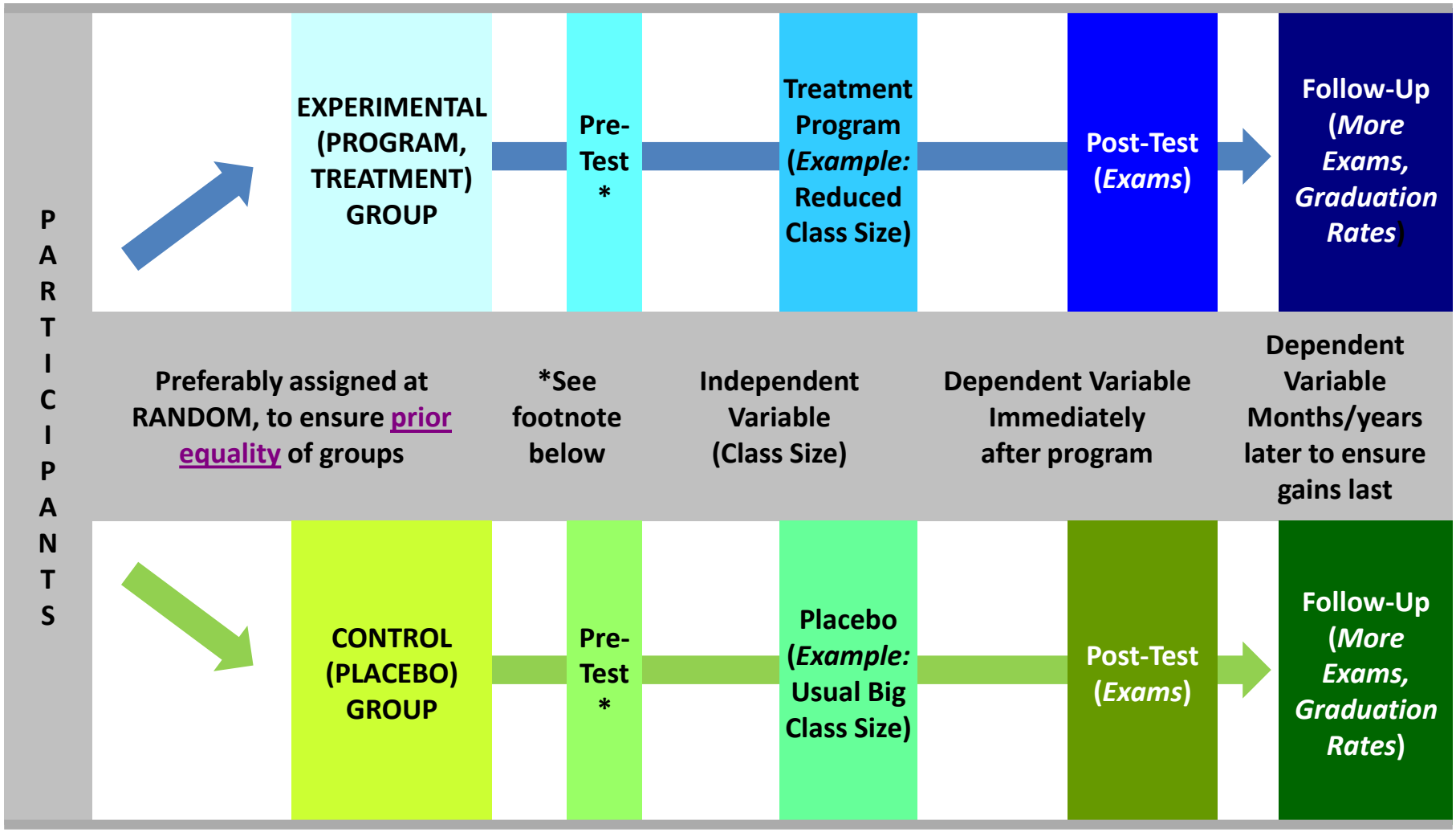
Research Methods (HDFS 3390),
Alan Reifman, Texas Tech University
Program Evaluation

Evaluating the impact of an educational or public policy change (experimenting in the "real world")

If the government or private sources are going to spend millions of dollars trying to improve something, shouldn't we be finding out if the program is actually working as expected? Consider the case of universal pre-K education.

