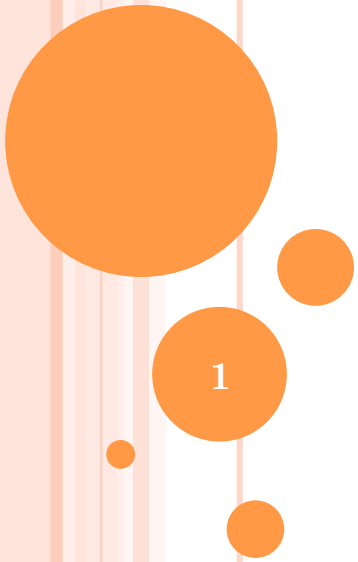
EITM



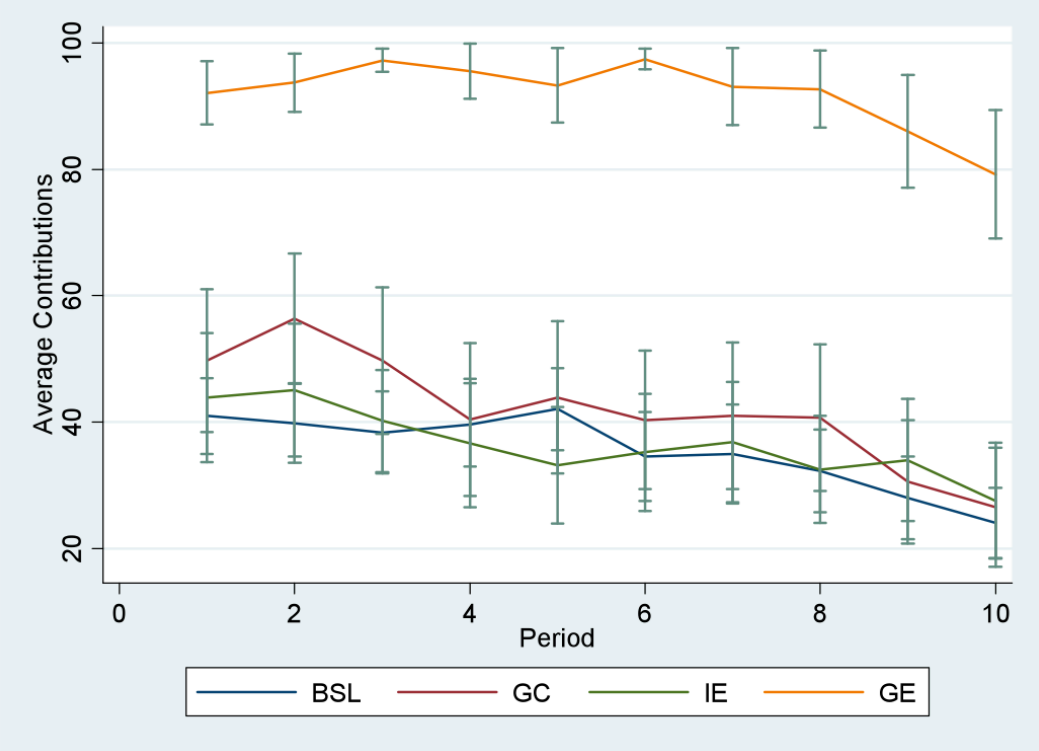
Experimental Design

AND NOW FOR SOMETHING COMPLETELY DIFFERENT

A WELL PLANNED EXPERIMENT ONLY NEEDS

$$\mu_T \neq \mu_C$$

OR



What Is an Experiment?

- Definition: A test of a hypothesis or demonstration of a fact under conditions manipulated by a researcher.
- Key elements:
  - Control, control, control
  - Simplify, simplify, simplify
  - Randomize, randomize, randomize

OBJECTIVES OF EXPERIMENTS

- Testing theories
- Establish empirical regularities as a basis for new theories
- Testing institutions and environments
- Policy advice and wind-tunnel experiments
- The elicitation of preferences
  - Goods, risk, fairness, time

What an experiment doesn't do

- Substitute for thinking
- Generate hypotheses
- Not an all-purpose tool

Objectives I: Theory and experiments

- Test a theory or discriminate between theories
  - Formal theory provides the basis for experimental design
  - Test a theory on its own domain:
    - Implement the conditions of the theory (e.g., preference assumptions, technology assumptions, institutional assumptions)
  - (Best to have an alternative hypothesis)
  - Compare the prediction(s) with the experimental outcome

THEORY, CONT'D

- What if the results reject theory?
- Explore the causes of a theory's failure
  - Check each of the assumptions
  - Explore parameter space
  - Find out when the theory fails and when it succeeds
  - Design proper control treatments that allows causal inferences about why the theory fails

Objectives II: Establishing Regularities

- Finding patterns
  - Lab is inexpensive
  - Manipulations are easy
- Pinpointing effects
  - GOTV, Interventions
- Fishing …
  - … in the right Lake

Objectives III: Institutions and environments

- Compare environments within the same institution
  - How robust are the results across different environments?
- Compare institutions within the same environment
  - Allows for comparisons even when no theory about the effects of the institution is available (Example: cheap talk vs. face-to-face communications in PG)
  - Usual aggregate welfare measure: Aggregate amount of money earned divided by the maximum that could be earned
  - Comparative statics in agenda setting

Objectives IV: Policy and wind-tunnel experiments

- Evaluate Policy Proposals
  - Does the reduction of entry barriers increase aggregate welfare?
  - Which auctions generate the higher revenue? (e.g., in arts auctions or broadband license auctions)
  - Do emission permits allow efficient pollution control?
  - How should airport slots be allocated?
- The laboratory as a wind tunnel for new institutions
  - What is the distributional consequence of eliminating the filibuster?
  - Do "at-large" or "single-member" districts leads to increase minority representation?

Objectives V: The elicitation of preferences

- Inform Policy
  - How much should the government spend on avoiding traffic injuries?
  - How much should be spend on the conservation of nature?
- Measuring people's values is hard
  - Are people risk seeking/averse?
  - Who is cooperative?
- Requires a theory of individual preferences and knowledge about the strength of particular "motives" (preferences).

