

# Business Analytics and Data Driven Decision Making

## Session#11: Lecture#22: Predictive Analytics: Experimental Design

### Ravi Vatrapu

*Director, Centre for Digital Enterprise Analytics and Leadership (DEAL)*

*Loretta Rogers Research Chair in Digital Enterprise*

*Professor, School of Information Technology Management*

*Ted Rogers School of Management*

*Toronto Metropolitan University, Canada*

*Founding Director, Centre for Business Data Analytics ([bda.cbs.dk](http://bda.cbs.dk))*

*Professor (on leave), Copenhagen Business School, Denmark*

*Visiting Professor, Indian Institute of Management Visakhapatnam, India*

*Honorary Visiting Professor, GITAM Deemed University, India*

*Adjunct Faculty, Indian Institute of Management Rohtak, India*

Email: [vatrapu@torontomu.ca](mailto:vatrapu@torontomu.ca)

DEAL Website: <https://www.torontomu.ca/tedrogersschool/digital-enterprise-analytics-and-leadership/>

Faculty Webpage: <https://www.torontomu.ca/information-technology-management/faculty-research/ravi-vatrapu/>





# 1. Research Design



# Steps in an Empirical Research

- Formulate a research question
  - Identify and define a research problem: variables and relationship
  - Formulate hypotheses on the basis of theory, research, or both
- Design a research
  - Participants or subjects
  - Procedures
  - Instrumentation
- Conduct the research and collect data
- Analyze the data
- Draw conclusions and interpret the results

# Variable

Name	Definition
Independent Variable Predictor Variable Explanatory Variable (X)	Variable manipulated, measured, or selected by the researcher in order to observe its relation to the subject's response.
Dependent Variable Predicted/Outcome Variable Response (Y)	Variable observed and measured in response to an independent variable.

# Independent or Dependent?

- A vocational trainer assigns students to one of three teaching methods, videotape, lecture, or interactive computer, to measure the influence of teaching method on student achievement. Which are the independent and dependent variables in this study?
- A researcher interested in comparing the empathy of psychologists in different areas of specialization administered a measure of empathy to clinical psychologists and experimental psychologists. Since there is some evidence that men and women differ in empathy, the researcher measured empathy of both males and females.



# Relationships

- Causal relationship

- The independent variable influences the value of the dependent variable.

Changes in  $X$  cause changes in  $Y$ .

- Noncausal relationship

- the independent variable does NOT influence the value of the dependent variable, but they are correlated.

Both  $X$  and  $Y$  respond to changes in some unobserved variable.

# Causal or Noncausal?

- Ice cream sales and shark attacks.
- People who attend a GRE training program obtained higher GRE scores than the people who did not attend the GRE training program.
- The more the cavities and the more children's vocabulary.
- The harder you work for a class, the better score you will get.
- Every time I eat chocolate, it gives me acne.
- People's height is strongly associated with people's salaries.



# Validity of Research

- **Internal Validity:** the extent to which the interpretation of a research study follows from the study itself.
- **External Validity:** the extent to which the findings of the particular study can be generalized to other people or other situations.

