Methods for Causal Inference in Educational Research

Friday 19 August 2022

UiO : CEMO – Centre for Educational Measurement
University of Oslo

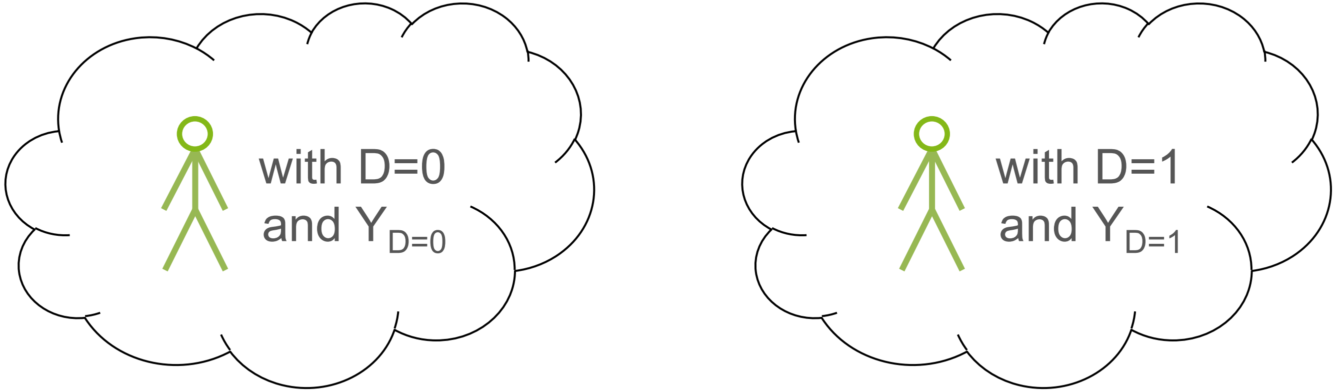
Take-away messages

- Causal inference possible if
  - plausible causal mechanism
  - treatment before outcome
  - comparison with counterfactual
  - ceteris paribus
- Very good counterfactuals and ceteris paribus are difficult to establish
- Multiple issues such as selection bias, reverse causation, and third-variable effects prohibit causal inference

Overview

- Recap: Rubin’s potential outcome framework
- The vocabulary of experiments
- Common experimental designs
- Exercise: causal and non-causal research questions
- Central issues of (quasi-)experimental designs

Rubin’s potential outcome framework



with D=0
and $Y_{D=0}$

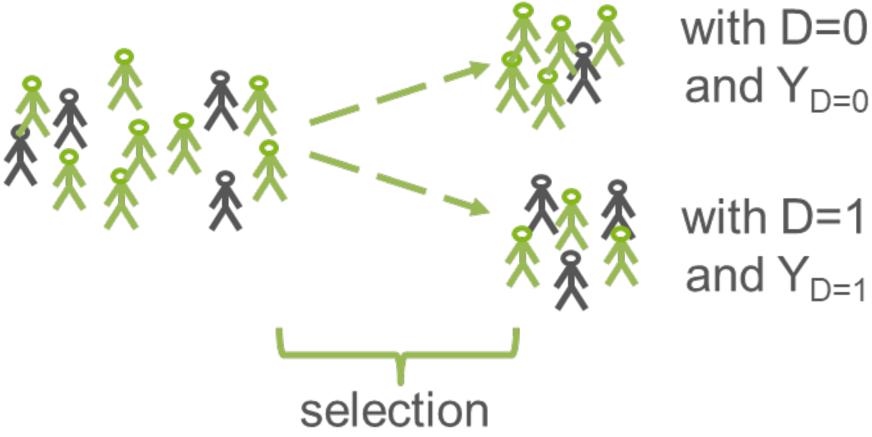
with D=1
and $Y_{D=1}$

The impossible, ideal experiment

- In *parallel universes* where the same individual is once treated and once not treated, nothing differs except the treatment and outcome (*ceteris paribus/other things equal*)
- What would have been (*potential outcome* or *counterfactual*)?
- Then, the difference $Y_{D=1}$ - $Y_{D=0}$ reflects the causal effect of D on Y

Rubin, 1974

Rubin’s potential outcome framework

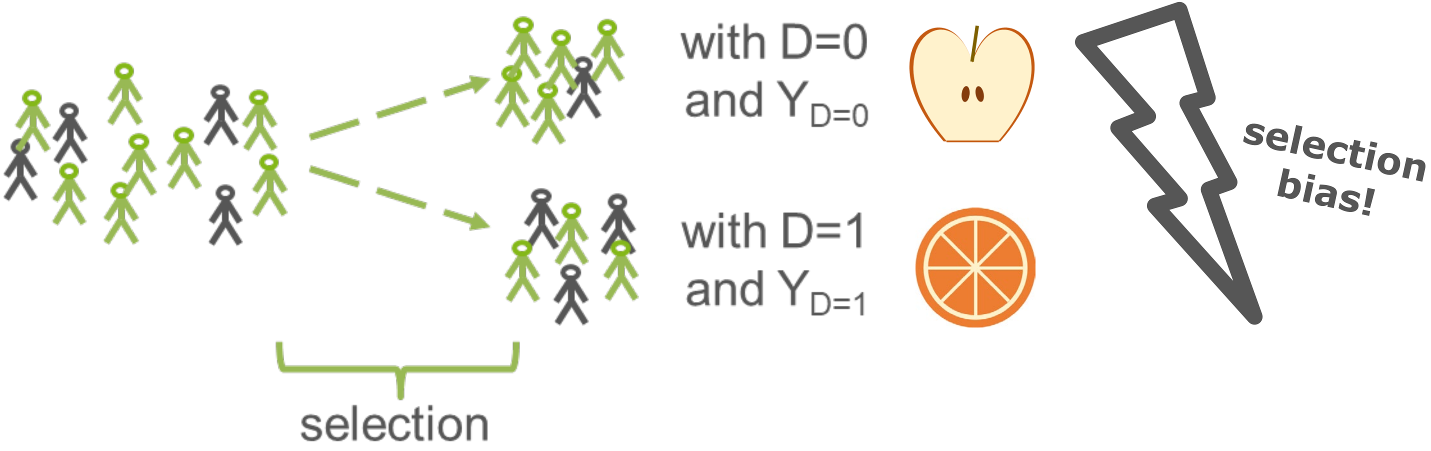


In reality, we need to compare groups

- $E[Y_{1,i}|D_i=1]$ - $E[Y_{0,i}|D_i=0]$
- Difference between treatment and control group only reflects effect of D, if other things are equal (*ceteris paribus*)
- If selection into groups relates to something that is also relevant for the outcome, difference between groups reflects causal effect and *selection bias*