Objections to experiments

- **Objection: *"Experiments are unrealistic."***
  - All models are unrealistic
    - They leave out many aspects of reality.
    - Simplicity is a virtue – focuses on critical aspects of a situation (a causal mechanism or logic of a complex relationship)
  - Experiments are like models
    - They leave out many aspects of reality
    - Focus on critical aspects (cause or precision of estimate)
  - Realism may be important but so is control.

Objections continued

- **Objection: "*Experiments are artificial.*"**
  - Biased subject pool (students)
  - Low stakes
  - Small number of participants
  - Inexperienced subjects
  - Anonymity
- All can be tested in the lab
- Such testing has never overthrown an important result

Objections continued

- **Objection: "*Experiments say nothing about the real world.*"**
  - External validity
  - Generalization
- The experiment, if properly designed, is "real" for the subjects
- What is the aim of the experiment?
  - Internal validity – ensuring that the causal inference is correct
  - Minimizing general claims

LIMITS OF EXPERIMENTS

- **Control is never perfect**
  - Weather, Laboratory environment
  - No real control about all other motives (no dominance)
  - Self-selection: who takes part in the experiment?
- **Randomization is difficult**
- **Experiments (like models) are never general, just examples**
- **Lab experiments compared to field experiments**
  - Difference in control
  - Difference in randomization
  - Problems with ITT

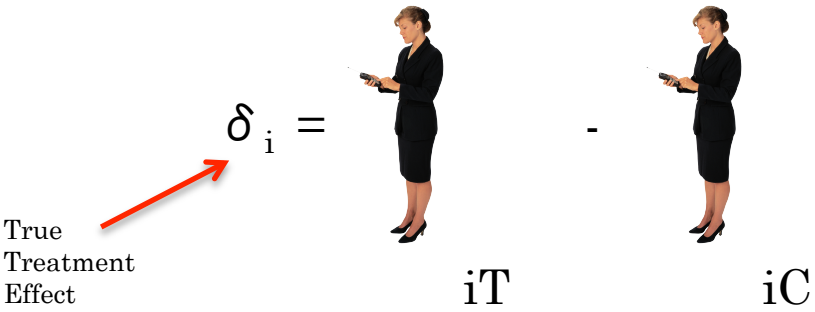
What Experiments Do Well

- Test for causal claims
- Inform theory
- Allow for replication
- Develop measures (problem of reliability and validity)
- Explore parameters of interest
- Develop counterfactuals

Causal Considerations

RUEBEN CAUSAL MODEL (RCM)

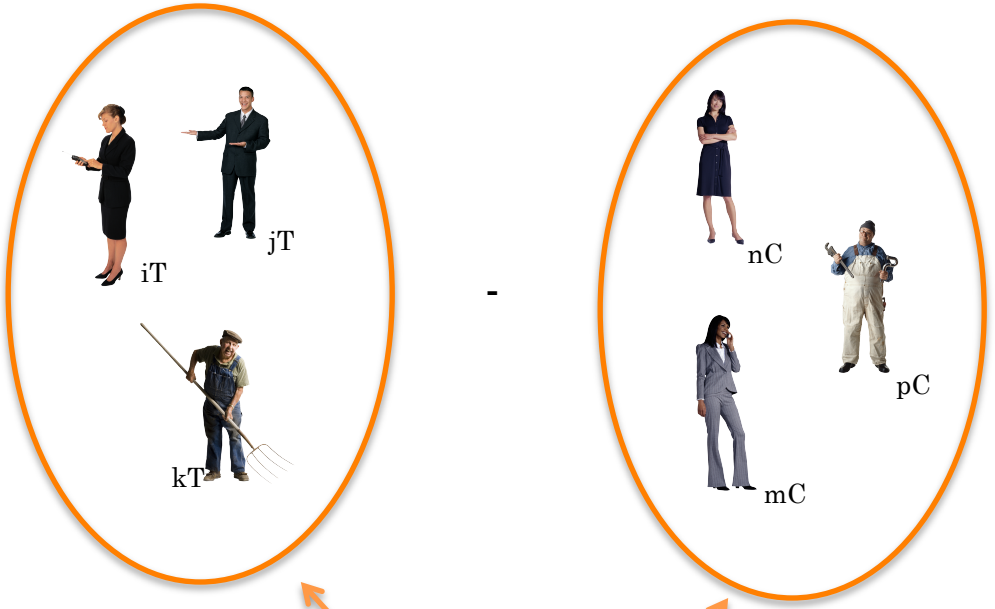
- The dilemma:



- The same "i" can't be in two states at the same time!

