha\gamma_{kl} + \varepsilon_{ijkl}$$

$\mu =$ grand mean

$\alpha_k =$ average effect of being in treatment k

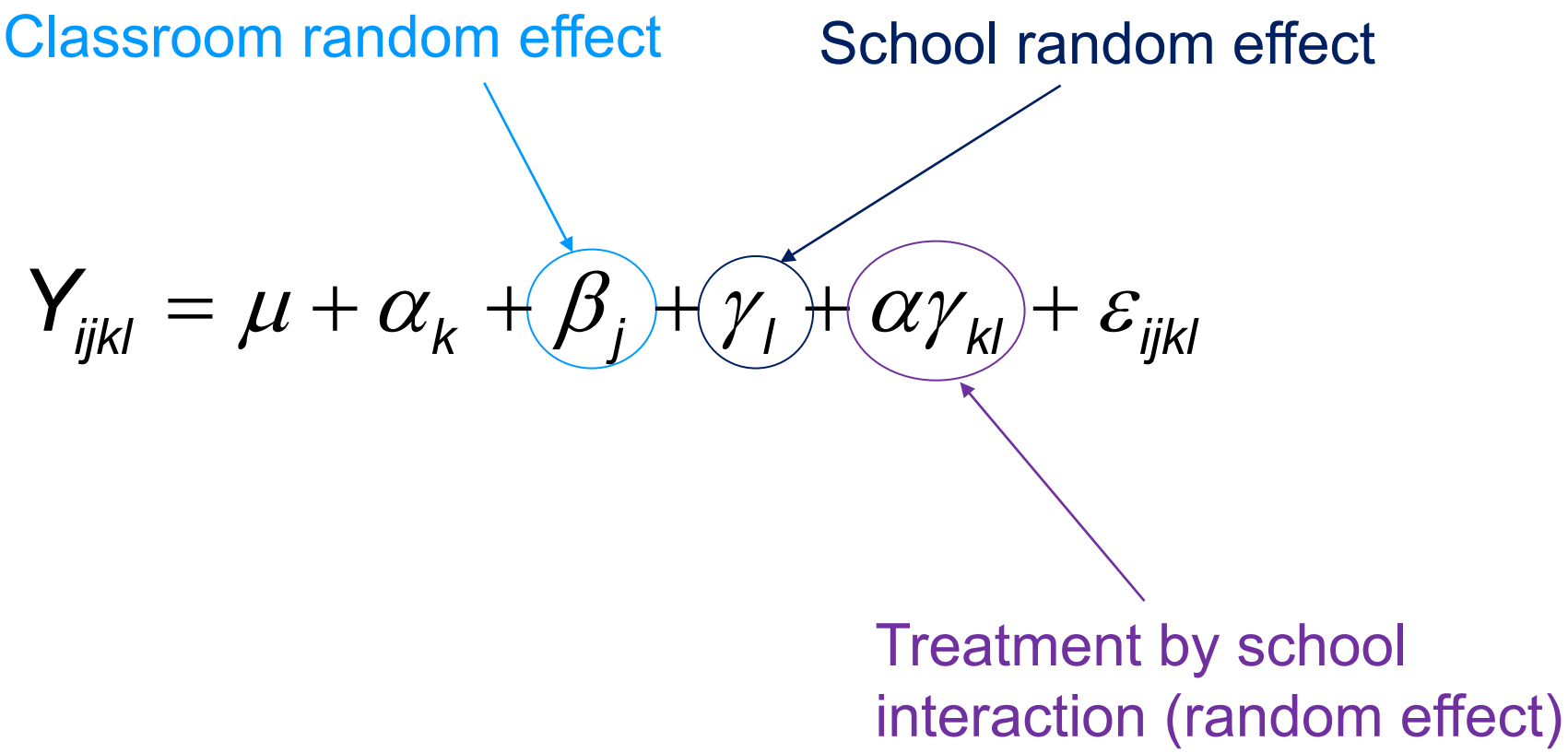
$\beta_j =$ average effect of being in class j

$\gamma_l =$ average effect of being in school l

$\alpha\gamma_{kl} =$ treatment by school interaction

$\varepsilon_{ijkl} =$ residual

Classroom and School Effects



Three-Level BRD: Conceptual MLF

No Covariates (student *i* in classroom *j* in school *k*)

Level 1 (individual level):

$$Y_{ijk} = \beta_{0jk} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

Level 2 (class level):

