PAI 705
McPeak
Lecture 8

Controlled Experiments.

- In the most basic form, an experiment breaks down into two steps:

1) Taking action.
2) Observing (and recording) the consequences of that action.

- Much of what we learn in life is based on some kind of experiment to see action leads to what outcome.
    - Wonder what happens if I lick the metal fence in Syracuse in January
    - My brother and the baseball (in Syracuse in January).
- This is the formal version of that process.
- Experiments are best suited for research topics where there are relatively limited and well defined concepts and propositions.
- It is great for hypothesis testing.
- They are also really good for looking at causation rather than correlation.
- We think of experiments as being in a lab, the fifth grade version of ourselves with a beaker in one hand and a loose leaf sheet of paper with the scientific method outlined in the other.

- This has been adapted to social science analysis to an increasing extent.
- There are also variants of it that are called ‘natural experiments’ where we try to use the sequential / spatial / differential impact of a given treatment in contrast to a control to identify causality.

Classic Experiment Design.

- Elements:
    - Independent and dependent variables defined clearly.
    - Experimental and control groups crisply defined.
    - Pre-testing and post-testing on the different groups, where the experimental group is administered the ‘treatment’ and the control is not.
- The independent variable is the treatment, the stimulus.
- Stimulus has the characteristic of being present or not present.
    - We administered the medicine.
    - We signed them up for the cash transfer scheme.
    - We had them watch a video.
- There are other independent variables (the regressors on the right hand side). Gender, age, education, income, assets, GPS location of their home, occupation, what kind of car they own, whether they have a horse drawn cart, what clan or sect they are in……
- Then we have the dependent variable, the outcome.

    - They got better from the medicine, the control did not.
    - They escaped poverty, the control did not.
    - They had less prejudicial attitudes towards immigrants, the control did not.
- Was the independent variable causative of a change in the dependent variable? That is our question.
- Experiment and control groups. We have a baseline on each.
    - Are they similar in profile making the contrast across the two groups reasonable?
    - What has occurred in the passage of time that impacted the control as well as the treatment that might lead to a change in the dependent variable NOT caused by the treatment, but by what has changed in the 'all else equal'.
    - Perhaps attitudes towards immigrants did not change due to the video but due to media reports of assaults.
    - What is the impact of having them be observed – the Hawthorne effect?   Trying to control for behavioral changes brought about by the mere fact of being observed.
    - Critical in medical research. How much of medicine is in our head? Treatment and 'placebo', the sugar pill. Control for the impact of the chemical compound in the treatment compared to the act of ingesting a pill

– where the psychological effect can lead to perceptions of changed physical symptoms.

- Pretesting and Post testing.
    - Subjects are measured in the dependent variable before the treatment takes place for the treatment sample, and at the same time for the control.
    - Following testing both groups are tested again for the dependent variable.
        - Before and After
        - With and Without

