Photo credit: Dave DiCello

17-803 Empirical Methods

Bogdan Vasilescu, Institute for Software Research

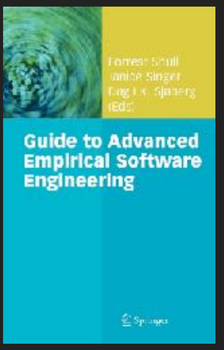
Designing Experiments (Part III)

Thursday, March 18, 2021

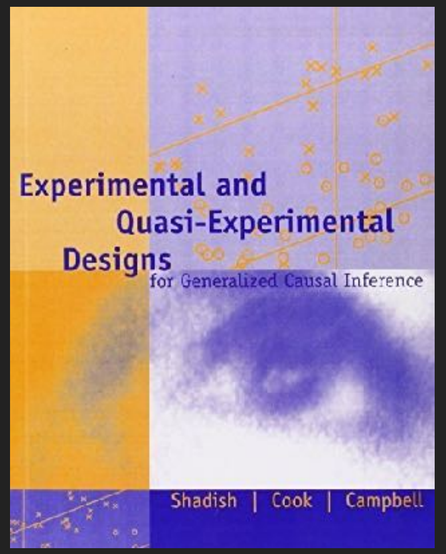
Outline for Today

- More on experimental design

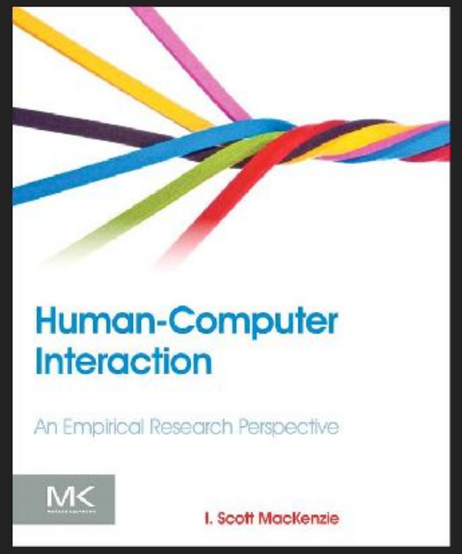
Ch 10 (Analysis and interpretation)



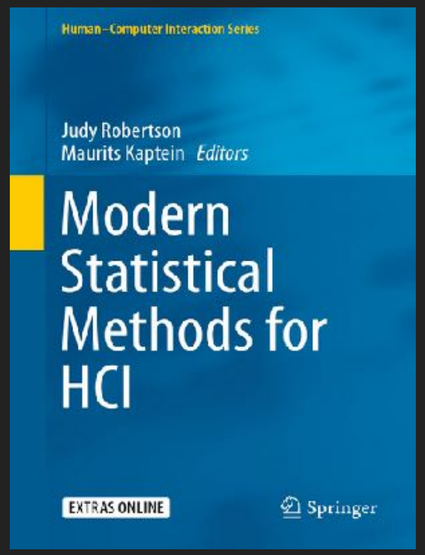
Ch 6 (Statistical methods and measurement)



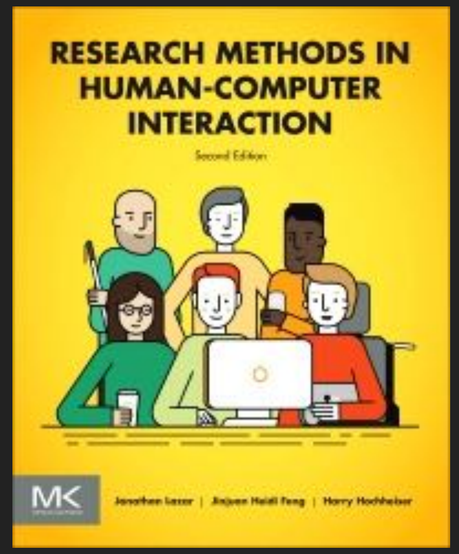
Ch 1 (Experiments and causality)
Ch 2 & 3 (Validity)
Ch 8 (Randomized experiments)



Ch 6 (Hypothesis testing)



Ch 5 (Effect sizes and power analysis)
Ch 13 (Fair statistical communication)
Ch 14 (Improving statistical practice)



Ch 3 (Experimental design)
Ch 4 (Statistical analysis)

The vocabulary of experiments

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

Quasi-Experiment

An experiment in which units are not assigned to conditions randomly

Natural Experiment

The cause usually can't be manipulated.