RCM – Best Correction

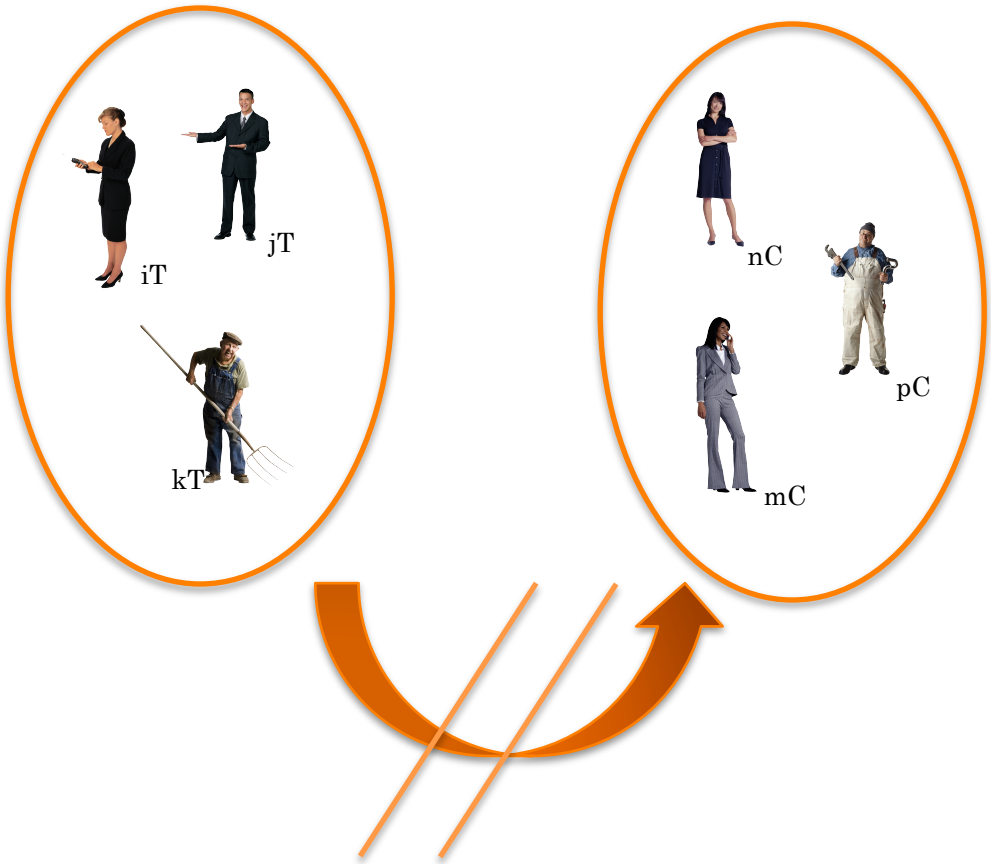
ATE =



Analog: $y = \alpha + \beta_1 X_1 + \varepsilon$

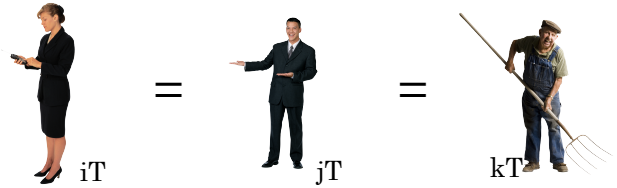
SUTVA – Stable Unit Treatment Value Assumption(s)

Assumption 1: Treatment ONLY affects the treated.



SUTVA – ASSUMPTION 2

Assumption 2: Average treatment effect is homogeneous across individuals

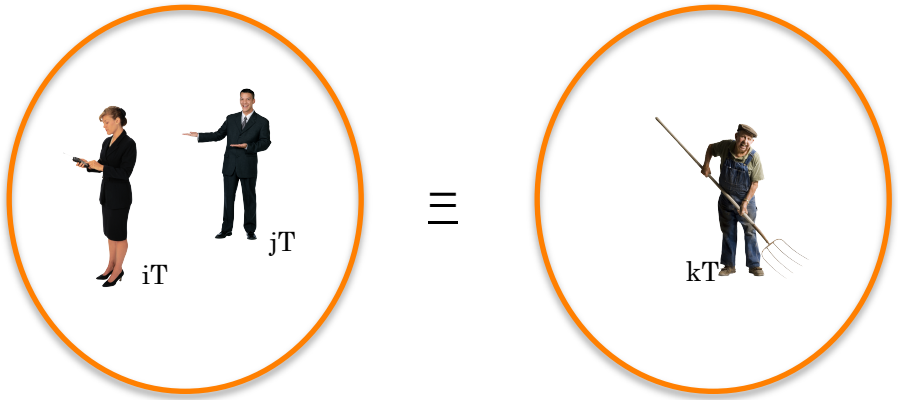


SUTVA – Assumption 3

Assumption 3: Treatment is invariant to manner delivered

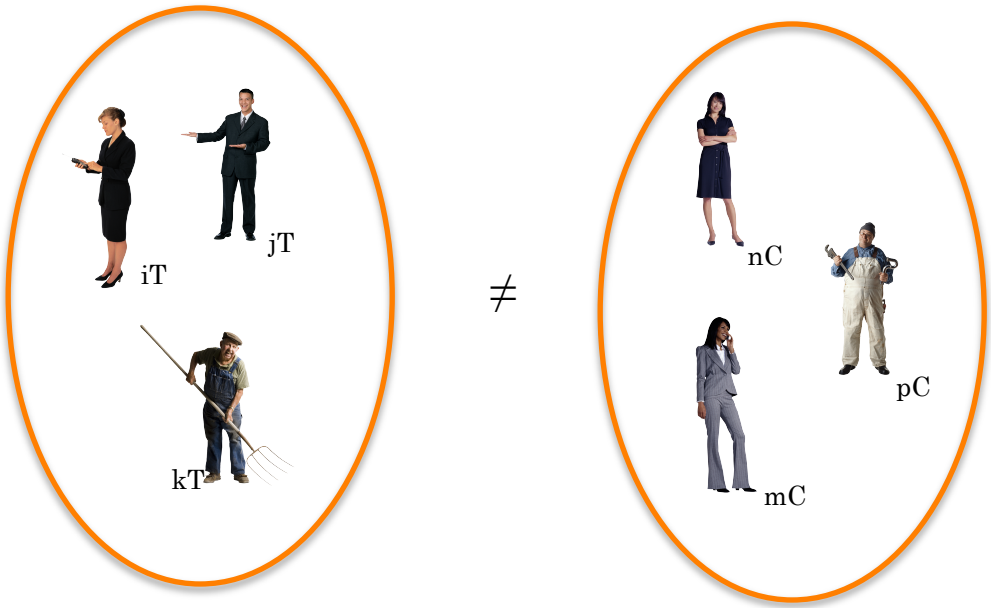
SUTVA – Assumption 4

Assumption 4: All possible states of the world are observed



SUTVA – ASSUMPTION 5

Assumption 5: Causal question of interest is historically bound to the data.