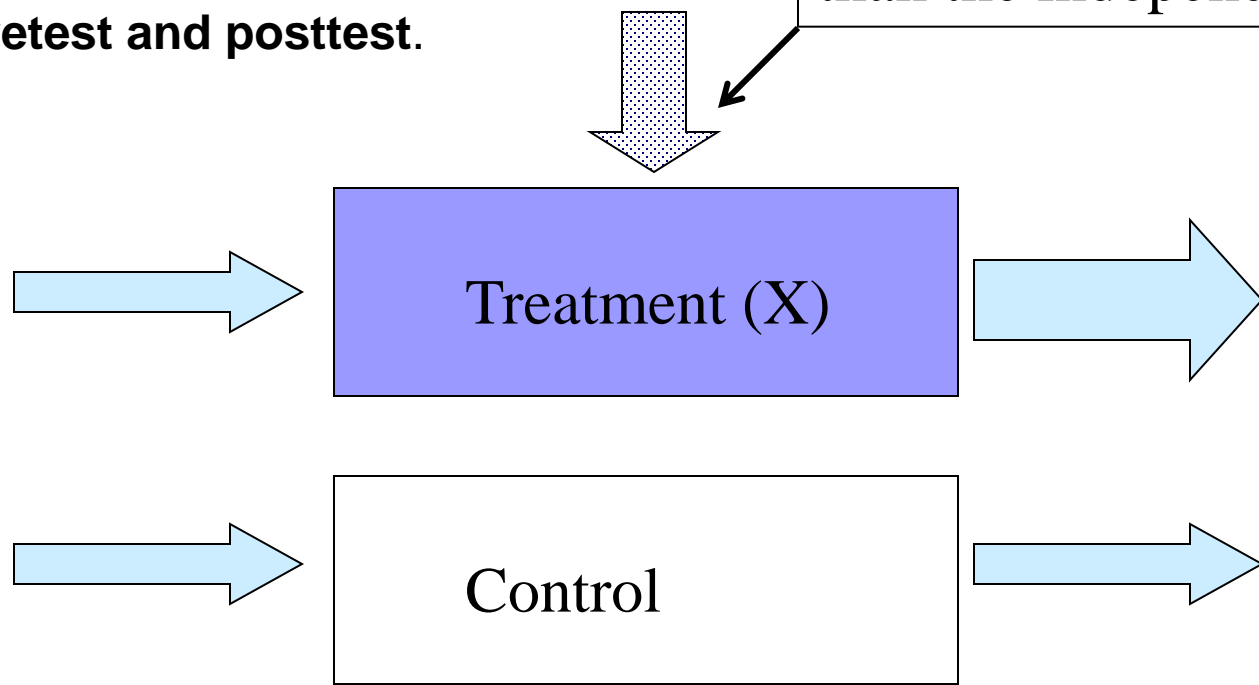
# Threats to Internal Validity

- **History** — something co-occurring with the treatment caused the outcome.
- **Selection** — the differences in outcomes existed before the treatments were given
- **Maturation** — maturation, not the treatment, caused the outcome
- **Mortality/Attrition** — loss of poorly performing subjects from a group caused the outcome
- **Statistical Regression** — extreme groups are likely to improve on retesting
- **Testing** — pretest cued subjects to outcome measure
- **Instrumentation** — outcome measure not reliable, valid, or both

# History Threat

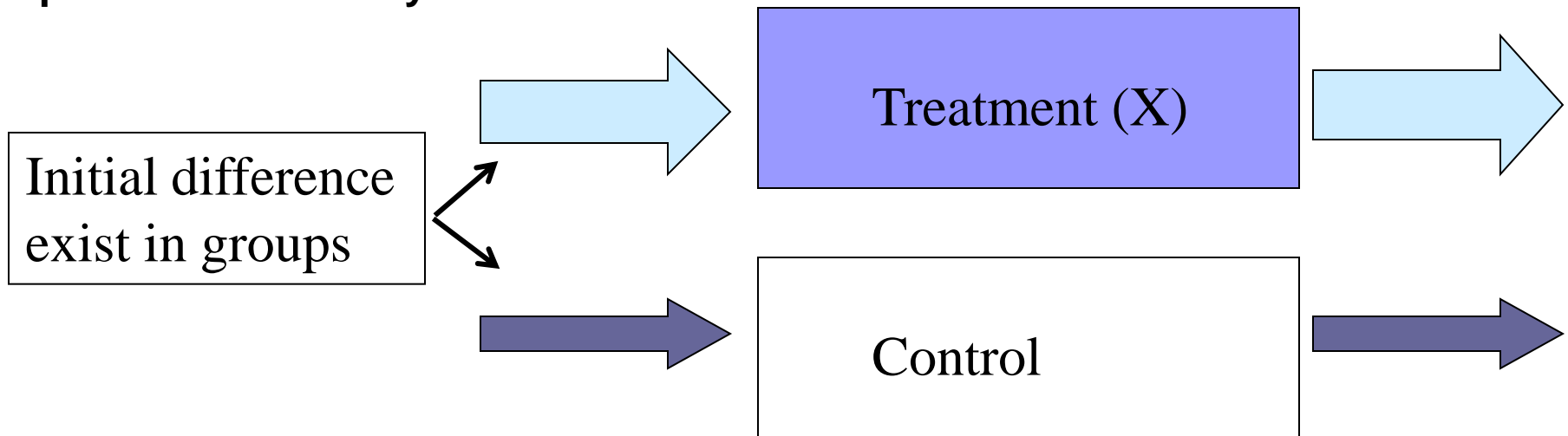
**History:** something (other than the treatment) occurred **between the pretest and posttest.**

Occurrence of events other than the independent variable.



