ECE1724H S2: Empirical Software Engineering

Experimentation 2

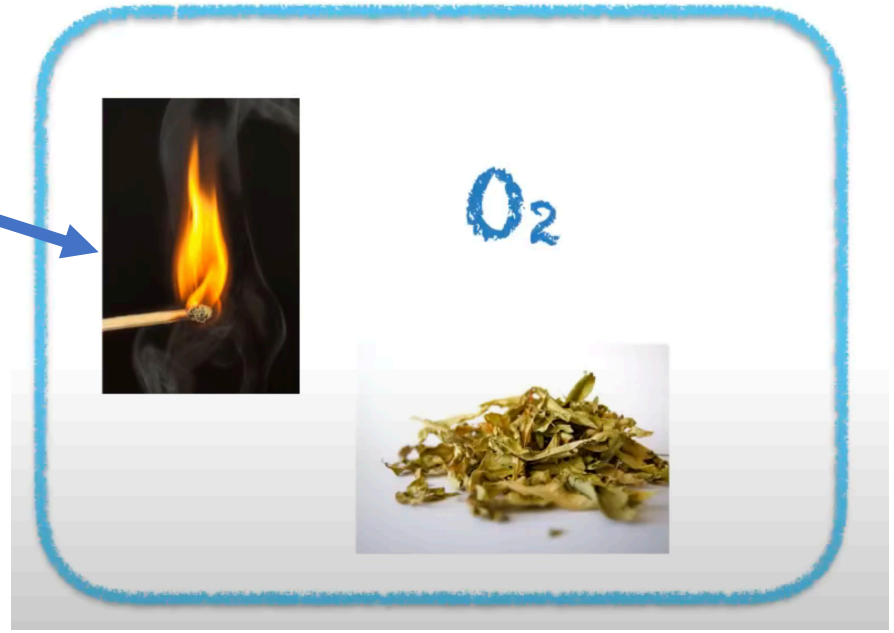
The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

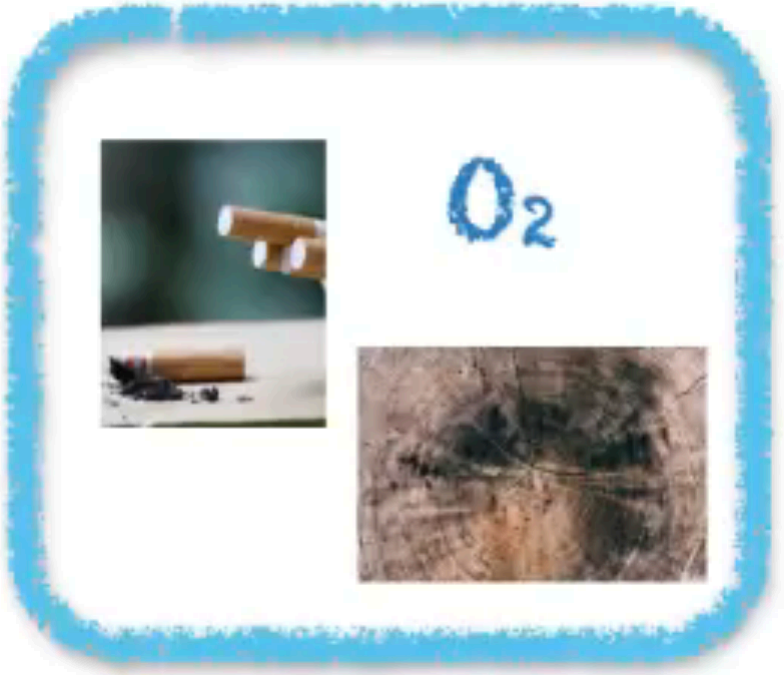
Experiments and Causation

Cause: Inus condition – insufficient but nonredundant part of an unnecessary but sufficient condition

Cause – Inus Condition

Inus Condition



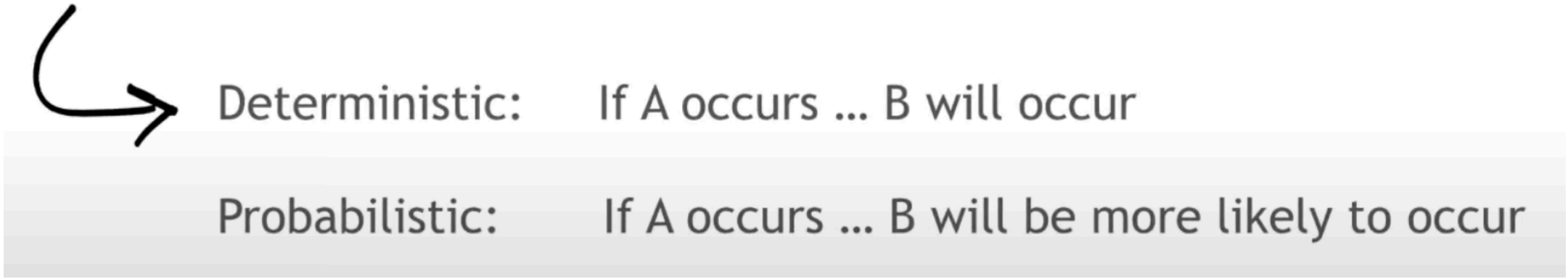


→ insufficient but nonredundant part of an unnecessary but sufficient condition

Experiments and Causation

Cause: Inus condition – insufficient but nonredundant part of an unnecessary but sufficient condition

- most causes are inus conditions
  - many factors are required for an effect to occur



Experiments and Causation

Cause: Inus condition – insufficient but nonredundant part of an unnecessary but sufficient condition

- most causes are inus conditions
  - many factors are required for an effect to occur

Effect: counterfactual

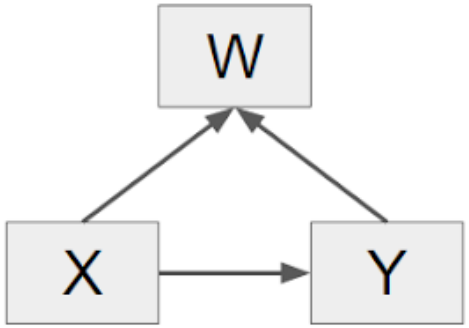
i.e., what would have happened to these subjects had the cause not been present?

Causal Relationships

- The cause preceded the effect
- The cause was related to the effect
- We can find no plausible alternative explanation for the effect other than the cause
  - ➔ Mirror what happens in experiments
  - ➔ No other scientific method regularly matches the characteristics of causal relationships so well

Correlation does not prove causation!