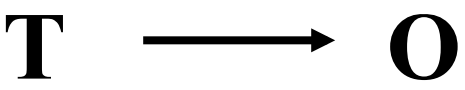


SUTVA – Assumption 6

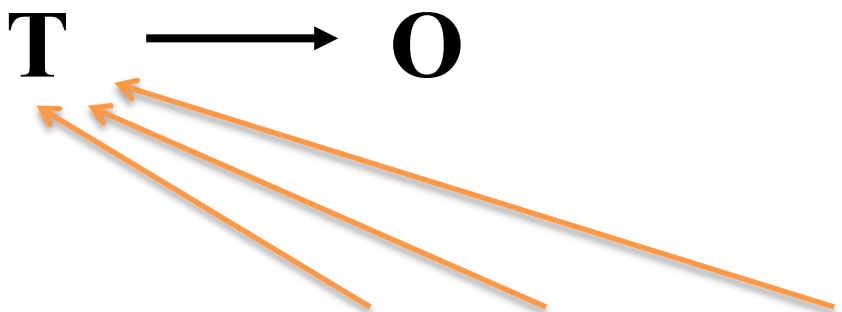
Assumption 6: The treatment precedes the action by subject – no simultaneity

Design Considerations

Common Designs: One-Shot Design



Common Designs: One-Shot Design



Inference: none

Statistics: descriptive or kitchen sink

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

SUTVA violations: Everyone is treated (maybe)?

Pre/Post-Test Design

$$O_1 \longrightarrow T \longrightarrow O_2$$

Pre/Post-Test Design

$$O_1 \longrightarrow T \longrightarrow O_2$$

Inference: Something might have caused a difference

Statistics: $O_1 \neq O_2$

SUTVA violations:

Everyone is treated (maybe)?

All possible states of the world are not observed.

Static Group Comparison

Group A $\quad T \longrightarrow O_{1A}$

Group B $\quad O_{1B}$

Static Group Comparison

Group A $\quad T \longrightarrow O_{1A}$

Group B $\quad O_{1B}$
(Control group?
Nope, not Randomized.)

Inference: Something might have caused a difference

Statistics: $O_{1A} \neq O_{1B}$ $\quad y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$

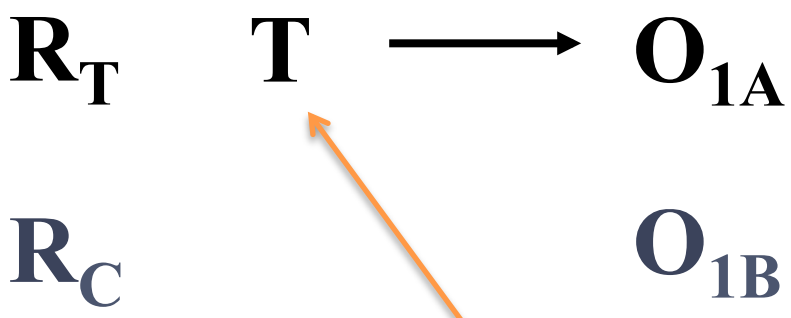
SUTVA violations:
Not clear that treatment only affects the treated
Average treatment effect is not homogeneous across individuals

Randomized Group Comparison

$$R_T \quad T \longrightarrow O_{1A}$$

$$R_C \qquad\qquad O_{1B}$$

Randomized Group Comparison



Inference: Very likely the treatment caused the difference

Statistics: $O_{1A} \neq O_{1B}$

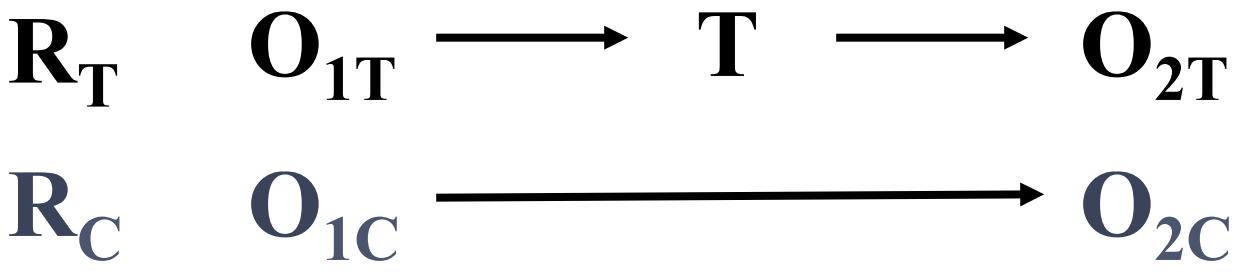
$$y = \alpha + \beta_1 X_1 + \varepsilon$$

SUTVA violations:

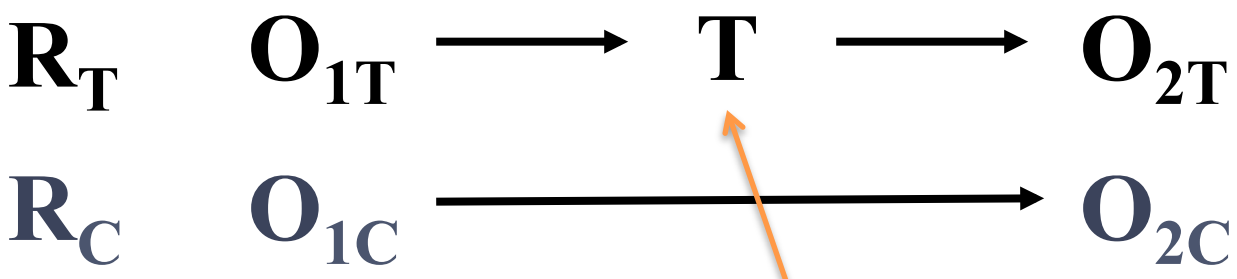
All possible states of the world may not be observed

Historical boundedness.