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k} T_{jk} + \xi_{0jk} \qquad \xi_{0jk} \sim N\left(0, \tau^2\right)$$

Level 3 (school level):

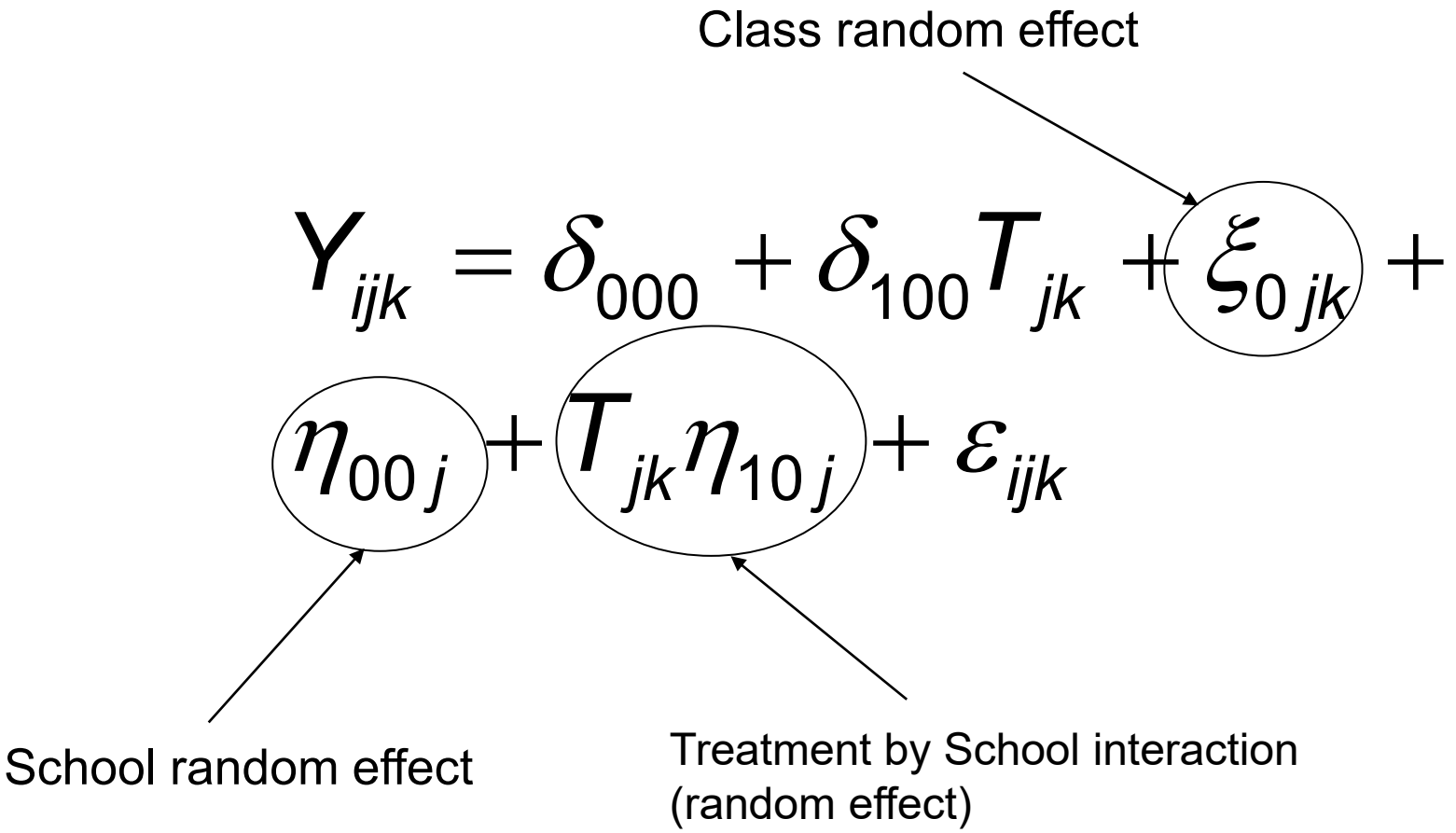
$$\gamma_{00k} = \delta_{000} + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega^2\right)$$

$$\gamma_{01k} = \delta_{010} + \eta_{10k} \qquad \eta_{10k} \sim N\left(0, \omega_T^2\right)$$

Subscript *T* indicates treatment.