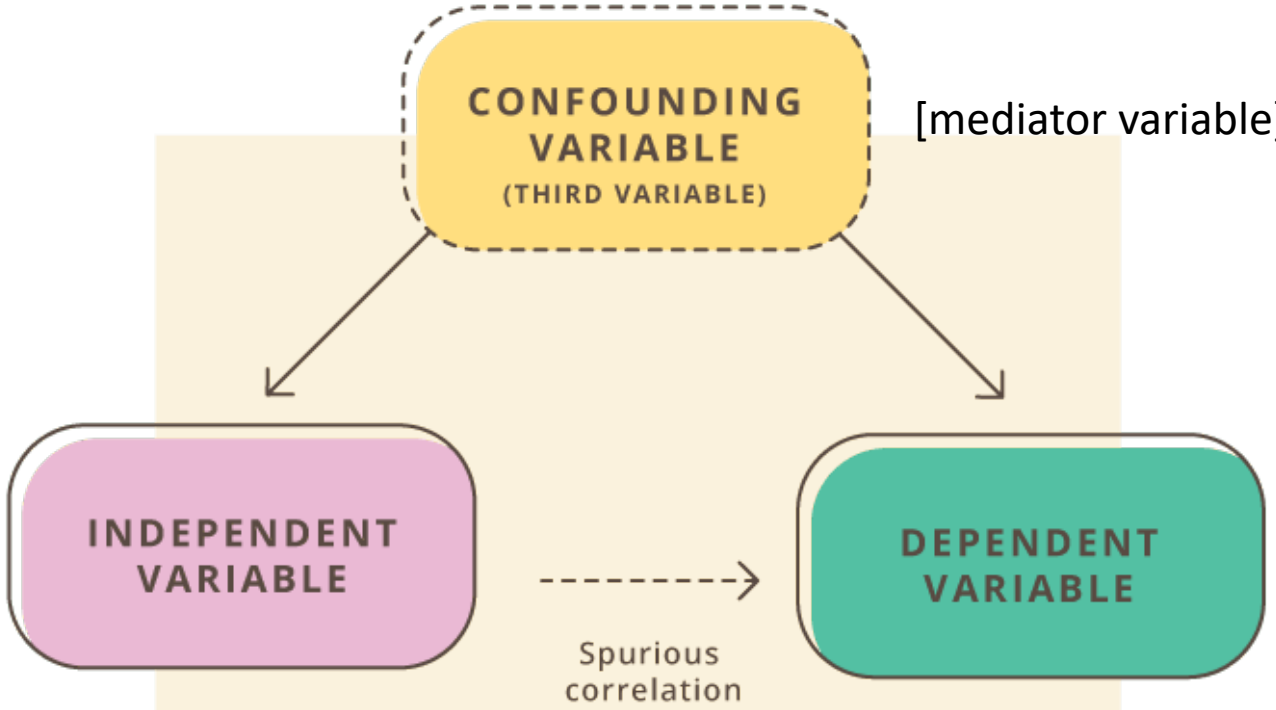
- Which variable came first?
- Are there alternative explanations for the presumed effect?



  - Example: income ~ education or education ~ income?
  - **Confounding variable:** intelligence, family socioeconomic status (causes both high education and high income)

Definitions

- Independent Variables
  - Variables (factors) that are manipulated to measure their effect
  - Typically select specific levels of each variable to test
- Dependent Variables
  - "output" variables - tested to see how the independent variables affect them
- Treatments
  - Each combination of values of the independent variables is a treatment
  - Simplest design: 1 independent variable x 2 levels = 2 treatments
    - E.g. tool A vs. tool B
- Subjects
  - Human participants who perform some task to which the treatments are applied
  - Note: subjects must be assigned to treatments randomly

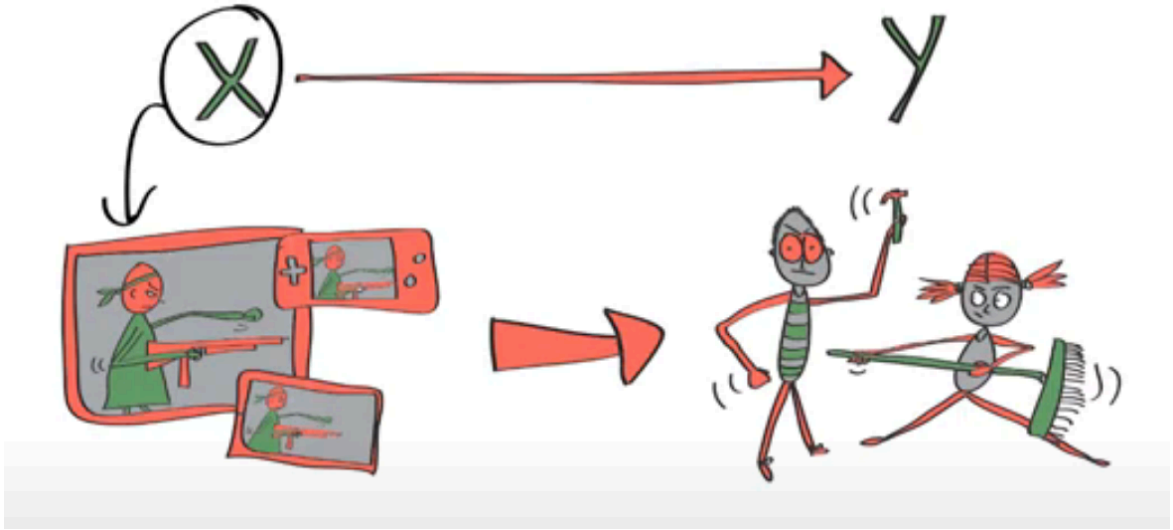


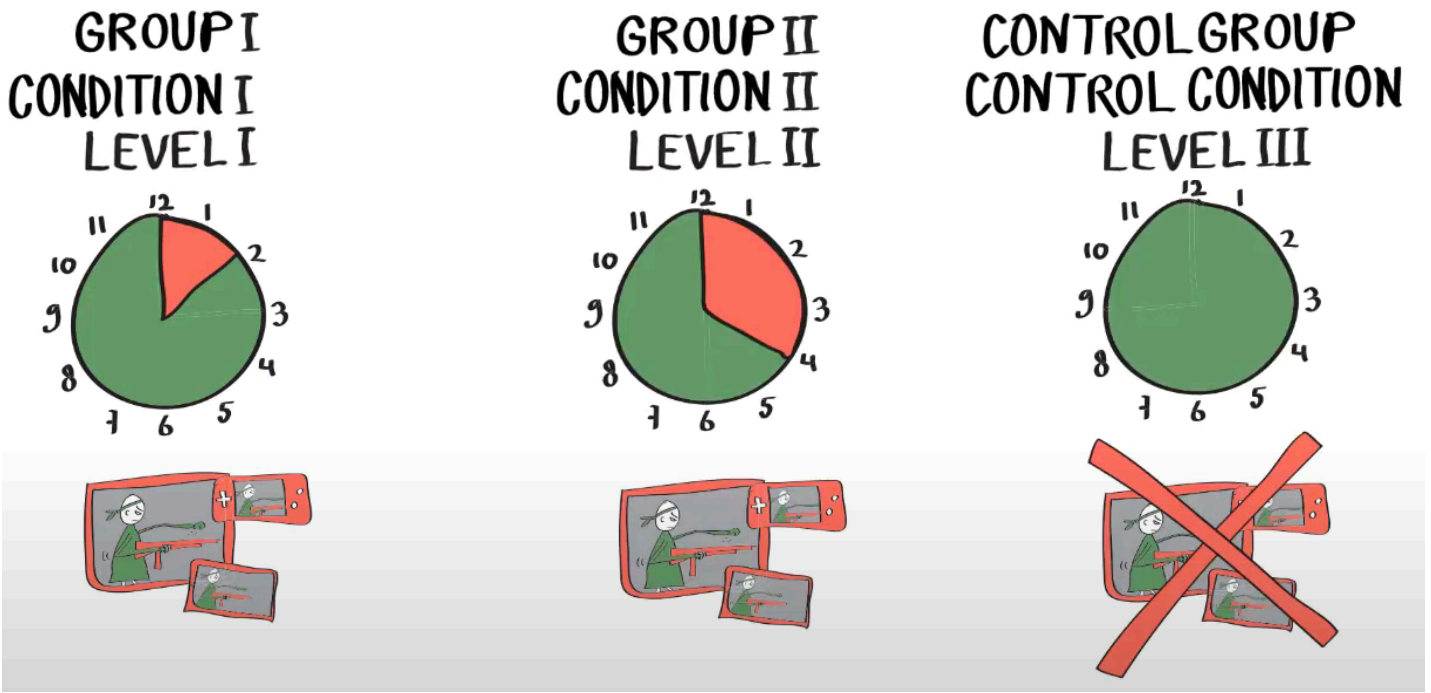
[mediator variable]

What at first looks like a causal relationship between IV and DV is ultimately spurious. The confounding variable is the hidden explanation.