Pre/Post Control



PRE/POST CONTROL



Inference: Treatment probably caused a difference

Statistics: $O_{1T} = O_{1C}$; $O_{2T} \neq O_{2C}$

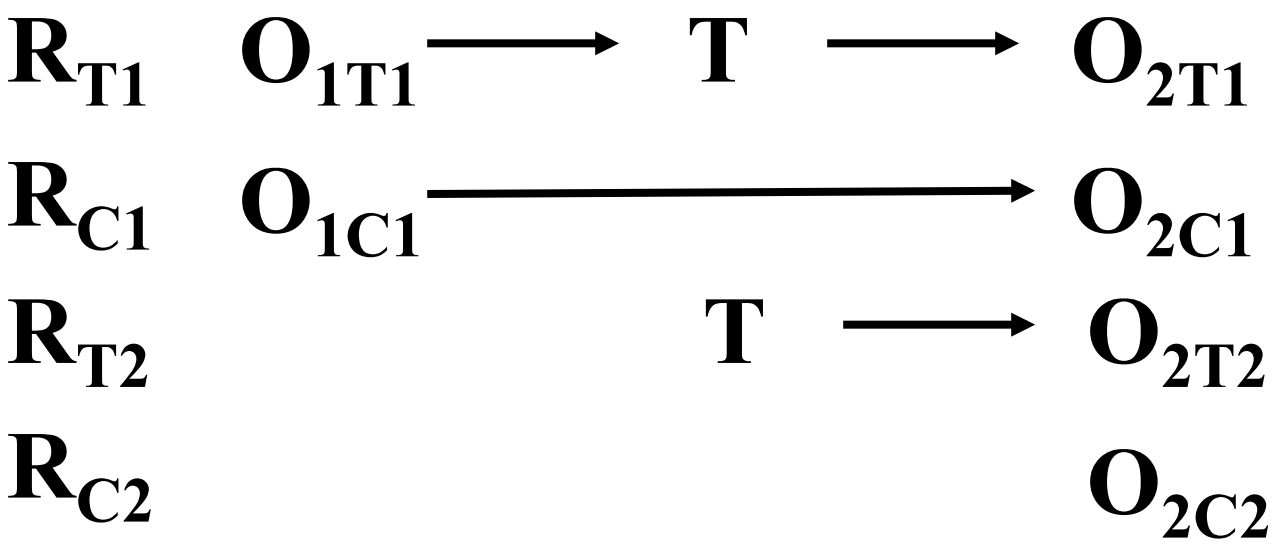
$$y = \alpha + \beta_1 X_1 + \varepsilon$$

where $y = O_2 - O_1$

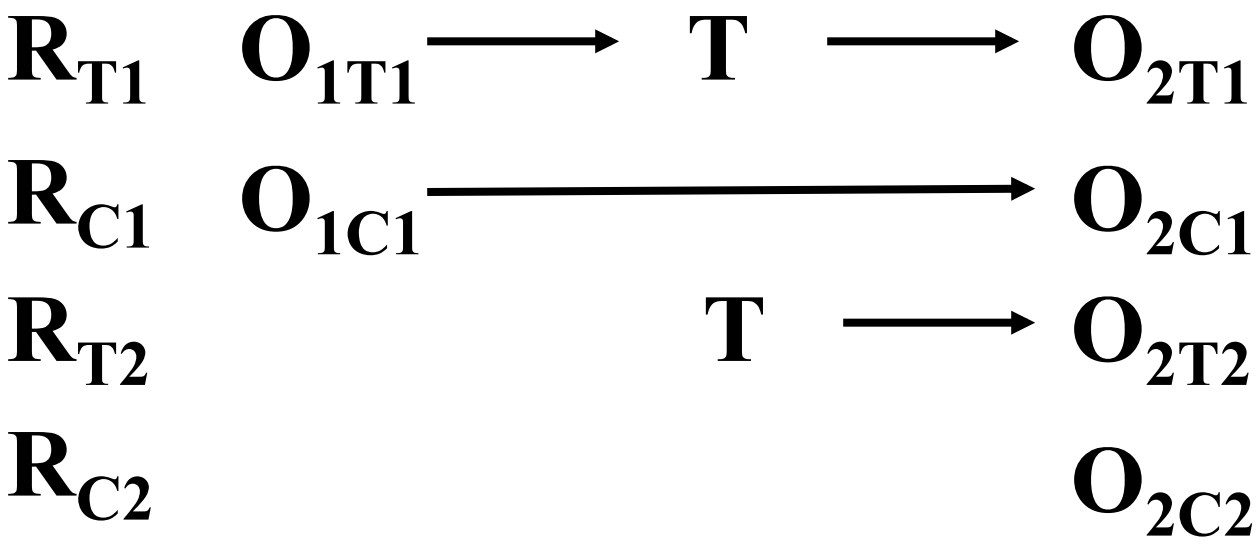
SUTVA violations:

- Treatment ONLY affect the treated?
- Treatment homogeneous across individuals?
- Treatment invariant to delivery method?

SOLOMON FOUR-GROUP



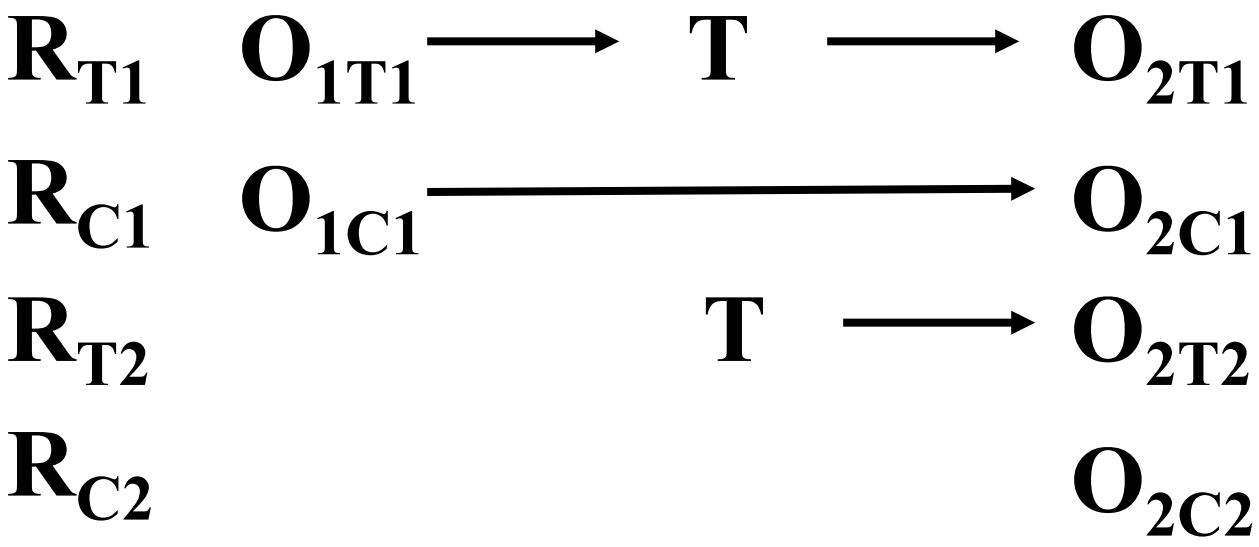
SOLOMON FOUR-GROUP



Inference: Treatment very likely caused a difference

Statistics: $O_{1T1} = O_{1C1} = O_{2C1} = O_{2C2}$;
$O_{2T1} = O_{2T2} \neq O_{2C1} = O_{2C2}$

