Basics of Experimental Design

Spyros Konstantopoulos
spyros@msu.edu

Michigan State University

Prepared for the IES Summer Research Training Institute 2026

Experimental Design

- "Experimental Design" encompasses:
  1. Strategies for organizing data collection
  2. Knowledge of data generating processes
  3. Data analysis procedures *linked* to those data collection strategies
- Suppose a researcher is interested in determining the effect of a treatment (e.g., school intervention) on an 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 treatment and groups)
- The effect is the change in the outcome of interest (e.g., increase math achievement from pre- to post-test) because of the intervention/treatment implemented

Experimental Design: Analysis

- Analysis of Variance (ANOVA) is a traditional analysis procedure used to analyze data from randomized experiments including Randomized Control Trials (RCTs)
- Other appropriate analytic procedures include:
  - Multiple linear regression
  - Multilevel or hierarchical linear models
  - Statistical estimation applied to aggregate data (classroom or school level data)
- All these procedures estimate the mean difference in an outcome between treatment and control groups
- Analytic procedures should *match* research hypotheses, the research design, and a priori power analysis

Why Do We Need Experimental Design?

- Aim to identify treatment effects in the presence of variation/differences of units and/or responses
- Variation/differences exists because:
  - Units (students, teachers, & schools) are not identical
  - Units respond in different ways to treatments
- We need experimental design to control differences (i.e., equate treatment and control groups on average at the beginning of the study) and then identify treatment effects on outcomes of interest (e.g., student achievement)
- It is viewed as the strongest design to identify what causes a change in an outcome of interest when threats to the internal validity of the study are minimized and randomization is kept intact

History

- The idea of controlling variation by creating similar groups through a research 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 (which are rich in vitamin C). 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 comparison group (standard treatment) according to the day they were admitted (i.e., every other day patients were assigned to different groups). Quasi-experiment

- 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 the trial of streptomycin for treating tuberculosis

- It was conducted by the British Medical Research Council 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
- Rothamsted agricultural station had a lot of observational data on crop yields and hoped a statistician could analyze it to find potential 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 (i.e., laid out the basic concept of randomization)
- RCTs are extensions of Fisher's pioneering work on experimental design

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 emphasis IES placed on rigorous research designs in education, and the availability of funding streams, there has been an abundance of RCTs
- IES has funded more than 350 RCTs since its inception

Randomized Experiment

- Experiment: deliberate interruption of an ongoing process to identify the effects of that interruption
- Randomized experiment: experiments that involve the creation of two (or more) groups, where participants are assigned randomly to these groups

Randomized Experiment

- Random assignment is a procedure that assigns units to treatment and control conditions based only on chance, where each unit has a nonzero probability of being assigned to a condition. Randomization is a key process for causal inference
- This random process of assignment to groups uses for example the toss of a fair coin or the table of random numbers or computer generated random numbers and assignment
- Because allocation to treatment and control groups is based solely on the luck of the draw the treatment and control groups are on average equivalent on all known and unknown variables at the beginning of the study (the baseline)

Randomized Experiment

- Because or randomization, the treatment and control groups are equivalent on average before the treatment starts and therefore we can compute the average outcome score of all inidividuals in the treatment group and then the average outcome score of all individuals in the control group and finally compute the mean difference in the two groups. This is an average treatment effect across all individuals in the treatment and control groups

Principles of Experimental Design

- Objective: Control variation 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
3. ***Statistical Adjustment*** ← Observational studies

Random Assignment

Controls variation of ***all*** characteristics:

- observables or non-observables
- known or unknown

  ⇨ Equates treatment and control groups *on* ***average*** on ***all*** characteristics at the baseline

- Differences in outcomes after the treatment has been applied 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 a control condition by chance (a random allocation mechanism)
- The treatment and control conditions are then alike. In particular, treatment and control groups are equivalent on average at the beginning of the study, and changes in outcomes are due to the treatment only. Reasonably large numbers are needed for random allocation to groups (works best in the long run)

Random Assignment

- Currently randomized experiments are used frequently in education

  ⇨ Strongest design for causal inference

- Notice that the assignment of units to treatment and control groups involves first randomization (the genesis of the unbiased random sequence) and second the unbiased (unaltered) implementation of randomization. The second component is very crucial in conducting infallible experiments

Random Assignment

- When using random assignment, we do not have to know a lot to use it effectively
- We simply conduct random assignment of sample units to treatment and control conditions. That is, the randomization aspect of the study is straightforward. The implementation of randomization needs careful monitoring of course to ensure the experiment is not compromised or broken
- It is good practice to measure important relevant covariates at the baseline of the experiment (e.g., in education it is crucial to measure prior achievement and SES) and include them in the analysis to achieve more precise estimation

Compliance and Non-Compliance

- Compliers are individuals who will take the treatment if assigned to the intervention group and will not take the treatment if assigned to the control group (i.e., participants that comply with the assignment according to randomization)
- There are three categories of non-compliers: (a) units that, regardless of random assignment, will not take the treatment (never takers); (b) units that, regardless of random assignment, will always take the treatment (always takers); and (c) units who defy random assignment and do the opposite of what their assignment suggests (defiers)

Compliance and Non-Compliance

- When individuals are not complying with randomization results in systematic ways, bias might be introduced in the treatment effect estimate. For example, suppose some students assigned randomly to the control group deliberately switch to the treatment group to receive the treatment (also called crossing over from one group to another). Alternatively, suppose some students in the treatment group intentionally decide not to take the treatment and switch to the control condition. Whenever switching between treatment and control groups is non-random, deliberate, the risk of treatment effect bias increases

Intention to Treat

- One analytic approach that can serve as a bulwark to bias due to non-compliance is the intention to treat analysis (ITT). The treatment effect is estimated according to the individuals' initial/original assignment to treatment or control groups through randomization, regardless of whether crossing over from one group to another took place (i.e., regardless of non-compliance)
- The rationale of the ITT analysis is that all participants who were part of the original sample of the RCT and were assigned via some random allocation mechanism to a treatment or a control group are included in the statistical analysis of post-test outcomes regardless of whether they actually complied with their initial random assignment or not

Intention to Treat

- The ITT analysis does not produce the treatment effect for compliers; it produces the treatment effect of the treatment offered through randomization. Intuitively, as non-compliance rates increase the ITT effect deviates from the anticipated treatment effect for compliers

Treatment on the Treated

- Analyses that examine the effect of “treatment on the treated” attempt to take into account whether, and often how much, of the treatment have participants received. This is about the effect of treatment actually received by participants (not offered)
- This effect could be biased because of selection. That is, if individuals who actually receive the treatment are systematically different (higher motivation, ability, SES, etc.) than those who did not receive it, the treatment effect is likely biased. Even controlling for important observed variables may not completely alleviate the selection issue (if unobservables are different between individuals who received and did not receive the treatment)

Instrumental Variables

- Instrumental Variables (IV) Estimation can be used to facilitate causal inference in this case
- Specifically, the IV procedure estimates the treatment effect for compliers. In experiments, a strong instrument is the initial random assignment to treatment or control groups. A two-stage approach can be applied to estimate the IV treatment effect for compliers. In the first stage, the binary variable that indicates whether a participant actually received the treatment (or not) is regressed on the binary variable of initial random assignment (randomization results)
- Covariates can potentially be included in the first-stage regression. The regression estimate of the initial random assignment binary variable in the first-stage regression captures the association between initial random assignment and actual receipt of the treatment and represents the degree of compliance

Instrumental Variables

- The first stage keeps the component of the binary variable that represents treatment actually received that is linked with the original random assignment process, and purges all non-compliance processes. The fitted/predicted values of the first-stage regression are used now as the treatment variable that predicts a dependent variable of interest (e.g., math achievement) in the second-stage outcome variable regression. The second stage regression may also include relevant covariates. This IV analysis offers a causal estimate of the treatment effect for compliers

Matching

- *Known* sources of variation may be eliminated by matching (i.e., matching is conducted using measured/observed relevant covariates)

- For example, eliminate district, school, or classroom effects before comparing students, that is, compare students in similar classrooms, schools or districts

- Matching can take place in the design phase of a study or in the data analysis stage. For example, propensity score methods is one post hoc statistical method that creates similar groups using observed covariates to estimate a treatment effect. Matching including propensity score methods is based solely on measured, observed sources of variation

Matching

- Matching methods including propensity score methods "mimic" random assignment (i.e., aim to balance baseline variables in treatment and comparison groups). Under the assumption that all relevant baseline covariates have been measured and used and there is no omitted variable bias, matching could be as good as random assignment. However, that is a strong assumption, the best-case scenario. In principle, it is always possible that an unmeasured variable could impact causal inference in matching

- Matching can be used before random assignment takes place. That is, create homogeneous groups of units and then assign them randomly to treatment and control conditions. One example is the matched pairs design