https://explorable.com/confounding-variables

Manipulable and Nonmanipulable Causes


X
Y


GROUP I
CONDITION I
LEVEL I
GROUP II
CONDITION II
LEVEL II
CONTROL GROUP
CONTROL CONDITION
LEVEL III

Manipulable and Nonmanipulable Causes

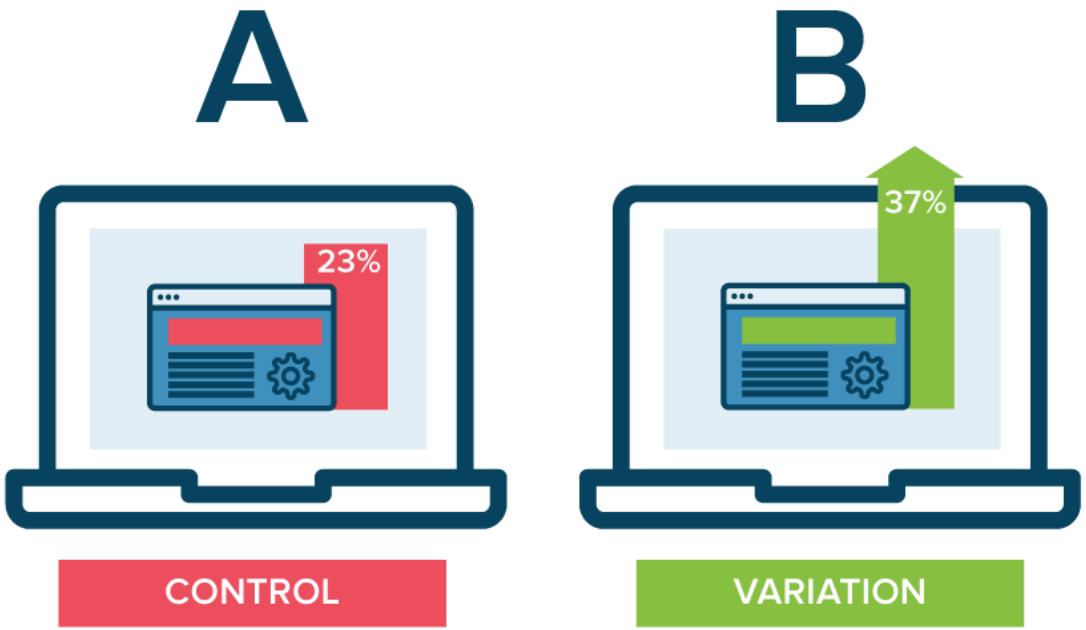
- Experimental variables
- Individual difference variables

Intrinsic property of the participant

Disadvantages of experiments ?

Disadvantages

- Tell nothing about how and why effects occurred
- Cannot deal with cases when we first observe effect and need to look for causes



Advantages of experiments ?

Unique strength

- Causal description: describe consequences attributable to deliberately varying a treatment
- Not causal explanation (mechanisms)

If so limited, then why so central to science?

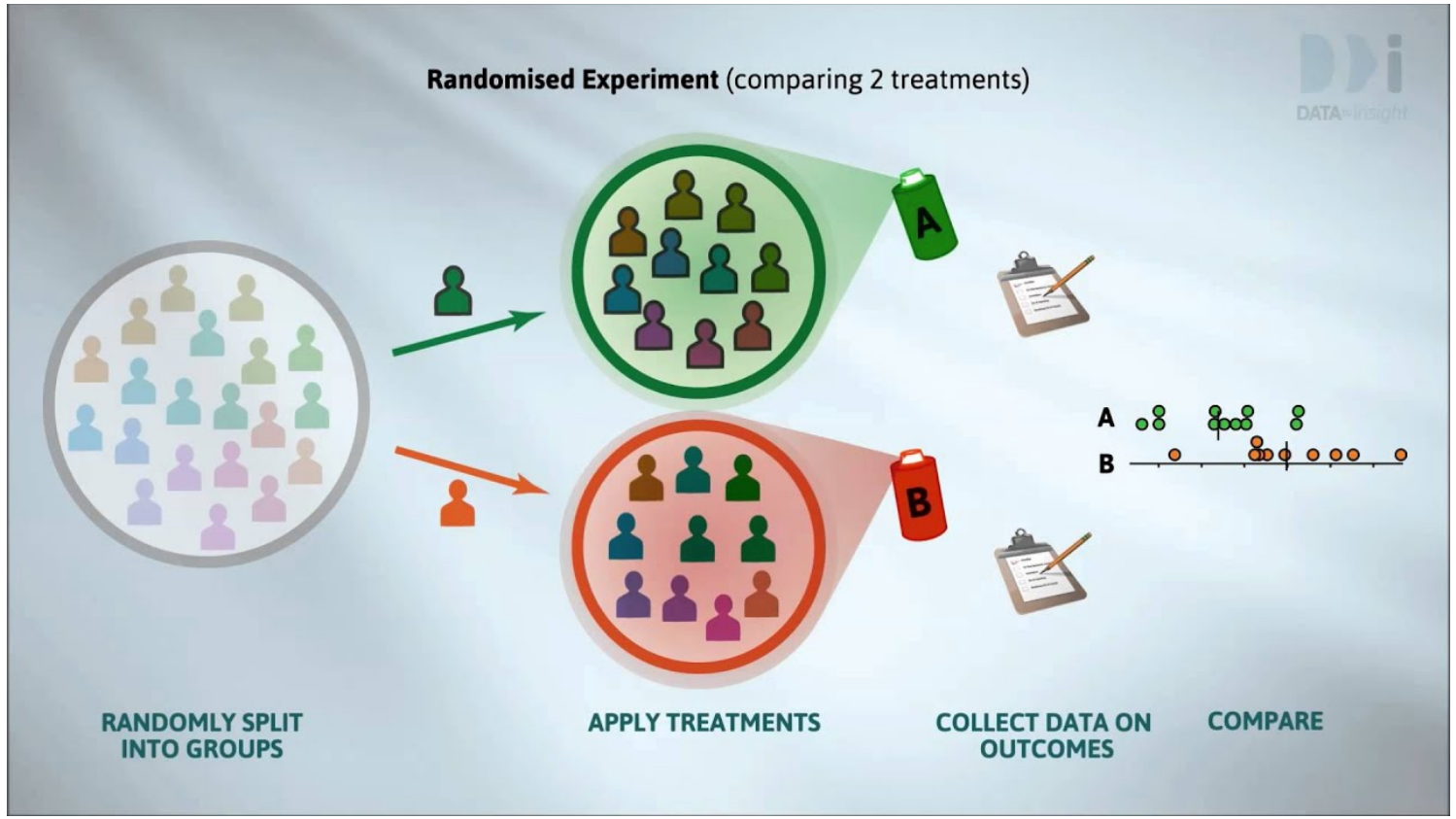
Why is experiments so central to science?