A study that contrasts a naturally occurring event such as an earthquake with a comparison condition

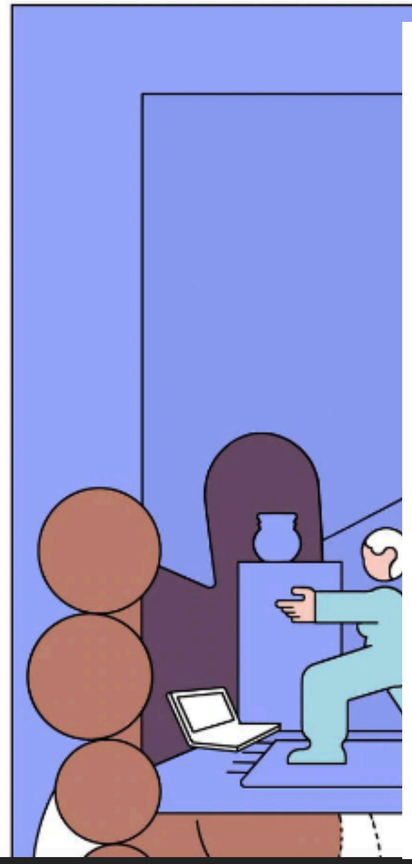
Correlational Study

Aka "observational study."

A study that simply observes the size and direction of a relationship among variables

The great experiment

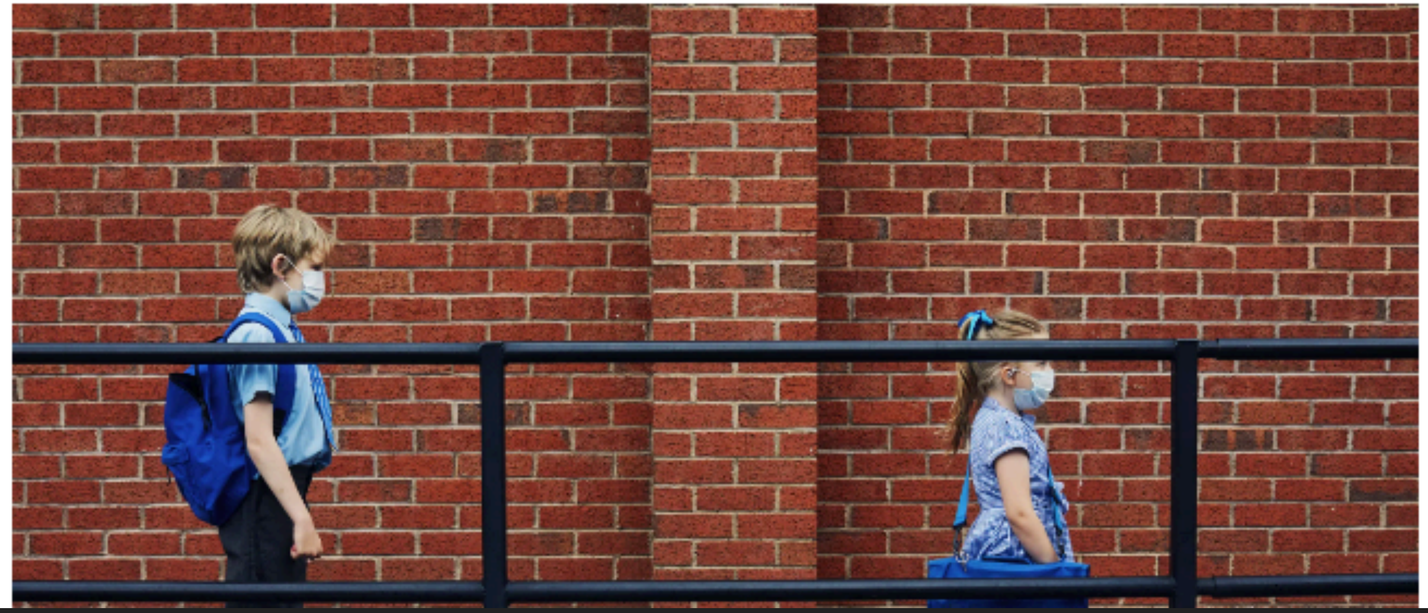
The pandemic is tragic. It's also an incredible chance to study human behavior.



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

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

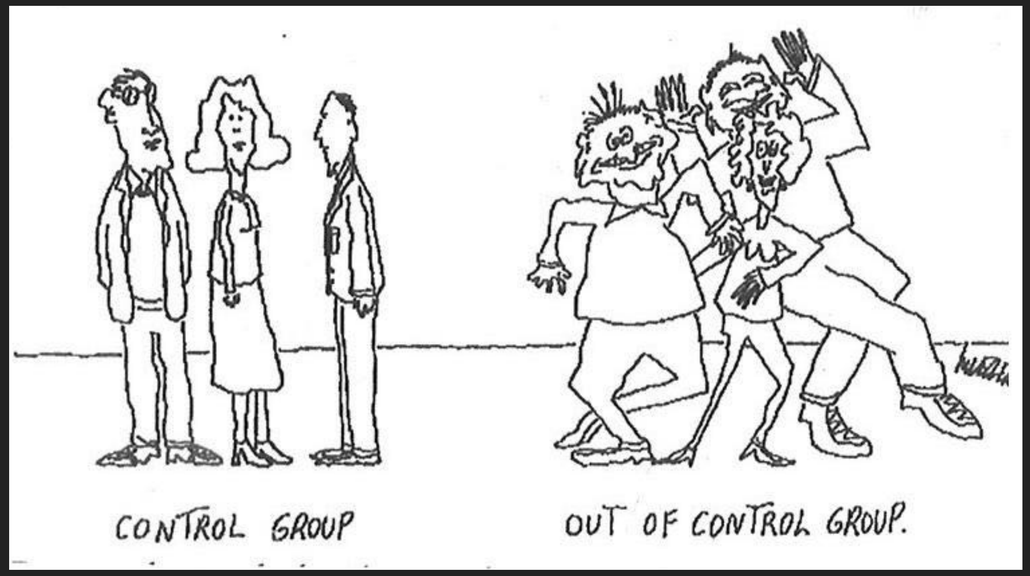
As K-12 students head back to school, epidemiologists are watching for clues about how kids spread the virus, and what can stop it.