Matching

- Matching can only be performed on **known and observable** characteristics that have been measured
- Perfect matching is not always possible
- It is critical to measure the right/relevant variables that will minimize variation and create more homogeneous groups (e.g., prior achievement, SES)
- May limit 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 (i.e., improves statistical power)

Statistical Adjustment

- A form of post-hoc pseudo-matching that also mimics random assignment
- Uses statistical associations between outcomes and controls/covariates to simulate matching
- Reduces variation of outcomes in regression and ANCOVA
- Controlling for covariates increases the precision of the 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 and included in the model. If the model is specified correctly, the treatment effect could be unbiased. But that is a strong assumption

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
- That is, the randomization aspect of the study is straightforward. The implementation of randomization needs careful monitoring
- Nonetheless, it is good practice to measure important relevant covariates at the baseline of the experiment (e.g., prior achievement, SES) and include them in the analysis to achieve more precise estimation

Principles of Experimental Design

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

- Some thorough thinking, when designing a quasi-experiment or an observational study, is necessary in order to measure all relevant variables and include them in the analyses to produce equivalent groups and reduce bias

Principles of Experimental Design

- Random assignment per se may not be as efficient as matching or statistical control (i.e., may require larger sample sizes for the same power) because it does not reduce, it controls variation

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

- Including covariates in a regression model would reduce variation in the outcome and result in a more precise estimation (higher statistical power of the test)

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 by the researchers (e.g., reduce class size) and can cause an event we wish to measure
  - ***Blocks*** (e.g., classrooms, schools) and ***covariates*** (e.g., gender, race) are independent variables that cannot be manipulated by the researchers
- 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 are classes created for the purposes of forming homogeneous groups

- Blocks can be naturally formed groups (e.g., regions, states, cities, school districts, schools, grades, classrooms)

- Blocks can be known variables/factors (e.g., age, ability, health status)

- Blocks reduce variation (similar to matching and statistical control)

Blocks

- In education
  - We can assign randomly schools to treatment conditions within school districts (the blocks)
  - Or we can assign randomly students or classrooms to treatment conditions within schools (the blocks)

- Or individuals with similar age can be grouped in homogeneous blocks and then random assignment to conditions may take place
- Block effects should be taken into account in a priori power computations and in statistical analyses. Blocks could be random or fixed effects

Basic Ideas of Design: Nesting & Crossing

- *Example:* schools are randomly assigned to treatment conditions (treatment is at a higher level than schools)
  - ⇨ ***schools are then nested within each treatment condition***

Schools

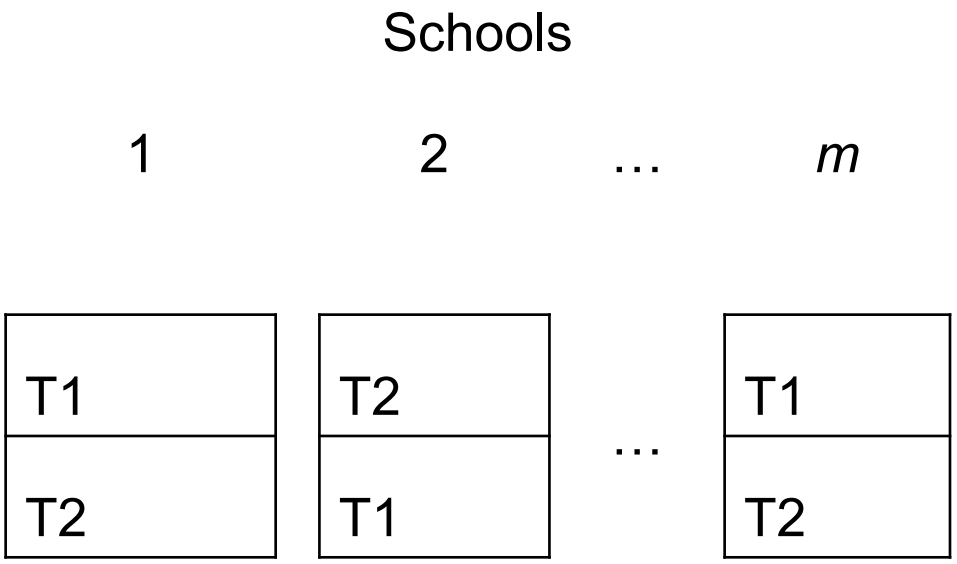
1, 2, … ,m m + 1, … , 2m

Treatments

1 2

Basic Ideas of Design: Nesting & Crossing

- *Example:* classrooms or students are randomly assigned to treatment or control conditions within schools (treatment is at a lower level than schools/blocks)



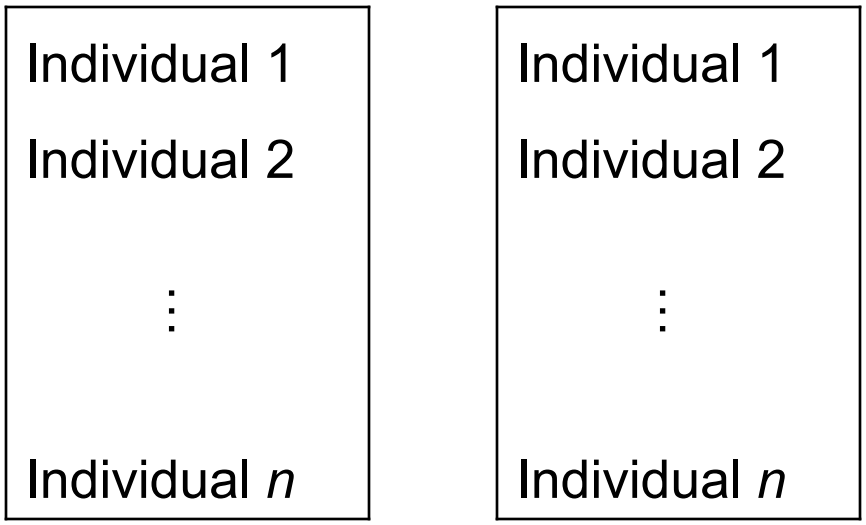
⇨ ***treatments are then crossed with schools (blocks)***

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 (classrooms and students)
- Randomized Block Design
  - For example, students are assigned randomly to treatment and control conditions *within* schools or grades (the blocks)
  - Larger units such as classrooms can also be assigned randomly to treatments *within* schools or grades (the blocks)
  - This design is also known as a multisite design where blocks are the sites

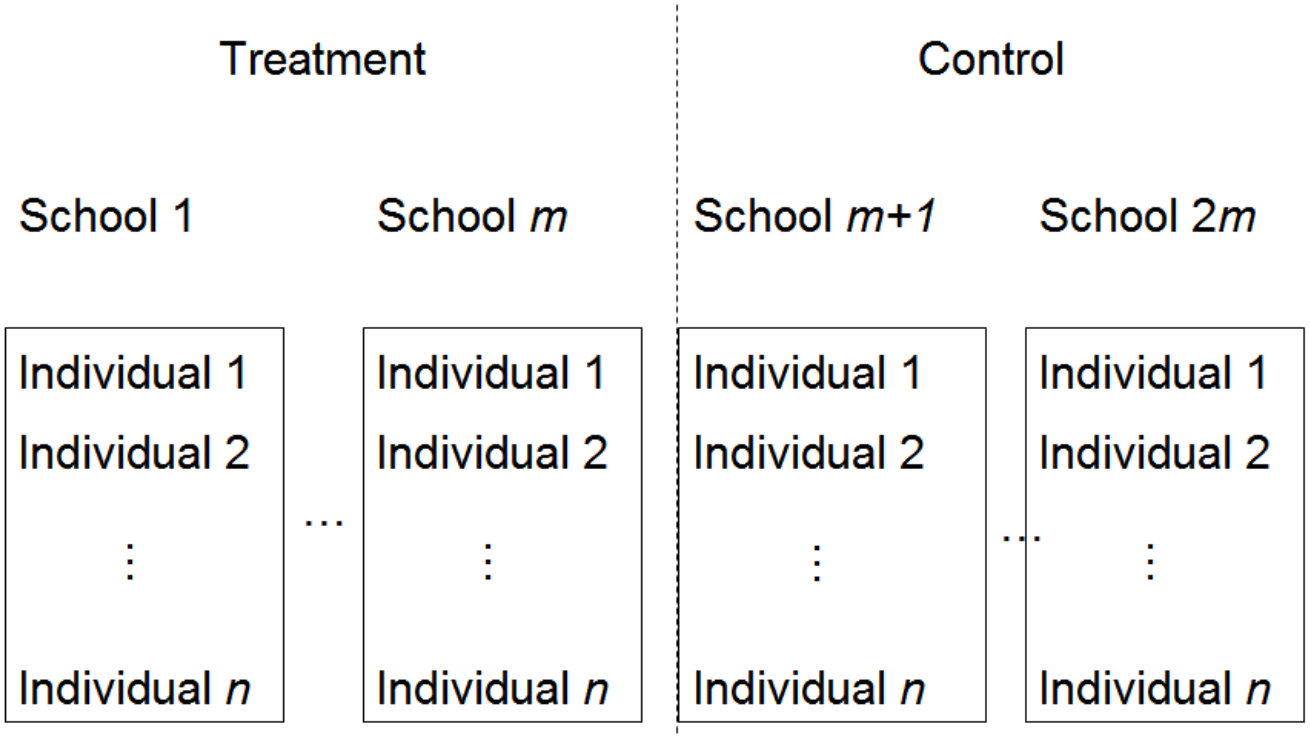
Completely Randomized Design

- Individuals are randomly assigned to one of two treatments:



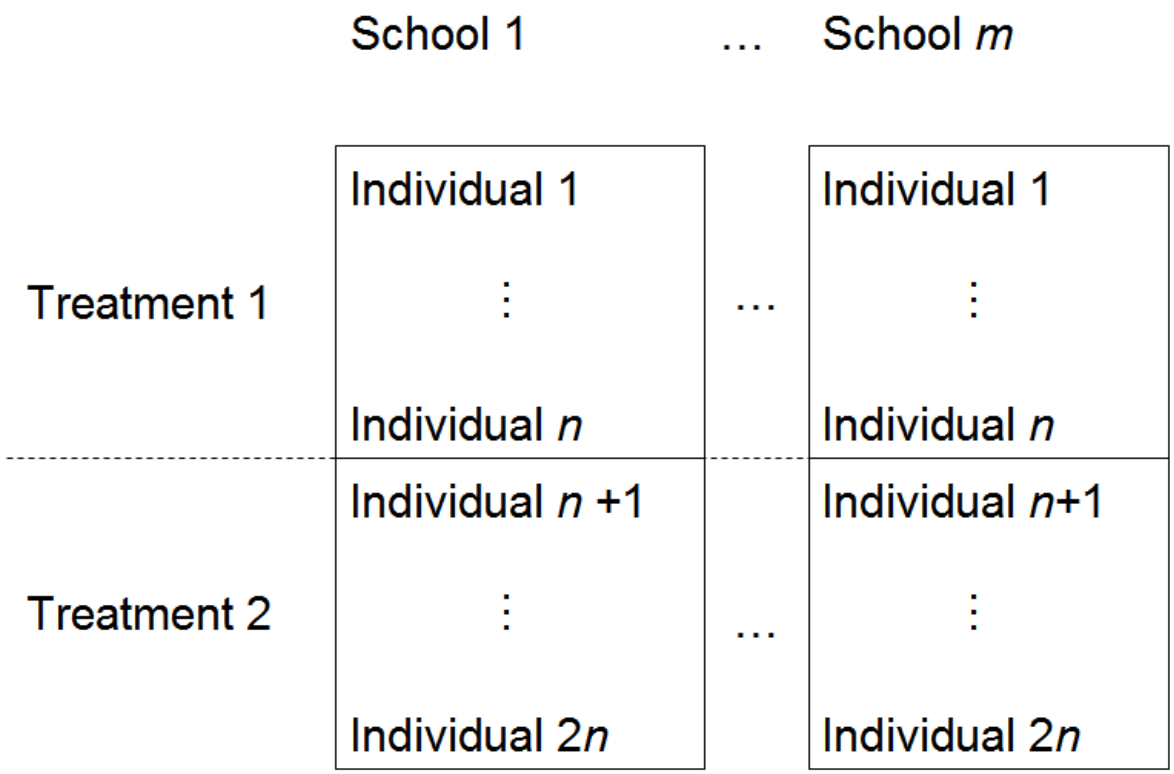
Cluster or Group Randomized Design

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



Randomized Block Design

- Individuals are randomly assigned to one of two treatments within their school (the block or site):



Randomization Procedures

- Could use a table of random numbers. *Be sure to pick an arbitrary starting point each time*
- Could use random number generators in statistical software packages. *Be sure the seed value varies each time*
- *Lottery (random picks)*
- *Flipping a fair coin*

Post Hoc Test to Check Randomization

- It is common practice to check whether random assignment was successful using observed variables at baseline (i.e., check baseline equivalence of measured 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 procedure that can identify variables where random assignment did not work as expected by design (i.e., systematic mean differences of baseline covariates in the treatment)

Post Hoc Test to Check Randomization

- This procedure cannot discredit randomization per se (e.g., a mean difference may be observed by chance). However, when there is systematic evidence about mean differences, this may indicate that the implementation of random assignment might have been flawed

- Mean differences should not be significant (but that depends on the sample size). More importantly, the magnitude of the mean difference should not exceed 0.25 standard deviations (according to WWC)

- Regression or ANCOVA can be used to check imbalance in baseline covariates. The model should include all relevant measured covariates identified and used in the outcome variables regressions (except the covariate in question each time)

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/Final_WWC-HandbookVer5_0-0-508.pdf

A Useful Framework for Clustering: 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

- Education populations have nested structures (multiple levels, units of different sizes – students, classes, schools, districts)
  - Students at the first level, classrooms at the second level, schools at the third level, school districts at the fourth level

Sampling Models in Educational Research

- Survey research in education often exploits this multilevel 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 within schools, then students within classrooms

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

Variance of the Mean of Clustered Samples: Two Levels

- Suppose we have *n* level-1 units within each *level-2* unit and *m* level-2 units overall
- Assume a sample size $N = mn$ and a total population variance σ_T^2 defined as

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

where $\tau^2 =$ Level−2 variance, $\sigma^2 =$ Level−1 variance

- If the sampling strategy had been *simple* (e.g., simple random sampling of students across schools) the variance of the mean would be: $\frac{(\sigma_T^2)}{mn}$

Variance of the Mean of Clustered Sample: Two Levels

- When cluster sampling is also involved however, the variance of the mean is

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

Variance of the Mean of Clustered Sample: Two Levels

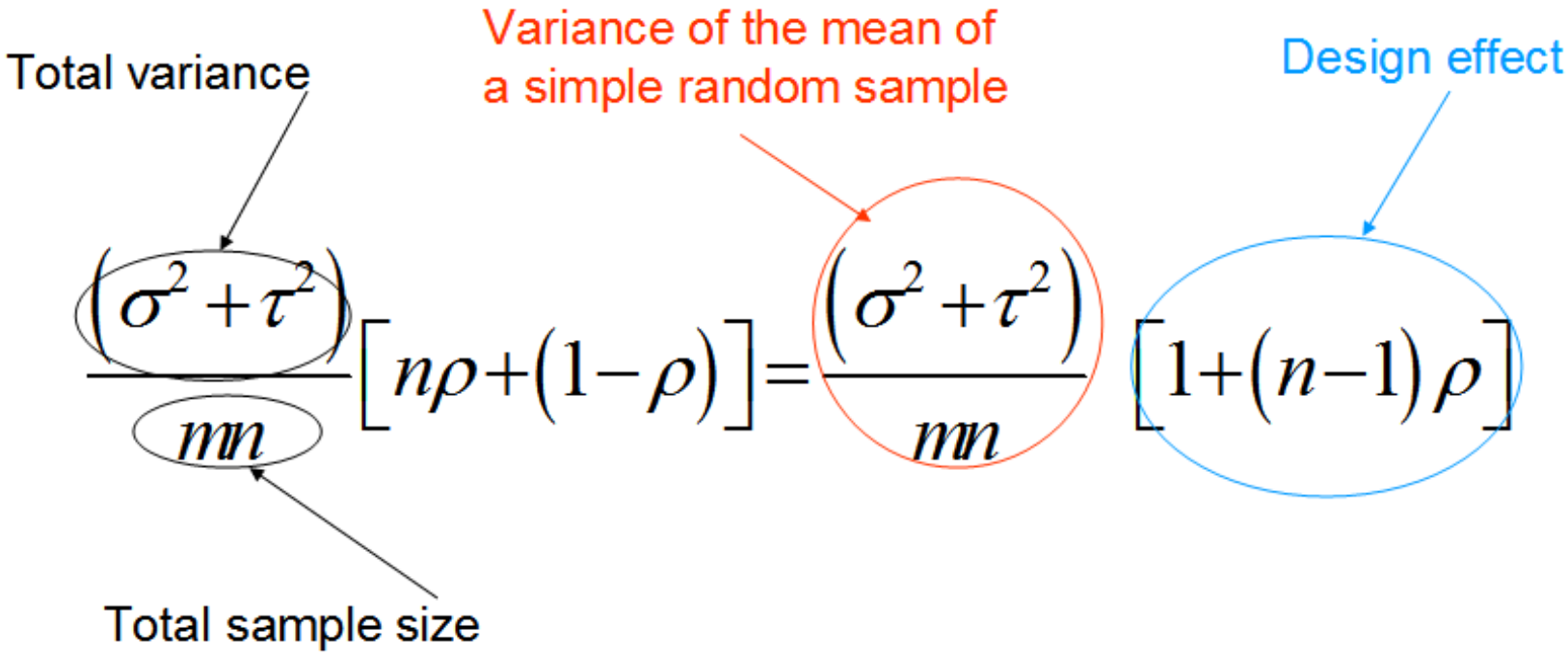
- The intraclass correlation coefficient (ICC), ρ, defined as the proportion of the second level variance to the total variance represents clustering
- If we write $\rho = \tau^2/(\sigma^2 + \tau^2)$, the variance of the mean becomes:

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

- 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 clustering

Variance of the Mean of Clustered Sample: Two Levels

- Specifically:



- Or $\frac{(\sigma_T^2)}{mn}[1+(n-1)\rho]$

Variance of the Mean of Clustered Sample: Three Levels

- Suppose now we have *n* students (level-1 units) in *p* classes (level-2 units) in each of *m* schools (level-3 units)
- Assume a sample size $N = mpn$, and a total population variance $\sigma_T^2 = \sigma^2 + \tau^2 + \omega^2$ where σ^2, τ^2 and ω^2 are the first, second and third level variances respectively

If the sampling strategy had been *simple* (e.g., simple random sampling of students across classrooms and schools) then the variance of the mean would be: $\dfrac{\left(\sigma_T^2\right)}{mpn}$

Variance of the Mean of Clustered Sample: Three Levels

- When cluster sampling is also involved however in the first and second stages of sampling (e.g., cluster sampling of schools and then cluster sampling of classrooms) two ICCs, ρ_3 *(third or school level) and* ρ_2 *(second or classroom level)* can be defined to capture clustering at both levels
- 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 clustering at the second and third levels

Design Effect

- In two-stage sampling the design effect depends on *n* and the ICC. When both are small the design effect should be close to one. When both are large the design effect could be much larger than one. In practice the square root of the design effect can be used to correct standard errors of regression estimates produced from typical regression models (when cluster sampling is assumed). Specifically, one can multiply the square root of the design effect with the standard error of the regression estimate produced from a typical regression model

Design Effect

- In three-stage sampling the design effect depends on *n, p* and the ICC's. When *n, p* and the ICC's are small the design effect should be close to one. When *n, p* and the ICC's are large the design effect could be much larger than one. Again, to correct the standard errors for cluster sampling at two levels (classes, schools), one can multiply the square root of the design effect with the standard error of the regression estimate produced from a typical regression model

The ICC

- The ICC is defined as a variance ratio. It is the proportion of total variance in the dependent variable that is attributable to clusters (the larger units)
- For example, suppose students are nested within schools and the outcome variable is at the student level (achievement scores). Then, the ICC is the proportion of the variance in achievement scores attributed to schools and indicates school differences. That is, the ICC is the ratio of the between-cluster variance to the total variance in the dependent variable

The ICC

- If the total variance is 1 and the between cluster variance is 0.2, the ICC = 0.2. This means 20% of the total variance in the outcome variable is attributed to the variance between clusters and 80% of the total variance is attributed to the variance within clusters. The ICC ranges from 0 to 1. Zero indicates no between cluster variance (no clustering) and 1 indicates no within cluster variance. Increases in ICC indicate differences between clusters (more heterogeneity between clusters and more homogeneity within clusters). Smaller ICCs indicate more homogeneity between clusters (reduced differences between clusters) and more heterogeneity within clusters

Variance of the Mean of Clustered Sample

- The sampling model used dictates the variance structure and estimation

- Variance impacts:
  - Precision of the treatment effect estimates (standard errors)
  - Statistical power (inverse relationship)

Nesting in Multilevel Models

- Nesting is a similar notion to clustering. That is, a related useful framework capitalizes on data dependencies because of the nesting of lower-level units within higher-level units
- For example, students grouped in the same classroom are more alike than students grouped in a different classroom. It is possible that the outcomes of students in the same classroom covary or are correlated to some degree and this data dependency (covariance or correlation) should be taken into account when estimating the variance of the mean

Inferential Population and Inference Models

- The inference model has implications for analyses 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:* The *i*nference in **conditional** is different than that in **unconditional** models

Inferential Population and Inference Models

- In a conditional inference, we are estimating the mean treatment effect in the observed schools in the sample
- 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?*

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

Specifying Analyses

Know the design

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

- ***Treatment effects*** should also be fixed effects unless the design permits the treatment to be random such as, a randomized block design. For instance, a classroom intervention may have different effects from school to school

Applications to Experimental Design

- We will look in detail at the two most widely used experimental designs in large-scale education research
  - Cluster randomized designs
  - Randomized block designs

Cluster Randomized Design

The Cluster Randomized Design

- Clusters are naturally occurring groups (large units) within which smaller units are grouped

- In education, schools are naturally occurring clusters. Teachers, classrooms and students are grouped within schools (the clusters). School districts are larger clusters than schools and classrooms are smaller clusters than schools

- Assignment to groups is made to whole clusters (e.g., schools) randomly. Clusters are nested within treatment and control conditions

The Cluster Randomized Design

- We are typically interested in comparing means of two different conditions (a treatment and a control group)
- Assignment to groups is made to whole clusters (e.g., schools) randomly. Clusters are nested within treatment and control conditions
- Assign *2m* schools with *n* students in each school (typically assume balanced design)
- There are *m* schools in each treatment condition
- Assign *all* students in each school to the *same* treatment condition

The Cluster Randomized Design

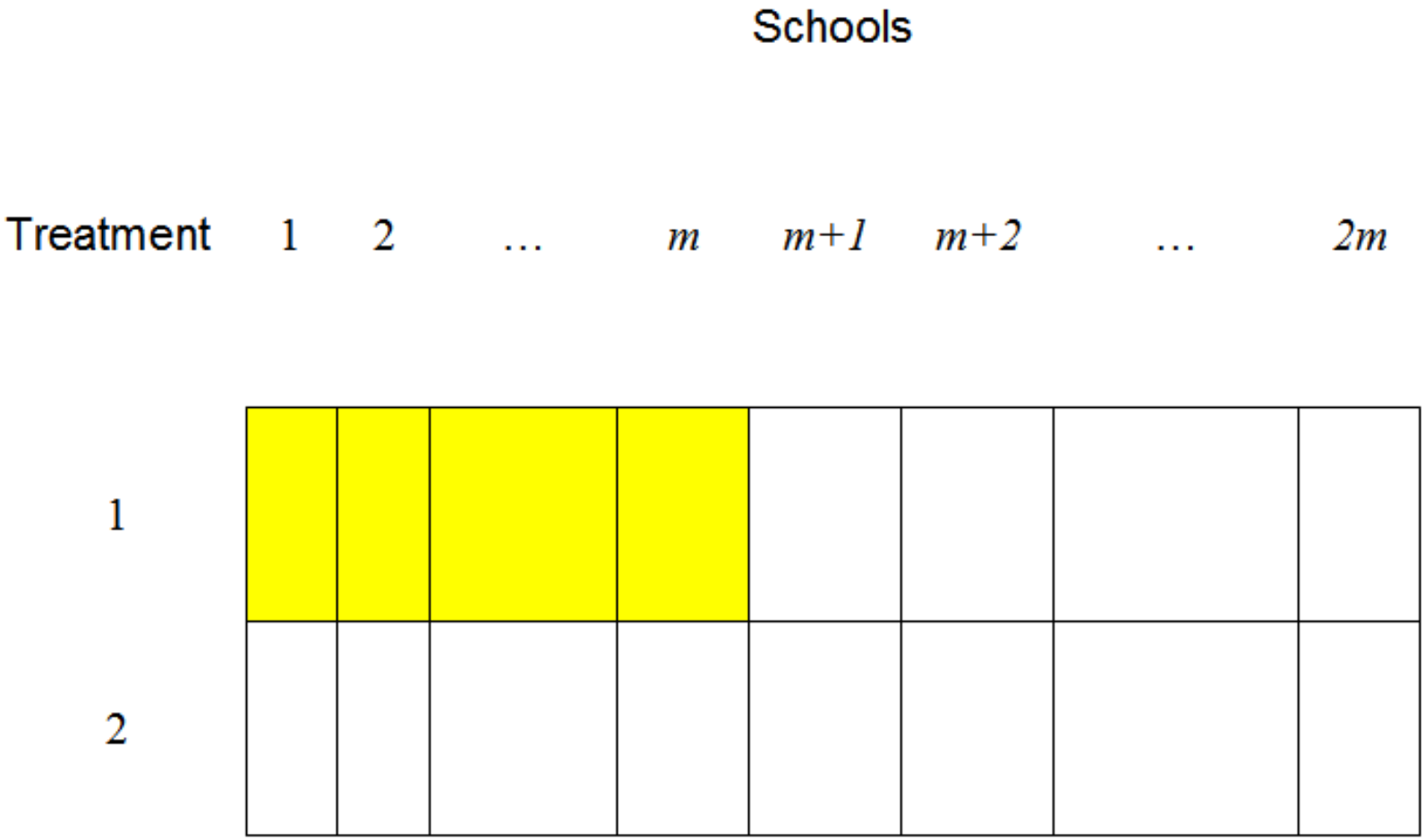
- Diagram of the Experiment:

Schools

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

The Cluster Randomized Design

- Treatment 1 Schools:

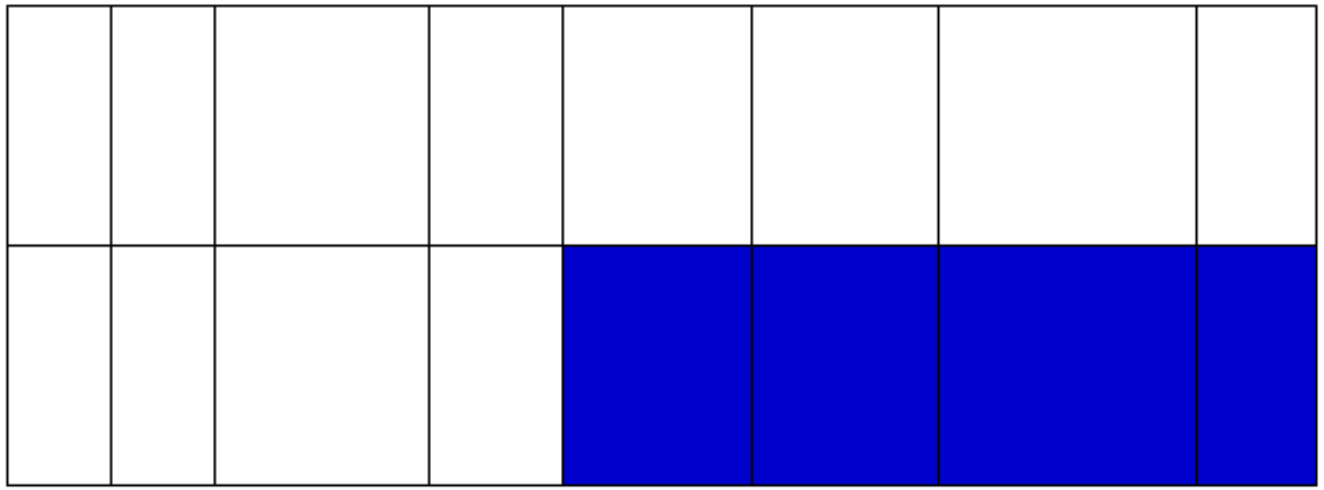


The Cluster Randomized Design

- 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 Specific Equations

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 and *T* is the treatment

Two-Level CRT Design with Covariates: Conceptual Multilevel Model – Level Specific Equations

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):

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

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

where *X* is a level-1 covariate (e.g., student SES), *W* is a level-2 covariate (e.g., school size) and *T* is the treatment. The first level covariate effect is modeled as fixed in the second level. Also, the first and second level variances are now residual variances (subscript A indicates adjustment). The first level intercept is random at the second level

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

The previous multilevel equations can be written as single level regression equations (***mixed effects models***) with a complicated error term

No covariates:

Error

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

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 Specific Equations

Level 1 (individual level):

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

Level 2 (classroom level):

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

Level 3 (school level):

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

Two ICCs:

Third level: $\rho_3 = \omega^2 \Big/ \left(\sigma^2 + \tau^2 + \omega^2\right) = \omega^2 / \sigma_T^2$ (School)

Second level: $\rho_2 = \tau^2 \Big/ \left(\sigma^2 + \tau^2 + \omega^2\right) = \tau^2 / \sigma_T^2$ (Classroom)

Three-Level CRT Design with Covariates: Conceptual Multilevel Model - Level Specific Equations

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 with Covariates: Conceptual Multilevel Model - Level Specific Equations

X is a level-1 covariate (e.g., student SES), *Z* is a level-2 covariate (e.g., class size), *W* is a level-3 covariate (e.g., school sector) and *T* is the treatment. The first level covariate effect is modeled as fixed at the second and third levels. Similarly, the second level covariate is fixed at the third level. All three variances are now residual because of covariates (subscript A indicates adjustment). The first and second level intercepts are random at the second and third levels. In cluster designs the treatment is always at the top level

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

The previous multilevel equations can be written as single level regression equations (***mixed effects models***) with a more complicated error term

No covariates:

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

Complicated 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 (nesting or clustering)
- Otherwise, the standard errors of the regression estimates and statistical tests are incorrect
- The standard errors of treatment effect estimates are typically smaller when clustering is ignored, especially in higher levels (cluster levels)
- This results in a higher value of a *t*-test and a higher probability of rejecting the null hypothesis when it is true (committing a Type I error)

Standard Errors of Regression Estimates and Clustering

- There are different ways of adjusting standard errors for clustering
  - Conduct the analysis using multilevel models (e.g., SAS proc MIXED, SPSS linear mixed models, HLM, Mlwin, 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 regression estimate
    - Use clustered standard errors (e.g., Stata) that adjust for clustering (typically clustered robust standard errors are computed that account for clustering and heteroscedasticity)

Randomized Block (or Multisite) Design

The Randomized Block Design (RBD)

- We wish to compare the means between 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 variation is taken into account). The block is a cluster or sub-cluster and thus cluster sampling is assumed at the top level (and the middle level in three-level cases)

Two-Level RBD

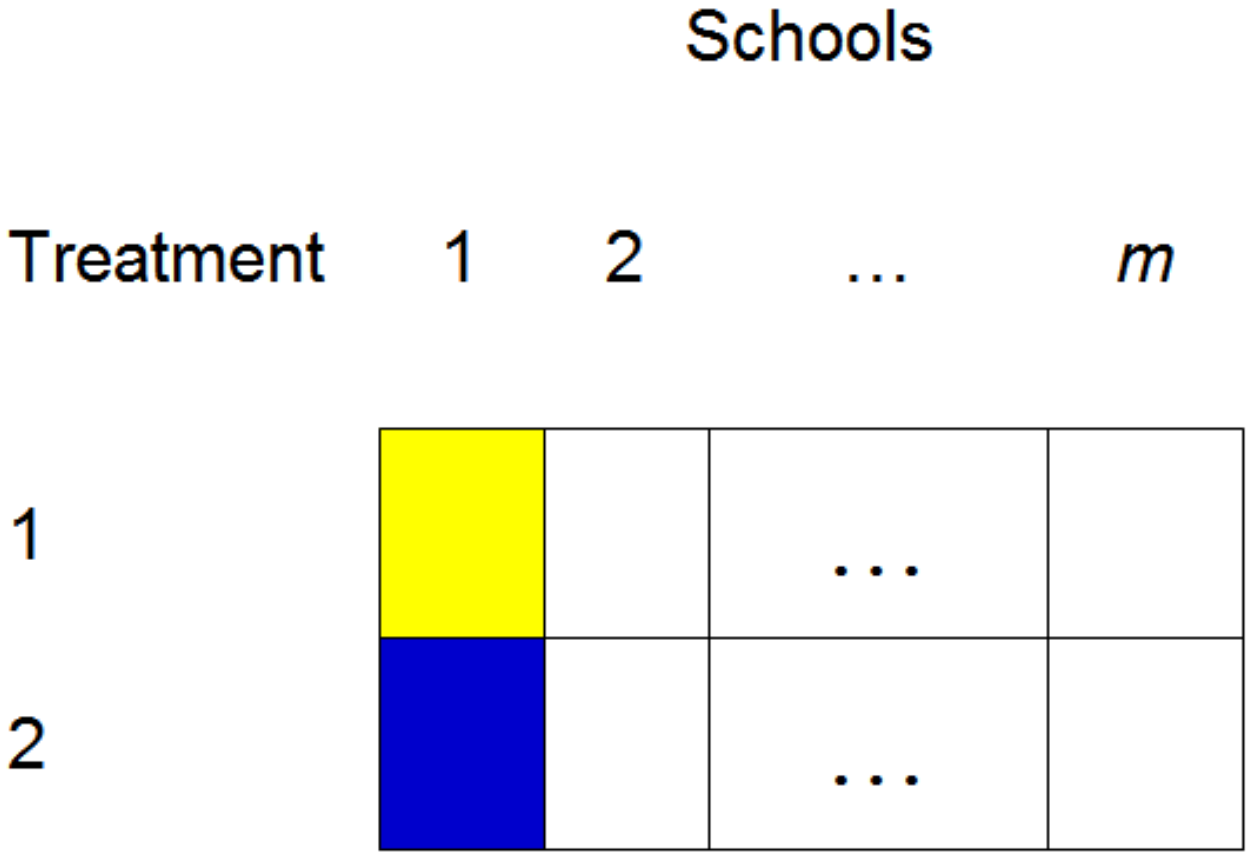
- Diagram of the Experiment:

Schools

Treatment	1	2	…	*m*
1			…	
2			…	

Two-Level RBD

- Diagram of the Experiment:



- *m* smaller scale experiments are overall conducted (one per block)

Two-Level RBD: Conceptual Multilevel Framework (MLF) – Level Specific Equations

- 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

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

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

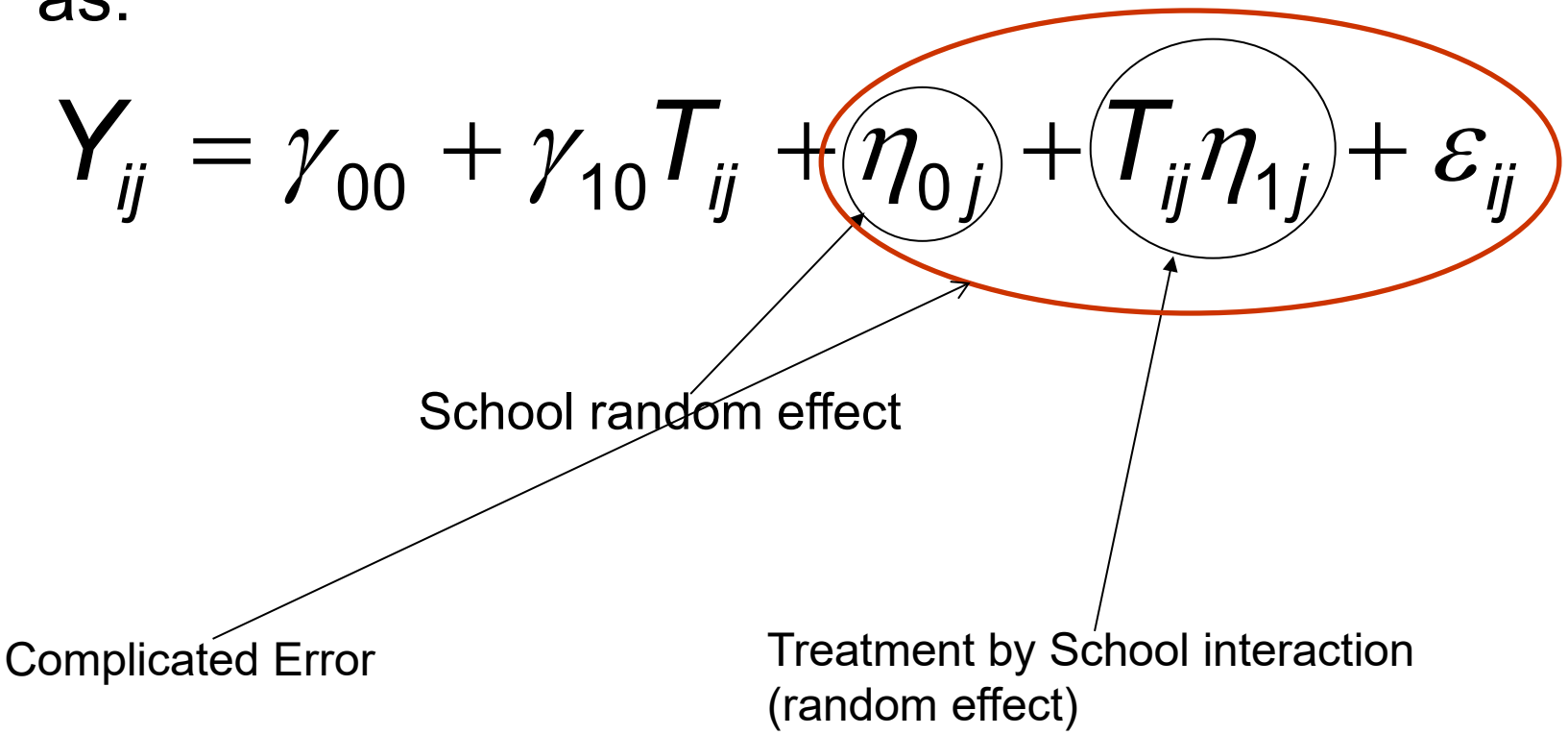
Treatment by School interaction (random effect)

Subscript *T* indicates treatment. The first level treatment effect is modeled as random at the second level

Two-Level RBD: Single Level Equation

- The previous two-level model can be expressed in a single level equation (mixed effects model) as:

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



Two-Level RBD with Covariates: Conceptual MLF - Level Specific Equations

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):

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

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

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

where *T* is the treatment, *X* is a level-1 covariate (e.g., student SES) and *W* is a level-2 covariate (e.g., school size). The first level covariate effect is fixed at the second level. In block designs the treatment is always below the top level

Two-Level RBD with Covariates: Single Level Equation

The single level equation (mixed effects model) is:

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

Complicated error

The term $T_{ij}\eta_{1j}$ is a cross-level random effects interaction

Fixed and Random Effects

- Should blocks be fixed or random?
- Fixed Effects
  - If the inference focuses on the blocks (e.g., schools) in the sample, then the blocks can be treated as fixed effects (e.g., include a set of school dummies in the regression model). In this case the two-level model collapses in a single-level regression model
  - Specifically, a model with one fixed level-1 covariate is

$$Y_i = \gamma_0 + \gamma_1 T_i + \gamma_2 X_i + \boldsymbol{SC}_j \boldsymbol{\Gamma}_3 + T_i \boldsymbol{SC}_j \boldsymbol{\Gamma}_4 + \varepsilon_i$$

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

Three-Level RBD: Conceptual MLF = Level Specific Equations

Treatment is at the first level. 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 (classroom 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):

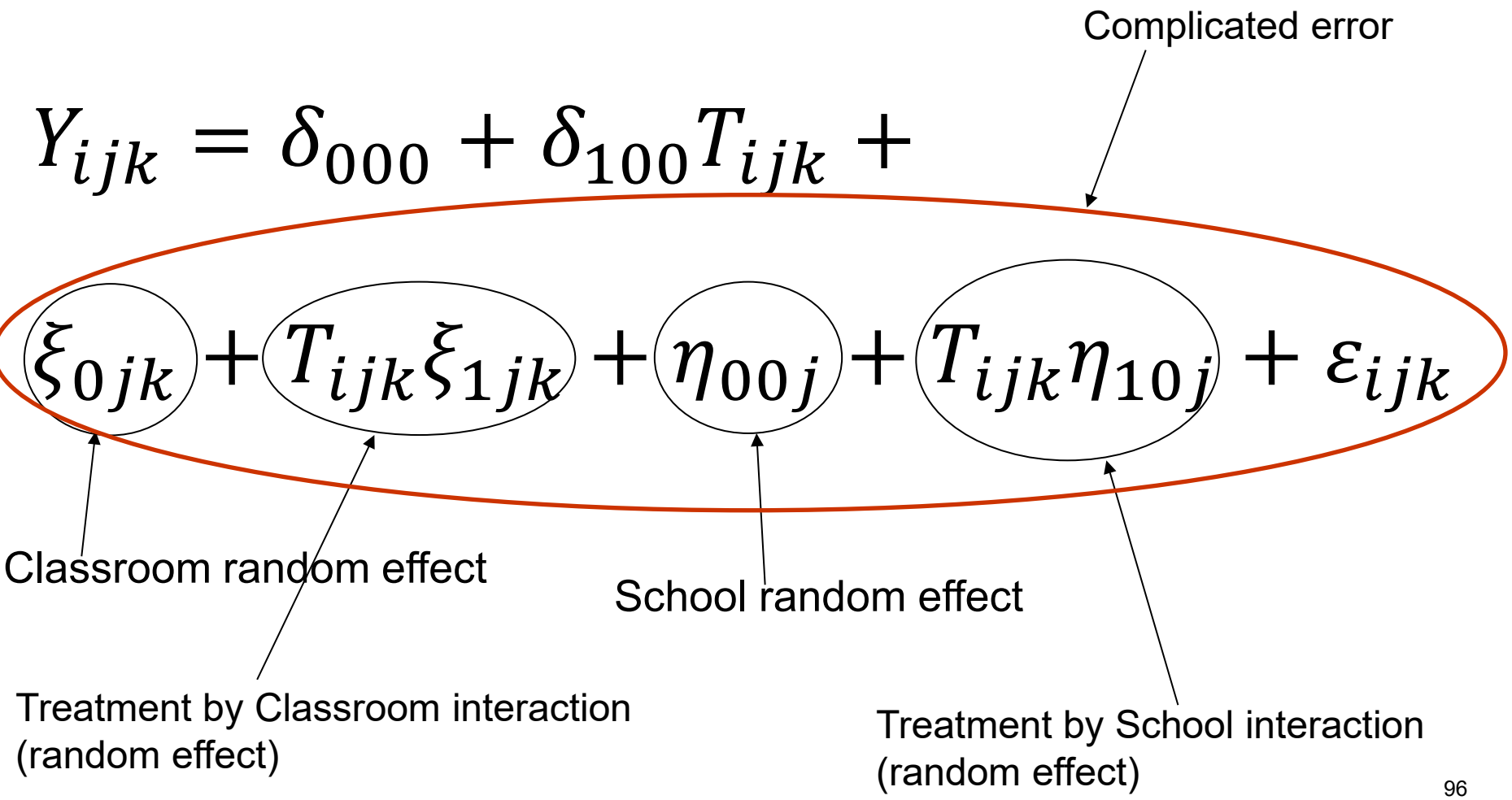
$$\gamma_{00k} = \delta_{000} + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega^2\right)$$