- The dichotomy between descriptive and explanatory causation is less clear in scientific practice than in abstract discussions about causation.
  - many causal explanations consist of chains of descriptive causal links
  - experiments help distinguish between the validity of competing explanatory theories
  - some experiments test whether a descriptive causal relationship varies in strength or direction under Condition A versus Condition B
  - some experiments add quantitative or qualitative observations of the links in the explanatory chain

The Vocabulary of Experiments

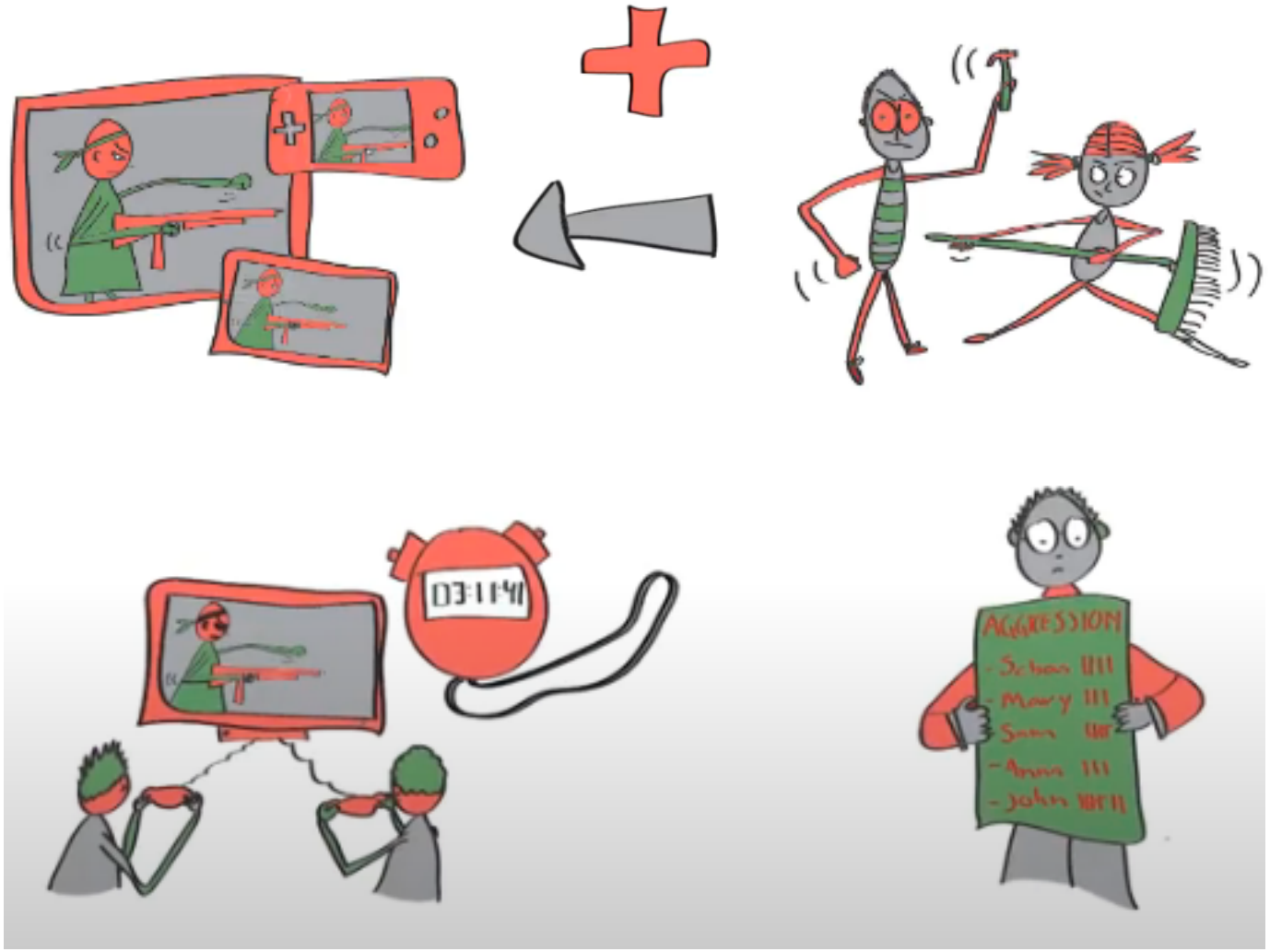
- **Experiment**: A study in which an intervention is deliberately introduced to observe its effects.
- **Randomized Experiment**: An experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers.
- **Quasi-Experiment**: An experiment in which units are not assigned to conditions randomly.
- **Natural Experiment**: Not really an experiment because the cause usually cannot be manipulated; a study that contrasts a naturally occurring event such as an earthquake with a comparison condition.
- **Correlational Study**: Usually synonymous with nonexperimental or observational study; a study that simply observes the size and direction of a relationship among variables.

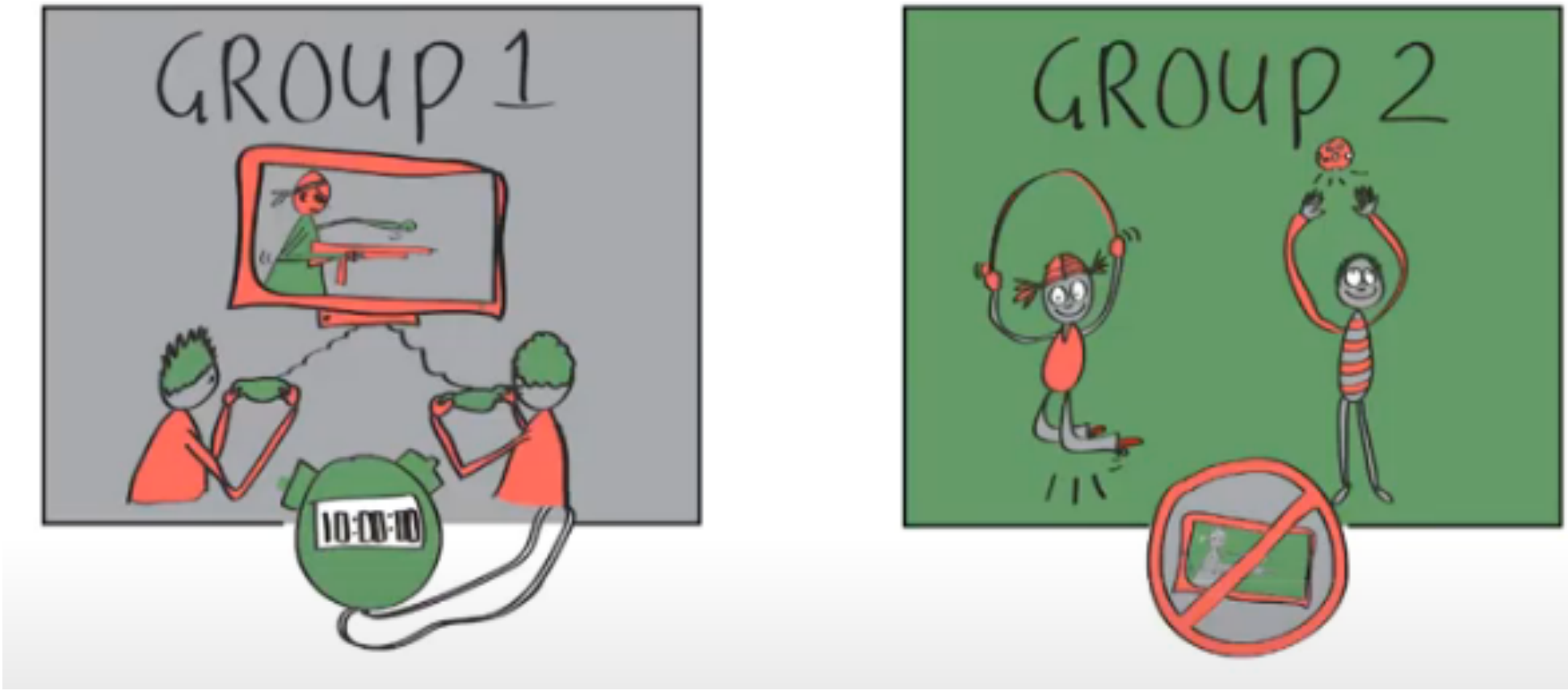
Randomized experiment



Three requirements:

- manipulation
- comparison
- random assignment

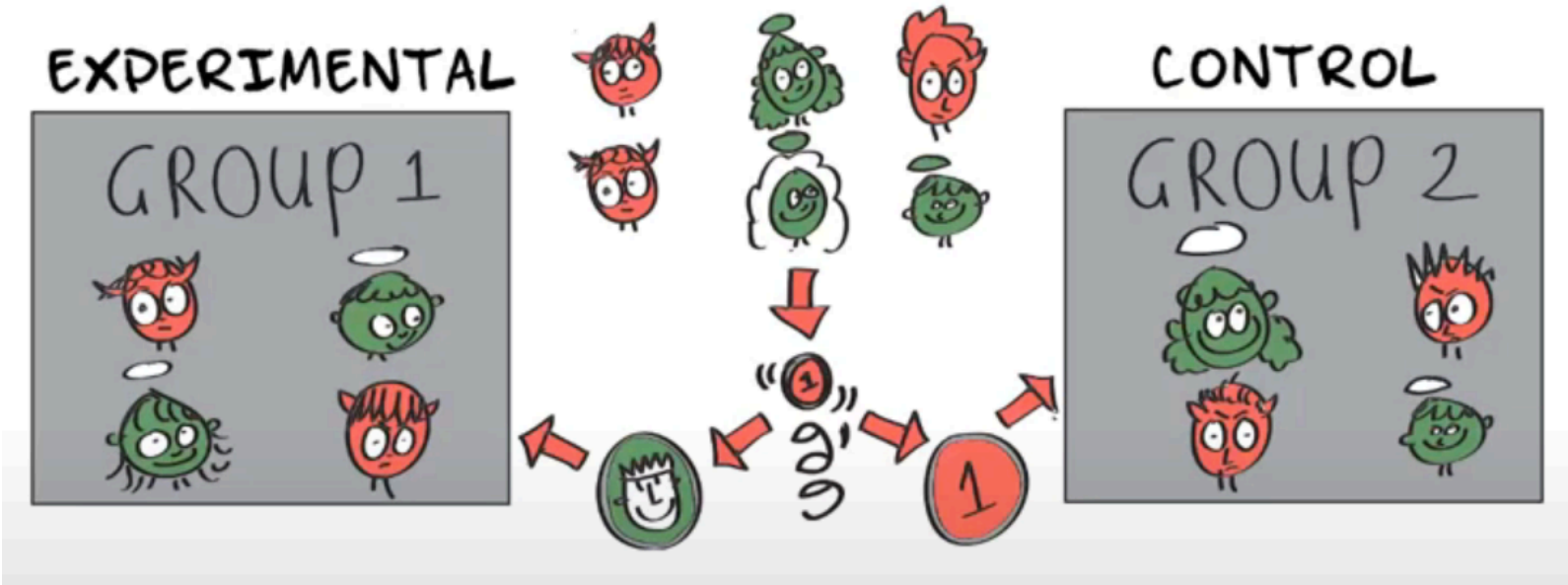

AGGRESSION



score aggression > score aggression

Random assignment

All participants in the sample have an equal chance of being assigned to the various experimental conditions.



How do we know if the random assignment works?

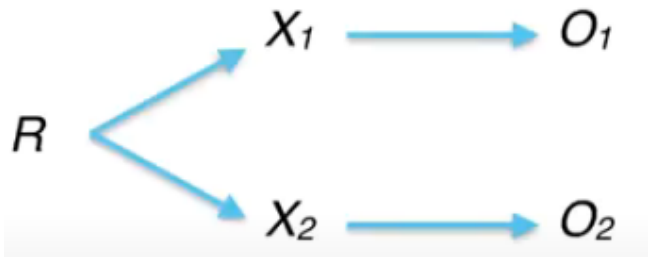
Randomized Sampling
vs
Randomized Assignment

Randomized experiment

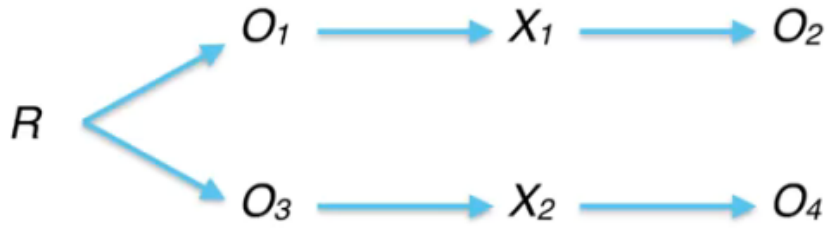
- various treatments being contrasted (including no treatment at all) are assigned to experimental units by chance;
- resulting 2+ groups of units are probabilistically similar to each other on the average
- outcome differences are likely due to treatment

Randomized Experiment Design

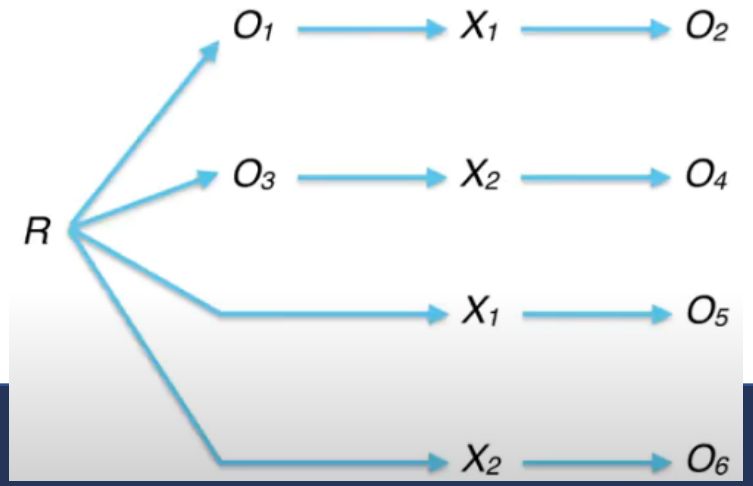
- Randomized two-group design



- Pretest-Posttest two-group design

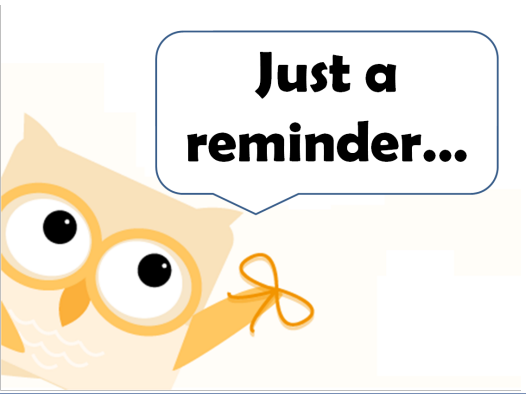


- Solomon four-group design



Planning Checklist

- [x] Pick a topic
- [x] Identify the research question(s)
- [x] Check the literature
- [x] Identify your philosophical stance
- [x] Identify appropriate theories
- [x] Choose the method(s)
- [ ] Design the study
  - [x] Unit of analysis?
  - [x] Target population?
  - [x] Sampling technique?
  - [x] Data collection techniques?
  - [x] Metrics for key variables?
  - [x] Handle confounding factors
- [ ] Critically appraise the design for threats to validity
- [ ] Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- [ ] Recruit subjects / field sites
- [ ] Conduct the study
- [ ] Analyze the data
- [ ] Write up the results and publish them
- [ ] Iterate



Four criteria to evaluate research design

- Construct validity
- **Internal validity**
- External validity
- Conclusion validity

The degree to which the design of a study allows to draw causal conclusions about the effect of one or more independent variables on one or more dependent variables.

Validity

- Construct Validity
  - Are we measuring the construct we intended to measure?
  - Did we translate these constructs correctly into observable measures?
  - Did the metrics we use have suitable discriminatory power?
- Internal Validity
  - Do the results really follow from the data?
  - Have we properly eliminated any confounding variables?
- External Validity
  - Are the findings generalizable beyond the immediate study?
  - Do the results support the claims of generalizability?
- Empirical Reliability
  - If the study was repeated, would we get the same results?
  - Did we eliminate all researcher biases?

