Quasi-Experimental Workshop

Tom Cook and Will Shadish

Supported by Institute for Educational Sciences

Introductions: People

Workshop Staff

Your names and affiliations

Logic for selecting you

Introduction: Purposes

- Briefly describe the still improving current state of causal research in education
- Briefly learn why the randomized experiment is so preferred when it is feasible
- Learn better quasi-experimental designs and principles for creating novel designs
- Improve your own causal research projects through discussion with Will and Tom
- Disseminate better practices when go home
- Have fun, eat well, meet interesting new people

Decorum

No dress code

Please interrupt for clarification

Engage instructors in side conversations at breaks and meals

No titles, only first names

Schedule by Day

- Morning session, lunch period, afternoon session, rest or talk 4 to 6, dinner
- Lunch to meet others, let off steam, or discuss your own projects with Tom or Will
- Rest period to catch up on home, meet Tom or Will, discuss material with others
- Dinners are to have fun and socialize

And at the End of the Workshop

- We will provide a web site and email addresses for followup
- Provide you with finalized sets of powerpoint slides (since we do tend to change them a little at each workshop).

Framing the Week's Substance:

- The Current State of the Causal Art in Education Research

Education leads the other Social Sciences in Methods for:

- Meta-Analysis
- Hierarchical Linear Modeling
- Psychometrics
- Analysis of Individual Change
- Next 5 days do not entail an indictment of education research methods writ large
- Only of its methods for identifying what works, identifying causal relationships

State of Practice in Causal Research in Education

- Theory and research suggest best methods for causal purposes. Yet
- Low prevalence of randomized experiments--Ehri; Mosteller; but now improving
- Low prevalence of Regression-discontinuity
- Low prevalence of interrupted time series
- Low prevalence of studies combining control groups, pretests and sophisticated matching
- High prevalence of weakest designs: pretest-posttest only; non-equivalent control group without a pretest; and even with it.

Forces Impelling Change in Causal Research Practice

- General dissatisfaction with knowledge of “what works” in education
- IES’ experimental agenda in Bush 43 years and its control over research funds
- Role of some foundations, esp. W.T. Grant
- Growth of applied micro-economists in education research in USA and abroad
- Better causal research practice in early childhood education and school-based prevention -- why?

RA in American Educational Research Today

Heavily promoted at IES, NIH and in some Foundations

- Normative in pre-school research (ASPE) and in research on behavior problems in schools (IES and NIJJ)
- Reality in terms of IES Funding Decisions
- Growing reality in terms of publication decisions and univ. teaching practices

IES Causal Programs in Bush Administration

- National Evaluations mandated by Congress or some Title and done by contract research firms
- Program Announcements for Field-Initiated Studies mostly done in universities
- Training and Center Grants to universities
- What Works Clearinghouse
- Regional Labs
- SREE - Society for Research in Educational Effectiveness
- Unusual degree of focus on a single method

Institutionalizing the Agenda

- Depends on more persons able and willing to assign at random entering into ed research
- Degree to which the opposition is mobilized to fight random assignment; seems dispirited now
- Emerging results show few large effects - shall we shoot the messenger?
- Bush priority for RA is being pursued less monolithically in Obama administration
- But still, there has been clear change towards RA in education research that may lose some saliency in future but not all of it.

Our Main Purpose

- To raise quality of causal research in ed, we will:
- Do one afternoon on randomized experiments, though some content also applies to quasi-exps
- A day on Regression-discontinuity
- A half-day on short interrupted time-series, with some material on value-added analyses
- A day on various sample- and individual case-matching practices, good and bad
- A day on other causal design principles that are not predicated on matching for grp comparability

Terminology to Keep on Track

- Experimentation - deliberate intrusion into an ongoing process to identify effects of that intrusion – role of the exogenous shock

- Randomized experiments involve assignment to treatment and comparison groups based on chance—unbiased in expectation

- Natural experiment denotes some sudden and non-researcher controlled intrusion into ongoing process—examples with and without RA

Terminology

- Quasi-experiments involve exogenous shocks but control groups not randomly assigned—examples look like experiments in structure except for assignment process
- A non-experiment deals with causal agent not deliberately manipulated nor suddenly intruding into an ongoing process – say, a longitudinal survey relating attention to learning gains

Framing Today’s Substance

- Discuss what we mean by causation
- Discuss threats to validity, esp. internal validity
- Analyze the randomized experiment as the archetypal causal study
- Discuss the limitations to doing experiments in real school settings
- Discuss ways of circumventing these limitations

Some Working Conceptions of Causation

Activity or Manipulability Theory from Philosophy of Science

- What is it?
- Some examples from daily life and science
- Why it is important for practice and policy
- How it relates to experimentation
- Illustrating its major limitations through confrontation with other theories of causation

Mackie’s INUS Conditional

Causal agents as Insufficient but Necessary **Parts** of Unnecessary but Sufficient conditions for an effect

- Example of all the “hidden” factors it takes for a matchstick to cause fire or for class size to cause learning DEPENDABLY
- Experimentation is causally incomplete cos it teaches us about very few causal contingencies
- Full causal knowledge requires knowing the causal role of multiple contingency variables
- So the conclusion from any one study may be unstable - causal heterogeneity.