Rubin, 1974

Rubin’s potential outcome framework



with D=0
and $Y_{D=0}$

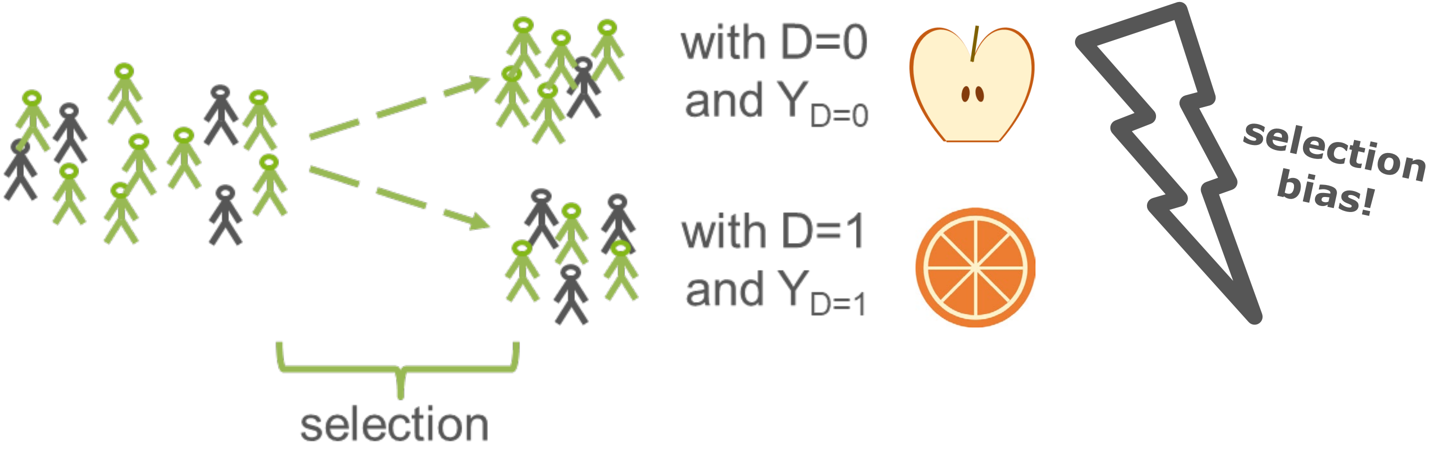
with D=1
and $Y_{D=1}$

selection

selection
bias!

In other words, selection bias means
that we compare apples and oranges!

Rubin’s potential outcome framework



with D=0 and $Y_{D=0}$

with D=1 and $Y_{D=1}$

selection

selection bias!

Example:

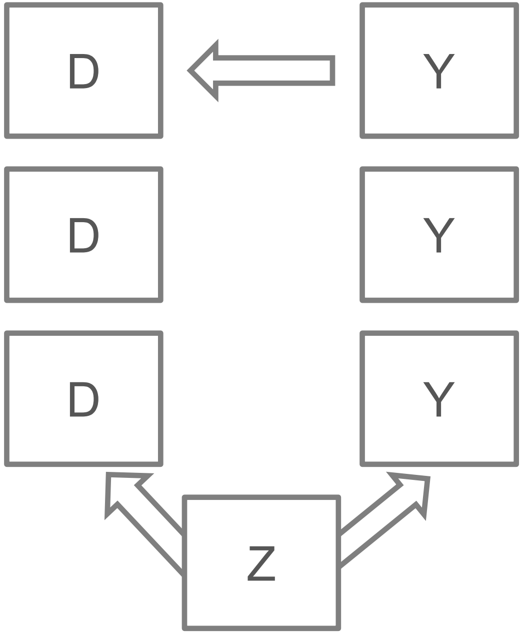
Negative correlation between private tutoring and mathematics achievement

Correlation ≠ Causation (!)

Why does correlation not prove causation?

- maybe we don’t know which variable came first (e.g., *reverse causation*)
- maybe, there is no true association, at all (e.g., *spurious correlation*)
- maybe, there are other explanations (e.g., *third-variable effect*, *confounding*, *selection bias*)

Campfire example



The vocabulary of experiments

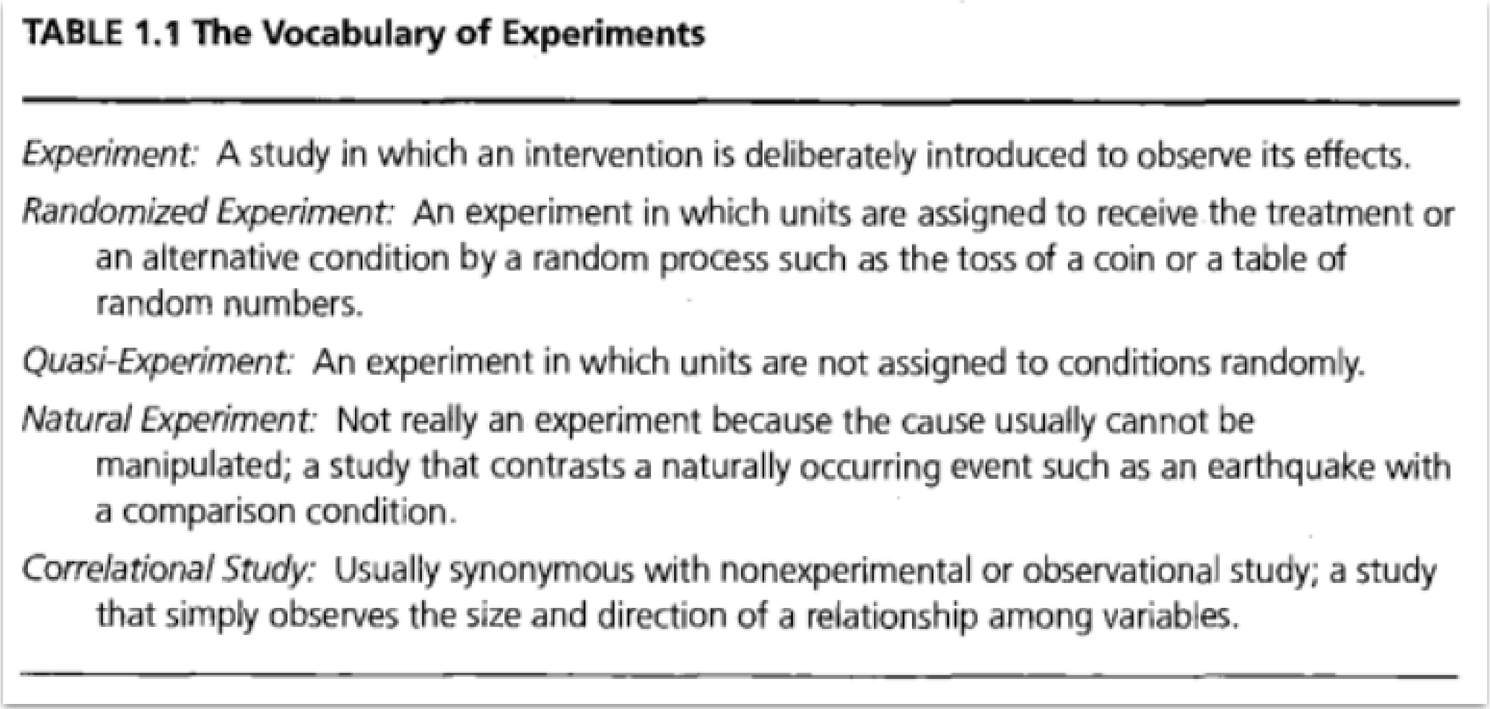


TABLE 1.1 The Vocabulary of Experiments

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.