Three-Level BRD

- Without covariates the mixed model is:



Three-Level BRD: Conceptual MLF

With covariates (student *i* in classroom *j* in school *k*)

Level 1 (individual level):

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} X_{ijk} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N\left(0, \sigma_A^2\right)$$

Level 2 (class level):

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k} T_{jk} + \gamma_{02k} Z_{jk} + \xi_{0jk} \qquad \xi_{0jk} \sim N\left(0, \tau_A^2\right)$$

$$\beta_{1jk} = \gamma_{10k}$$

Level 3 (school level):

$$\gamma_{00k} = \delta_{000} + \delta_{001} W_k + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega_A^2\right)$$

$$\gamma_{01k} = \delta_{010} + \eta_{01k} \qquad \eta_{01k} \sim N\left(0, \omega_T^2\right)$$

$$\gamma_{02k} = \delta_{020}$$

$$\gamma_{10k} = \delta_{100}$$

Fixed and Random Effects

- Random Effects
  - Notice that in this design a whole classroom is randomly assigned to a treatment or a control group and is a sub-cluster
  - If the inference targets a larger population of schools, then the schools are also treated as random effects
  - The mixed model is:

$$Y_{ijk} = \delta_{000} + \delta_{010} T_{jk} + \delta_{100} X_{ijk} + \delta_{020} Z_{jk} + \delta_{001} W_k + \xi_{0jk} + \eta_{00k} + T_{jk} \eta_{01k} + \varepsilon_{ijk}$$

What Determines Fixed or Random Effects

- The underlying assumptions about the sampling scheme involved are crucial in dictating whether effects should be fixed or random
- Similarly, the underlying assumptions about the inference (conditional or unconditional) are crucial in dictating whether effects should be fixed or random
- In small-scale empirical research sampling at the individual level (e.g., simple random or convenient sampling) is frequent. In this case the number of larger units such as classrooms or schools is small and thus classes or schools don't necessarily need to be random. They can be modeled as fixed instead. The inference is about a population of students

What Determines Fixed or Random Effects

- In large-scale empirical research larger units such as schools are often sampled. This is called cluster sampling and schools are modeled as random. If classes are sampled at the second stage, then classes are modeled as random as well. The inference is about a population of schools or classrooms
- When a larger unit such as a classroom or a school is the unit of random assignment then we would want to capture the clustering at that level (i.e., use random effects)
- Notice that clustering matters the most when we analyze outcomes at the individual level and the assignment is at a higher level (e.g., classroom or school)

Fixed Vs Random Effects

- When the assignment is at the student level although students are nested within classes and schools the sampling frame does not always have to follow a cluster sampling. That is, perhaps the sampling involves individuals only and the inference is about populations of individuals (not classes or schools)
- In such cases of block designs classes and schools do not need to be random. Instead, they can be modeled as fixed either as observed variables or as fixed effects (block of dummies)
- In multilevel models treating larger units (e.g., schools) as random adds levels in the hierarchy (and variances). Modeling larger units (e.g., schools) as fixed however, reduces levels in the hierarchy

Power Analysis
In BRD

Statistical Power: Two-Level BRD

- The power of the t-test of the treatment effect is a function of the degrees of freedom (*df*) and the non-centrality parameter λ of the test
- The *df* are a function of the number of level-2 units (e.g., schools): $df = m - q - 1$ (*m* = *total* number of schools, q = number of level-2 covariates)
- The non-centrality parameter is a function of the population treatment effect, the variance of the treatment effect τ_T^2, and the number of level-1 (e.g., students) and level-2 units (e.g., schools)

Statistical Power: Two-Level BRD

- Power increases as the variance of the treatment effect decreases
- Power increases as the effect size increases
- Power increases as the number of schools increases
- Power increases as the number of students increases

Statistical Power: Two-Level BRD

- Covariates that explain variance at the first or second level increase power. Covariates at the second level reduce the *df* of the t-test (use fewer powerful school predictors)

- Power is typically higher in block randomized designs. One main reason is that the variance of the treatment effect across level-2 units is typically smaller than the variance of the outcome across level-2 units (between-school variance)

Statistical Power of Treatment Effect: Three-Level BRD

- Suppose students (level-1 units) are nested within classes (level-2 units) and classes are nested within schools (level-3 units)
- The power of the t-test of the treatment effect is a function of the degrees of freedom (*df*) and the non-centrality parameter λ of the test
- The *df* are a function of the number of level-3 units (e.g., schools): $df = m - q - 1$ (*m* = *total* number of schools, *q* = number of level-3 covariates)
- The non-centrality parameter is a function of the population treatment effect, the variance of the treatment effect, and the number of level-1, level-2, and level-3 units