The Classical Experiment ■ 227

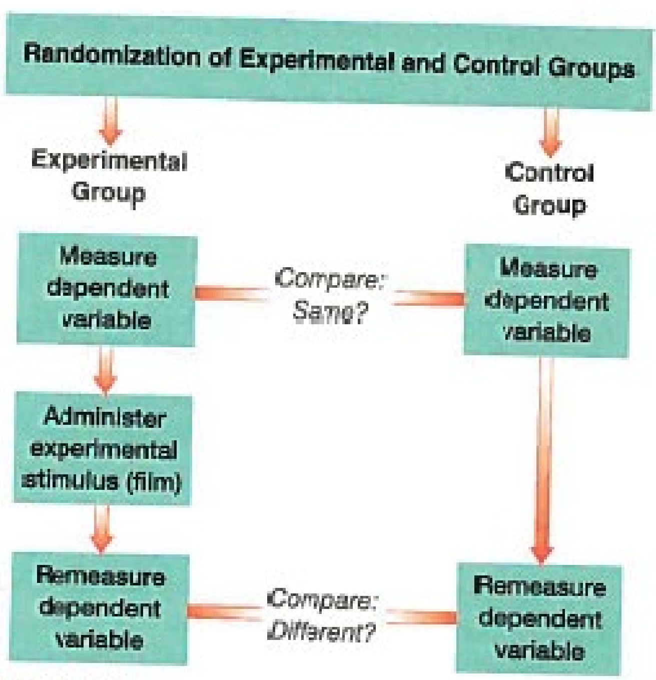


FIGURE 8-1

Diagram of Basic Experimental Design. The fundamental purpose of an experiment is to isolate the possible effect of an independent variable (called the *stimulus* in experiments) on a dependent variable. Members of the experimental group(s) are exposed to the stimulus, whereas those in the control group(s) are not.

Double Blind Experiment (note in passing double blind review).

- Control for the fact that not only may experimental subjects be swayed by the act of treatment, but that the experimental research team may give off clues to the treatment group to undermine the integrity of the treatment and control sample.
- Neither the researcher nor the subject know whether the subject is in the treatment or control sample at the time of the treatment.
- Control for issues like:
  - Researcher asks the treatment group a detailed list of symptoms, the control a much less detailed list.
  - Research asks leading questions to the treatment, not to the control.
    - Researchers may have an interest in the outcome, and even if they try to hide it may not be able to be completely unbiased.
    - If we don't find a significant treatment effect, we can't publish. If we cannot publish, the lab will not get more money / I will not get tenure.

Selecting Subjects.

- Research at a university, it is natural to use students as a population from which to draw research subjects.
  - Are students at a university representative of a larger population?
  - In what ways might they not be representative?
- In some cases, it might not matter that much.
  - Does the treatment lead to a change? The interest is in the change, the delta Δ.
    - The level of variables at the start might be off from the overall population, but the size and significance of the Δ is what we are after.
      - Is the nature of the Δ different in the overall population however?

Probability Sampling

- Each sample (treatment and control) represents the overall population. As such, each resembles the other, at least at some aggregate level.
  - Note the degree to which a sample represents the population has something to do with sample size.
  - Given that samples of less than 100 are not likely to be representative of populations we are interested in for social science research (note this is a ballpark figure- the representativeness of n depends on N), we don't tend towards probability sampling in experiments in the strict sense of an experiment for social science purposes.
    - It is too expensive and there is generally a lower cost, acceptable alternative.

Randomization.

- We may want to give up on the idea of the treatment and control sample representing the larger population.
- We may be more concerned that there is no systematic bias in selecting who is in the treatment group and who is in the control group.
- List everybody who showed up for the trial. Select them:
  - Flip a coin.
  - Odd – even
  - Random number table / excel random numbers.

- For the purposes of the experiment, both samples (control and treatment) are representative of who showed up for the experiment, not necessarily the population. That may be OK for the purposes we have in mind. But there is an element of self-selection to consider at the initial level.

Matching

- If household A is in the treatment, we look for a household with the same general profile to be in the control.
    - Stratify by size, education, asset level, income diversification profile, gender composition, health history….
    - Note that was the example in the treatment and control for the Borana sites and the Guji sites in the earlier lecture with the conflict map from Ethiopia.
        - Made it not representative of Borana overall, but made it so the treatment in Borana was compared to a similar set of households in the control in Guji.
    - Inventory your volunteers, form pairs, randomly assign one to the treatment and one to the control.
    - Treatment group and control group look about the same for all 'relevant' characteristics.
    - 'Relevant' characteristics reflect your 'priors'. Priors are your assumptions about what will matter

for determining the impact of the treatment on people in the treatment group.

- So if 'prior' indicates gender, age, and race might influence the size of the impact, we can use matching at the sub-group level.