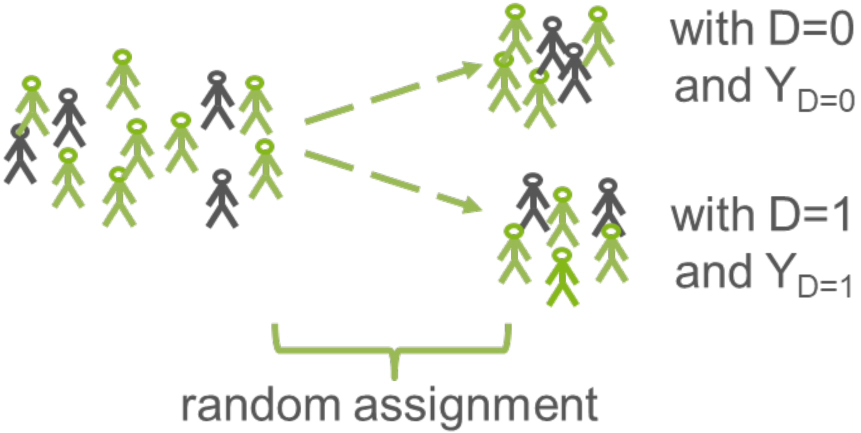
Quasi-Experiment: An experiment in which units are not assigned to conditions randomly.

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.

Shadish, Cook, & Campbell, 2002, p. 12

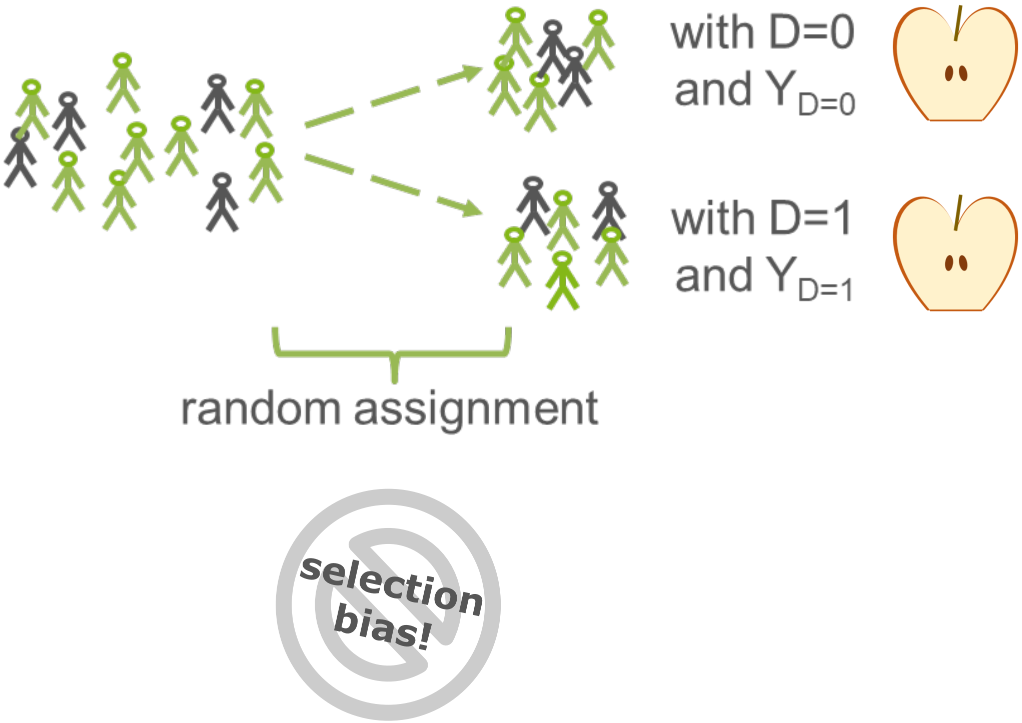
The vocabulary of experiments: randomized experiment



- creation of manipulable treatment and control condition
- randomized allocation to treatment vs. control
- if number of units large enough, randomization effectively balances the groups
- if randomization is successful, the two groups should not differ in anything but the treatment → *ceteris paribus*

Shadish, Cook, & Campbell, 2002

The vocabulary of experiments: randomized experiment



with D=0 and $Y_{D=0}$

with D=1 and $Y_{D=1}$

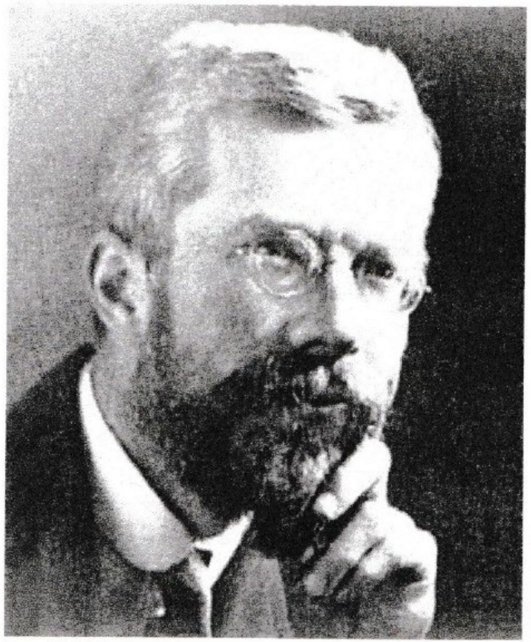
random assignment

selection bias!

- creation of manipulable treatment and control condition
- randomized allocation to treatment vs. control
- if number of units large enough, randomization effectively balances the groups
- if randomization is successful, the two groups should not differ in anything but the treatment → *ceteris paribus*

Shadish, Cook, & Campbell, 2002

The vocabulary of experiments: randomized experiment



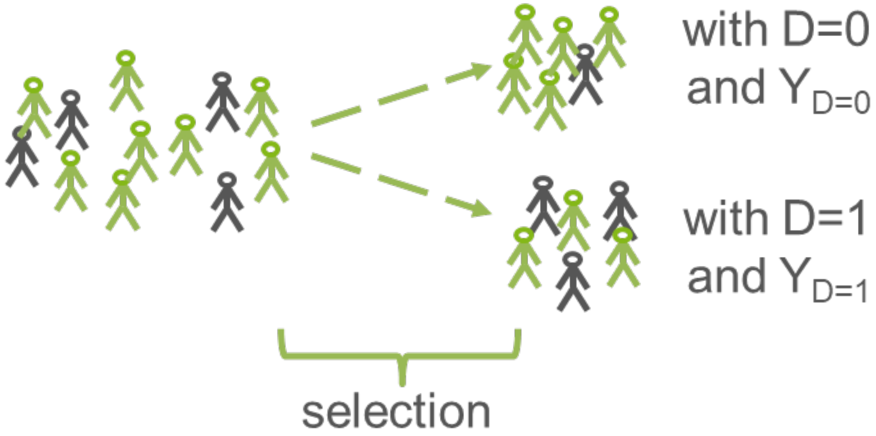
Ronald Fisher

A famous example: Fields and fertilizers

- Controlled randomized trials to find best fertilizers
- Ronald Aylmer Fisher, agricultural scientist, 1920s and 1930s
- Before Fisher, Rothamsted Agricultural Experimental Station in England tested one fertilizer each year → confounding with weather etc.
- Fisher revolutionized this practice
  - division of fields into sections
  - randomized allocation of fertilizers to sections
  - controlled repeated measures of effects
  - statistical tests for significance
- Founding father of randomized experiments (and ANOVA, F-test, p-values, etc.)