Typical Problems

- Construct Validity
  - Using things that are easy to measure instead of the intended concept
  - Wrong scale; insufficient discriminatory power
- Internal Validity
  - Confounding variables: Familiarity and learning;
  - Unmeasured variables: time to complete task, quality of result, etc.
- External Validity
  - Task representativeness: toy problem?
  - Subject representativeness: students for professional developers!
- Theoretical Reliability
  - Researcher bias: subjects know what outcome you prefer

Four criteria to evaluate research design

- Construct validity
- Internal validity
- External validity
- Conclusion validity

can we generalize results to the theoretical constructs that the units, treatments, observations, and settings are supposed to represent?

Construct Validity

- Are we measuring what we intend to measure?
  - Akin to the requirements problem: are we building the right system?
  - If we don't get this right, the rest doesn't matter
  - Helps if concepts in the theory have been precisely defined!

Four criteria to evaluate research design

Construct validity

Internal validity

External validity

Conclusion validity

Does the causal relationship hold over variations in persons, settings, treatments, and outcomes?

External Validity

- Two issues:
  - Results will generalize beyond the specific situations studied
    - E.g. do results on students generalize to professionals?
  - Do the results *support the claims* of generalizability?
    - E.g. if the effect size is small, will it be swamped/masked in other settings?
    - E.g. will other (unstudied) phenomena dominate?
- Two strategies:
  - Provide arguments in favour of generalizability
  - Replicate the finding in further studies:
    - Literal replication - repeat study using the same design
    - Empirical Induction - related studies test additional aspects of the theory

Strategies for constructivists

- Triangulation
  - Different sources of data used to confirm findings
- Member checking
  - Research participants confirm that results make sense from their perspective
- Rich, thick descriptions
  - As much detail as possible on the setting and the data collected
- Clarify bias
  - Be honest about researcher's bias
  - Self-reflection when reporting findings
- Report discrepant information
  - Include data that contradicts findings as well as that which confirms
- Prolonged contact with participants
  - Spend long enough to ensure researcher really understands the situation being studied
- Peer debriefing
  - A colleague critically reviews the study and tests assumptions
- External Auditor
  - Independent expert reviews procedures and findings

Four criteria to evaluate research design

Construct validity

Internal validity

External validity

Conclusion validity

(aka. Statistical conclusion validity) the validity of inferences about the correlation between treatment and outcome

[type I and type II error]

5. THREATS TO VALIDITY

The five bugs and ten generated patches we used to construct debugging tasks may not be representative of all bugs and generated patches ...itigate this threat by selecting ...defect types and using real bugs with varied

3.3 Threat to validity

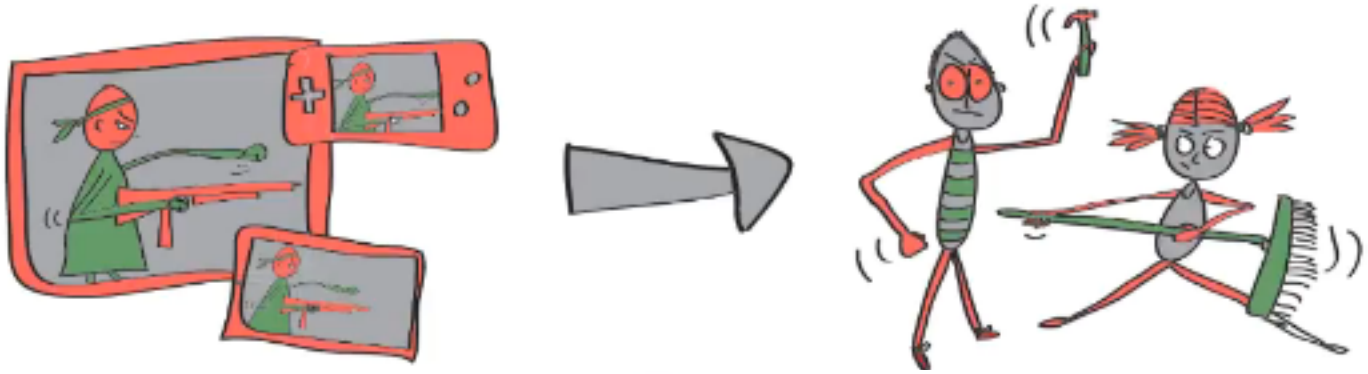
The internal threats to validity include the possible errors of our manual inspection. To reduce the threat, we ask two students to inspect our bugs. When they encounter controversial cases, they discuss them with others on our group meeting, until they reach an agreement. The threat can be mitigated with more researchers, so we release our inspection results on our website. The internal threats to external validity include our subject, since we analyzed the bugs inside only TensorFlow. Although our analyzed bugs are comparable with the prior studies and other studies (*e.g.* [53]) also analyzed only TensorFlow bugs, they are limited. The threat can be reduced by analyzing more libraries in future.

Planning Checklist

- [x] Pick a topic
- [x] Identify the research question(s)
- [x] Check the literature
- [x] Identify your philosophical stance
- [x] Identify appropriate theories
- [x] Choose the method(s)
- [ ] Design the study
  - [x] Unit of analysis?
  - [x] Target population?
  - [x] Sampling technique?
  - [x] Data collection techniques?
  - [x] Metrics for key variables?
  - [x] Handle confounding factors
- [x] Critically appraise the design for threats to validity
- [ ] Get IRB approval
  - Informed consent?
  - Benefits outweigh risks?
- [ ] Recruit subjects / field sites
- [ ] Conduct the study
- [ ] Analyze the data
- [ ] Write up the results and publish them
- [ ] Iterate

The Vocabulary of Experiments

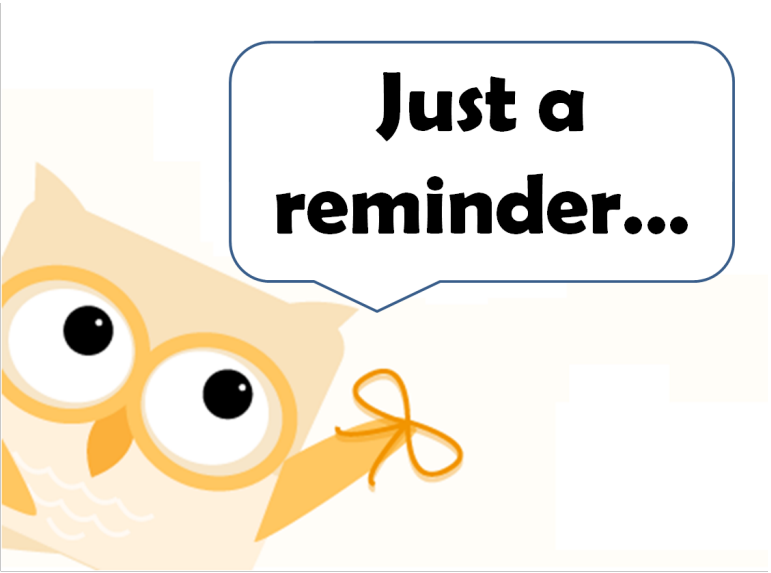
- **Experiment**: A study in which an intervention is deliberately introduced to observe its effects.
- **Randomized Experiment**: An experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers.
- **Quasi-Experiment**: An experiment in which units are not assigned to conditions randomly.
- **Natural Experiment**: Not really an experiment because the cause usually cannot be manipulated; a study that contrasts a naturally occurring event such as an earthquake with a comparison condition.
- **Correlational Study**: Usually synonymous with nonexperimental or observational study; a study that simply observes the size and direction of a relationship among variables.