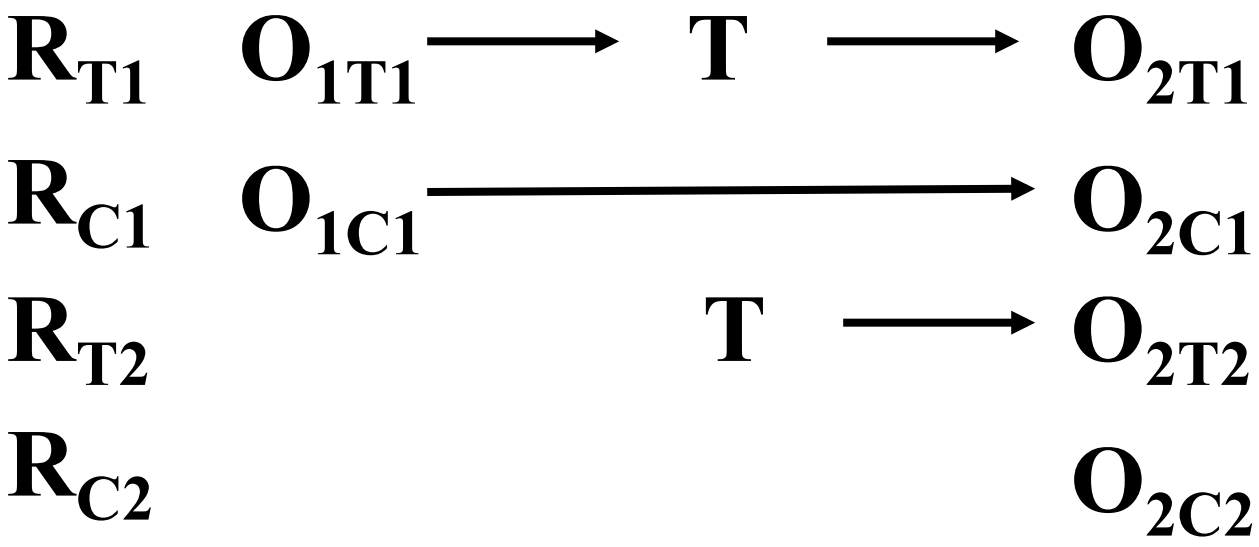
Solomon Four-Group



SUTVA violations:

Treatment ONLY affect the treated?

SOLOMON FOUR-GROUP



NOTE:

This is easy to accomplish in the Lab. It is a nightmare in the field.

Practical Design Considerations

WHAT MUST BE DESIGNED?

- "Laboratory experimental design involves designing a microeconomic system"
  - Vernon Smith, AER, December, 1982
- Environment:
  - Agents (Number, type, motivation)
  - Commodities -- what do decisions get made over?
  - Endowments -- what do the decision-makers have at the outset?
  - Mechanism by which learning can occur (search opportunities, practice)
- Institution:
  - Decisions available to subjects
    - Rules about choices
    - Rules about communication
  - Connection between decisions and payoffs

Fatal Errors in Design

- Inadequate or inappropriate incentive
- Nonstandardized instructions
- Inappropriate context
- Uncontrolled effects of psychological biases
- Insufficient statistical power
- Loss of control due to deception or biased terminology
- Failure to provide a calibrated baseline
- Change in more than one factor at a time

INCENTIVES: INDUCED VALUE THEORY
SMITH (AER 1976; AER 1982)

- In many experiments the experimenter wants to **control** subjects' preferences. How can this be achieved?
- Subjects' homegrown preferences must be "neutralized" and the experimenter "induces" new preferences. Subjects' actions should be driven by the induced preferences.
- Reward Medium: Money
- Assumption: People care about money and some other motives.
  - Note 1: money may function as the "price" of other motives
  - Note 2: sometimes you are interested in "homegrown preferences." But be willing to adjust for heterogeneous treatment effects (Imai et al. APSR 2011). Example: partisan preferences.

Incentives Continued
Minimal Conditions for Control

- **Monotonicity/nonsatiation:** Subjects must prefer more of the reward medium to less and not become satiated.
- **Salience:** The reward depends on a subject's actions (note: show up fee is not salient).
- **Dominance:** Changes in a subject's utility from the experiment come predominantly from the reward medium and the influence of the other motives is negligible (this assumption is the most critical).
- If these conditions are satisfied, the experimenter has control of the subjects' preferences, i.e., there is an incentive to perform actions that are paid.

Incentives Continued: Qualifications

- Subjective costs
  - Solution: a trading commission or raising stakes
- Utility of winning or earning points
  - Usually this is no problem, and may enhance incentives
  - But in others it can look like risk aversion (overbidding in common value auctions)
  - Solution: raise stakes
- Payoffs to others may matter
  - Envy, egalitarianism
  - Solution: do not reveal others' payoffs
- Desire to please the experimenter
  - Solution: conceal the purpose of the experiment
  - Debrief

INCENTIVES CONTINUED: EFFECTS?

Analysis of 74 studies about different topics with no, low and high financial incentives. Camerer & Hogarth, 1999.

	Number of studies in which incentives…		
Experimental job in a particular study:	help	have no effect	are damaging
Evaluation - and decision experiments	15	5	8
Markets and strategic interactions	7	15	0
Individual decision experiments	1	7	1

In 13 studie s there is no efficiency standard, but no effects of incentives.

Incentives Continued: Other considerations

- In experiments in which incentives have an effect, the difference between no and low incentives is often bigger than the difference between low and high incentives.
- Higher incentives often lead to a reduction of the variance of decisions (Smith&Walker, IntJGameTheory 1993)
- Treatment effects are often at least as high as incentive effects.
- Payment of subjects necessary for getting published (Econ -- Poli?).