Cronbach's UTOS Formulation

- Studies require Units, Treatments, Outcomes (Observations), Settings -- and also Times
- These condition the results of any one causal claim from an experiment -- some examples
- Implies Unit of progress is review not single study; and identifying **general** causal mediating processes should be main goal. BUT
- Both causal explanation and causal robustness require having some studies whose causal conclusions we can trust! Hence this workshop.

Another Way of Saying this (1)

- More than study-specific causal descriptions from A to B, Science values (a) explanatory causal knowledge of why A affects B and (b) causal descriptions that robustly replicate across multiple, heterogeneous studies
- Aspirations of science should also animate public policy - each requires **stable** knowledge
- Experimentation is useful because causal explanation always contains causal descriptions that are better if stable. Why explain causal phenomena that are wrong or weakly replicable?

Another Way of Saying this (2)

- Reviews allow us to establish dependability of a causal connection IF the UTOS sampling frame is heterogeneous
- Reviews allow us to identify some specific moderator and mediator variables
- But reviews require at least some individual causal conclusions we trust. Why review many studies if they are biased in same direction?
- Hence this workshop. Good knowledge of descriptive causal connections facilitates both explanations and reviews that are dependable and so less dependent on unknown conditions

- Now we turn to the best explicated theory of descriptive causal practice for the social sciences that introduces a notational system and a vocabulary:

Rubin’s Causal Model

Rubin's Counterfactual Model

- At a conceptual level, this is a counterfactual model of causation.
  - An observed treatment given to a person. The outcome of that treatment is Y(1)
  - The counterfactual is the outcome that would have happened Y(0) if the person had not received the treatment.
  - An effect is the difference between what did happen and what would have happened:

$$\text{Effect} = Y(1) – Y(0).$$

- Unfortunately, it is impossible to observe the counterfactual, so much of experimental design is about finding a credible source of counterfactual inference.

Rubin’s Model: Potential Outcomes

- Rubin often refers to this model as a “potential outcomes” model.
- Before an experiment starts, each participant has two potential outcomes,
  - Y(1): Their outcome given treatment
  - Y(0): Their outcome without treatment
- This can be diagrammed as follows:

Rubin’s Potential Outcomes Model

Units	Potential Outcomes		Causal Effects
	Treatment	Control	
1	$Y_1(1)$	$Y_1(0)$	$Y_1(1) – Y_1(0)$
⋮			
i	$Y_i(1)$	$Yi_i(0)$	$Y_i(1) – Y_i(0)$
⋮			
N	$Y_N(1)$	$Y_N(0)$	$Y_N(1) – Y_N(0)$
⋮			
	$\overline{Y(1)}$	$\overline{Y(0)}$	$\overline{Y(1)}-\overline{Y(0)}$

Under this model, we can get a causal effect for each person.

And we can get an average causal effect as the difference between group means.

Rubin's Potential Outcomes Model

Units	Potential Outcomes		Causal Effects
	Treatment	Control	
1	$Y_1(1)$	$Y_1(0)$	$Y_1(1) – Y_1(0)$
• • •			
i	$Y_i(1)$	$Yi_i(0)$	$Y_i(1) – Y_i(0)$
• • •			
N	$\underline{Y_N(1)}$	$\underline{Y_N(0)}$	$Y_N(1) – Y_N(0)$
	$\overline{Y(1)}$	$\overline{Y(0)}$	$\overline{Y(1)}$ - $\overline{Y(0)}$

Unfortunately, we can only observe one of the two potential outcomes for each unit. Rubin proposed that we do so randomly, which we accomplish by random assignment:

Rubin’s Potential Outcomes Model

Units	Potential Outcomes		Causal Effects
	Treatment	Control	
1	$Y_1(1)$		
⋮		⋮	⋮
i		$Yi_i(0)$	
⋮		⋮	⋮
N	$\underline{Y_N(1)}$		
	$\overline{Y(1)}$	$\overline{Y(0)}$	$\overline{Y(1)}$ - $\overline{Y(0)}$

The cost of doing this is that we can no longer estimate individual causal effects. But we can still estimate Average Causal Effect (ACE) as the difference between the two group means. This estimate is unbiased because the potential outcomes are missing completely at random.

Rubin’s Model and Quasi-Experiments

- The aim is to construct a good source of counterfactual inference given that we cannot assign randomly, for example
  - Well-matched groups
  - Persons as their own controls
- Rubin has also created statistical methods for helping in this task:
  - Propensity scores
  - Hidden bias analysis

Is Rubin's Model Universally Applicable?

- Natural Sciences invoke causation and they experiment, but they rarely use comparison groups for matching purposes
- They pattern-match instead, creating either a
- Very specific hypothesis as a point prediction; or
- Very elaborate hypothesis that is then tested via re-application and removal of treatment under experimenter control
- We will later use insights from this notion to construct a non-matching approach to causal inference in quasi-experiments to complement matching approaches

Very Brief Exigesis of Validity

- This goes over some well known ground
- But it forces us to be explicit about the issues on which we prioritize in this workshop

Validity

- We do (or read about) a quasi-experiment that gathered (or reported) data
- Then we make all sorts of inferences from the data
  - About whether the treatment worked
  - About whether it might work elsewhere
- The question of validity is the question of the truth of those inferences.
- Campbell’s validity typology is one way to organize our thinking about inferences.