Program Evaluation (A Form of Experimental Design)



*Pre-test not necessary with random assignment to groups, but essential with non-random assignment.

One state that conducted a proper research study in conjunction with an educational innovation:

Do smaller class sizes improve student learning?

Project STAR (Student/Teacher Achievement Ratio), done in Tennessee

- **EXPERIMENTAL GROUP: Small classes in grades K-3, then regular-size classes from then on**
- **CONTROL GROUP: Regular-size classes throughout**

Dr. Reifman's 5E Principle:
Everything Equal Except Essential Element
Using Project STAR, Tennessee's Class-Size Experiment, to Illustrate

	Experimental	Control
Independent Variable: Class Size in Kindergarten Through 3rd Grade (Essential element on which groups are made to differ)		
Teacher Quality (Random assignment of teachers to E or C groups should equalize teacher ability in the two groups, on average)	 =	
Student Ability Prior to Study (Random assignment of students to E or C groups should equalize student ability in the two groups, on average)	 =	
Assume other factors (e.g., how well maintained the classroom facilities are and the availability of school supplies) have been held constant between E & C groups.		
Dependent Variable: Student Achievement from End of 3rd Grade - 12th Grade If (as found) students in small classes K-3 perform better in later grades than students from large classes, it can be attributed only to class size (independent variable).		

Article showing that benefits of being in small classes in grades K-3 last through grades 11-12;
yet another, more detailed, article on evaluating the effects of class size on student achievement

Quasi-Experiments
(quasi = "as if," "seemingly," "in part")

Babbie definition (p. 357) focuses on presence/absence of random assignment as the distinguishing feature between true and quasi-experiments

Breakdown of Types of Quasi-Experiments

Some people are in a group (e.g., organization or school) that receives a special program (like an Experimental Group), whereas others are in a group that does not receive the special program (like a Control Group). Groups are thought to be similar demographically or in other ways. However, groups are NOT created by randomly assigning individuals. Membership in these groups may have been established months or years before the study was ever thought up. Babbie refers to this as a "Nonequivalent Control Group" design (pp. 350-351).

Hypothetical Example: An anti-smoking campaign is launched throughout the Texas Tech campus (Experimental Group), but not at Texas A&M (Control Group). Smoking rates at the two schools are then compared afterwards. (A pre-test to check if the two schools had similar smoking rates at the start of the project would have been a good idea.)

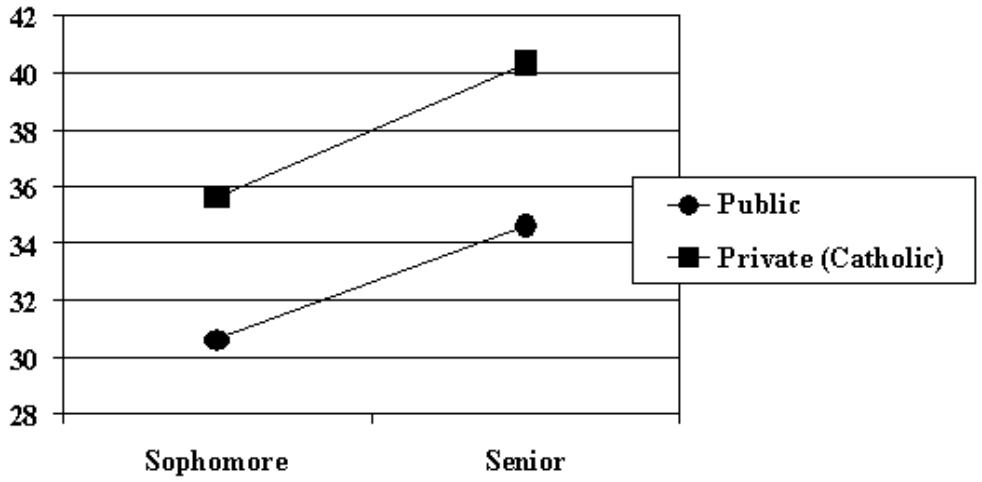
Before-after (pre-post) comparison within a single setting of how often a behavior occurred before some program or policy was implemented and after. Babbie refers to this as a "Time-Series" design (pp. 349-350).