Selecting Subjects ■ 231

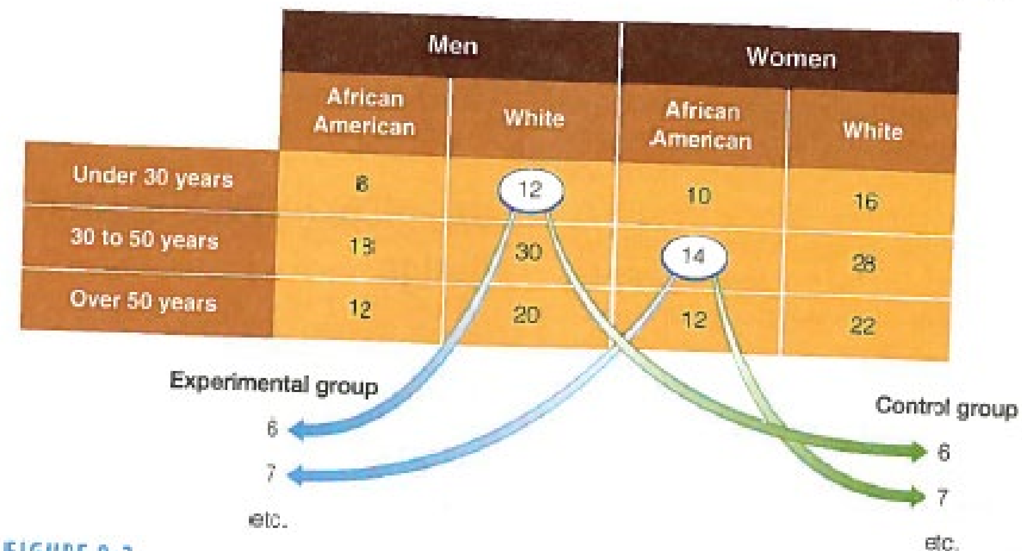


FIGURE 8-2

Quota Matrix Illustration. Sometimes the experimental and control groups are created by finding pairs of matching subjects and assigning one to the experimental group and the other to the control group.

© Cengage Learning®

Matching or Randomization.

- Why is randomization better?
  - First, who knows if you priors are right? Is your theory right about what matters in terms of characteristics that shape the outcome?
    - Again, how do you know *ex ante* – it is a guess and it might be wrong as illustrated by the evidence you gathered.

    - Second, statistical inference is based on random sampling from a large population. Matching is not the basis of the basics of probability theory.
- Why might you still go with matching?
    - Defining a matching sample costs a lot less. I need a one to one match. Randomization requires large numbers to appeal to the law of large numbers.
        - Large numbers cost large money.
- Mixed method noted in the book. A stratified sampling procedure for an experiment, with some matching up front on strata, then randomization within.

Variations on Experimental Design

- Pre-experimental Research Design. Pre-experimental in the sense of people use these in a quasi-experimental way, but they are not up to the standards of pure scientific inquiry.
    - They may have some merit at least as far as adding <u>some</u> information.
- When we talk about this class being designed to help you develop discernment over what is pretty good research and what is questionable research, these are good examples of the kinds of things we want you to be on the lookout for in your careers going forward.
- One-shot case study. The researcher studies one group of subjects on a dependent variable following the administration of some treatment stimulus.

    - I treated a class of MAIR students to a stimulus of super bowl ads.
    - I administered a survey that captures psychological characteristics such as hopelessness, despondency, and despair.
    - I found abnormally high readings in the sample for symptoms of hopelessness, despondency, and despair following viewing of the ad.
    - I conclude the ad leads to these emotions.
        - I then consider that midterms were coming up, it was February in Syracuse, and people are starting to turn their thoughts to the job market.
            - However, now, flawed though my method was, I at least have a competing hypothesis that I can test with a better study.
- One-group pretest-posttest design.
    - Now I do have a ‘before’ and ‘after’ sampling frame (as compared to the last example where I only had an ‘after’ sampling frame).
    - However, I remain with the approach that I am looking at the ‘before’ and ‘after’ for the treatment group without having some kind of control sample in place.
    - I am caught in the potential trap of using ‘before’ and ‘after’ on the treatment group as a proxy for ‘with’ and ‘without’ which is the underlying concept with treatment and control design.