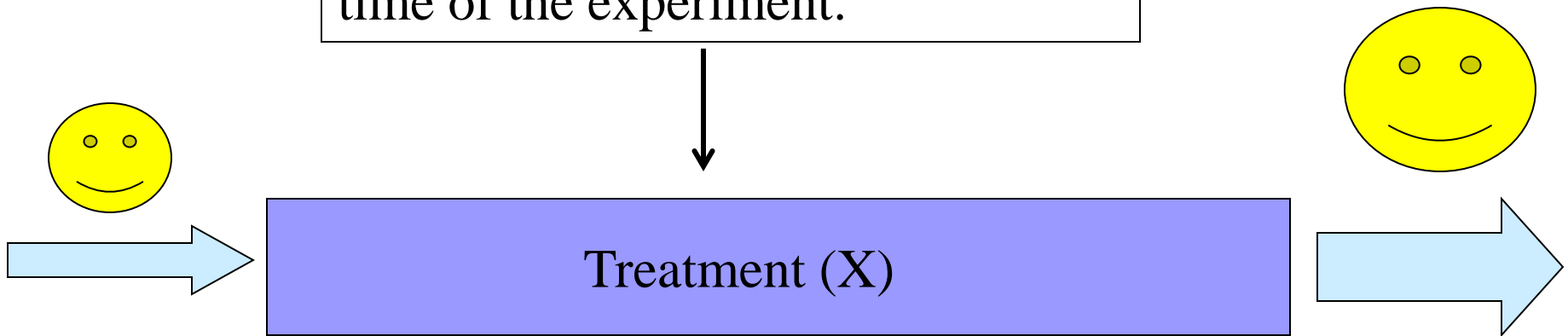
# Selection Threat

**Selection:** something (other than the treatment) occurred **before the pretest or the study.**

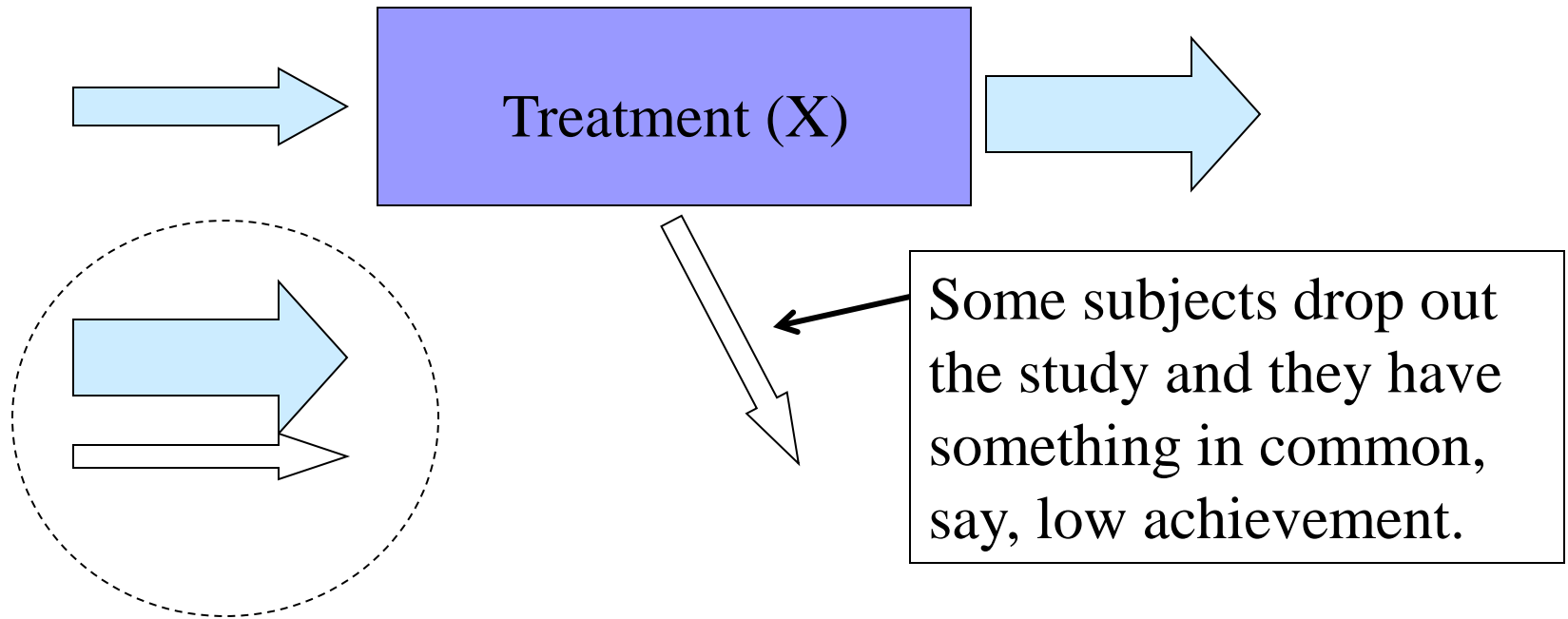