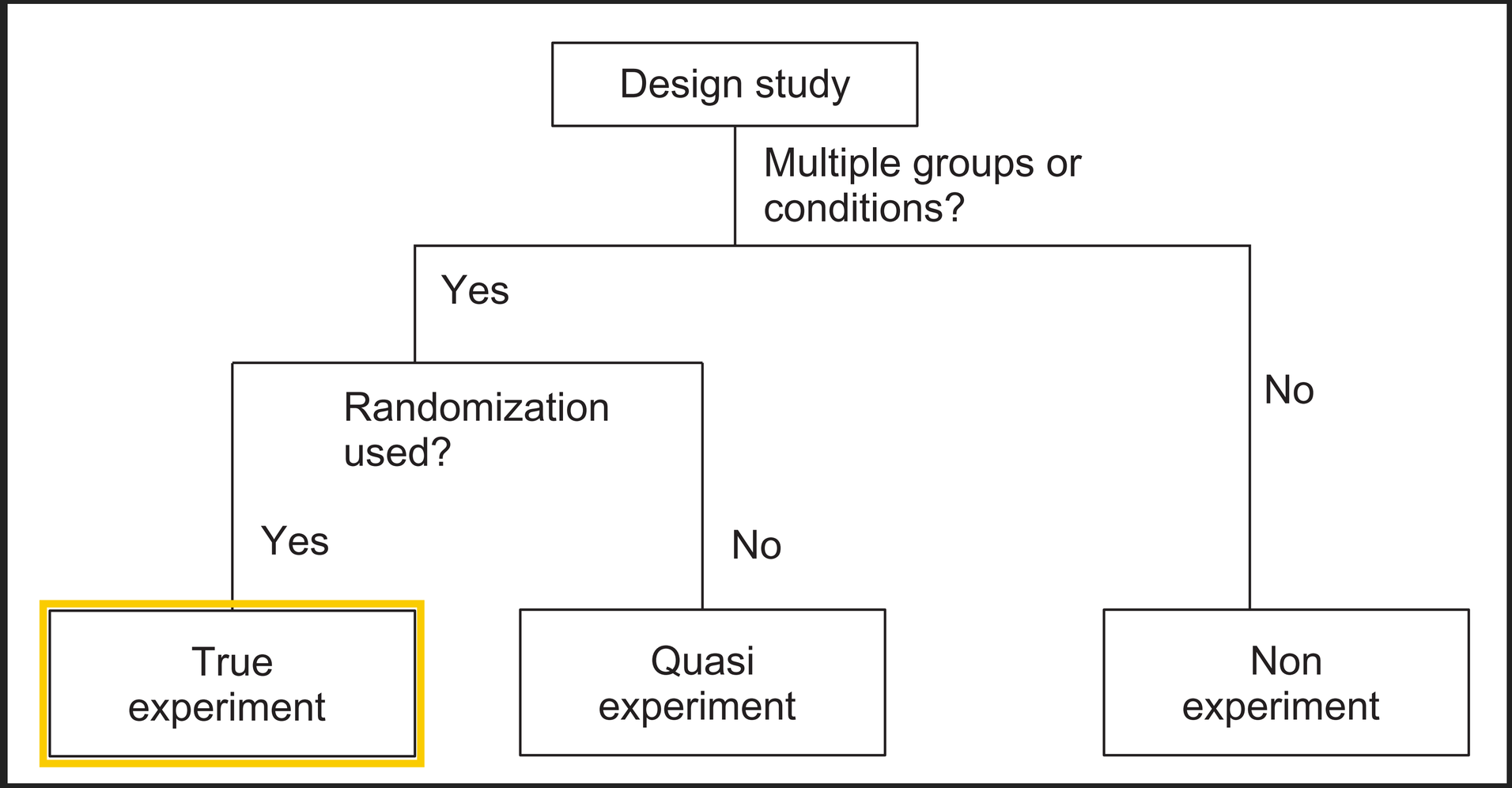
https://www.wired.com/story/a-huge-covid-19-natural-experiment-is-underway-in-classrooms/

Randomized Experiment (Sometimes "True 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.



Are You Really Doing an "Experiment"?