- I run a baseline survey of households' reports on the amount of grain as measured in months of home consumption they have stored over the past year.
- I also ask them to report their resilience on a Likert like scale.
- From this year to next year, my NGO supports the establishment of a grain storage facility in the community.
- Next year, I run the same set of questions about grain storage and resilience. They have 62% more in terms of months covered and an increase of 0.2 points on the resilience Likert scale.
    - I go to USAID and with great fanfare announce my fantastic development success.
    - The agricultural expert at USAID points out that the year of the baseline was a below average rainfall year and this year is an above average rainfall year.
    - Grain harvests are 70% higher in the area in which I am working.
    - My grain storage project actually made them sell grain to cover storage fees at the facility so rather focusing on the 62% increase, USAID asks why my 'beneficiaries'

are 8% behind the average for this area (assuming such information / estimates exist and are accurate).

In the BRACED example (Mali)

	months food	resilience
Treatment 2017	11.73	2.65
Control 2017	11.32	2.59
Baseline 2015	7.25	2.45
T test DiffBC	***	*
T test DiffBT	***	***
T test DiffTC	***	

***=1%, *=10%

- Static-group comparison.
  - This does have a treatment and control, but no baseline.
  - So consider a micro-finance example.
  - I come in and do the treatment (microfinance) in one village, call is Sare Balde.
  - In the next village, called Sare Sintian, I do not have a micro-finance intervention.
  - I survey households in Sare Balde (treatment) and Sare Sintian (control) one year after the microfinance institution opens its doors in Sare Balde.
  - In this survey, I find:
    - Child malnutrition in Sare Balde (T) is 45% of the level found in Sare Sintian (C).
    - Cash incomes in Sare Balde (T) are 124% of the cash incomes found in Sare Sintian (C).

        - Maternal mortality rates in Sare Balde (T) are 78% of those found in Sare Sintian (C).
        - Primary enrollment levels in the treatment site of girls aged 5 to 13 are 146% of those found in Sare Sintian (C).
    - Again, I go rocketing up the road to trumpet to USAID my amazing success.
        - The meeting takes an unforeseen turn when the USAID field officer points out that my treatment site is on a paved road with a weekly market, a health center, a maternity, and two primary schools.
        - My control site is on a dirt road which is not accessible in much of the rainy season (which is why I did not go there to do the microfinance work) with no health center, no maternity, no primary school, and no weekly market.
            - My control site residents access all these things in the treatment site which is a 6 kilometer walk away.
        - The question arises, what was the comparison in child malnutrition, cash incomes, maternal mortality rates, and enrollment levels <u>before</u> the intervention if these two sites are compared.
            - I could actually have made things worse if the relative advantage of Sare Balde was greater before the intervention than after.

        - My survey approach does not allow me to appeal to evidence to answer this question.

Validity Issues in Experimental Research

- Problems of Internal Validity.
    - The conclusions drawn from the experimental results may not actually reflect what went on in the experiment itself.
    - What might confound our internal validity?
        - External events in the political / historical domain that might interfere with our results. Exogenous factors outside of our experimental design.
            - Unless we somehow sequester our survey subjects, events exogenous to our experiment may impact variables of interest.
                - North Korea launches a missile.
                - Another NGO shows up in Sare Balde and starts distributing unconditional cash transfers.
        - Maturation / passage of time.
            - The kids that we are monitoring were young and eating the foods mom prepared in the 'before' survey.

- Our health intervention was to extend techniques of cooking vegetables to moms to maximize nutrient retention.
- Now the kids we are monitoring have learned to cook for themselves over the study period, and are surviving on hot dogs, ramen, and peanut butter sandwiches.
- The kids have gone from healthy to pale shadows of themselves. Our intervention has failed?
    - Their younger brothers and sisters eating what mom makes are qualifying for the junior Olympics.