Manipulable and Nonmanipulable Causes

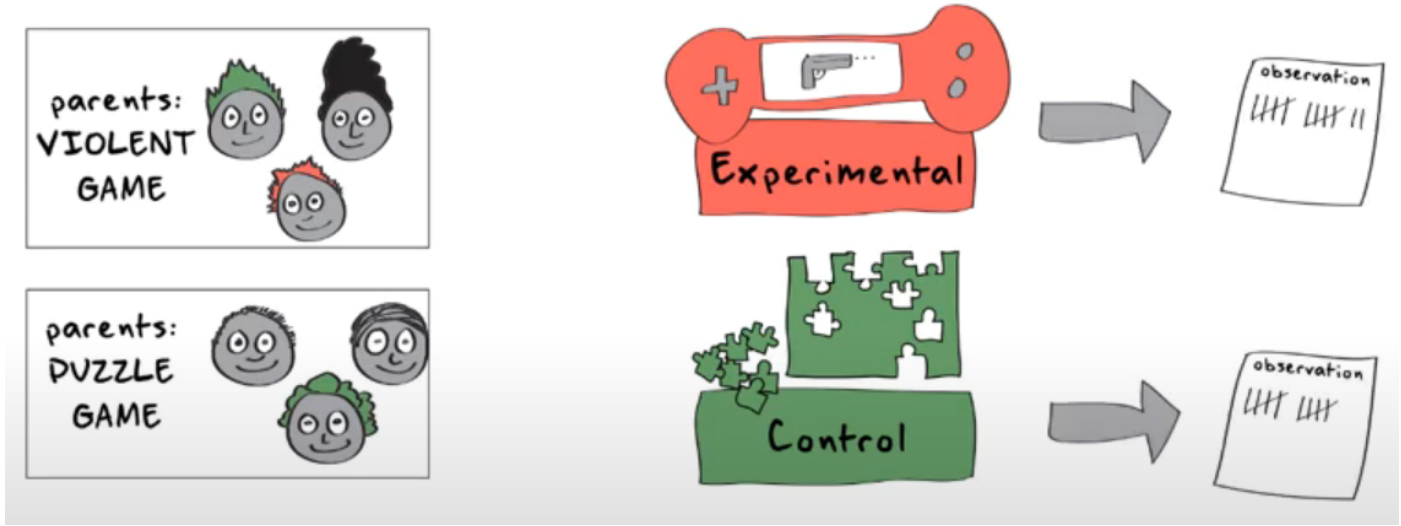
- Experimental variables
- Individual difference variables