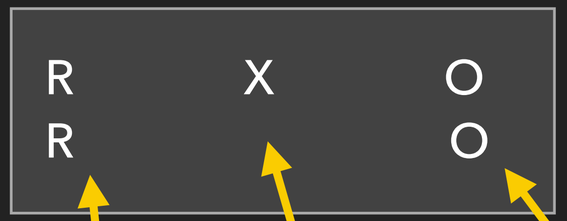


Some designs used with random assignment

Basic X vs C

R	X	O
R		O

Basic X vs C



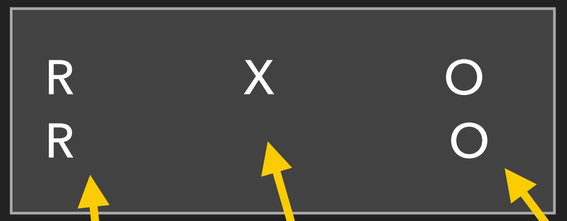
Two conditions

Posttest assessment

Treatment / Intervention

Random assignment of participants to conditions

Basic X vs C



Two conditions

Posttest assessment

Treatment / Intervention

Random assignment of participants to conditions

- Limitation:

Can’t separate active ingredients in treatment from the experience of being treated

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

- Innovative treatment vs gold standard
- Limitation:
  - If no effect, can't distinguish if both treatments were equally effective or equally ineffective

- Innovative treatment vs gold standard vs control

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

- Common limitation: Lack of pretest
  - Especially if attrition
  - But not always undesirable
    - E.g., unwanted sensitization effect from pretest, physically impossible to collect, constant (all alive)

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

- Some extra statistical analysis advantages, besides robustness to attrition.

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

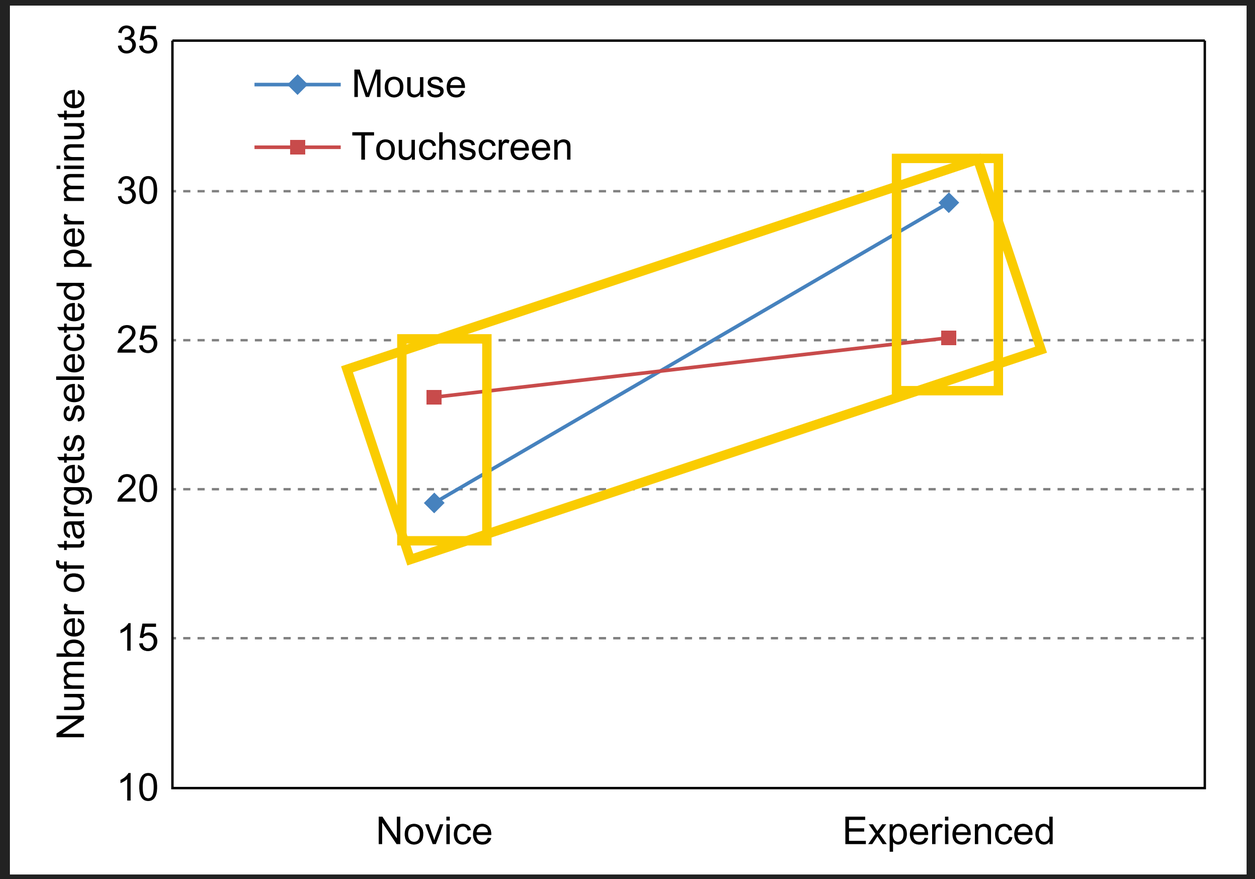
Factorial

R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

- Three major advantages:
  - They often require fewer units.
  - They allow testing combinations of treatments more easily.
  - They allow testing interactions.