Coleman, 2019, pp. 28-30

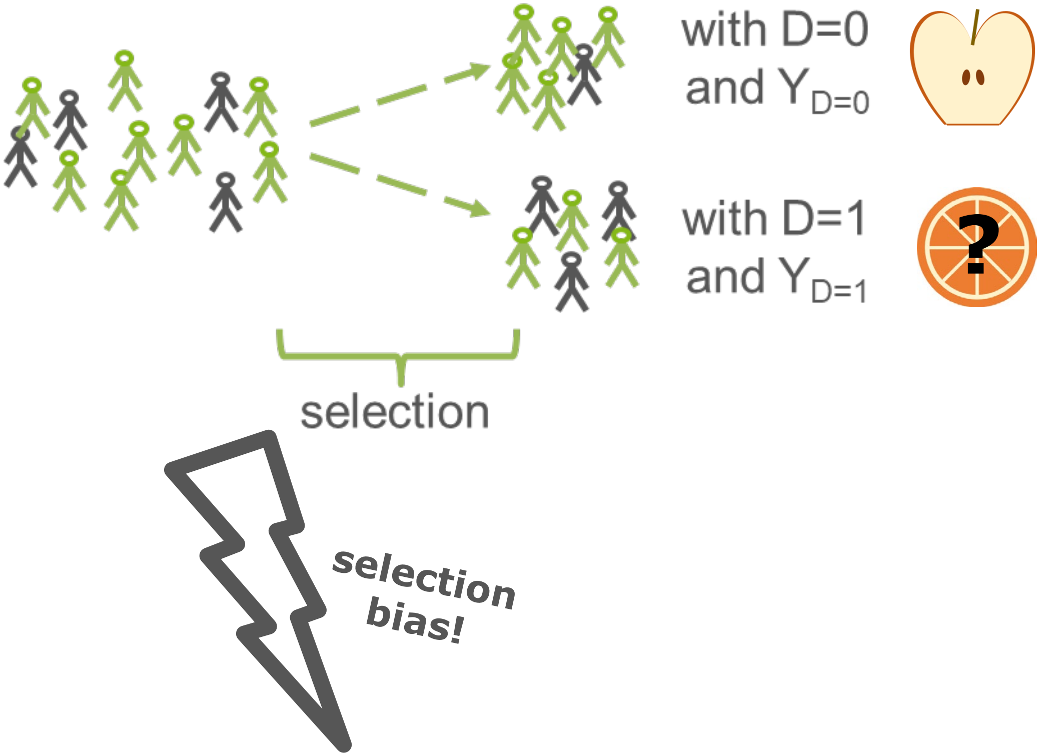
The vocabulary of experiments: quasi-experiment



- creation of manipulable treatment and control condition
- but no randomized allocation to treatment vs. control
- instead, self-selection or administrator selection (e.g., teachers, bureaucrats,...) → *selection bias threat*
- however, one can still try to make treatment and control group comparable

Shadish, Cook, & Campbell, 2002

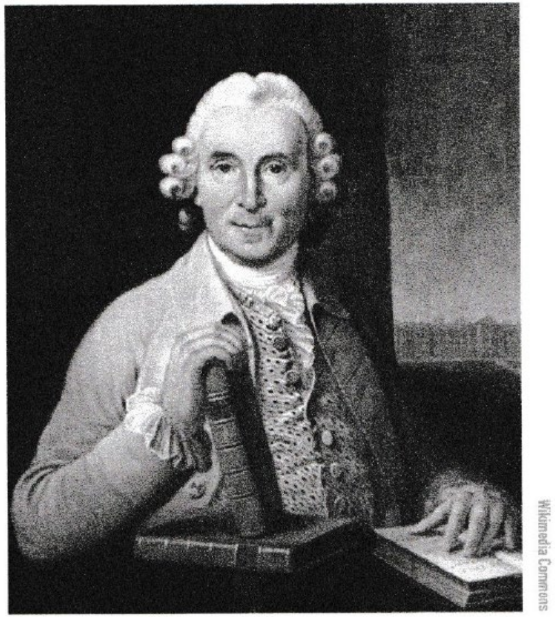
The vocabulary of experiments: quasi-experiment



- creation of manipulable treatment and control condition
- but no randomized allocation to treatment vs. control
- instead, self-selection or administrator selection (e.g., teachers, bureaucrats,...) → *selection bias threat*
- however, one can still try to make treatment and control group comparable

Shadish, Cook, & Campbell, 2002

The vocabulary of experiments: quasi-experiment



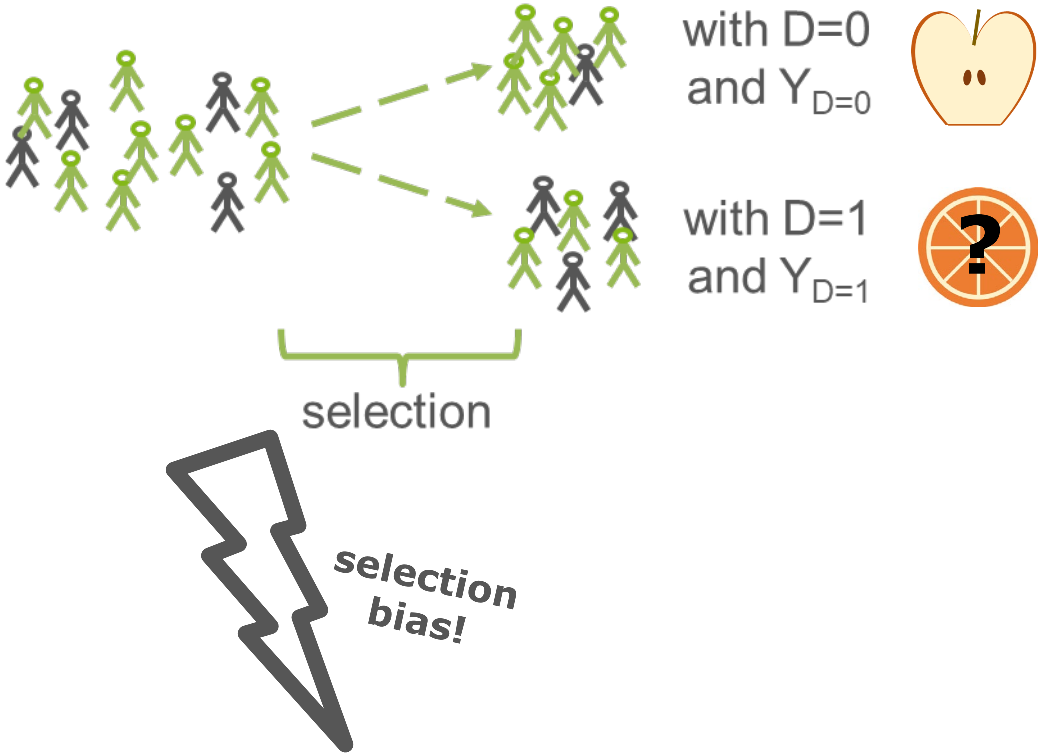
James Lind

A famous example: The scurvy studies

- Controlled experiment to find cure against scurvy
- James Lind, ship surgeon, 1747
- 12 sailors with scurvy allocated to 6 treatment groups (*matching* based on age, health, etc.):
  - quart of cider
  - sulphuric acid
  - half pint of seawater
  - mixture of garlic, mustard and horseradish
  - vinegar
  - two oranges and a lemon
- Otherwise similar conditions
- Both men in fruit group recovered after 6 days
- Conclusion that vitamin C cures scurvy