- Testing.
    - Our pretest was a survey of household income and asset levels. The people in the community only knew we were there to do a survey.
    - After the survey was run, we explained that the survey results were to be used in a conditional cash transfer program, where cash transfers are targeted at the lowest income and lowest asset households.
    - In our repeat survey, we find that the entire community has completely collapsed in terms of reported assets and incomes. When pressed, households report livestock

        assets were sucked up in the air during a windstorm never to be seen again, and mysterious ‘glowing night locusts’ ate all of the crops in fields.
            - They then ask when the targeted cash transfers will be available for them in their newly destitute, asset poor, income disappeared status.
            - I look at a herd of cattle happily munching stalks from harvested fields and wonder what in the world went wrong with my carefully designed, well-targeted, cash transfer scheme?

- Instrumentation.
    - In my treatment site, I ran a survey pre-intervention to record cash income levels in an agro-pastoral context.
        - It was a not so good rainfall year. The maize harvest failed and milk supplies from the animals are relatively low. They sold many goats to buy food so cash income was capturing these sales.
    - My intervention is to train people in fattening goats to sell in the market at higher prices than they get for unfattened goats.

- I run my post intervention survey to record cash-incomes following the goat fattening training.
  - It was a good rainfall year, so people are harvesting maize from their fields and eating it so they don't have to sell animals to buy grains.
  - Average income as measured by cash has gone down dramatically following my intervention designed to increase their cash revenues from livestock sales.
  - What is wrong with my instrumentation of cash income as a measure of well-being?

▪ Statistical Regression.

- Statistically, there is an idea of 'regression to the mean'. An outlier, or a value outside one or two standard deviations from the mean is probabilistically likely to score closer to the mean in the next round of observations since the likelihood of being way out in the 'tail' is pretty small.
- Boru and Galgallo have about the same size camel herd. Boru's camel herd has milk production that is far below the mean for

that size herd, which is represented by production from Galgallo's herd.

- I suggest Boru buy anti tick medicine, deworm his animals, buy some salt licks and vaccinate his camels. Boru does.
- I come back in 9 months and see Boru's herd has now the same milk production as Galgallo's.
- I quickly sit down and fire off a research brief for USAID to demonstrate the impact we are having on livelihoods with our livestock input package.
- Boru and Galgallo sit under the tree and discuss whether to tell me that Boru had a string of bad luck just before the baseline where his female camels either did not conceive or miscarried. That patch of bad luck has passed and now things in Boru's herd and Galgallo's herd were about the same in terms of female camel conception and birth.
    - Boru is annoyed that I recommended he buy all this stuff that did nothing; all that was needed was the passage of time for the patch of bad luck to pass.

▪ Selection bias.

- I have come up with a great idea for conflict resolution. It involves trust exercises and making people from different sides of a conflict work together to site and hang a tire swing.
    - I am going to take existing environmental management committees (EMCs) that I have been working with for the past five years and ask them to tackle issues of conflict management.
    - The EMC from area A is going to meet with the EMC from area B.
    - They will go through my trust exercises and tire swing treatment.
- As a control, I am going to conduct conflict assessments in areas C and D which are also areas where there is the same kind of conflict that A and B find themselves in.
- A is the same ethnic group as C and B is the same ethnic group as D.
- Nobody has worked in areas C and D with environmental management committees over the past 5 years.
- After my trust exercises and tire swing adventure with A and B conflict is reduced between A and B.

        - Over the same time period, conflict is not reduced between C and D.
        - I once again start firing off a research brief to USAID on my success, and ask for funding in the coming year for more ropes and tires for further conflict reduction training.
            - What did I perhaps miss?
    - Experimental mortality, attrition.
        - A kind of selection bias.
        - My treatment site is a rural community that has a long history of labor out-migration for the literate and numerate population.
        - My control site is a community that has very little labor out-migration and is not well connected to these networks.
        - My training is numeracy and literacy to help create conditions for local small enterprise development.
        - I am not able to find half of my treatment sample when I do the follow up since many of the households migrated out of the community and found work in distant urban areas.
        - My control community has set up more small businesses in their community than the treatment has in their community, even

though I did not train for numeracy and literacy in the control site.