Power in Three-Level BRD: Treatment at First Level

- Power increases as the variance of the treatment effect decreases at the second or third levels
- Power increases as the effect size increases
- Power increases as the number of schools increases
- Power increases as the number of classes increases
- Power increases as the number of students increases

Power in Three-Level BRD: Treatment at First Level

- Covariates that explain much variance at the first, second, or third levels increase power. But level-3 covariates reduce the *df* of the t-test
- Power is typically higher in this block randomized design. One main reason is that the variance of the treatment effect across level-2 or level-3 units is typically smaller than the variance of the outcome across level-2 or level-3 units (e.g., between-classroom or between-school variance)
- When treatment is at the first level the design produces typically the highest power other things being equal

Power in Three-Level BRD: Treatment at Second Level

- Power increases as the variance of the treatment effect decreases at the third level
- Power increases as the effect size increases
- Power increases as the number of schools increases
- Power increases as the number of classes increases
- Power increases as the number of students increases

Power in Three-Level BRD: Treatment at Second Level

- Covariates that explain much variance at the first, second, or third levels increase power. But level-3 covariates reduce the *df* of the t-test

Cost Considerations in BRD: Optimal Design

- One can incorporate cost functions to maximize the non-centrality parameter and in turn the power estimates

- The idea is to conduct optimal sampling of units at each level given a budget

- Two-level designs (Raudenbush & Liu, 2000)
- Three-level designs (Konstantopoulos, 2013)
- Four-level designs (Hedges & Borenstein, 2014)

Centering

Centering

- Centering is a transformation applied typically to the independent variables
- In simple random sample designs, a variable is centered by subtracting the mean from each value
- If X_i is the independent variable, the centered variable is:

$$X_i^c = \left(X_i - \bar{X}_\bullet\right)$$

where $\bar{X}_\bullet$ is the mean of the X_i's in the sample

- The mean of the new centered variable is zero

Centering

- Centering changes the value and the meaning of the intercept (it's the mean of the outcome)
- Centering also changes the standard error of the intercept
- Centering ***does not*** change the value or the meaning of the regression coefficient
- Centering ***does not*** change the standard error of the regression coefficient

Centering: Two-Level Case

- In two-level designs (e.g., students within schools, there are two kinds of centering:
  - Grand mean centering of student predictors
  - Group mean centering of student predictors
- Grand mean centering is subtracting the grand mean:

$$X_{ij}^{Grand} = \left(X_{ij} - \bar{X}_{\bullet\bullet}\right)$$

- Group mean centering is subtracting the group/school mean:

$$X_{ij}^{Group} = \left(X_{ij} - \bar{X}_{i\bullet}\right)$$

- These centering methods affect the interpretation of the school intercept

Grand Mean Centering

- Grand mean centering changes the meaning of the intercept in the j^{th} school
- The school intercept is now the mean outcome in the school minus an adjustment due to the student predictors
- With Grand Mean Centering:
  - Student predictors can explain school variance
  - Student predictors are not independent of school predictors
- **Centering changes the precision of the intercept only (as in regression)**

Group Mean Centering

- Group mean centering changes the meaning of the intercept in the i^{th} school
- The school intercept is now the mean outcome in the school *not* adjusted by student predictors
- With Group Mean Centering:
  - Student predictors *cannot* explain school variance
  - Student predictors are *independent* of school predictors
  - Can use aggregate variables at the school level to reduce school-level variance
  - Student predictors are adjusted for school differences (school effects)
- **Centering changes the precision of all estimates**