Campbell's Validity Typology

- As developed by Campbell (1957), Campbell & Stanley (1963), Cook & Campbell (1979), with very minor changes in Shadish, Cook & Campbell (2002)
  - Internal Validity
  - Statistical Conclusion Validity
  - Construct Validity
  - External Validity
- Each of the validity types has prototypical threats to validity—common reasons why we are often wrong about each of the four inferences.

Internal Validity

- *Internal Validity*: The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B, as those variables were manipulated or measured.
- Or more simply—did the treatment affect the outcome?
- This will be the main priority in this workshop.

Threats to Internal Validity

1. Ambiguous Temporal Precedence
2. Selection
3. History
4. Maturation
5. Regression
6. Attrition
7. Testing
8. Instrumentation
9. Additive and Interactive Effects of Threats to Internal Validity

Think of these threats as specific kinds of counterfactuals—things that might have happened to the participants if they had not received treatment.

Statistical Conclusion Validity

- Statistical Conclusion Validity: The validity of inferences about the correlation (covariation) between treatment and outcome.
- Closely tied to Internal Validity
  - SCV asks if the two variables are correlated
  - IV asks if that correlation is due to causation

Threats to Statistical Conclusion Validity

1. **Low Statistical Power (very common)**
2. **Violated Assumptions of Statistical Tests (especially problems of nesting—students nested in classes)**
3. Fishing and the Error Rate Problem
4. Unreliability of Measures
5. Restriction of Range
6. Unreliability of Treatment Implementation
7. Extraneous Variance in the Experimental Setting
8. Heterogeneity of Units
9. Inaccurate Effect Size Estimation

Construct Validity

- Construct Validity: The validity of inferences about the higher-order constructs that represent sampling particulars.
  - We *do* things in experiments
  - We *talk about* the things we did in our reports
  - One way to think about construct validity is that it is about how accurately our *talk* matches what we actually *did*.

External Validity

- External Validity: The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.
- Always the “stepchild” in Campbell’s work, Cook has developed a theory of causal generalization addressing both construct and external validity.
- But that is another workshop.

Validity Priorities for This Workshop

Main Focus is Internal Validity

Statistical Conclusion Validity: Because it is so closely tied to Internal Validity

Relatively little focus

Construct Validity

External Validity

Randomized Experiments

with Individual Students and

with Clusters of Classrooms or Schools

Randomized Control Trials: Some Selective Issues

1. Logic of random assignment
2. Clarification of Assumptions of RCTs
3. Recent Advances for Dealing with Partial and not Full Implementation of Treatment
4. Recent Advances in Dealing with Sample Size Needs when assigning Schools or Classrooms rather than Students

What is an Experiment?

- The key feature common to all experiments is to deliberately *manipulate* a cause in order to *discover* its effects
- Note this differentiates experiments from
  - Case control studies, which first identify an effect, and then try to discover causes, a much harder task

Random Assignment

- Any procedure that *assigns units to conditions based only on chance*, where each unit has a nonzero probability of being assigned to a condition
  - Coin toss
  - Dice roll
  - Lottery
  - More formal methods (more shortly)

What Random Assignment Is Not

- Random assignment is not random sampling
  - Random sampling is rarely feasible in experiments
- Random assignment does not require that every unit have an *equal* probability of being assigned to conditions
  - You can assign unequal proportions to conditions

Equating on Expectation

- Randomization equates groups on *expectation* for all *observed and unobserved* variables, not in each experiment
  - In quasi-experiments matching only equates on *observed* variables.
- *Expectation*: the mean of the distribution of all possible sample means resulting from all possible random assignments of units to conditions
  - In cards, some get good hands and some don't (luck of the draw)
  - But over time, you get your share of good hands

Estimates are Unbiased and Consistent

- Estimates of effect from randomized experiments are *unbiased*: the expectation equals the population parameter.
  - So the average of many randomized experiments is a good estimate of the parameter (e.g., Meta-analysis)
- Estimates from randomized experiments are *consistent*: as the sample size increases in an experiment, the sample estimate approaches the population parameter.
  - So large sample sizes are good
- Quasi-experiments have neither of these characteristics.

Randomized Experiments and The Logic of Causal Relationships

- Logic of Causal Relationships
  - Cause must precede effect
  - Cause must covary with effect
  - Must rule out alternative causes
- Randomized Experiments Do All This
  - They give treatment, then measure effect
  - Can easily measure covariation
  - Randomization makes most other causes less likely
- Quasi-experiments are problematic on the third criterion.
- But no method matches this logic perfectly (e.g., attrition in randomized experiments).

Assumptions on which a Treatment Main Effect depends

- Posttest group means will differ, but they are causally interpretable only if:
- The assignment is proper, so that pretest and other covariate means do not differ on observables on expectation (and in theory on unobservables)
- There is no differential attrition, and so the attrition rate and profile of remaining units is constant across treatment groups
- There is no contamination across groups, which is relevant for answering questions about treatment-on-treated but not about intent to treat.