Uncontrolled Psychological biases

- Loss aversion
  - Avoid losses or zero payoff options
- Status quo bias
  - Avoid accidentally anchoring subjects
  - Experimenter demand: experimenter can accidentally set the status quo by signaling expected behavior
- Endowment effect
  - Willingness to accept v. willingness to pay
- Emotion
  - Ss may vary in their mood

Insufficient Statistical Power

- You must have enough data to do a statistical test
- Plan ahead – decide what test you want to do and run the experiment that will let you do it
  - "Decide what regression you want to run and then design the experiment to give you what you need to run it."
    - Ernst Fehr, January, 2005.
- Avoid too many treatments
  - Complete Factorial Designs
  - (# factors)*(#factors)*(#factors)
- Calculate your power test

Insufficient Statistical Power: Power Tests, I

- Need three elements:
  - Significance criterion – specify the trade off between Type I and Type II errors (both α and β). (Even a Bayesian has to worry about low power for updating beliefs)
  - Magnitude of the effect
    - ATE or LATE: $(MEAN_T - MEAN_C)$
    - Standardized Effect Size (with common variance) $(MEAN_T - MEAN_C)/\sigma$
    - Maximize the expected difference in effects!
    - Pretest Data can inform you about means and variance
  - Sample size
    - Obviously related to the sample error – as sample size goes up, sampling error goes down
    - Measurement precision helps here as well – decrease variance

Insufficient Statistical Power: Power Tests, II

- Many tools available
  - in r: power.t.test(n, delta, sd, sig.level, power, type, alternative) – omit n and it will be calculated.
  - in STATA: sampsi mean1 mean2, sd1(value) sd2(value)
- Examples
  - **Between** Ss

```
. sampsi 5.1 4.3, sd1(3.3) sd2(3.0)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.9000
           m1 =      5.1
           m2 =      4.3
          sd1 =      3.3
          sd2 =        3
        n2/n1 =     1.00

Estimated required sample sizes:

           n1 =      327
           n2 =      327
```

```
. sampsi 5.1 4.3, sd1(1.3) sd2(1.0)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.9000
           m1 =      5.1
           m2 =      4.3
          sd1 =      1.3
          sd2 =        1
        n2/n1 =     1.00

Estimated required sample sizes:

           n1 =       45
           n2 =       45
```

Insufficient Statistical Power: Power Tests, III

- **Within** Subjects

```
. sampsi 4.3 5.1, sd1(4.3) onesample

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    4.3, where m is the mean in the population

Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.9000
 alternative m =      5.1
           sd =      4.3

Estimated required sample size:

            n =      304
```

```
. sampsi 4.3 5.1, sd1(1.3) onesample

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    4.3, where m is the mean in the population

Assumptions:

        alpha =   0.0500  (two-sided)
        power =   0.9000
 alternative m =      5.1
           sd =      1.3

Estimated required sample size:

            n =       28
```

Efficiency in Design: Counter-balance/Within Subject

- Counter-balanced designs
  - $O_1\ T_A\ O_2\ T_B\ O_3$ and $O_1\ T_B\ O_2\ T_A\ O_3$
    - Builds within subject design
    - Decreases the number of trials
    - Accounts for treatment ordering effect
- Cross-over designs
  - $O_1\ T_A\ O_2\ T_B\ O_3\ T_A\ O_4$
    - Accounts for treatment plus learning

Efficiency In Design: Blocking (and Matching)

- Randomized Blocking Designs
  - Suppose a 2(H,L)x2(B,S) within subject design with a 4 game ordering effect (HB, HS, LB,LS).
    - Would require 16 cells under complete factorial design
    - Blocking allows 4 cells: (1) HB,HS,LB,LS (2) HS,LB,LS,HB (3) LB,LS,HB,HS (4) LS,HB,HS,LB
  - Assumptions
    - Blocks must be homogeneous
    - Blocks must be randomly assigned
- Blocking on "nuisance" variables
  - Sex is not randomly assigned

Note: Similar to Imai et al. APSR 2011 "parallel encouragement design"

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

CALIBRATION

- Keep in mind that you are producing a data set
- Include a "baseline" in the experimental design
- Set parameters so you can be sure to tell if hypothesis is supported
- Ideally, you need a competing hypothesis that is "far away" in the design space.
- Try to factor in "noise" in behavior – variability in the performance of the subjects. Lots of noise means results are hard to determine.
- Develop criteria for rejection

Advice on Controls

- Control everything that might be controlled (and collect data on everything that might be uncontrolled)
- Randomize everything else
- Maximize contrasts with treatments (do you really need to use low, medium and high manipulations?)
- If a "nuisance" variable is suspected to interact with a treatment, then make it a separate treatment.
- If a "nuisance" variable is too much of a problem, then make it a constant (blocking) – comparative statics will then play out.

Nuts and Bolts

First Steps (Practical Advice)

- Begin with Theory. Translate theory to lab.
- Begin with phenomenon. Design experiments to dissect
- Begin with something you want to measure. Design experiment to measure it.

Second Steps: (After the Question/Theory)

- Instrumentation
  - Construct Validity – how will I test what I want to test?
  - Paper/Pencil or Computer?
  - Timeline of experiment
  - Instructions
- Sampling/Randomization
  - What subject pool?
  - How will Treatment be randomized?
- Analysis Plan
  - What are the units of analysis
  - Power tests

TimeLine Example – Eckel/Wilson

- Subject Check-in
- General Instructions
- Risk Task (Everyone)
  - Public Officials Risk Choice for Citizens
- Time Discounting Task
- Within Group Trust Task
- Public Official/Citizen Trust Task
- Charitable Giving Task (social distance)
- Charitable Giving Task (social distance + choice of charity)
- Choice of Task to Pay
- Questionnaire
- Payment

Second Steps: (After the Question/Theory)

- Instrumentation
  - Construct Validity – how will I test what I want to test?
  - Paper/Pencil or Computer?
  - Timeline of experiment
  - Instructions
- Sampling/Randomization
  - What subject pool?
  - How will Treatment be randomized?
- Analysis Plan
  - What are the units of analysis
  - Power tests

Subject Selection, I

- Convenience Samples: students
  - Students advantages:
    - Convenient, inexpensive and relatively homogeneous
  - Student disadvantages:
    - May behave differently from target population, young, educated, and talk to each other (diffusion)
  - Classroom:
    - Representative sample of students
    - Environment might affect behavior:
  - Lab:
    - May select certain students
    - Neutral environment
  - Data: Eckel and Grossman ExEc:
    - Students give more to charity in the classroom than in the lab
    - Why?