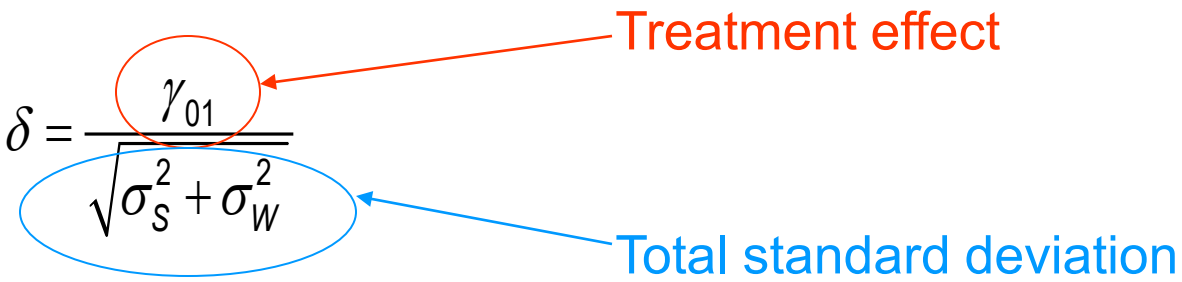
Effect Sizes

Effect Sizes

- Effect sizes can be defined in more than one way in multilevel designs
- The effect size is a typically defined a standardized mean difference
- The numerator is the mean difference
- The key difference is which standard deviation is used to standardize the mean difference
- The easiest one to use is the total standard deviation

Effect Sizes

- In two-level cluster randomized designs, this leads to:

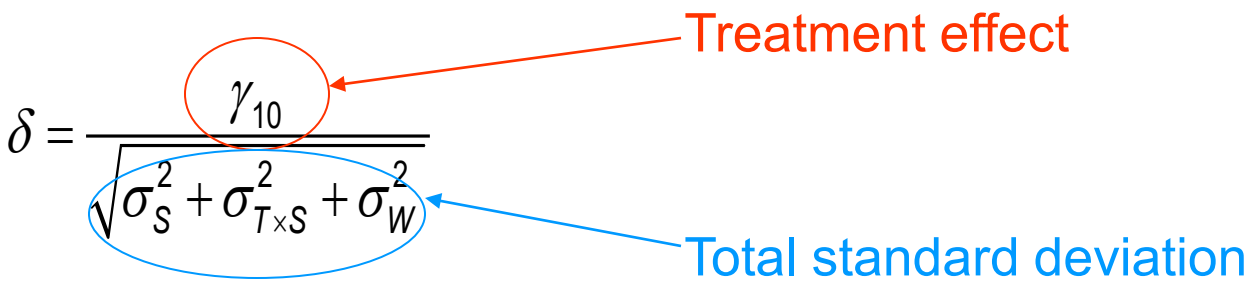


- In three-level cluster randomized designs, this leads to:

$$\delta = \frac{\gamma_{001}}{\sqrt{\sigma_S^2 + \sigma_C^2 + \sigma_W^2}}$$

Effect Sizes

- In two-level block randomized designs, this leads to:



- In three-level block randomized designs, this leads to:

$$\delta = \frac{\gamma_{010}}{\sqrt{\sigma_S^2 + \sigma_{T\times S}^2 + \sigma_C^2 + \sigma_W^2}}$$

Questions About Analyses

Questions About Analyses

Q. My schools all come from two districts, but I am randomly assigning the schools. Do I have to take district into account some way?

A. In this case district is a block and a district dummy can be created and included as a school level variable to capture possible district differences. Perhaps districts shouldn't be modeled as random effects in this case (very few units). Inferences are most likely made for these two districts only. An interaction term between treatment and district will also be informative

Questions About Analyses

Q. Why can't I use regression to analyze experiments? What's the advantage of multilevel models?

A. Of course regression can be used to analyze experimental data. However, if there is clustering in the data, the standard errors of the estimates need to be adjusted. This can be done using either a design effect or clustered standard errors. Multilevel models correct the standard errors instantly via the estimation

Questions About Analyses

Q. Can I use “school fixed effects” to analyze data from a randomized block design?

A. If cluster sampling is not assumed for schools, then one can use a regression model that controls for school fixed effects (differences between blocks). It is also recommended however, that one includes interactions between the treatment and schools. In practice, If there are 81 schools in the sample that suggests 160 dummies in the regression model.

B. If cluster sampling is assumed for schools, one only needs to introduce two random effects (one to capture between-school variability and one to capture the treatment by school random effect interaction). This is a simpler model

Questions About Analyses

Q. We randomly assigned, but our assignment was corrupted by treatment switchers. What do we do?

A. One could run an intention to treat (ITT) analysis that estimates the effect of the initial assignment (make sure you have collected such data). A robustness check would be to also run a treatment on the treated (TOT) analysis and compare the estimates. If the estimates are very similar, then switching was likely not a threat. One could also use initial random assignment as an instrument for treatment received. This is called an instrumental variables (IV) analysis and involves two stages

Questions About Analyses

Q. We randomly assigned, but our assignment was corrupted by attrition. What do we do?

A. Report the overall attrition as well as the attrition in the treatment and control schools separately (differential attrition). Show baseline equivalence in observed covariates in the analytic sample (after attrition) for the two groups especially if attrition is not low. If data on the initial sample are available, one could run an ITT analysis and compare these estimates to estimates of the TOT analysis. Alternatively, one could run an IV analysis using initial random assignment as the instrument

Questions About Analyses

Q. We randomly assigned but got a big imbalance on characteristics we care about (gender, race, language, SES, pretest scores). What do we do?