Example of Interaction Effects

- Novice users can select targets faster with a touchscreen than with a mouse.
- Experienced users can select targets faster with a mouse than with a touchscreen.
- The target selection speeds for both the mouse and the touchscreen increase as the user gains more experience with the device.
- However, the increase in speed is much larger for the mouse than for the touchscreen.



Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

Factorial

R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

Longitudinal

R	O … O	X	O … O
R	O … O		O … O

- Examine how effects change over time

Basic X vs C

R	X	O
R		O

Basic X_A vs X_B

R	X_A	O
R	X_B	O

Basic X_A vs X_B vs C

R	X_A	O
R	X_B	O
R		O

Pretest-posttest

R	O	X	O
R	O		O

Alternative Xs with pretest

R	O	X_A	O
R	O	X_B	O

Factorial

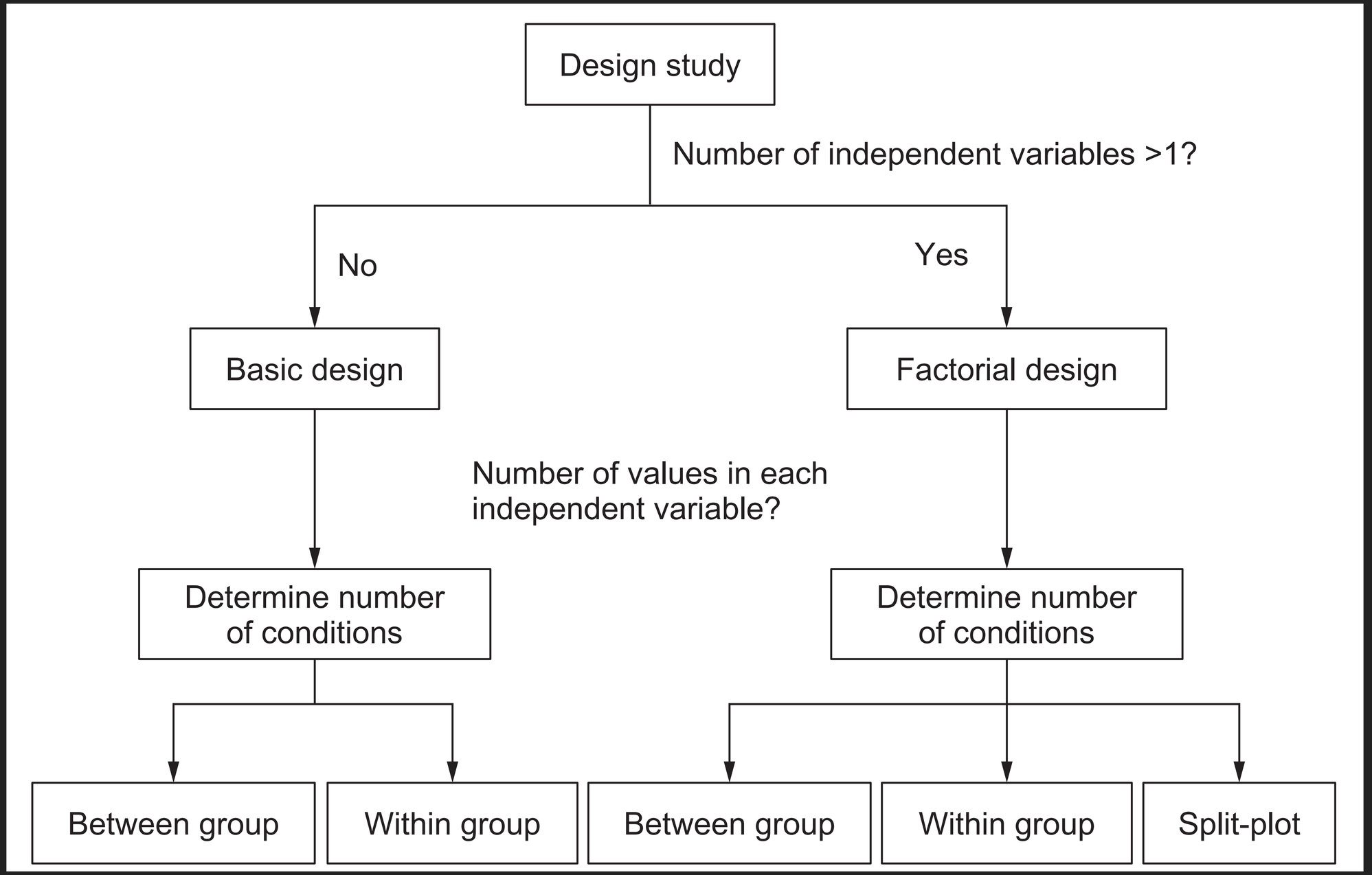
R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