Coleman, 2019, pp. 24-27

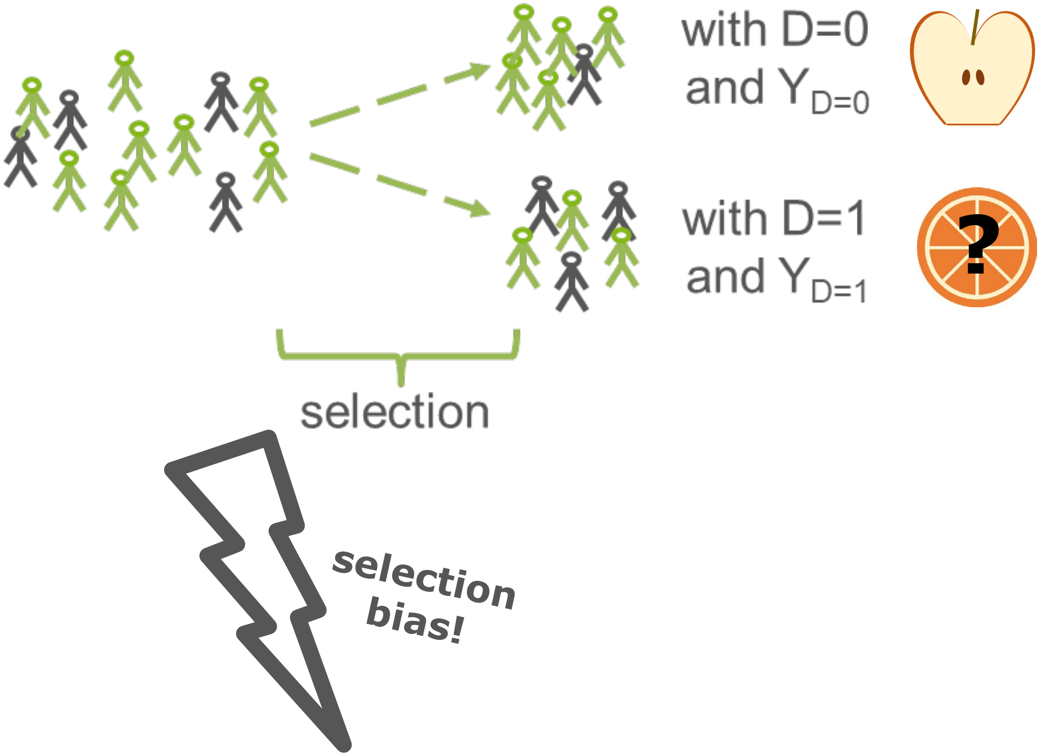
The vocabulary of experiments: natural experiment



- natural treatment and comparison conditions; often not manipulable
- complex and often intransparent selection mechanisms → *selection bias threat*
- however, meaningful comparisons can be made depending on the comparison conditions

Shadish, Cook, & Campbell, 2002

The vocabulary of experiments: non-experimental designs



- correlational design, passive observational design, or non-experimental design → *selection bias threat*
- often, all variables assessed at the same time, so unclear if treatment precedes outcome
- assumptions about causal mechanisms, but difficult to rule out other explanations for correlations

Shadish, Cook, & Campbell, 2002

Common experimental designs

- What are treatment and outcome of interest?
  - What is the exact population of interest?
  - When and how often should an outcome be observed?
  - What are treatment and control/comparison conditions?
  - …

Shadish, Cook, & Campbell, 2002

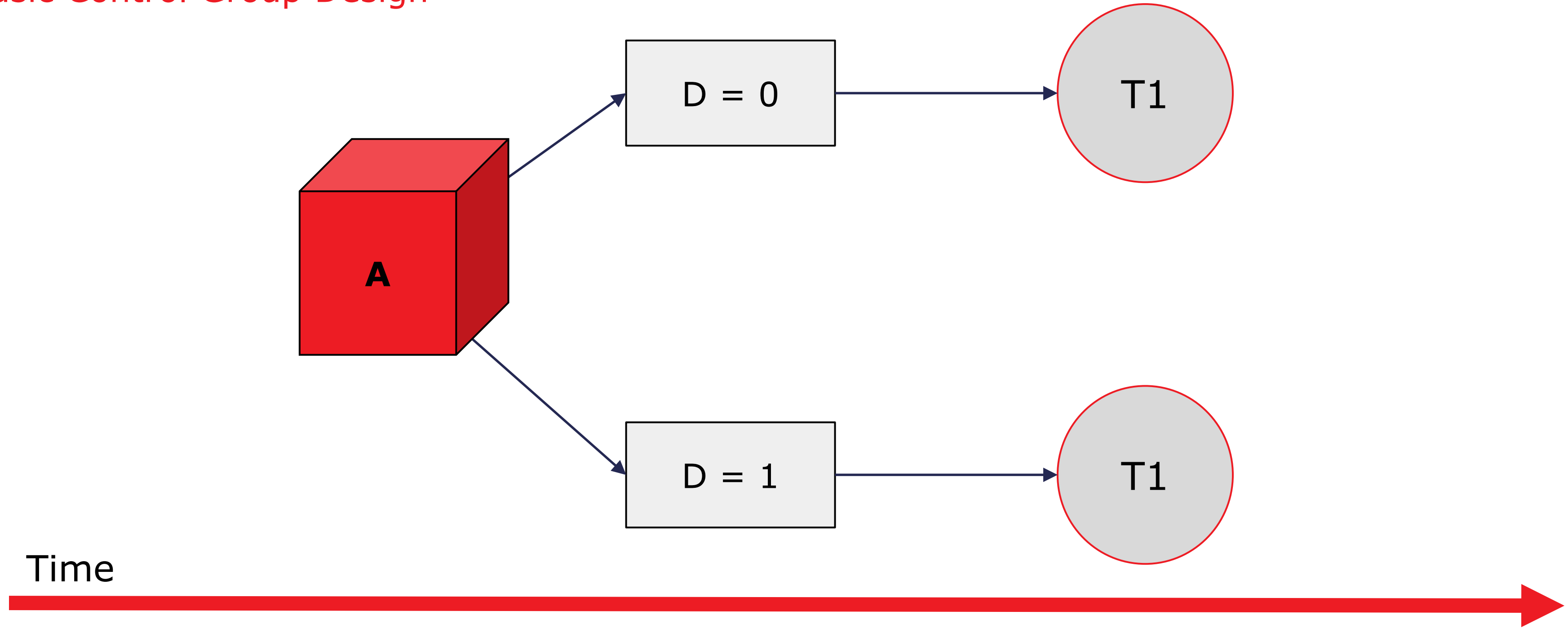
Common experimental designs

- What are treatment and outcome of interest?
  - What is the exact population of interest?
  - When and how often should an outcome be observed?
  - What are treatment and control/comparison conditions?
  - …
- What is the assumed causal mechanism, i.e., which conditions do we have to observe?
  - How can we assign or how are units assigned to groups?
  - How large should my sample size be (depends on expected effect sizes, number of compared groups, quality and reliability of measures, lower or higher order units for assignment, etc.)?
  - …