A. First, is there credible information about whether the experiment was compromised or not? Second, one could check whether baseline differences are non-significant once strata are taken into account. Another possibility is to use these variables in propensity score methods that aim to create similar groups for the two treatment conditions. Or in the regression or multilevel model one could include these covariates as statistical controls to correct for selection. Perhaps interactions between treatment and characteristics should also be included in the model. Note that when student characteristics are not part of the random assignment process imbalance is possible.

Questions About Analyses

Q. We want to use student covariates to improve precision, but we find that they act somewhat differently in different schools (have different slopes). What do we do?

A. In a multilevel model one can model student covariates as random effects at the school level (i.e., cross-level interaction random effects) and compute their variances across schools. In a regression model one would need to create student covariate by school interactions (that may produce a large number of interaction effects if many schools)

Questions About Analyses

Q. We get somewhat different variances in different schools. Should we use robust standard errors?

A. Non-constant variance needs to be taken into account in the computation of standard errors. A common way to address the problem of variance heterogeneity is to compute robust standard errors

References

- Boruch, R. F. (1997). Randomized experiments for planning and evaluation. Thousand Oaks, CA: Sage.
- Boruch, R., Weisburd, D., & Berk, R. (2010). Place randomized trials. In A. Piquero & D. Weisburd (Eds.), Handbook of quantitative criminology (pp. 481-502). New York, NY: Springer.
- Cochran, W. G. (1977). Sampling techniques. New York, NY: Wiley.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd ed. New York: Academic Press.
- Donner,A., & Klar, N. (2000).Design and analysis of cluster randomization trials in health research. London, UK: Arnold.
- Hedges L. V., & Borenstein, M. (2014). Conditional Optimal Design in Three- and Four-Level Experiments. Journal of Educational and Behavioral Statistics, 39 (4), 257-281.
- Hedges, L. V., & Hedberg, E. (2007). Intraclass correlation values for planning group randomized trials in education. Educational Evaluation and Policy Analysis, 29, 60-87.
- Hedges L. V., & Rhoads, C. (2010). *Statistical power analysis in education research*. U.S. Department of Education

- Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage Publishing.
- Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1,* 66-88.
- Konstantopoulos, S. (2008c). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1,* 265-288.
- Konstantopoulos, S. (2009). Using Power Tables to Compute Power in Multilevel Experimental Designs. *Practical Assessment Research and Evaluation, 14(10)*, 1-9.
- Konstantopoulos, S. (2011). Optimal Sampling of Units in Three-Level Cluster Randomized Designs: An ANCOVA Framework. *Educational and Psychological Measurement, 71*, 798-813.
- Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research, 47,* 392-420.
- Konstantopoulos, S. (2013). Optimal Design in Three-Level Block Randomized Designs with two Levels of Nesting: An ANOVA Framework with Random Effects. *Educational and Psychological Measurement, 73(5),* 784-802

- Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effects of different forms of centering in hierarchical linear models. Multivariate Behavioral Research, 30, 1–21.
- Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. Educational and Psychological Measurement, 57, 800-812.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. Psychological Methods, 2, 173-185.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models.* Thousand Oaks, CA: Sage.
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trails. *Psychological Methods*, *5*, 199-213.

- Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi- experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Schneider, B., & McDonald, S.K. (2006). *Scale up in education: Ideas in principle.* Lanham, MD: Rowman & Littlefield.
- Turpin, R. S., & Sinacore, J. M. (1991). *Multisite evaluations*. San Francisco, CA: Jossey-Bass.Verma, V., & Lee, T. (1996). An analysis of sampling errors for demographic and Health surveys. *International Statistical Review*, *64*, 265-294.
- Weisberg, H. I., Hayden, V. C., & Pontes, V. P. (2009). Selection criteria and generalizability within a counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? *Clinical Trials, 6,* 109-118.