Intrinsic property of the participant



Quasi-experiments Design

- Static group comparison design

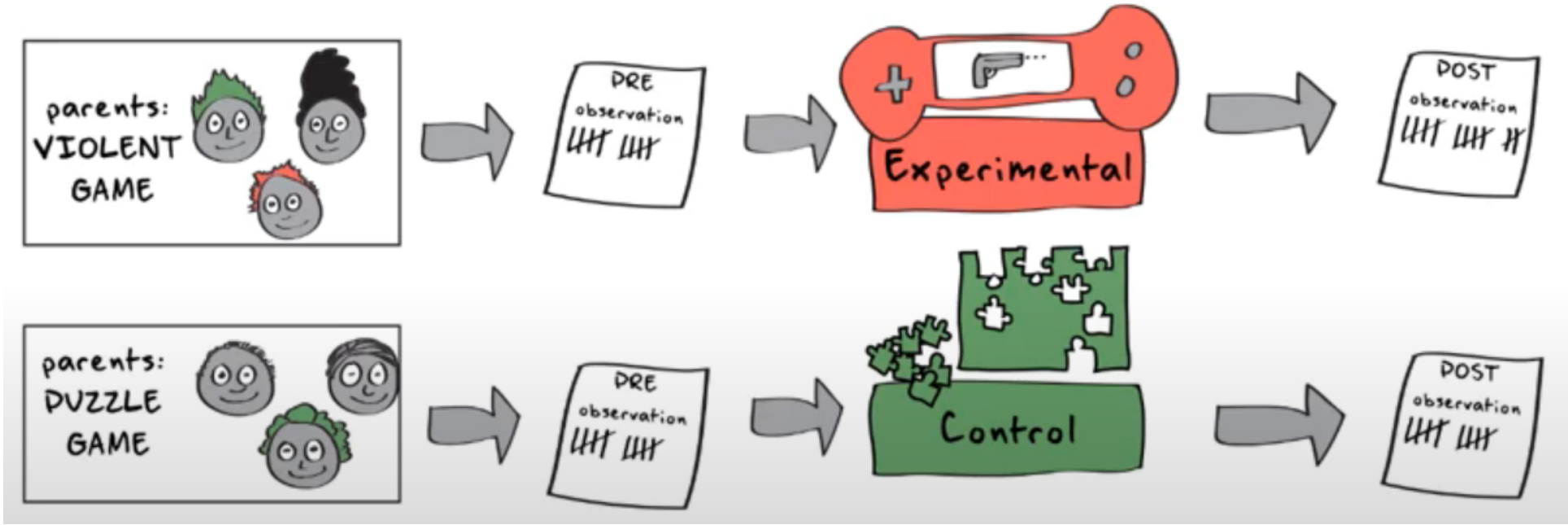


Group 1 X_1 O_1

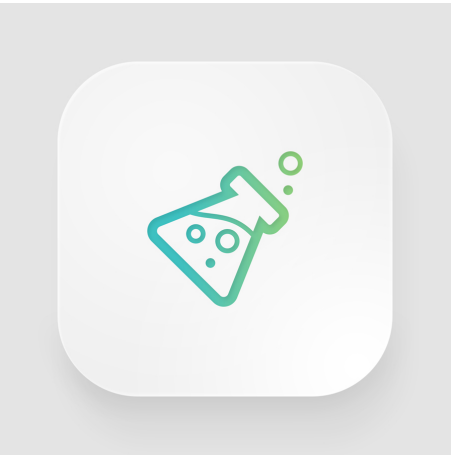
Group 2 X_2 O_2

Quasi-experiments Design

- Pretest-Posttest non-equivalent control group design

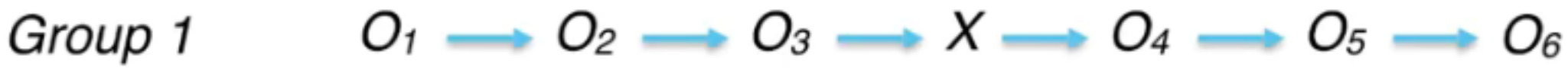


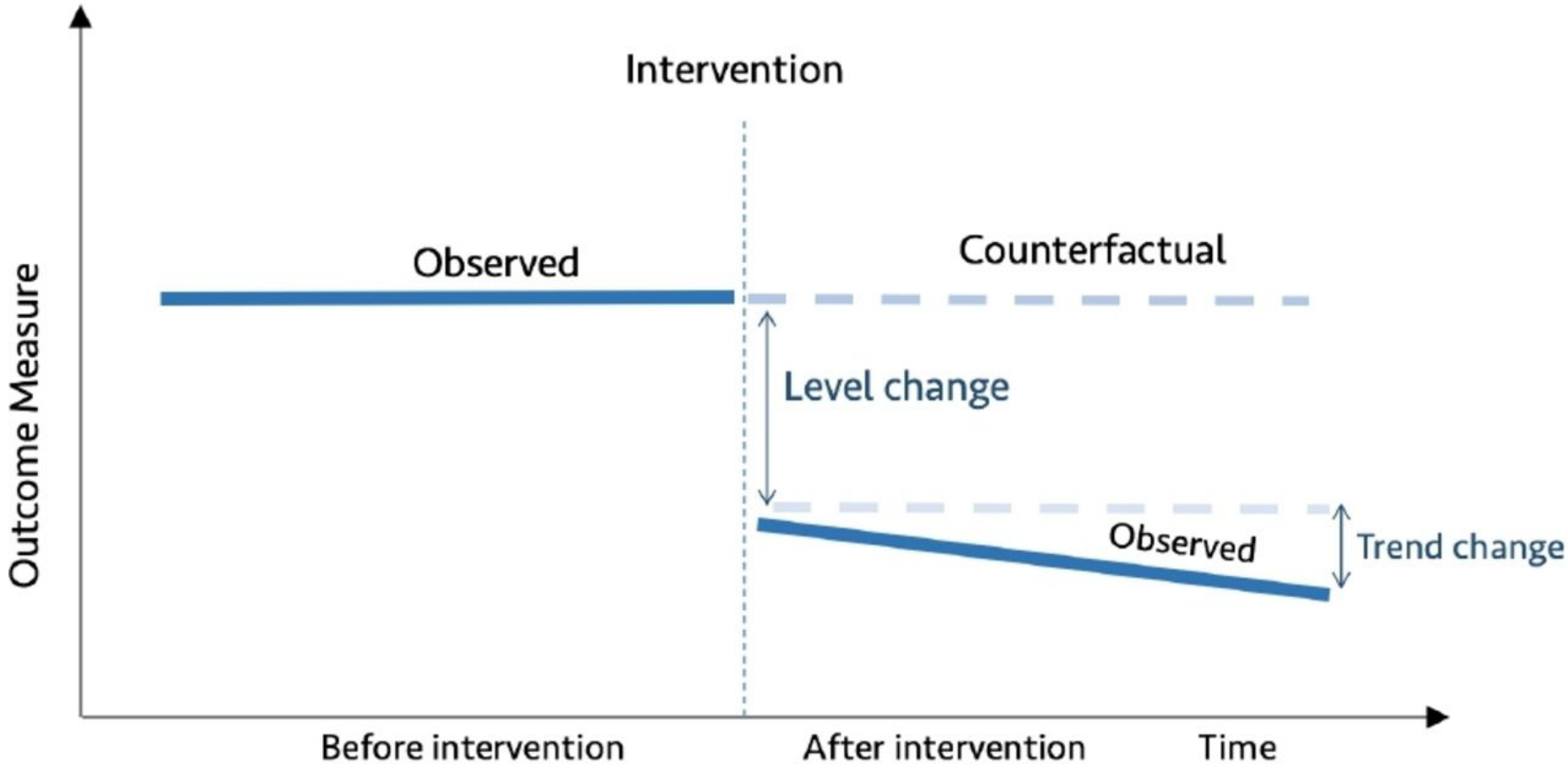
Quasi-experiments Design



- One-group Pretest-Posttest design

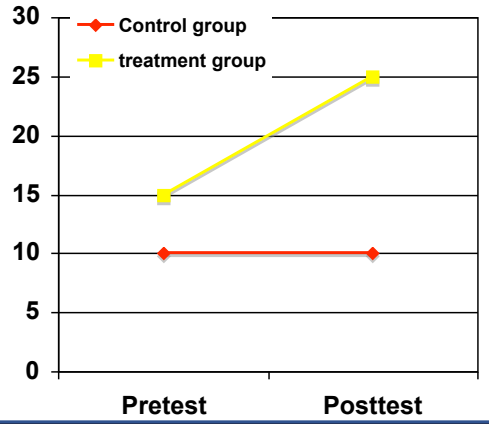
- Interrupted time series design

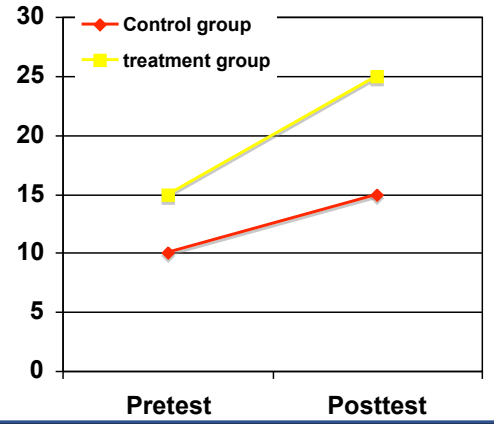


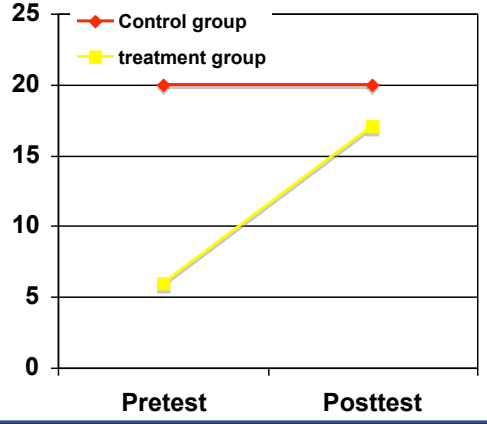

Intervention
Observed
Counterfactual
Level change
Observed
Trend change
Outcome Measure
Before intervention
After intervention
Time

Quasi-experiments

- When subjects are not assigned to treatments randomly:
  - Because particular skills/experience needed for some treatments
  - Because ethical reasons dictate that subjects get to choose
  - Because the experiment is conducted on a real project
- e.g. A Non-equivalent Groups Design
  - Pretest-posttest measurements, but without randomized assignment
  - E.g. two pre-existing teams, one using a tool, the other not
  - Compare groups' improvement from pre-test to post-test





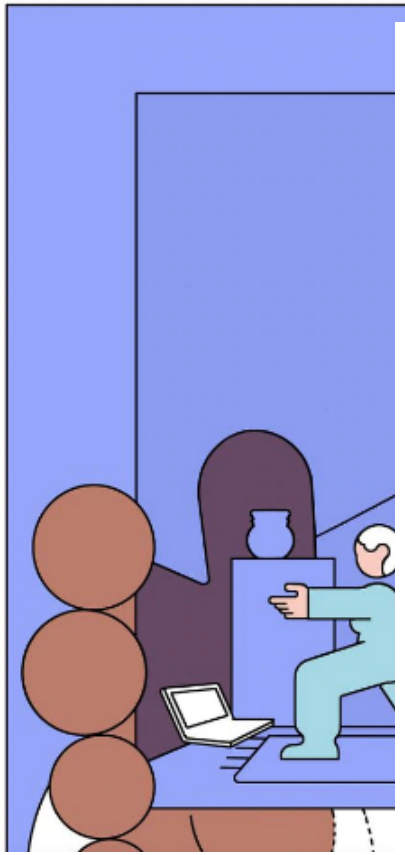


The Vocabulary of Experiments

- **Experiment**: A study in which an intervention is deliberately introduced to observe its effects.
- **Randomized Experiment**: An experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers.
- **Quasi-Experiment**: An experiment in which units are not assigned to conditions randomly.
- **Natural Experiment**: Not really an experiment because the cause usually cannot be manipulated; a study that contrasts a naturally occurring event such as an earthquake with a comparison condition.
- **Correlational Study**: Usually synonymous with nonexperimental or observational study; a study that simply observes the size and direction of a relationship among variables.

The great experiment

The pandemic is tragic. It's also an incredible chance to study human behavior.



A Huge Covid-19 Natural Experiment Is Underway—in Classrooms

As K-12 students head back to school, epidemiologists are watching for clues about how kids spread the virus, and what can stop it.

https://www.wired.com/story/a-huge-covid-19-natural-experiment-is-underway-in-classrooms/

https://www.washingtonpost.com/outlook/2020/09/10/coronavirus-research-experiment-behavior/?arc404=true

When not to use experiments?

When not to use experiments

- When you can't control the variables
- When there are many more variables than data points
- When you cannot separate phenomena from context
  - Phenomena that don't occur in a lab setting
  - E.g. large scale, complex software projects
  - Effects can be wide-ranging.
  - Effects can take a long time to appear (weeks, months, years!)
- When the context is important
  - E.g. When you need to know how context affects the phenomena
- When you need to know whether your theory applies to a specific real world setting

Briefly summarize your course project