- Used to counterbalance and assess order effects with multiple treatments

Crossover

R	O	X_A	O	X_B	O
R	O	X_B	O	X_A	O

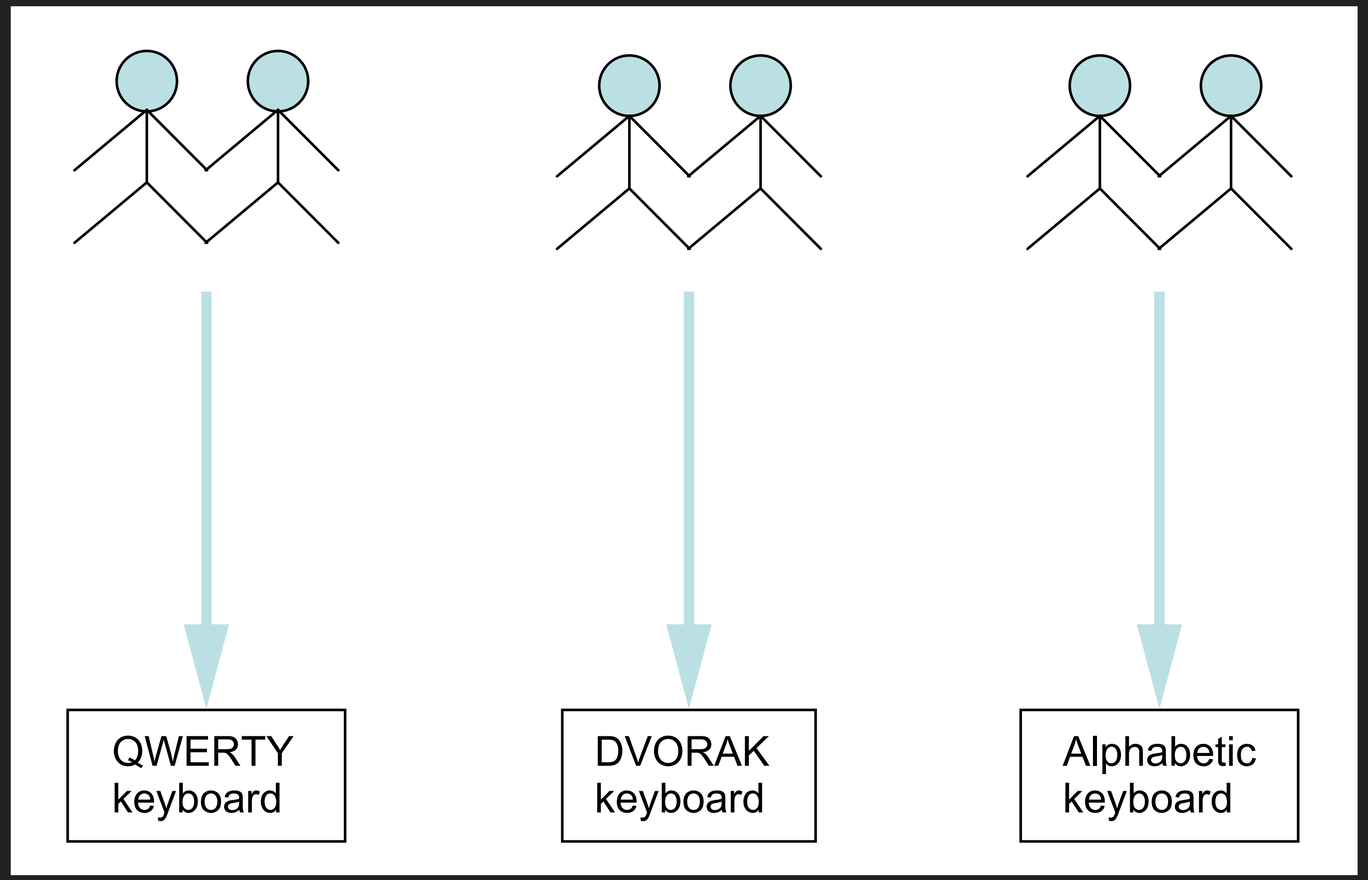
Another way to think about designs


Design study
Number of independent variables >1?
No
Yes
Basic design
Factorial design
Number of values in each independent variable?
Determine number of conditions
Determine number of conditions
Between group
Within group
Between group
Within group
Split-plot

Between-Group Design

- Aka “between-subject design.”
- Each participant is only exposed to one experimental condition.