Advantages of Experiments

- Unbiased estimates of effects
- Relatively few, transparent and testable assumptions
- More statistical power than alternatives
- Long history of implementation in health, and in some areas of education
- Credibility in science and policy circles

Disadvantages attributed to Experiments we must discuss

- Not always feasible for reasons of ethics, politics, logistics and ignorance
- Experience is limited in education, especially with higher order units like whole schools
- Limited generality of results - voluntarism and INUS conditionals revisited
- Danger that the method alone will determine types of causal questions asked and not asked and crowd out other types of knowledge
- Asks intent-to-treat questions that have limited yield for theory and program developers

Analyses Taking Degree or Quality of Implementation into Account

- An intent-to-treat analysis (ITT)
- An analysis by amount of treatment actually received (TOT)
- Need to construct studies that give unbiased inference about each type of treatment effect
- We have seen how to do ITT. What about TOT?

Partial Treatment Implementation

Intent to Treat

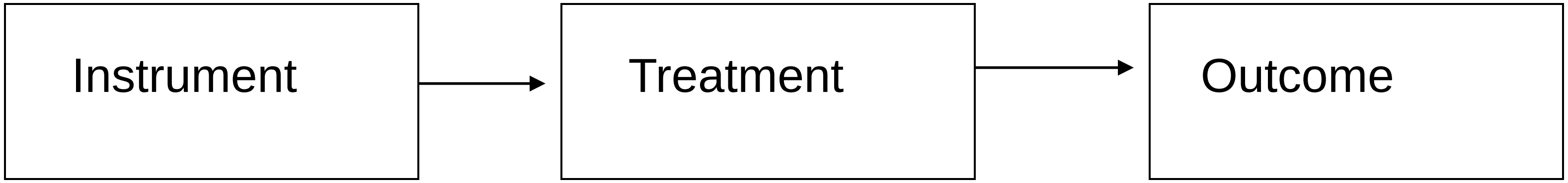
- Participants analyzed in condition to which they were assigned
- Preserves internal validity
- Yields unbiased estimate about effects of being assigned to treatment, not of receiving treatment
- May be of policy interest
- But should be complemented by other analyses

Analysis by Treatment Received

- Compare outcomes for those who received treatment to outcomes for those who did not
- Estimates effects of treatment receipt
- But is quasi-experimental
- Rarely a good option by itself

Instrumental Variables Analysis

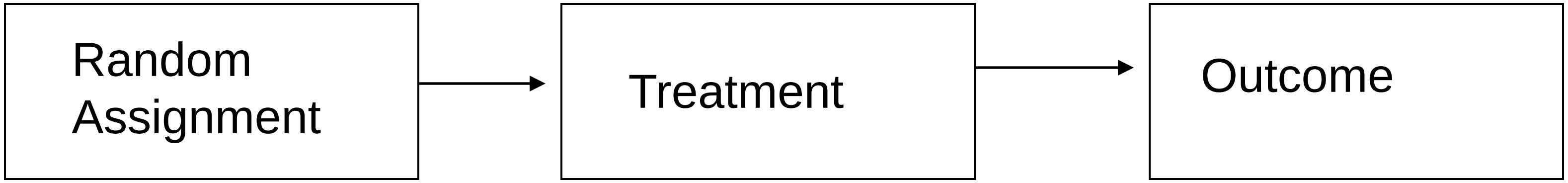
- Angrist, Imbens, Rubin JASA 1996
- In economics, an instrument is a variable or set of variables is correlated with outcome only through an effect on other variables (in this case, on treatment)



- Can use the instrument to obtain an unbiased estimate of effect

Instrumental Variables Analysis

- Use random assignment as an instrument for incomplete treatment implementation
- Yields unbiased estimate of the effects of receipt of treatment



Random Assignment

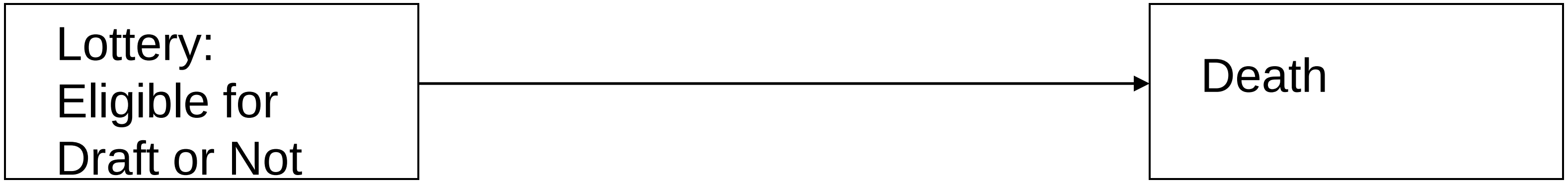
Treatment

Outcome

- Random assignment is certainly related to treatment, but it is unrelated to outcome except through the treatment.