    - The control sample appears to be doing better than my treatment sample in spite of the fact that I had a successful numeracy and literacy program in the treatment but not the control.
        - In the treatment site has the numerate / literate population moved out of the community and are out of the post-test sample?
        - Also the control sample is different from the treatment sample in terms of the connection to outmigration networks. The control may invest more locally.

- Another issue for the general topic is that mortality could be an issue when your treatment and control samples are elderly.
- This problem becomes more pronounced the longer the time horizon of the repeated observations / time gap between observations. https://www.bls.gov/nls/

- Demoralization.
  - Index Based Livestock Insurance. Much fanfare, people talking about it, here I come to target your community for survey work to identify the impact.
    - Bad news, people. You are the control, not the treatment. You won't get any support in getting access to the insurance. We are interested in you to see what happens in the absence of insurance. Over in the treatment sites, we are doing skits, passing out comic books, giving out discount coupons, having field days,…
    - For the next three years, I am going to show up and ask you questions for four hours on your livelihood and management practices.
    - Maybe in year four we will be able to bring the insurance here if it proves to be effective.
      - If it does not, we won't get funded and there will be no insurance.
    - Do you want to spend quality time with me and answer all kinds of

intrusive questions over the next few years with no reward here while your neighbors get this shiny new product?

      - Could I measure your kids upper arm circumference and head circumference please?

  - Proper experimental design gives you some ability to control for all these different threats to internal validity.
  - A lack of awareness / concern about these kinds of issues is not uncommon, and can undermine your research.

- Sources of External Validity.
  - How generalizable are the research results to the real world?
  - An experiment is a contrived, controlled world for the purposes of the experiment. What does that tell us about behavior and outcomes in the much messier world as people live it outside the confines of the experiment.
  - Recall my earlier presentation of the sources of knowledge about the Index Based Livestock Insurance Product. One of the responses was 'the survey'.
    - I am testing the impact of index insurance, but one of the main ways they learn about index insurance is us asking about it every year in our baseline and repeat survey.

    - This very act contaminates the external validity to a degree I need to think about.
        - If I introduced index insurance to a population where I was not doing the baseline and repeat survey, would it have the same outcome?

236 ■ Chapter 8: Experiments

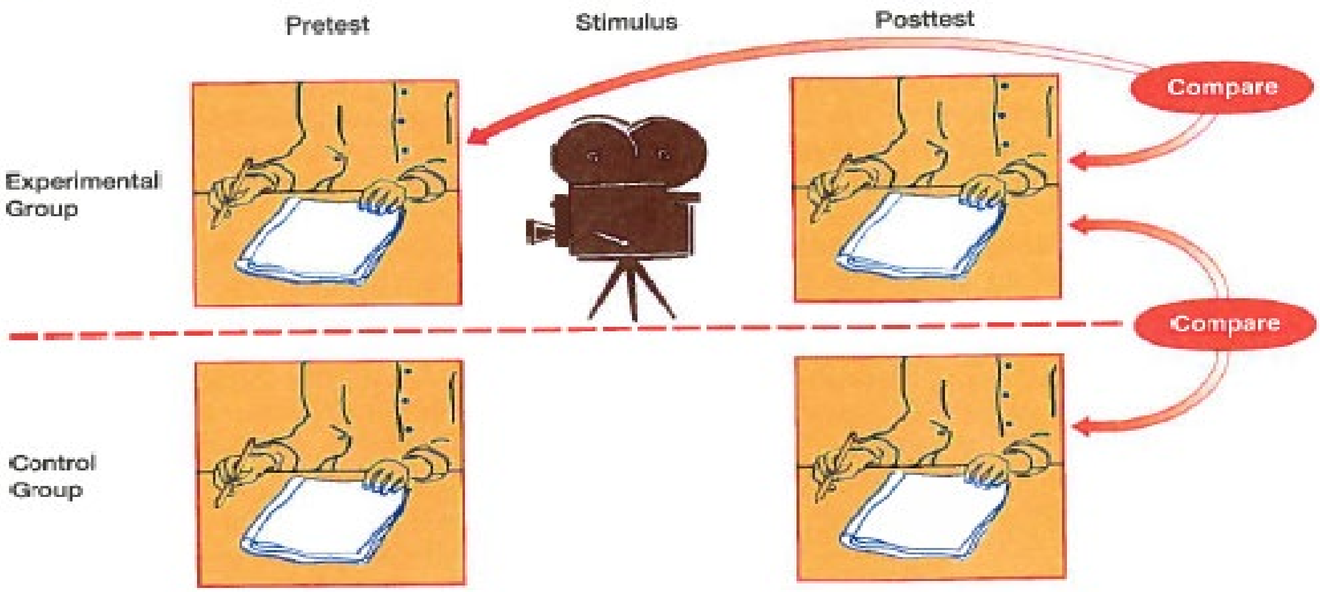
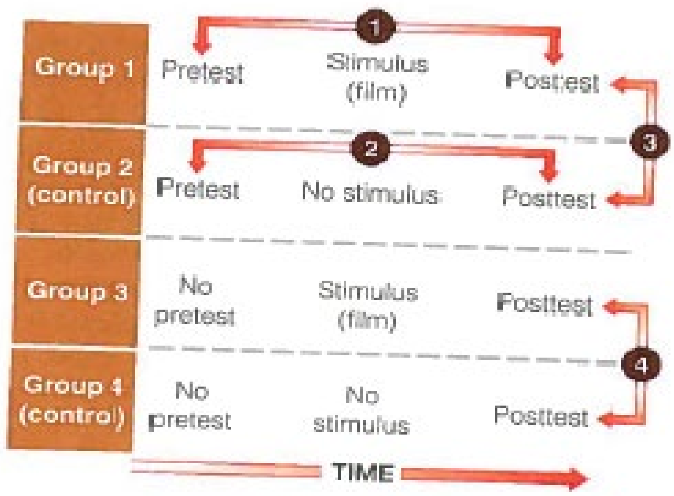