Real Example: Looking at violence/rowdiness at University of Colorado football games before and after stadium ban on beer sales (Bormann & Stone, 2001).

In pre-post studies lacking random assignment to conditions (e.g., a campaign to increase lottery-ticket sales), there are ways to rule out alternative explanations such as customers having more disposable income at post-test than at pre-test. See the "Illustration" paragraph on p. 27 and Figure 1B of:

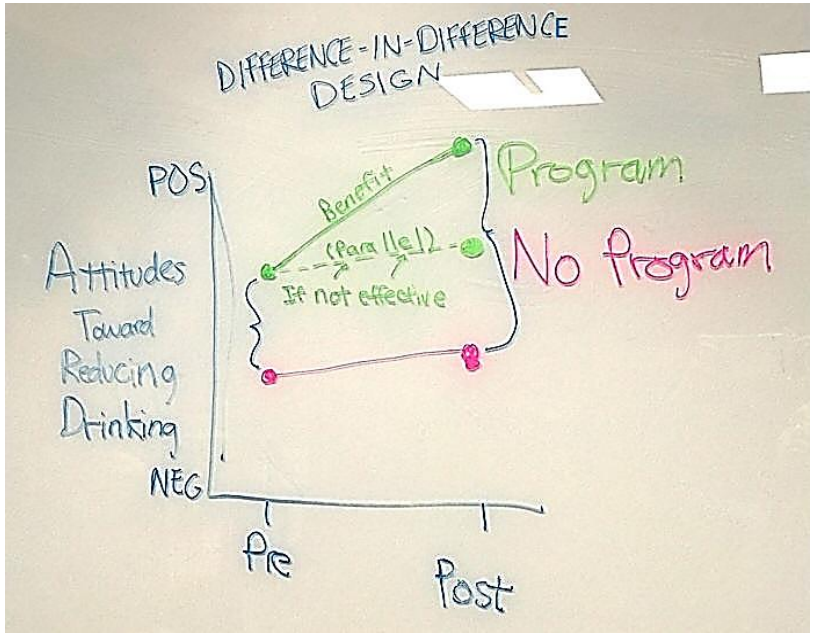
West, S.G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal Inference. *Psychological Methods, 15*, 18-37.

Mean Number of Test Questions Answered Correctly
High School and Beyond Study,
(Reported by N. Pearson, Lubbock A-J, 10/18/98)



Importance of a Pre-Test When Groups are Not Created by Random Assignment:

Can Take Pre-Test Difference into Account When Looking at Final Difference



(For graduate students: Technically, this is known as a difference-in-difference design. The dotted green line shows the assumption of what would have happened in the program group if the program were completely ineffective [i.e., following the trend of the control group]. See equation on slide 7 of this slideshow.)

Program Evaluation (Quasi-Experiment) of AmeriCorps, a National Service Program for Young People

From the Full Report of "Serving Country and Community":

This study was designed to address three objectives

- **Describe AmeriCorps programs**
- **Describe AmeriCorps members**
- **Describe the impact of AmeriCorps on members' attitudes and behaviors**

…To address these objectives, the Corporation undertook the current longitudinal study of the long-term effects of participation in AmeriCorps. Impact evaluations measure the degree to which a particular program, service, or intervention affects its intended target group. The ideal strategy for assessing program impacts is to employ an experimental design in which program applicants are randomly assigned into two groups: treatment (enrolled in the program) and control (excluded from enrollment in the program). However, during the 1999–2000 program year, when this study was implemented, AmeriCorps was still in the process of building national awareness and many local programs were struggling to recruit enough qualified candidates to fill their enrollment targets. Therefore, the Corporation determined that implementation of random assignment would not be feasible. In order to assess impacts, the study relied upon a quasi-experimental design that used a comparison group of individuals similar to the individuals enrolled in AmeriCorps…

In selecting comparison groups for this study, our goal was to identify individuals who demonstrated both an awareness of AmeriCorps and some interest in participation in service. The State and National comparison group comprised individuals who had indicated knowledge of, and interest in, AmeriCorps by contacting the Corporation's toll-free information line and requesting information about the program, but who did not actually enroll during the study period.