# Maturity Threat

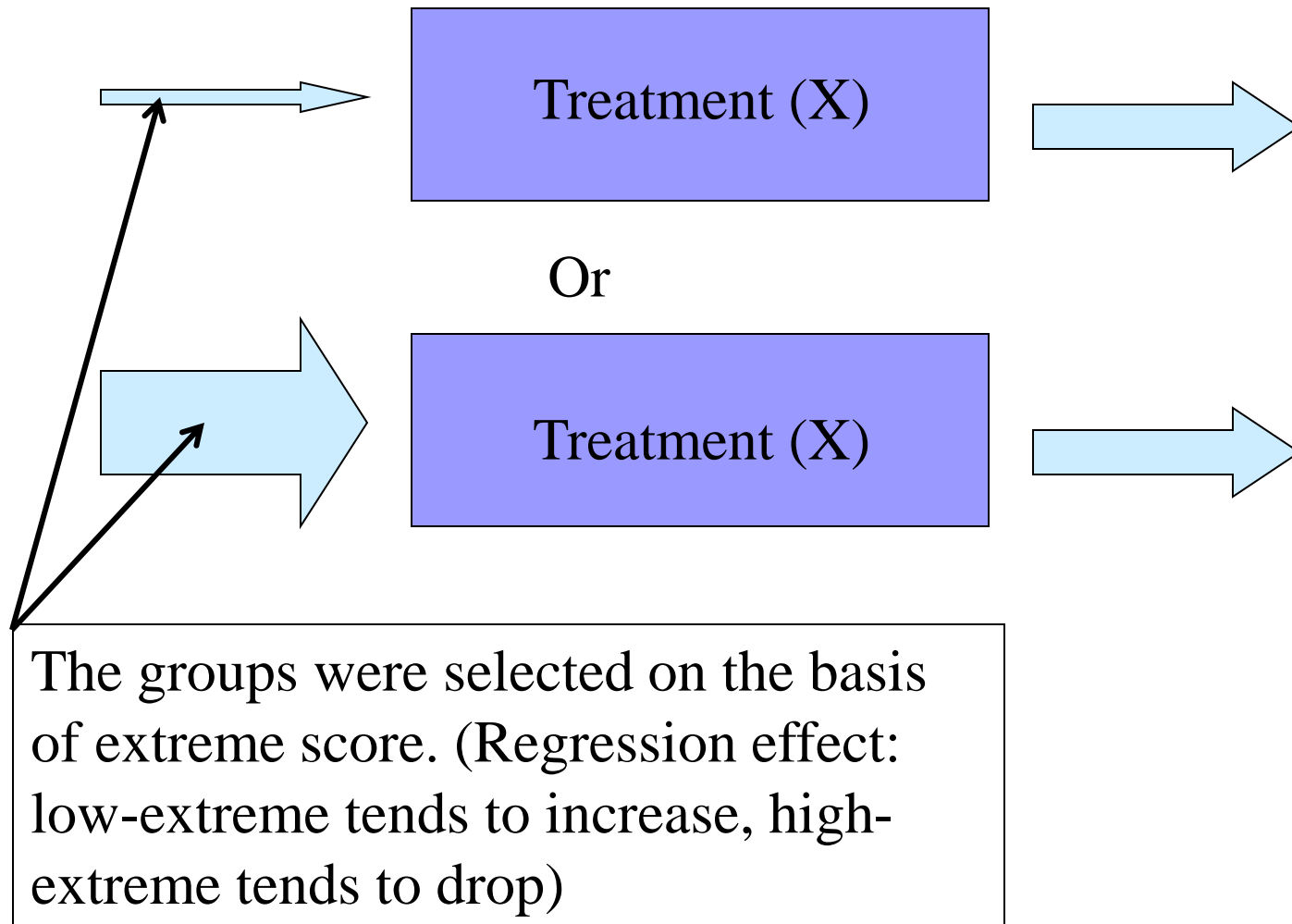
There may be developmental (physical or mental) changes occurring to the subjects during the time of the experiment.