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

Subscript *T* indicates treatment. The treatment effect varies across level-2 and level-3 units

Three-Level RBD – Single Level Equation

- The single level equation (mixed effects model) is:



Three-Level RBD with Covariates: Conceptual MLF – Level Specific Equations

Treatment is at the first level, covariates included (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 (classroom 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):

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

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

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

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

Three-Level RBD with Covariates: Conceptual MLF – Level Specific Equations

X is a level-1 covariate (e.g., student SES), *Z* is a level-2 covariate (e.g., class size), *W* is a level-3 covariate (e.g., school sector) and *T* is the treatment. The first level covariate effect is modeled as fixed at the second and third levels. Similarly, the second level covariate is fixed at the third level. The treatment effect varies across level-2 and level-3 units. In block designs the treatment is always below the top level

Three-Level RBD – Single Level Equation

- The single level equation (mixed effects model) is:

$$Y_{ijk} = \delta_{000} + \delta_{100}T_{ijk} + \delta_{200}X_{ijk} + \delta_{010}Z_{jk} + \delta_{001}W_k +$$

$$\xi_{0jk} + T_{ijk}\xi_{1jk} + \eta_{00j} + T_{ijk}\eta_{10j} + \varepsilon_{ijk}$$

Classroom random effect

School random effect

Complicated error

Treatment by Classroom interaction (random effect)

Treatment by School interaction (random effect)

Fixed Effects

- Fixed Effects
  - If the inference focuses on the blocks in the sample, then the blocks are treated as fixed effects (e.g., a set of school dummies is included in the model at the second level). The third level can be eliminated, and the model represents then a two-level RBD with blocks included as covariates at the second level
  - The model becomes:

$$Y_{ij} = \delta_0 + \delta_1 T_{ij} + \delta_2 X_{ij} + \delta_3 Z_j + \boldsymbol{SC}_j \boldsymbol{\Delta}_4 + T_{ij} \boldsymbol{SC}_\mathbf{j} \boldsymbol{\Delta}_5 + \xi_{0j} + T_{ij} \xi_{1j} + \varepsilon_{ij}$$

where ***SC*** represent block fixed effects (e.g., a set of school dummies)

Three-Level RBD: Conceptual MLF – Level Specific Equations

Treatment is at the second level. No Covariates

Level 1 (individual level):

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

Level 2 (classroom 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):

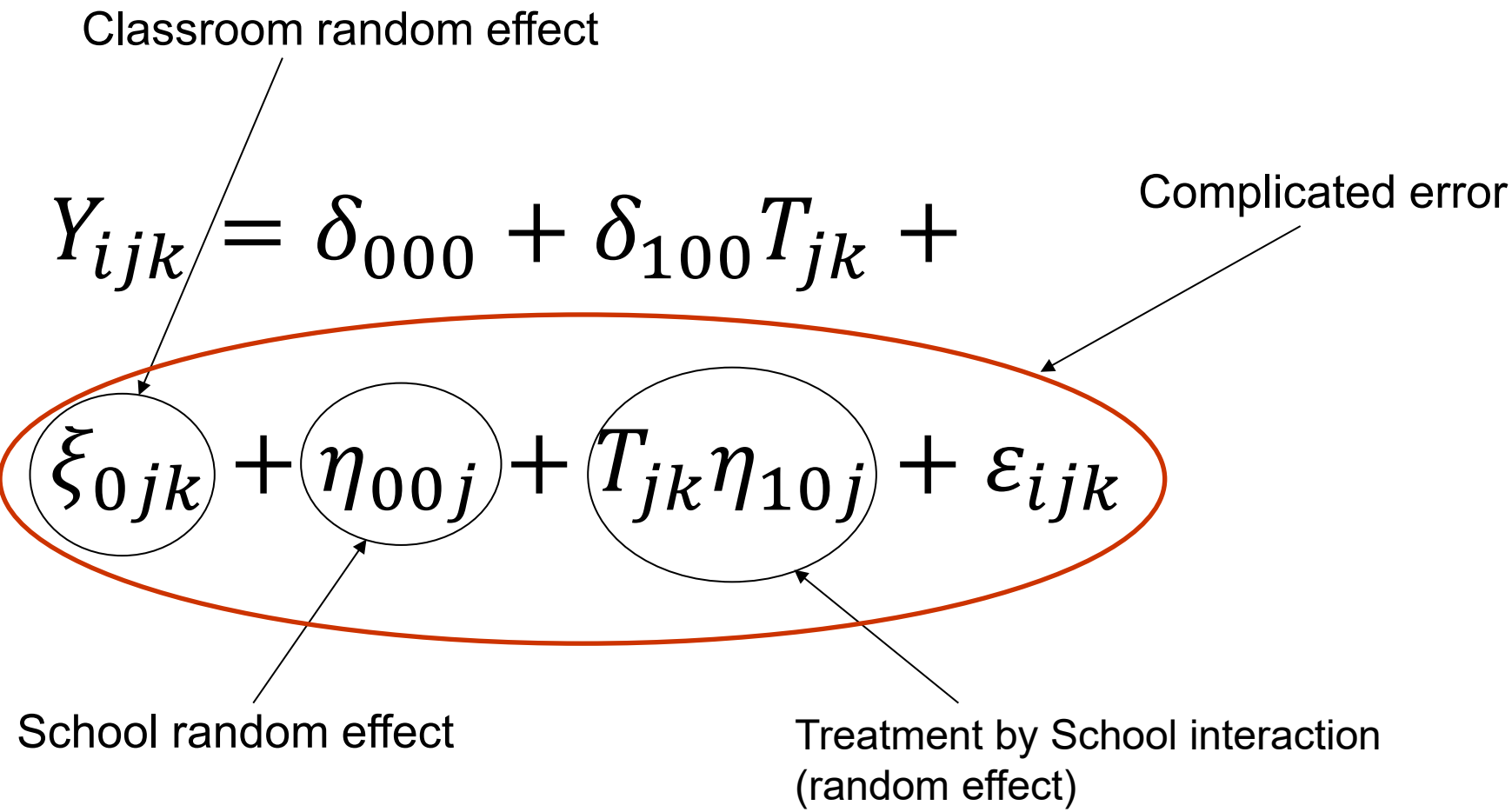
$$\gamma_{00k} = \delta_{000} + \eta_{00k} \qquad \eta_{00k} \sim N\left(0, \omega^2\right)$$

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

Subscript *T* indicates treatment. The treatment effect varies across level-3 units

Three-Level RBD: Single Level Equation

- Without covariates the mixed model is:



Three-Level RBD: Conceptual MLF – Level Specific Equations

Treatment is at the second level (covariates included)

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 (classroom 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):