Practice Exercise: Name That Method -- Experimental, Quasi-Experimental, or Correlational? *(From Paul Fuglestad, via Social Psychology Course Resources On the Web, CROW)*

To get in the proper frame of mind for this practice exercise, let's think about how the question of whether smaller class sizes in schools produce better learning could be addressed by the three approaches:

True Experimental	Quasi-Experimental	Correlational
Randomly assign kids (and teachers) to small (experimental) or large (control) class size (with "class size" being the IV); then compare on later test scores (DV); this is what Tennessee's Project STAR actually did	Find one school that has small class sizes and another that has large class sizes ("IV"); then compare on later test scores ("DV"); no random assignment, so can't infer causality, but could do pre-test to compare degree of improvement in the two schools; you could get an idea of *possible* causation, which would justify conducting a true experiment in the future	Conduct large survey of people who attended many different schools, ask them to estimate what the typical class size was in their elementary school and what their SAT/ACT score was; use data points to calculate correlation; can't infer causation, but results could pave the way for conducting a true experiment in the future

Note that, in real-life, if a true experiment (highest-quality method for causation) existed, there would be no need to conduct further studies using the less-effective quasi-experimental and correlational designs.

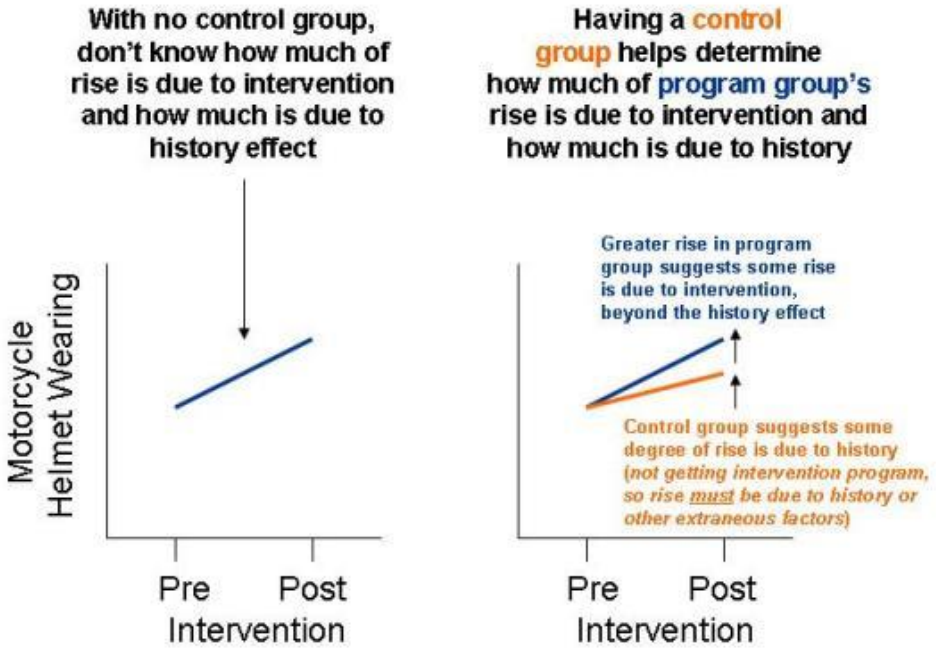
Here's another document with a good summary of the different kind of research designs

Additional Threats to Internal Validity in Long-Term Research

(Based on Campbell & Stanley, Cook & Campbell; many of the threats are discussed here, here, and here; whereas a videotaped talk on the threats is available here)

1. History Effect

Hypothetically, if researchers were implementing a program to increase motorcycle helmet use, this event could possibly create a history effect.