Subject Selection, II

- Specialized Groups:
  - Elderly
  - Professionals
  - Medical cases
  - Poor
  - Residents of hurricane-vulnerable areas
  - Public officials
- Population Samples
  - Pluses: External validity, Heterogeneity
  - Minuses: Costly, decreased control, heterogeneity

Subject selection III

- Subject selection should suit the question you are asking
- Theory testing:
  - Independent of subjects?
- Measurement for policy:
  - Target group subjects
- Institutional design
  - Targeted participants

Second Steps: (After the Question/Theory)

- Instrumentation
  - Construct Validity – how will I test what I want to test?
  - Paper/Pencil or Computer?
  - Timeline of experiment
  - Instructions
- Sampling/Randomization
  - What subject pool?
  - How will Treatment be randomized?
- Analysis Plan
  - What are the units of analysis
  - Power tests

Nuts and Bolts, I

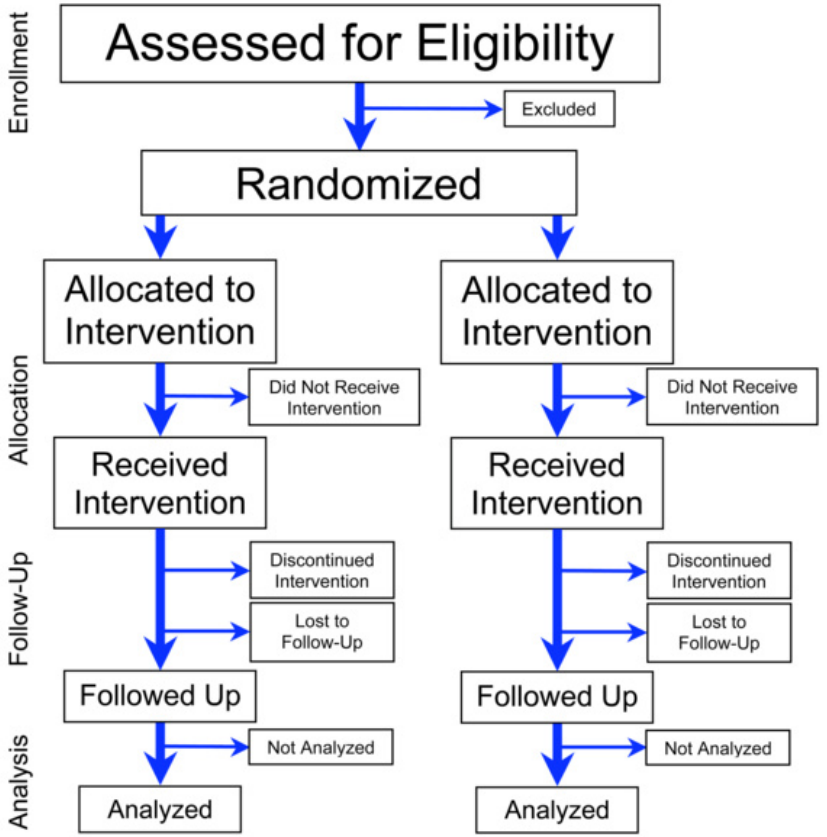
- Lab log.
- IRB and Ethics
- Pilot experiments.
- Lab set-up
- Subject registration
- Experimenter(s)
- Monitor(s)
- Randomizing Devices
- Instructions
- Subject confidence (non-deception)

Lab Book (Lupia & Varian 2010)

- 1. State your objectives.
- 2. State a theory.
- 3. Explain how focal hypotheses are derived from the theory if the correspondence between a focal hypothesis and a theory is not 1:1.
- 4. Explain the criteria by which data for evaluating the focal hypotheses were selected or created.
- 5. Record all steps that convert human energy and dollars into datapoints.
- 6. State the empirical model to be used for leveraging the data in the service of evaluating the focal hypothesis. (a) All procedures for interpreting data require an explicit defense. (b) When doing more than simply offering raw comparisons of observed differences between treatment and control groups, offer an explicit defense of why a given structural relationship between observed outcomes and experimental variables and/or set of control variables is included.
- 7. Report the findings of the initial observation.
- 8. If the findings cause a change to the theory, data, or model, explain why the changes were necessary or sufficient to generate a more reliable inference.
- 9. Do this for every subsequent observation so that lab members and other scholars can trace the path from hypothesis to data collection to analytic method to every published empirical claim.
- ELNs: OneNote in Microsoft or Growlybird Notes for the Mac (http://www.growlybird.com/GrowlyBird/Notes.html)

CONSORT/REGISTRATION

- CONSORT Statement: improve the reporting of a randomized controlled trial (RCT), enabling readers to understand a trial's design, conduct, analysis and interpretation, and to assess the validity of its results.
  - http://www.consort-statement.org/



Nuts and Bolts, I

- Lab log.
- IRB and Ethics
- Pilot experiments.
- Lab set-up
- Subject registration
- Experimenter(s)
- Monitor(s)
- Randomizing Devices
- Instructions
- Subject confidence (non-deception)

Nuts and Bolts, II

- Subject questions
- "Learning periods"
- Experiment
- Recording data
- Termination of experiment
- Debriefing
- Subject payment
- Bankruptcy
- Backup plan

LAB TOOLS

- Handrun experiments
  - Pluses
  - Minuses
- Computerized experiments
  - Pluses
  - Minuses

General Remark

- Whether the conditions implemented in the laboratory are also present in reality will probably always be subject to some uncertainty.
- Therefore, laboratory experiments are no substitute
  - for the analysis of field happenstance data
  - for the conduct and the analysis of field experiments
  - and for survey data.
- We support use of a combination of all these empirical methods.