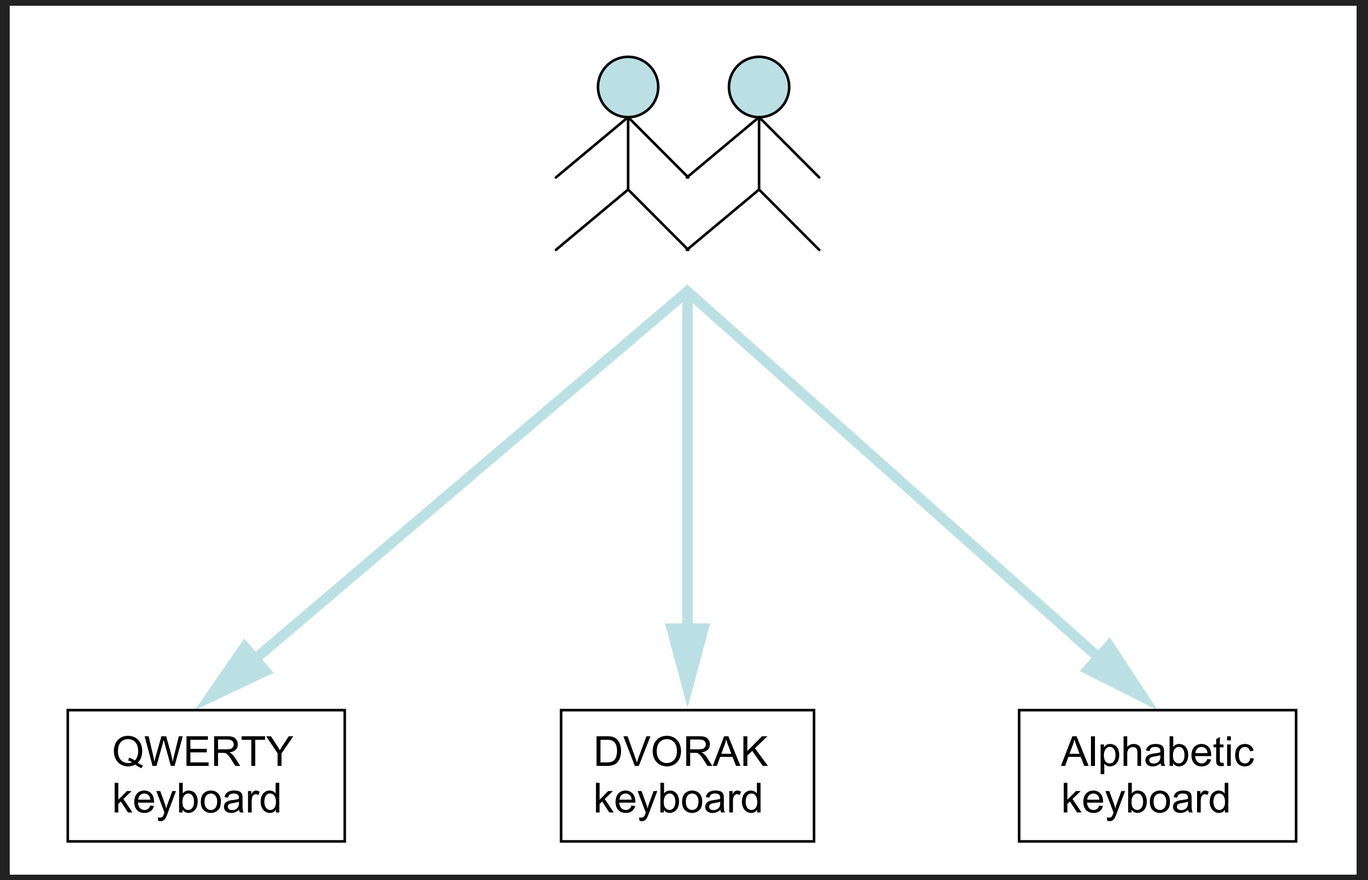
- E.g., if the task is to type a 500-word doc, each participant types one doc using one of the keyboards.



Within-Group Design

- Aka “within-subject design.”
- Each participant is exposed to multiple experimental conditions.

- E.g., each participant types three docs, using each of the three keyboards for one doc.



Comparison of Between-Group and Within-Group Designs

Table 3.1 Advantages and Disadvantages of Between-Group Design and Within-Group Design

	Type of Experiment Design	
	Between-Group Design	**Within-Group Design**
Advantages	Cleaner Avoids learning effect Better control of confounding factors, such as fatigue	Smaller sample size Effective isolation of individual differences More powerful tests
Limitations	Larger sample size Large impact of individual differences Harder to get statistically significant results	Hard to control learning effect Large impact of fatigue

The generalization of causal connections

Four Types of Validity

Statistical Conclusion Validity

The validity of inferences about the correlation (covariation) between treatment and outcome.

Internal Validity

The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.

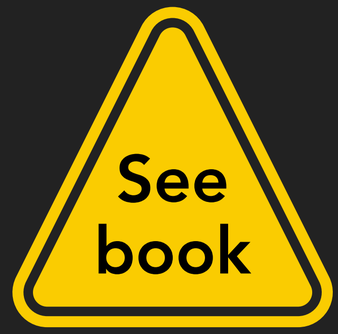
Construct Validity

The validity of inferences about the higher order constructs that represent sampling particulars.

External Validity

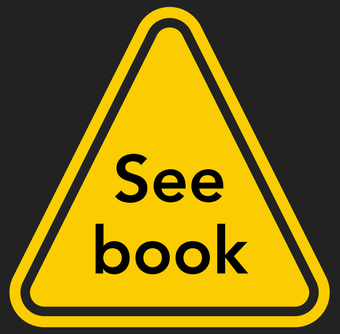
The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

Construct Validity



- Can we generalize results to the theoretical constructs that the units, treatments, observations, and settings are supposed to represent?
- E.g., whether
  - patient education (the target cause)
  - promotes physical recovery (the target effect)
  - among surgical patients (the target population of units)
  - in hospitals (the target universe of settings)
- Do the actual manipulations and measures used in the experiment really tap into the specific cause and effect constructs specified by the theory?