Historical Event (Accident of Famous Person) Occurs Around the Same Time as Intervention (Motorcycle Safety Program)

Threats to Internal Validity (Continued)

2. Maturation (improvement in children's test performance due simply to their increasing cognitive abilities)

3. Testing (with re-testing, participants become sensitive to what study is about)

- For maturation and testing, same control-group logic as above applies, just substitute "maturation" or "testing" for "history effect"

4. Instrumentation (possible example)

5. Statistical regression (visual depictions: here and here)
   - The "Sports Illustrated Jinx" (i.e., people on the cover experiencing misfortune shortly thereafter) could just be an example of regression to the mean. Also here, a control group can overcome problem. This is Dr. Reifman's attempt to analyze whether there's an SI Jinx, using a quasi-experimental design with control group.

6. Selection biases
   - Babbie: "Comparisons don't have any meaning unless the groups are comparable at the start of an experiment."

7. Experimental mortality/attrition (overview; example)

8. Causal time-order

Threats to Internal Validity (Continued)

9. Diffusion and imitation of treatments

(*Diffusion* is when word about the nature of the program leaks out of the experimental group, and *imitation* is when the control group adopts the behavior; possible example)

10. Compensatory equalization of treatment

"Were those administering the setting pressured, or did they decide on their own, to compensate the control group's lack of the benefits of treatment by providing some other benefit for the control group? Parents may pressure school administrators, for instance, to provide alternative learning experiences to compensate for their children in the control group not receiving the special test curriculum being studied in the experimental group." (Garson)

11. Compensatory rivalry

12. (Resentful) demoralization

Program Evaluation Examples

- "The Tamale Lesson" (Example of using story form, rather than purely factual presentation, for health education; video, report)
- Did the television show "16 and Pregnant" reduce the U.S. teen-pregnancy rate?
- Early Childhood Intervention: Abecedarian Project, Head Start, and others (summary of evaluative results)
- Effectiveness of programs to prevent child maltreatment
- Evaluation of abstinence-only sex education in Texas
- Evaluation of AlcoholEdu's effectiveness in preventing college drinking
- Example of group peer-based intervention that backfired: Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist, 54*, 755–764.
- University of Michigan program that appears to raise kids' science test scores
- UCLA evaluation study of California's policy of putting more emphasis on treatment and less on incarceration, for non-violent drug offenders
- Do pink jail cells calm prisoners and reduce their violent tendencies? (Thanks to Janis Henderson for the tip!)

Overviews of Conducting Program Evaluations

- American Evaluation Association's blog on tips for conducting evaluations
- Article from *The Economist* magazine on program evaluation
- Overview of formative evaluation (getting feedback to improve the process of implementing the program, i.e., helping "form" the program) and summative evaluation (doing a final evaluation of the program's effectiveness, i.e., getting a final "summary"). Nearly all of the examples in our class (e.g., class-size reduction, rowdiness/violence at football games) are *summative*.
- Staff's reflections on conducting a program evaluation (formative evaluation)
- Article on matching for quasi-experimental designs (some parts of the article are beyond the scope of an introductory methods course, but it provides a good overview in the initial sections)

More Advanced Quasi-Experimental Designs for Graduate Students

- **Regression-Discontinuity Design**
  - University of Michigan study looking at criminal "offense rates a handful of weeks before and after the 18th birthday," so that individuals' cognitive and physical characteristics would be similar between groups, but the juvenile vs. adult sentences would differ

- **Natural Experiments** (taking advantage of when exposure to a condition in a real-world setting is done by lottery)
  - In federal appeals-court cases, three-judge panels are selected randomly from a larger number of available judges. Sunstein and colleagues show some interesting effects on judges' voting, based on who they serve with on a case (see Figure 1 of this document).
  - In districts allowing school choice, when a given school gets more applicants than there is room for, the decision of whom to admit is sometimes made via random lottery. Here is one research example stemming from a lottery system.
  - In Olympic "combat sports" (e.g., wrestling, taekwondo), competitors are randomly assigned to wear either a red or blue uniform. The results may have some athletes seeing red!

For an overview, see:

Rockers, P. C., Røttingen, J-A., Shemilt, I., Tugwell, P., & Bärnighausen, T. (2015). Inclusion of quasi-experimental studies in systematic reviews of health systems research. *Health Policy, 119,* 511-521.