Example: The Effects of Serving in the Military on Death

- Lottery randomly assigned people to being eligible for the draft.
  – Intent to treat analysis would assess the effects of being eligible for the draft on death



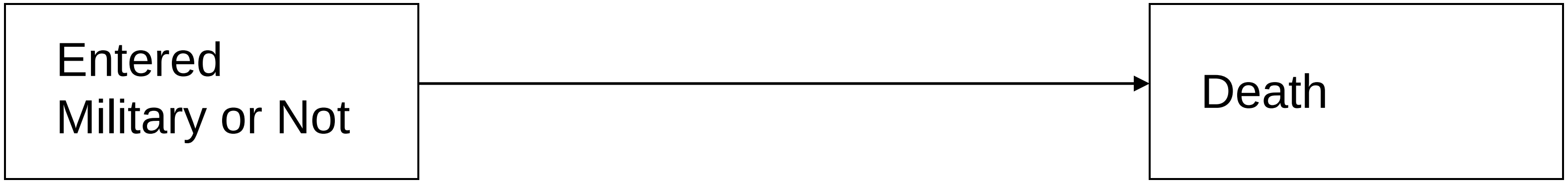
Lottery: Eligible for Draft or Not

Death

- This is a good randomized experiment yielding an unbiased estimate

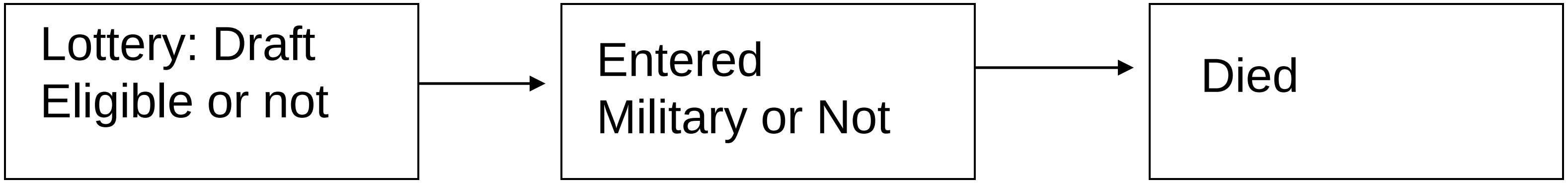
Example continued

- But that is not the question of interest
  - Not all those eligible for the draft entered the military (not all those assigned to treatment received it).
    - Some who were draft eligible were never drafted
    - Some who were not eligible chose to enlist
  - We could compare the death rates for those who actually entered the military with those who did not (we could compare those who received treatment to those who did not):



  - But this design is quasi-experimental

Example: Instrumental Variable Analysis



Lottery: Draft Eligible or not

Entered Military or Not

Died

- Random assignment is an instrument because it can only affect the outcome through the treatment.
- That is, being randomly assigned to being draft eligible only affects death if the person actually joins the military.

Analysis for binary outcome and binary treatment implementation

- 35.3% draft eligible served in military
- 19.4% not eligible served in military
- The lottery (random assignment) caused 15.9% to serve in the military (normal randomized experiment)
- 2.04% draft eligible died
- 1.95% not eligible died
- Draft eligible caused 0.09% to die
- Causal effect of serving in military on death among those participating in the lottery is .0009/.159 = .0058 = .56%

Assumptions of IV Strategy

1. One person’s outcomes do not vary depending on the treatment someone else is assigned
2. The causal effects of assignment both on receipt and on outcome can be estimated using standard intent-to-treat analyses
3. Assignment to treatment has a nonzero effect on receipt of treatment

Assumptions, continued

4. Random assignment (the instrumental variable) affects outcome only through its effects on receipt of treatment
   - a potential draftee's knowledge that he was now eligible for the draft might cause him to stay in school to gain a deferment, which might improve mortality rates through education and income
5. There are no "oppositional" participants who would always refuse treatment if assigned to it, but take treatment if not assigned to it
   - A person whose family history would have encouraged him to volunteer for the military in the absence of being drafted but who objected to the government draft and so refused to serve in protest

More on Angrist et al.

- Extensions to
  - Variable treatment intensity
  - Quasi-experiments of all kinds, but regression-discontinuity in particular
  - Continuous outcomes
- An area rapidly developing
- But still limited to analyses of a single mediator variable. In many substantive applications, there are many mediators of a treatment’s effects, as in a causal or structural equation model.

Issues of Nesting and Clusters, most of which is also relevant to Quasi-Experiments

Units and Aggregate Units

- Can randomly assign:
  - Units (e.g., children, households)
  - Aggregates (e.g., classrooms, neighborhoods)
- Why we use aggregates:
  - When the aggregate is of intrinsic interest (e.g., effects of whole school reform)
  - To avoid treatment contamination effects within aggregates.
  - When treatment cannot be restricted to individual units (e.g., city wide media campaigns)

The Problem with Aggregates

- Most statistical procedures assume (and require) that observations (errors) be independent of each other.
- When units are nested within aggregates, units are probably not independent
    - If units are analyzed as if they were independent, Type I error skyrockets
        - E.g., an intraclass correlation of .001 can lead to a Type I error rate of $\alpha > .20$!
- Further, degrees of freedom for tests of the treatment effect should now be based on the number of aggregates, not the number of persons
- This means test of hypotheses about aggregates can be over-powered if analyzed wrongly and that the correct analysis might need “many” classrooms or schools, which is expensive

What Creates Dependence?

- Aggregates create dependence by
  - Participants interacting with each other
  - Exposure to common influences (e.g,. Patients nested within physician practices)
- Both these problems are greater the longer the group members have been interacting with each other.

Making an Unnecessary Independence Problem

- Individual treatment provided in groups *for convenience alone* creates dependence the more groups members interact and are exposed to same influences.
- For instance, Empirically Supported Treatments or Type I errors?
  - About of a third of ESTs provide treatment in groups
  - When properly reanalyzed, very few results were still significant.