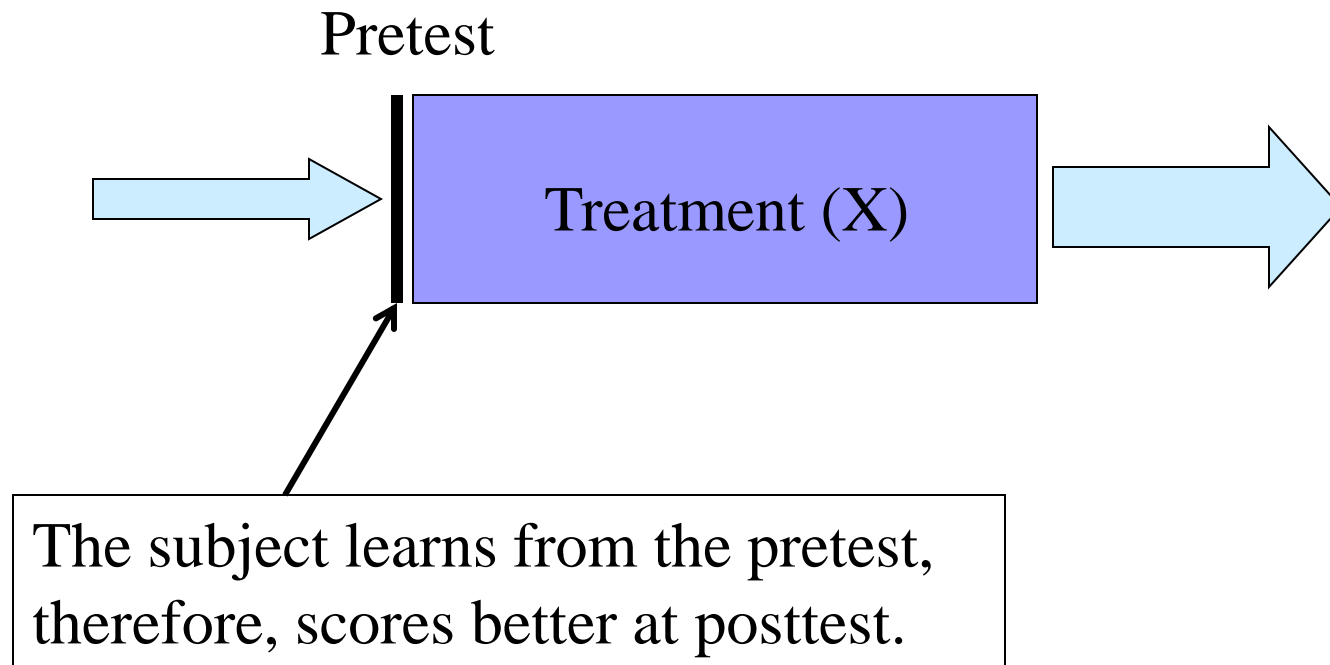
# Mortality/Attrition Threat



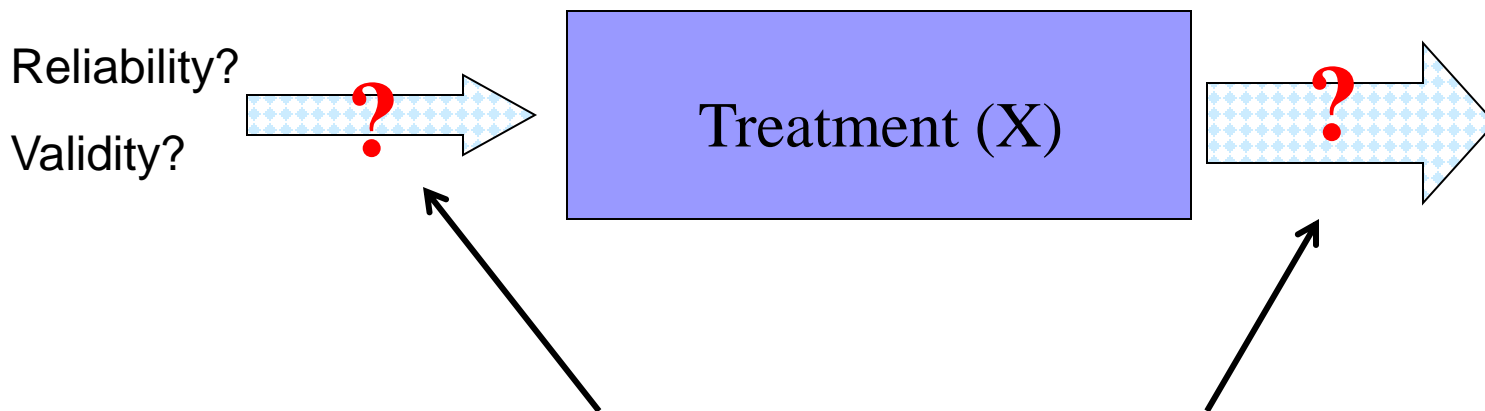
# Regression Threat



# Testing Threat



# Instrumentation Threat



The measuring instruments is not reliable or not valid, therefore, the score obtained by subjects could not be accurate. i.e. Alternative forms or different raters.

# Identify Validity Threat (1)

- In a study of maze learning, rats are administered either a drug to enhance learning or a placebo. Each day, over a 2-week period, the rats are tested in the maze. To ensure that the experimental rats recover from the drug, these rats are separated from the control rats and kept in a climate-controlled laboratory, whereas the controls are placed in a basement with other rats.

# Identify Validity Threat (2)

- During the Truman (Democratic)/Taft (Republican) presidential race, a pollster carried out a telephone survey to determine the front-runner. He took a random sample of names from the phone book and found overwhelming support for Taft.

# Identify Validity Threat (3)

- Within one kindergarten class, children who had had nursery school were compared on a readiness test at the end of kindergarten with other children having no such preparatory experience. The researcher attributed the superiority of the nursery school groups to their preparatory experience.

# Identify Validity Threat (4)

- Maybe they used easier tests at the end of the program than at the beginning.
- Treatment 1 was better than Treatment 2 because the kids in Treatment 1 were much better prepared prior to the treatments.
- Scores declined between pre-and posttest even though the subjects were in the top of the pretest ability distribution.
- All the kids did better in the posttest because they learned how to take these kinds of tests better, and not because they knew more.

# CNN's Disclaimer



CNN.com QuickVote - Microsoft Internet Explorer

CNN.com

**QUICKVOTE**

Created: Sunday, August 21, 2006, at 20:50:08 EDT

Would you go lava sledding?

Yeah, sounds cool		27%	5315 votes