Shadish, Cook, & Campbell, 2002

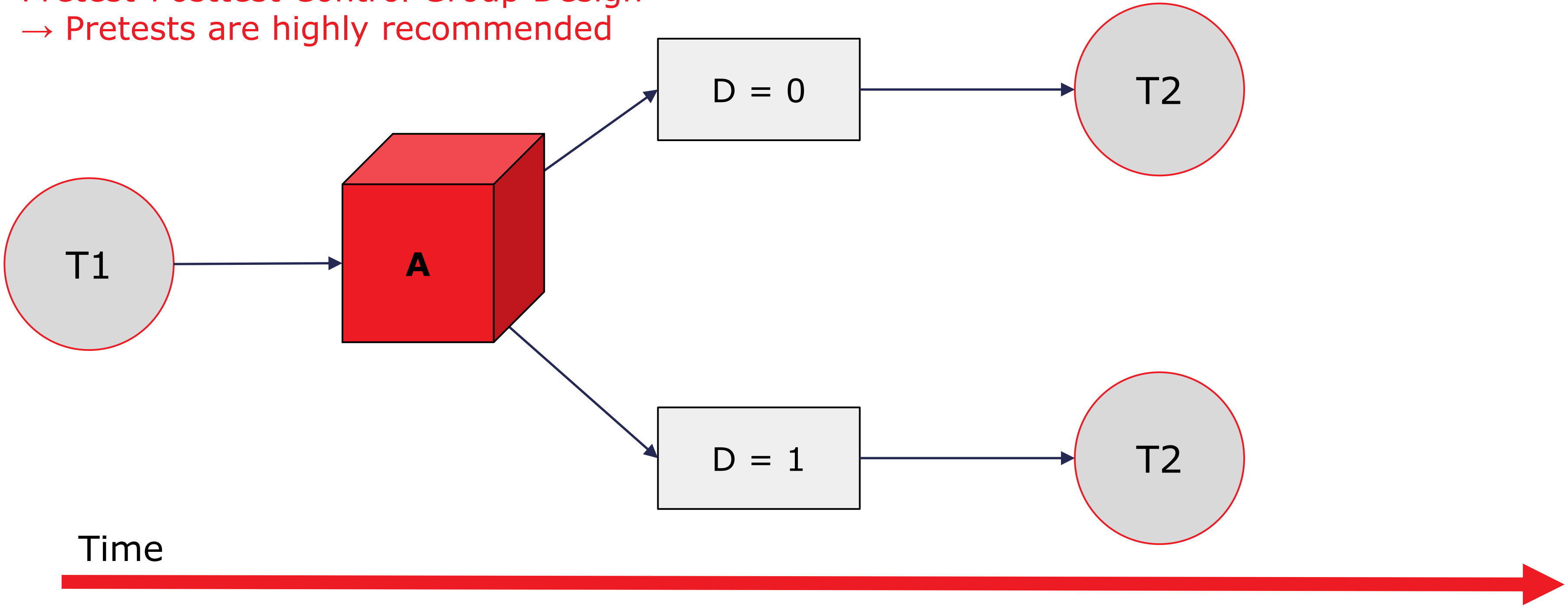
Common experimental designs

Basic Control Group Design



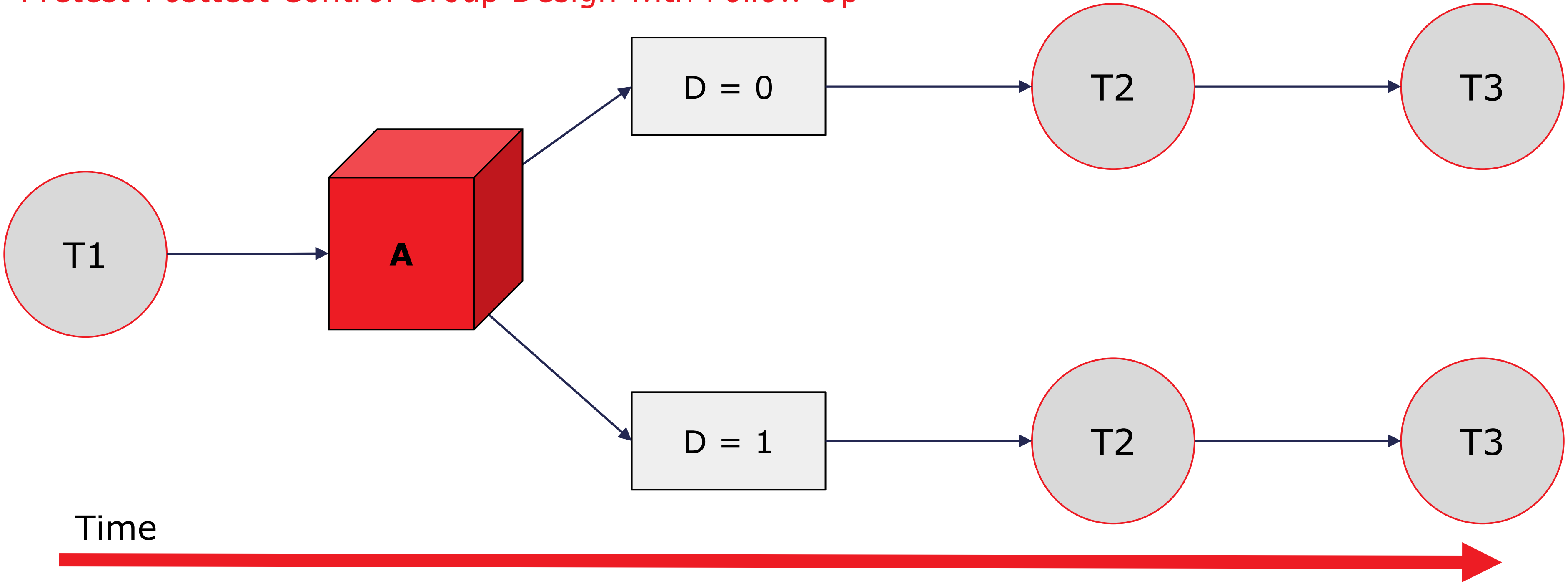
Common experimental designs

Pretest-Posttest Control Group Design
→ Pretests are highly recommended



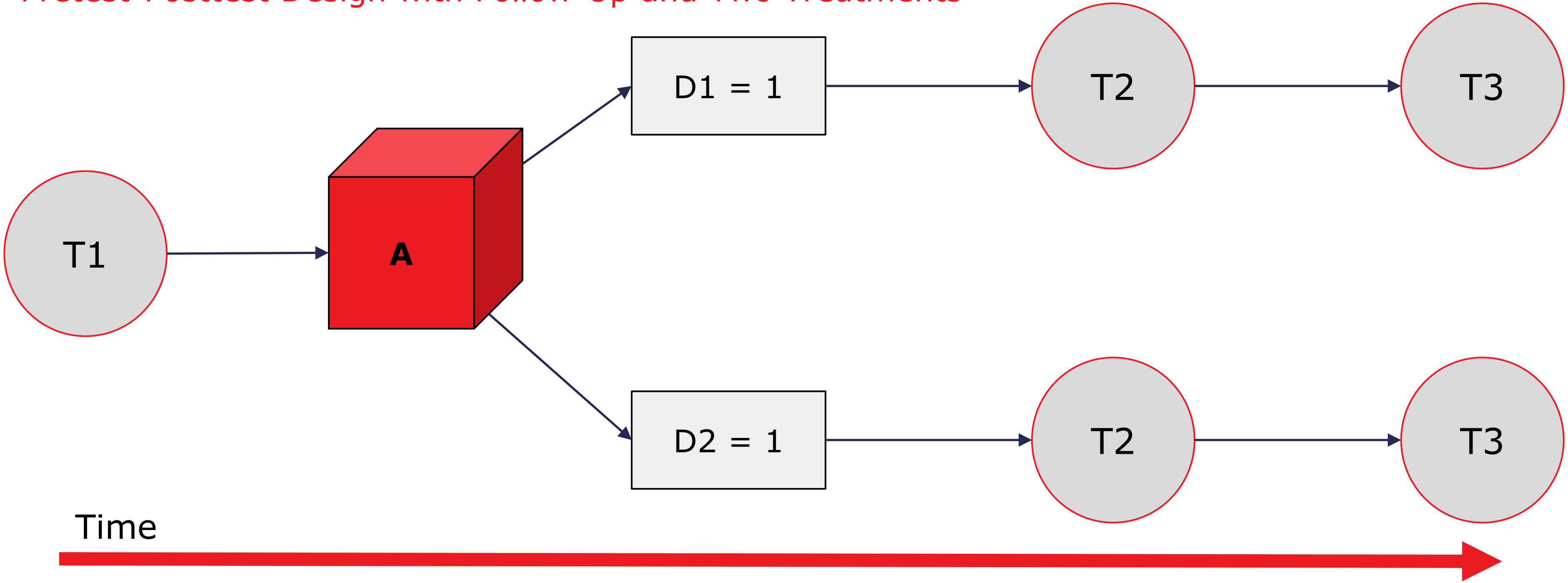
Common experimental designs

Pretest-Posttest Control Group Design with Follow-Up



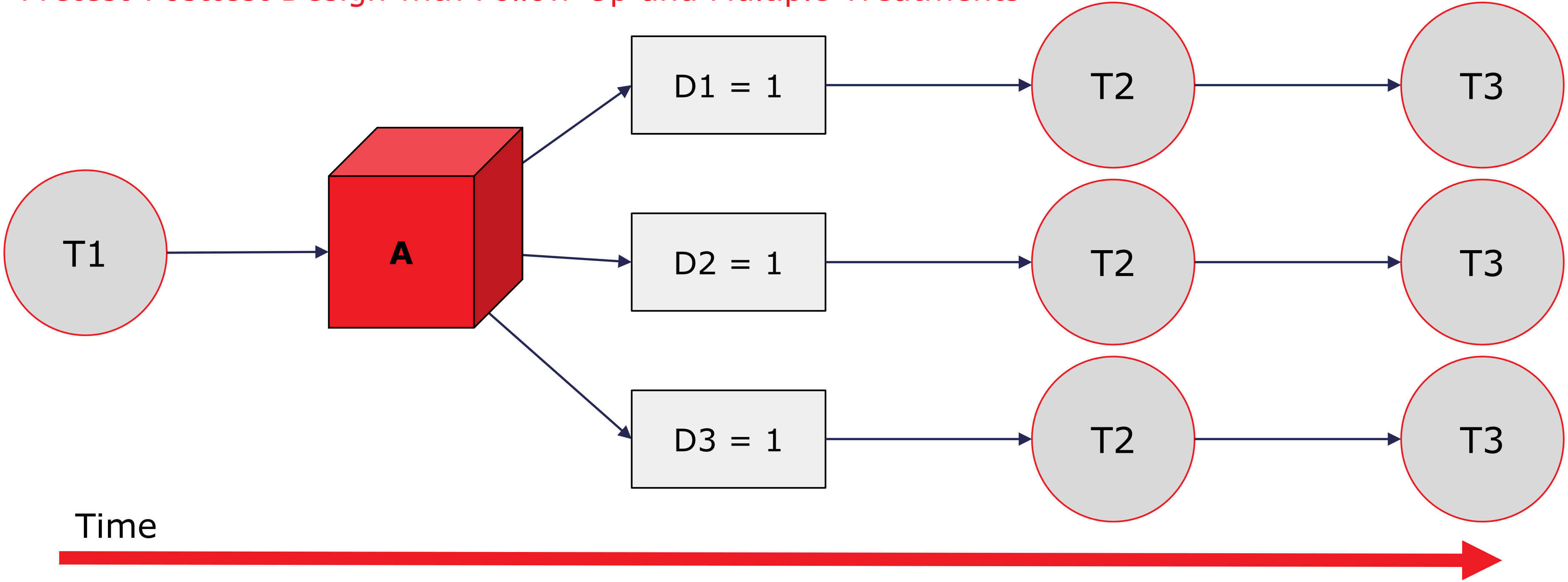
Common experimental designs

Pretest-Posttest Design with Follow-Up and Two Treatments



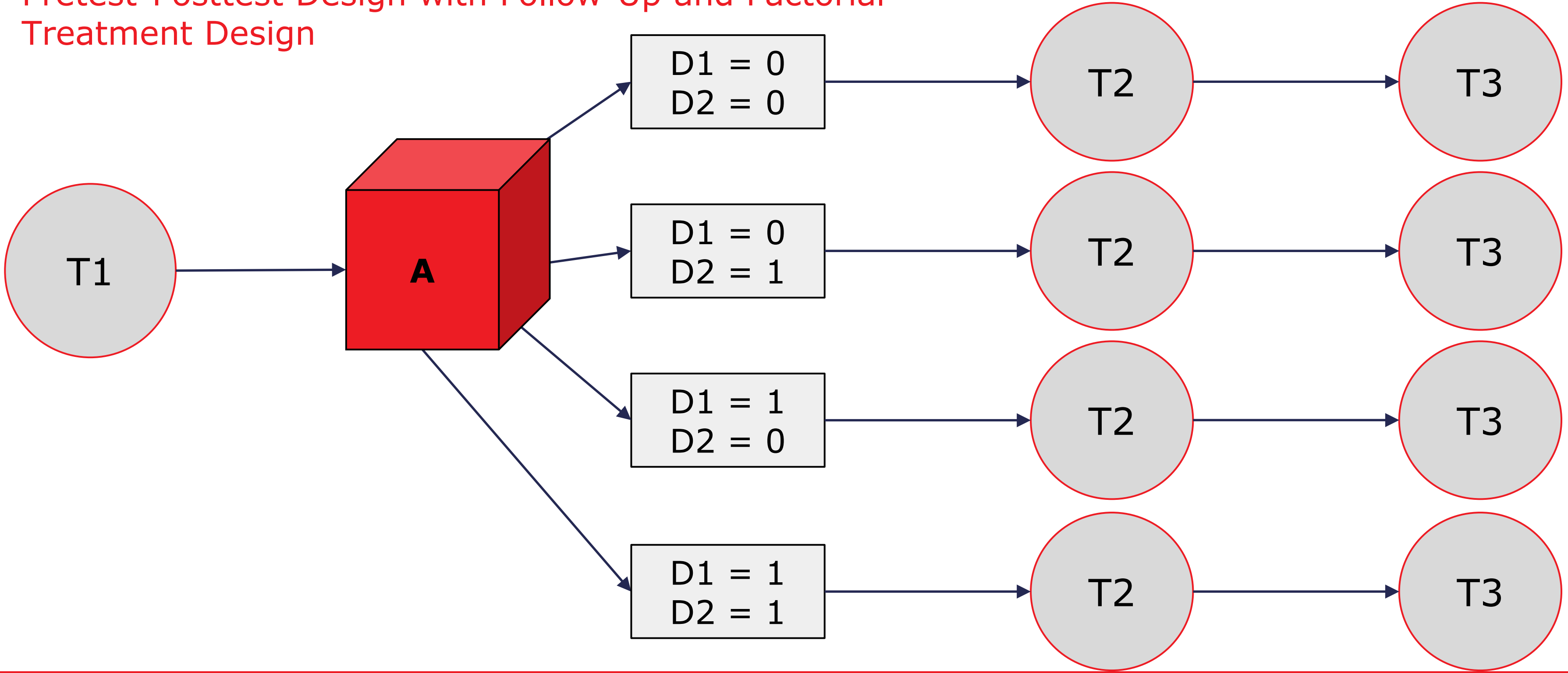
Common experimental designs

Pretest-Posttest Design with Follow-Up and Multiple Treatments



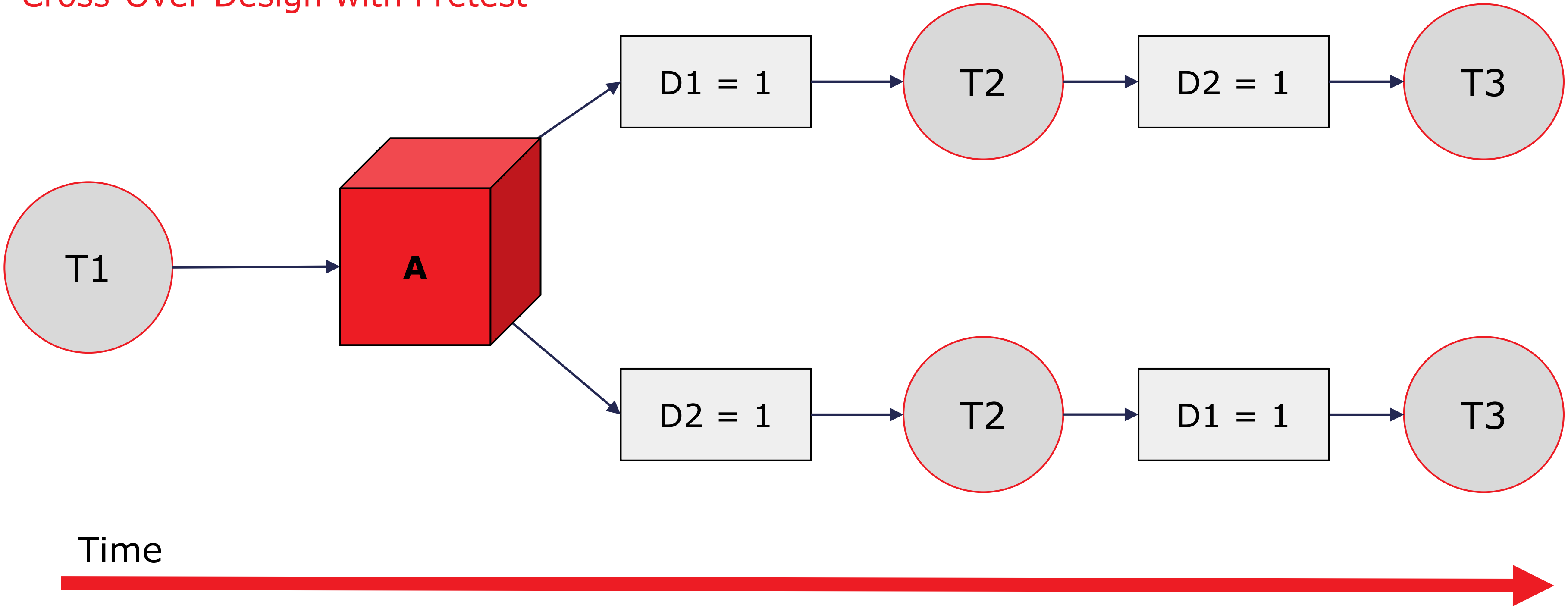
Common experimental designs

Pretest-Posttest Design with Follow-Up and Factorial Treatment Design



Common experimental designs

Cross-Over Design with Pretest



Group exercise

What are examples for causal research questions (i.e., that could be studied in (quasi-)experiments)?

What are examples for non-causal research questions (i.e., that could *not* be studied in (quasi-)experiments)?

Central issues of (quasi-)experiments

1. Only suitable for cause-and-effect research questions, not for questions like
   - what is the cause of a phenomenon?
   - why should there be an association between phenomena (i.e., theory-building)?
   - what is the nature of a phenomenon (i.e., descriptive questions)?
   - …

Shadish, Cook, & Campbell, 2002

Central issues of (quasi-)experiments

1. Only suitable for cause-and-effect research questions, not for questions like
   - what is the cause of a phenomenon?
   - why should there be an association between phenomena (i.e., theory-building)?