Some Myths about Nesting

- Myth: Random assignment to aggregates solves the problem.
  - This does not stop interacting or common influences
- Myth: All is OK if the unit of assignment is the same as the unit of analysis.
  - That is irrelevant if there is nesting.
- Myth: You can test if the ICC = 0, and if so, ignore aggregates.
  - That test is a low power test
- Myth: No problem if randomly assign students to two groups within one classroom.
  - Students are still interacting and exposed to same influences

The Worst Possible Case

- Random assignment of one aggregate (e.g., a class) per condition
  - The problem is that class and condition are completely confounded, leaving no degrees of freedom with which to estimate the effect of the class.
  - This is true even if you randomly assign students to classes first.

What to Do?

- Avoid using one aggregate per condition
- Design to ensure sufficient power--more to come later
  - have more aggregates with fewer units per aggregate
  - randomly assign from strata
  - use covariates or repeated measure
- Analyze correctly
  - On aggregate means (but low power, and loses individual data)
  - Using multilevel modeling (preferred)
- Increase degrees of freedom for the error term by borrowing information about ICCs from past studies

An Example: The Empirically Supported Treatments (EST) list.

- EST's touted as methodologically strong
  - But problem not limited to ESTs
- Includes 33 studies of group-administered treatment
  - Group therapies
  - Individual therapies administered in group settings for convenience
- None took nesting into account in analysis
- We estimated what proper analysis would have yielded, using various assumptions about ICC.
  - Adjust significance tests based on ICCs
  - Adjust df based on number of groups not individuals

Table 1

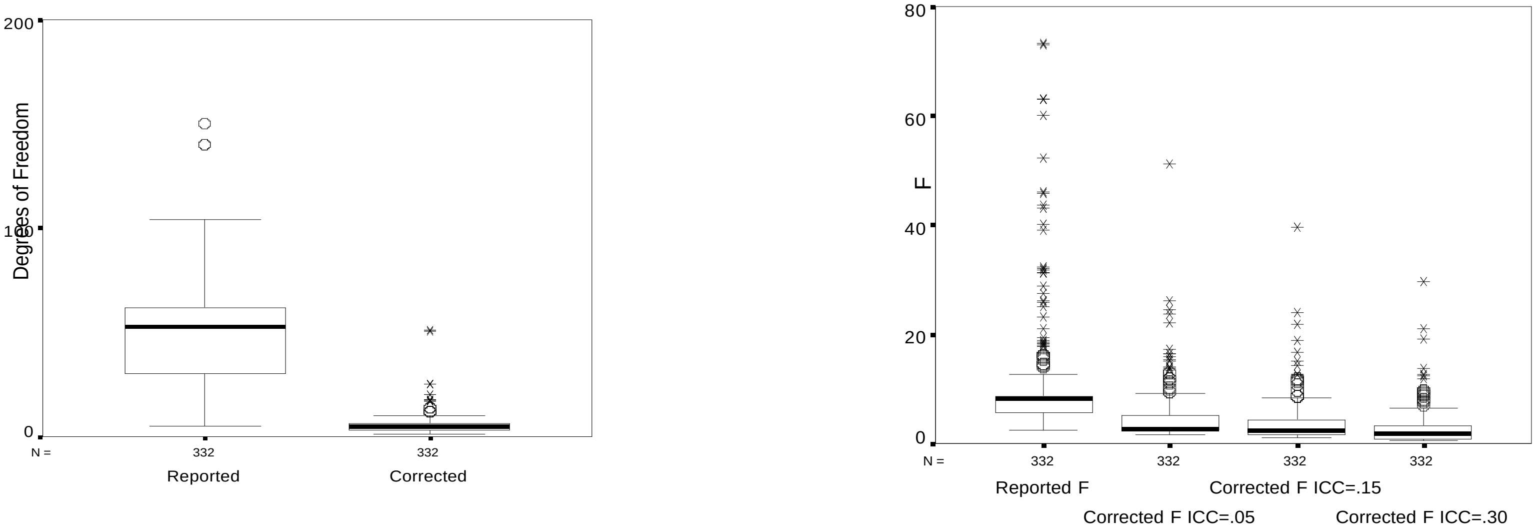
Equations for Adjusting Effects Estimators.

t-test	$t_{adj} = \dfrac{t_{unadj}}{\sqrt{1 + (m_o - 1)ICC}}$
F-test for ANOVA	$F_{adj} = \dfrac{F_{unadj}}{1 + (m_o - 1)ICC}$
Chi-Square	$\chi^{2}{}_{adj} = \dfrac{\chi^{2}{}_{unadj}}{1 + (m_o - 1)ICC}$