No, too hot to handle		73%	14253 votes

**Total: 19568 votes**

This QuickVote is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, nor the public as a whole. The QuickVote sponsor is not responsible for content, functionality or the opinions expressed therein.

**Related:**

- [Hawaiians revive ancient sport -- lava sledding](#)

# Terms Used in the Research Design

- **Random Sampling/Selection:** a method for selecting a sample of subjects from a population, in which each subject in the population of interest has an equal opportunity to be selected into the sample, so the results can be generalized to the entire population of which they represent.
- **Control Group:** a group of subjects whose selection and treatment are exactly the same as those of the experimental group except that the control group does not receive the experimental treatment.
- **Random Assignment:** a method for assigning subjects to control and experimental groups, in which each subject has an equal and independent chance of being assigned to any of the groups in the study.
- **Randomizer:** <http://www.randomizer.org/>

# Types of Research

- **Preexperimental designs (No control, no random assignment)**
  - One-shot case study
  - One group pretest to posttest design
  - Intact group comparison
- **Experimental designs (Both control and random assignment)**
  - Posttest only control group
  - Pretest-posttest control group design
  - Factorial experimental design
- **Quasiexperimental designs (Control, no random assignment)**
  - Nonequivalent-control group design
  - Time-series design
- **Ex post facto designs (no treatment administered)**
  - Correlation design
  - Criterion-group design

# Symbols of Donald Campbell and Julian Stanley (1963)

- **R = Randomization.** The process that the subjects are randomly assigned to either the experimental or control group.
- **X = Treatment or Independent Variable.** The treatment that is administered or the variable that may influence the dependent variable.
- **O = Observation.** An observation or test that has occurred during the design.