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

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

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

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

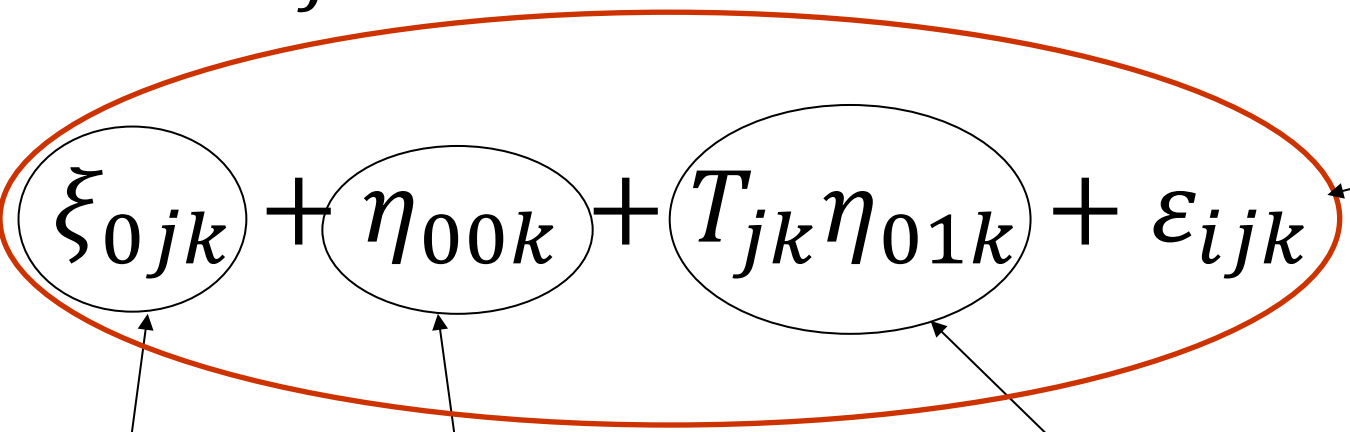
Three-Level RBD with Covariates: Conceptual MLF – Level Specific Equations

X is a level-1 covariate (e.g., student SES), *Z* is a level-2 covariate (e.g., class size), *W* is a level-3 covariate (e.g., school sector) and *T* is the treatment. The first level covariate effect is modeled as fixed at the second and third levels. Similarly, the second level covariate is fixed at the third level. The treatment effect varies across level-3 units. In block designs the treatment is always below the top level

Three-Level RBD: Single Level Equation

- With covariates the mixed model is:

$$Y_{ijk} = \delta_{000} + \delta_{010}T_{jk} + \delta_{100}X_{ijk} + \delta_{020}Z_{jk} + \delta_{001}W_k +$$



Fixed Effects

- Fixed Effects
  - If the inference focuses on the blocks in the sample, then the blocks are treated as fixed effects (e.g., a set of school dummies is included in the model at the second level). The third level can be eliminated, and the model represents then a two-level cluster design with blocks included as covariates at the second level
  - The model becomes:

$$Y_{ij} = \delta_0 + \delta_1 T_j + \delta_2 X_{ij} + \delta_3 Z_j + \boldsymbol{SC}_j \boldsymbol{\Delta}_4 + T_j \boldsymbol{SC}_j \boldsymbol{\Delta}_5 + \xi_{0j} + \varepsilon_{ij}$$

where $\boldsymbol{SC}$ represent block fixed effects (e.g., a set of school dummies)

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
- The mean of the new (centered) variable is zero

Centering

- Centering changes the value and the meaning of the intercept. In simple regression with a centered predictor the intercept is 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 nested within schools), there are three kinds of centering:
  - Grand-mean centering of a level-1 (student) predictor (using the overall mean of the predictor)
  - Group-mean centering of a level-1 (student) predictor (using the level-2 unit means of the predictor)
  - Grand-mean centering of a level-2 (school) predictor (using the overall mean of the predictor)
- Grand mean centering means subtracting the grand (overall) mean
- Group mean centering means subtracting the group or level-2 unit mean
- These two centering methods affect the interpretation of the cluster specific intercept

Grand-Mean Centering

- Grand-mean centering changes the meaning of the intercept in the j^{th} cluster (school)
- The intercept is now the mean outcome in the cluster (school) minus an adjustment due to the student predictors
- With Grand-Mean Centering:
  - Level-1 predictors can explain the level-2 variance
  - Level-1 predictors are not independent of the level-2 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} cluster (school)
- The cluster intercept is now the mean outcome in the cluster (school) *not* adjusted by student predictors
- With Group-Mean Centering:
  - Level-1 predictors *cannot* explain the level-2 variance (only the level-1 variance)
  - Level-1 predictors are *independent* of the level-2 predictors (no conditional effects across levels)
  - To reduce the level-2 variance one can create and use aggregate variables (of level-1 variables) at the second level
  - Level-1 effects are adjusted for level-2 (between cluster) differences (e.g., school effects)
- **Centering changes the precision of all estimates**

Centering

- Group-mean centering of level-1 predictors can be used in block designs to take into account potential differences among blocks (e.g., differences between schools or school effects)

- No centering or grand-mean centering can be used in cluster designs

- As in typical regression centering predictors is not a requirement

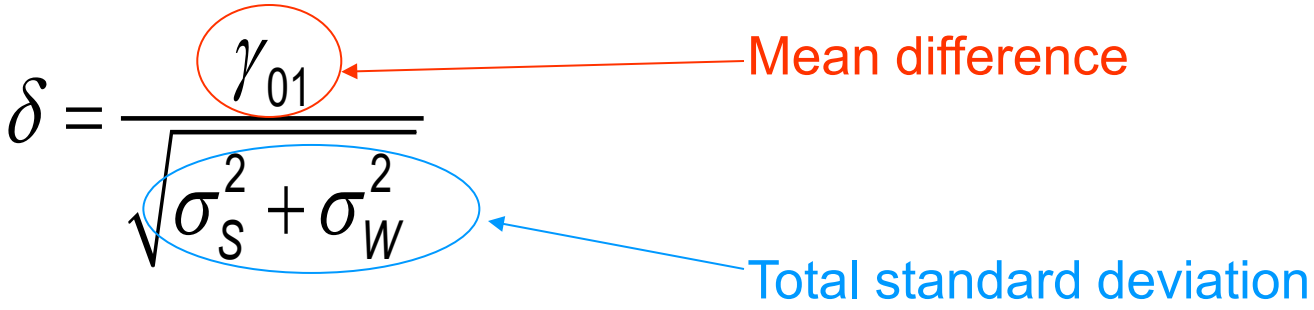
Effect Sizes

Effect Sizes

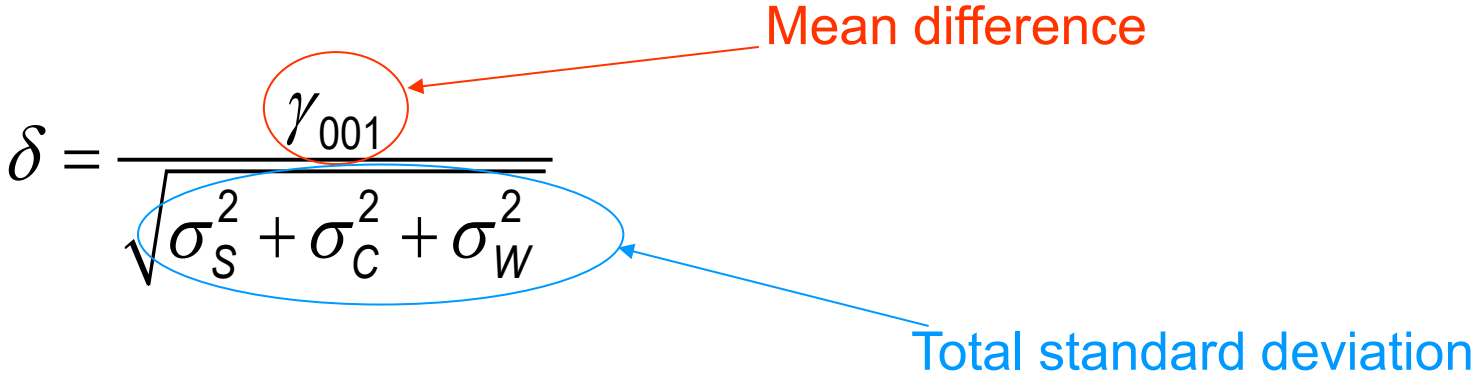
- Effect sizes can be defined in more than one way in multilevel designs
- A typical effect size is a standardized mean difference, which is relevant to experiments with a treatment and a control group
- The numerator of the effect size is naturally the mean difference
- The question is which standard deviation should be used in the denominator to standardize the mean difference
- One standardization procedure is to use the total standard deviation of the outcome

Effect Sizes

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



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



Effect Sizes

- In two-level randomized block designs, one could use the treatment effect variance. For example,

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma^2_{T\times S} + \sigma^2_W}}$$

- Similarly, in three-level randomized block designs where assignment is at the second level, the effect size can be defined as

$$\delta = \frac{\gamma_{010}}{\sqrt{\sigma^2_{T\times S} + \sigma^2_C + \sigma^2_W}}$$

References

- 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. & Schauer, J. (2018). The history of randomized trials in education in America. *Education Research, 60,* 265-275.
- Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage Publishing.
- Konstantopoulos, S. (2025). *Randomized Controlled Trials.* In L. Cohen-Vogel, J. Scott, & P. Youngs (Eds), *The Handbook of Education Policy Research* (2nd Ed), Washington DC: AERA.
- 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.

- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models.* Thousand Oaks, CA: Sage.
- 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.