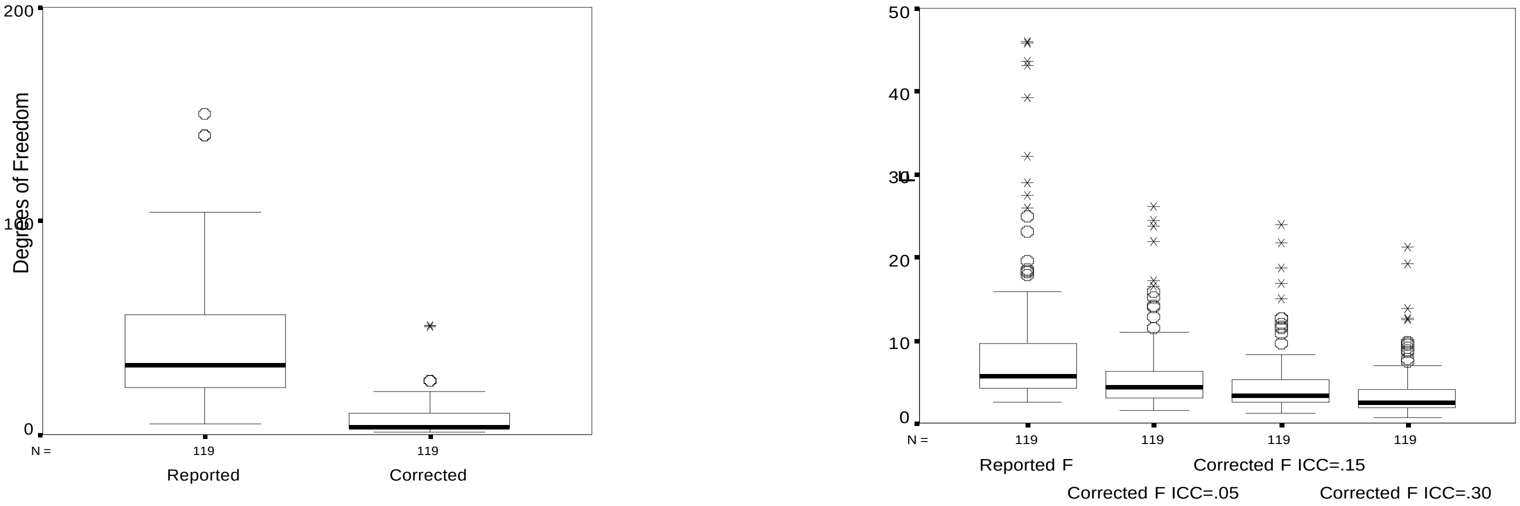
Note. Adapted from Rooney (1992). *m*=number of members per group, ICC=Intraclass Correlation.

Results

- After the corrections, only 12.4% to 68.2% of tests that were originally reported as significant remained significant
- When we considered all original tests, not just those that were significant, 7.3% to 40.2% of tests remained significant after correction
- The problem is even worse, because most of the studies tested multiple outcome variables without correcting for alpha inflation
- Of the 33 studies, 6-19 studies no longer had *any* significant results after correction, depending on assumptions



For all N = 332 *t*- and *F*-tests



For N = 119 omnibus *t*- and *F*-tests with exact information

Other Issues at the Cluster Level

- Sample Size and Power
- Contamination
- Getting Agreement to Participate

Estimating the Needed Sample Size

- We are not dealing here with statistical power in general, only at school level
- Question is: How many schools are needed for ES of .20, with $p < .05$, power .80, assuming a balanced design and > 50 students per school.
- Why .20? Why .05, why .80. Why balanced? What role does the N of students play?

Key Considerations

We estimate the cluster effect via the unconditional ICC, that part of the total

variation that is within schools

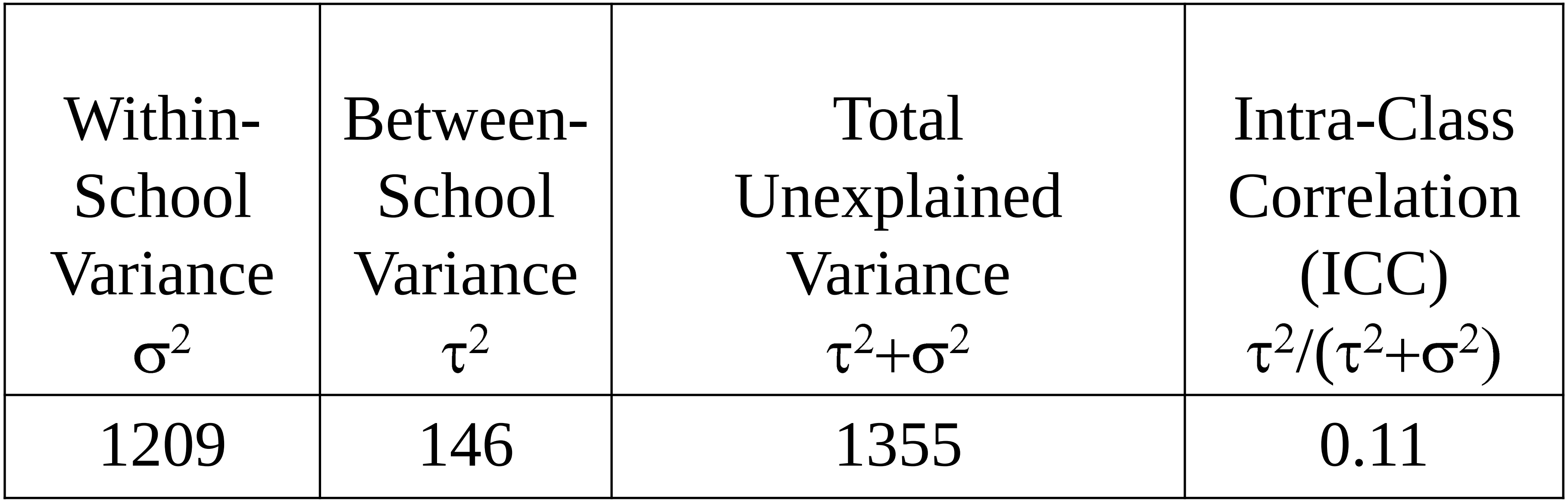
- But sample size needs are driven by the conditional ICC, the difference between schools after covariates are used to “explain” some of the between-school variation
- We want to use 2 examples, one local and one national, to illustrate how careful use of school-level covariates can reduce the N of schools needed