   - what is the nature of a phenomenon?
   - …

causal description (e.g., if I hold a match to dry leafs, they catch fire) vs. *causal explanation* (e.g., why and under which exact conditions does a match light leafs?)
→ importance of meaningful potential causal mechanisms
→ importance of consecutive fine-grained experiments and observations (incl. moderators, mediators)

Shadish, Cook, & Campbell, 2002

Central issues of (quasi-)experiments

2. Randomized experiments often not feasible for ethical (e.g., depriving students of schooling) or feasibility reasons (e.g., random assignment to school types). But even if they are
   - non-compliance issues (e.g., treatment units refuse treatment; control units get treatment anyway) → choice of comparison conditions and level of units
   - not all treatments are manipulable (e.g., gender) → careful choice of treatment (e.g., vignette studies)
   - attrition issues, especially in control groups → choice of comparison conditions, cross-over designs
   - treatment diffusion (e.g., less time with learning software than intended) → implementation checks
   - randomization is incorrectly or incompletely implemented, not enough cases (e.g., if deck of cards is shuffled, some players still get better cards) → balance checks, power analyses

Shadish, Cook, & Campbell, 2002

Central issues of (quasi-)experiments

3. Quasi-experimental designs can still suffer from selection bias (!)
   - is selection mechanism known?
   - can selection mechanism be methodologically accounted for?
   - are treatment and comparison conditions really comparable, except for the treatment (i.e., ceteris paribus)?
   - can we speculate about the size and directionality of unaccounted selection bias?