FIGURE 8-4

The Classical Experiment: Using a Muslim History Film to Reduce Prejudice. This diagram illustrates the basic structure of the classical experiment as a vehicle for testing the impact of a film on prejudice. Notice how the control group, the pretesting, and the posttesting function.

- Solomon four-group design as a way to deal with testing interacting with the treatment stimulus.



Expected Findings

1. In Group 1, posttest prejudice should be less than pretest prejudice.
2. In Group 2, prejudice should be the same in the pretest and the posttest.
3. The Group 1 posttest should show less prejudice than the Group 2 posttest does.
4. The Group 3 posttest should show less prejudice than the Group 4 posttest does.

FIGURE 8-5

The Solomon Four-Group Design. The classical experiment runs the risk that pretesting will have an effect on subjects, so the Solomon four-group design adds experimental and control groups that skip the pretest. Thus, it combines the classical experiment and the after-only design (with no pretest).

  - Randomly assign the subjects to four groups.
    - 1 and 2 are classic treatment and control.
    - 3 and 4 are a variant of treatment and control, but without the pretest.
    - In the figure, comparison 4 is an attempt to control for any impact the pretest may have

had via the treatment on the posttest outcome.

A variant of this is to just look at groups three and four, in what is called the posttest-only control-group design.

- You do away with groups one and two.
- As noted above, that could be a problem if group 3 differs from group 4 in systematic ways.
- However, if assignment is truly random to groups 3 and 4, this may not be an issue of major concern.
- Note that will be a problem in some of my 'village A and village B' kinds of example above unless you want to sponsor mass dislocation and random assignment to villages!
- There is a public good nature of some interventions that make it hard to randomly say in village A household 1 will benefit from the improved sanitation and household 2 will not while in village B household 3 will benefit while household 4 will not.

Alternative Experimental Settings.

- Factorial designs when you have more than one experimental variable.
    - Asset Insurance, Income transfer: both, neither, one, the other.
- Natural experiments.
    - Earthquake, hurricane, nuclear meltdown, drought, political disruption, policy change in one jurisdiction and not in another (classic case of minimum wage in NJ / Penn.).