Example 1: Kentucky

- An achievement study
- A school-level question
- A limited budget
- One year of prior achievement data at both the school and student levels
- Given these data, and traditional power assumptions, how many schools needed to detect an effect of .20?
- We use J for schools and N for students

Kentucky Cluster Table

Table 1: Estimates from Unconditional Model



Within-School Variance σ^2	Between-School Variance τ^2	Total Unexplained Variance $\tau^2+\sigma^2$	Intra-Class Correlation (ICC) $\tau^2/(\tau^2+\sigma^2)$
1209	146	1355	0.11

Table 2: Required J for the Unconditional Model

Unconditional Effect Size	Required J
0.20	94
0.25	61
0.30	43

What is the School Level Covariate like?

For reading, the obtained covariate-outcome r is .85--the usual range in other studies is .70 to .95

As corrected in HLM this value is .92

What happens when this pretest school-level covariate is used in the model?

Table 3: Estimates from Conditional Model (CTBS as Level-2 Covariate)

Within School Variance σ^2	Between School Variance T^2	Total Unexplained Variance $\tau^2+\sigma^2$	Intra-Class Correlation (ICC) $\tau^2/(\tau^2+\sigma^2)$
1210	21.6	1231.6	0.0175

What has happened?

- The total unexplained variation has shrunk from 1355 to 1232--why?
- The total between-school variation has shrunk from 146 to 26--why?
- So how many school are now needed for the same power?

Table 4: Required J for Two Level Unconditional and Conditional Models

Effect Size	Required J No Covariate	Required J With Covariate
0.20	94	22
0.25	61	15
0.30	43	12

How does these Values Compare?

- The work of Hedges and Hallberg with nationally representative data where m is his term for sample size at the school level (not J)

National Estimates from Hedges

Grade	Covariates	*m*=10	*m*=15	*m*=20	*m*=25	*m*=30
1	None	0.67	0.54	0.46	0.41	0.37
	Pretest	0.32	0.25	0.22	0.19	0.18
5	None	0.70	0.56	0.48	0.43	0.39
	pretest	0.30	0.24	0.21	0.19	0.17
12	None	0.58	0.46	0.40	0.36	0.32
	pretest	0.21	0.17	0.15	0.13	0.12

Conclusions about needed Sample Sizes

- Will vary by type of outcome, local setting and quality of the covariate structure
- With achievement outcomes, about 20 schools will often do, 10 per group in a two-group study
- But to protect against attrition, some more might be added
- Further gains accrue from several prior years of school-level achievement data, not difficult to get
- Since intervention groups can cost more, an unbalanced design with more control units will also help, though gain depends on harmonic n

Contamination Issues with Cluster-level Assignment

- To reduce contamination one can move to a higher level of analysis: from student to classroom to grade level to school to district
- Need to monitor type and level of contamination--PGC Comer as an example
- How to analyze if some: Instrumental Variables for dichotomously distributed contamination
- More problematic with more complex forms of contamination

Cluster Level Random Assignment- Getting Agreement

- High rate of RA in preschool studies of achievement and in school-based studies of prevention, but not in school-based studies of achievement. Why? Culture or Structure?
- Cook's war stories - PGC; Chicago; Detroit
- Grant Fdn. Resources
- Experiences at Mathematica
- IES experience generally positive that RA can be often achieved (and maintained). But difficult

Summary re RCTs

- For one understanding of cause, RCT is best
- Has its own assumptions that need to be tested
- Based on a marriage of statistical theory and an ad hoc “theory” of implementing RA
- RCTs not usable in all ed research practice
- Limited capacity to explore causal contingencies
- Results from single studies probabilistic rather than deterministic
- Philosophers of science might say: First rate method for second rate theory of cause

Summary 2

- Lower level at which assign the better; Higher order designs can be expensive
- Covariates help reduce sample size needs: Crucial role of pretest
- Value of description of implementation based on program theory and quality measurement
- Black box RCTs not a good idea, but ironic that current methods cannot yet support complex explanations of why A causes B – best for theories invoking a single mediator
- New frontier in RCT method studies

Remember, though…

- Binary causal descriptions of an A causes B form are "the cement of the universe" because:
- Each causal explanation of why A causes B requires that A causes B
- Explanatory models postulating C as a mediator assumes A to C + C to B binary causal links.
- Reviews of tests of binary causal relations identify **stable** causal knowledge and need to assume the validity of each binary reviewed.
- So testing binary A-B links is important even if it is rarely the end-goal of a generalizing science.