Shadish, Cook, & Campbell, 2002

Central issues of (quasi-)experiments

4. Generalizability
   - Making treatment and comparison conditions comparable implies that effects are investigated under very specific conditions → difficulty of broad interpretations
   - Would treatment be implemented similarly under other, more natural circumstances? Would treatment work in same way under other circumstances? → ecological validity questionable; importance of replication studies
   - Would we observe same effects under other circumstances? → importance of replication studies; importance of (non-convenience/probability) sampling strategies
   - Or good reasons to believe that we would observe something else under specific other circumstances? → "Grounded Theory of Causal Generalization"
     - To which degree are experimental and target generalized conditions comparable?
     - Which features do and do not matter for generalizability (probably)?
     - What are values in the variables that could have occured but did not occur?
     - Develop and test explanatory theories about the pattern of effects, causes, and mediational processes

Shadish, Cook, & Campbell, 2002

Take-away messages

- (Quasi-)experiments have the potential to answer causal research questions, but not others
- Selection bias, feasibility, implementation, and generalization are the most central issues in (quasi-)experiments

References

Angrist, J. & Pischke, J.-S. (2015). *Mastering `Metrics: The Path from Cause to Effect*. Princeton University Press

Coleman, R. (2019). *Designing experiments for the social sciences*. Sage

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Mifflin and Company