# Pre-experimental Designs

- **One-shot case study:** an experimental treatment is introduced and researchers look at effects on the dependent variable without comparing with a control group

X O

- **One-group pretest-posttest design:** a case study with an additional pretest so that subjects can serve as their own controls.

O X O

- **Intact-group comparison:** a design adding a control group, but the two groups are not known to be comparable.

X O

O

# Experimental Designs

- **Posttest-only control group:** Only posttesting subjects in a randomly assigned experimental group and control group

R X O  
R O

- **Pretest-posttest control group:** pretesting and posttesting subjects in a randomly assigned experimental group and control group

R O X O  
R O O

- **Factorial experimental design:** experimental designs that include more than one independent variable. e.g.,

R  $X_{A_1B_1}$  O  
R  $X_{A_1B_2}$  O  
R  $X_{A_2B_1}$  O  
R  $X_{A_2B_2}$  O

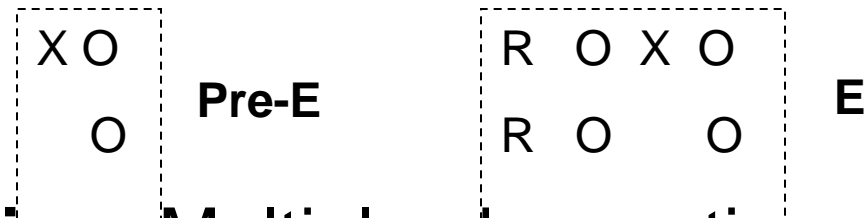
**A, B, C ... represents factors/independent variables**

**1, 2, 3, ... represents levels for each factor**

# Quasi-experimental Designs

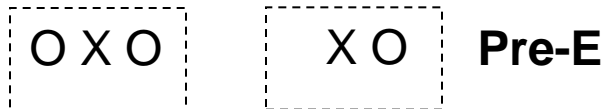
- **Nonequivalent-control group design:** subjects are not randomly assigned to the control and experimental groups, but receive a pretest and posttest.

O X O  
O O



- **Time-Series Design:** Multiple observations are taken before and after a treatment is administered.

O O X O O



# Ex Post Facto Designs

*Ex post facto: Late Latin, literally, from a thing done afterward: after the fact*

- **Correlation design:** measure one group with two measurements and examine the relationship between the scores obtained from the two measurements.

$O_x O_y$

- **Criterion-group design:** one group is the criterion and the other is compared to it. Two groups are selected intentionally.

$X_A O$

$X_B O$

# Type of Research Designs

	Control Group	Random Assignment	Causal Inferences
Experimental (E)	Yes	Yes	Yes
Pre-experimental (PE)	No or not adequate	No	No
Quasi-experimental (QE)	Yes	No	Better than PE, worse than E
Ex post facto	No	No	No, only measuring correlations

# Identify the Design Type (1)

- In a study of constructivist instruction, Mr. Brown's 3<sup>rd</sup> period class were taught math with manipulative while Brown's 4<sup>th</sup> period class was taught with worksheets. Pretest and posttest are given.
- Suppose you found out that students were randomly assigned to the 3<sup>rd</sup> and 4<sup>th</sup> period classes in Brown's school, then?
- A district compared students' performance before and immediately after a curriculum was implemented, announced that 8th grade students' performance improved by the new curriculum.



# Identify the Design Type? (2)

- In a study of the relationship between mathematics achievement and attitude towards mathematics, a researcher collected two measures of all freshmen enrolled in a major university: attitudes toward mathematics and scores of the quantitative section of the SAT were collected.
- In a study, men and women's empathy were measured and compared.

# Identify the Design Type? (3)

- Subjects who were afraid of heights were randomly assigned to one of 3 counseling treatments: control (talk with a therapist), Behavior Modification, or Rogerian therapy. At the end of 6 months treatment, the Students were flown to Pisa. They were placed on the top of the leaning tower and asked to walk as close to the edge as they dared. The distance to the edge was measured.
- 120 4th and 6th grade subjects,  $\frac{1}{2}$  M and  $\frac{1}{2}$  F at each grade level were randomly assigned to study at text either with or without an advanced organizer. After an hour's reading, they were given an achievement test to see how much of the text they had learned.



# Interesting, Confusing, or Overwhelming ?

- You can read more about research design at

<http://www.socialresearchmethods.net/kb/design.htm>