External Validity



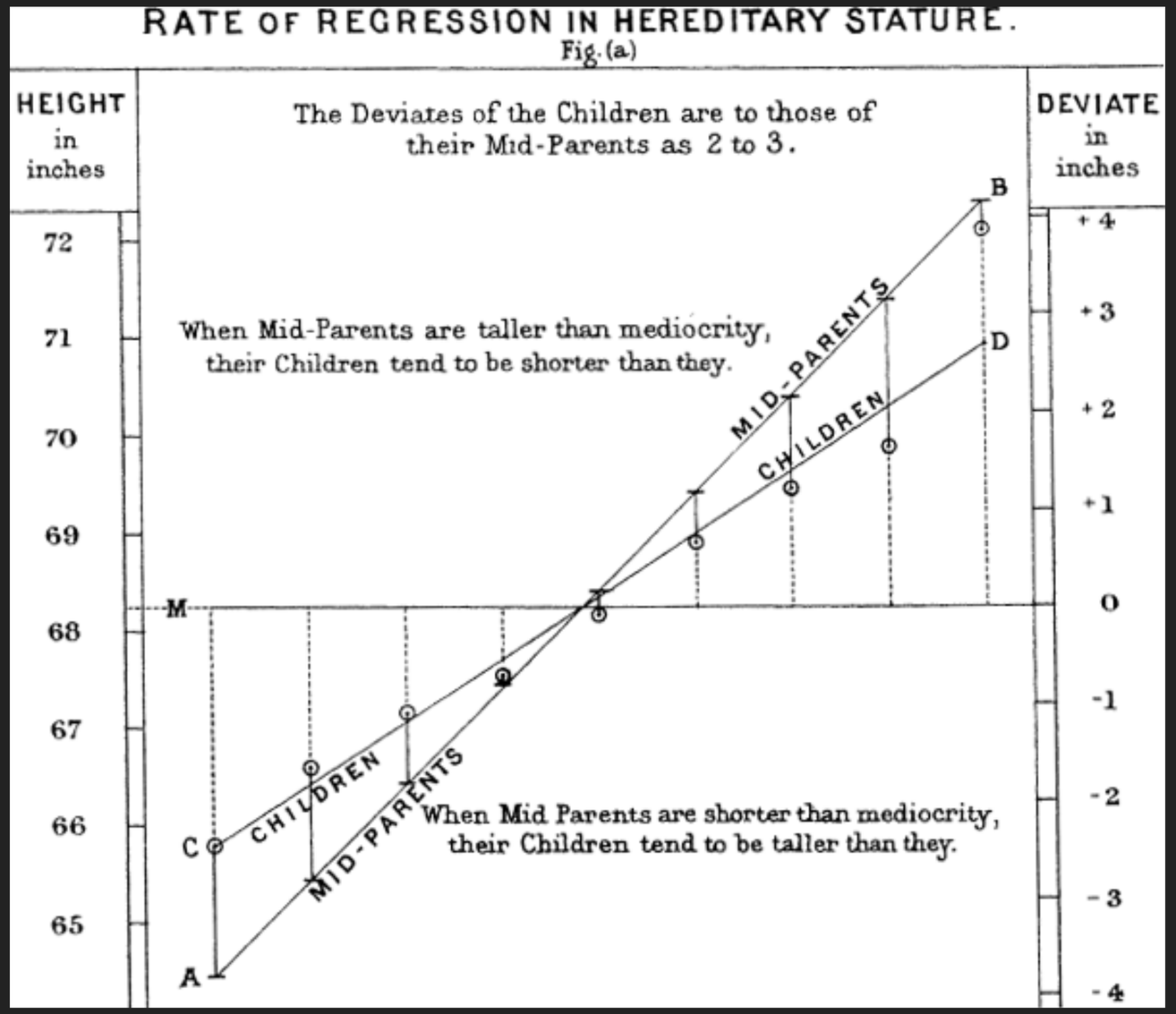
- Does the causal relationship hold over variations in persons, settings, treatments, and outcomes?
  - Narrow to broad?
  - Broad to narrow?
  - Across units at the same level of aggregation?

A Few Threats to Internal Validity

- Ambiguous Temporal Precedence:
  - Which variable occurred first?
- Selection:
  - Systematic differences over conditions in respondent characteristics.
- History:
  - Events occurring concurrently with treatment.
- Maturation:
  - Naturally occurring changes over time confused with a treatment effect.
- Regression:
  - When units are selected for their extreme scores, they will often have less extreme scores on other variables.

Regression to the Mean

- Phenomenon involving successive measurements on a given variable.
- Extreme observations tend to be followed by more central ones.
  - E.g., the children of extremely tall men tend not to be as tall as their father [Galton-1886].



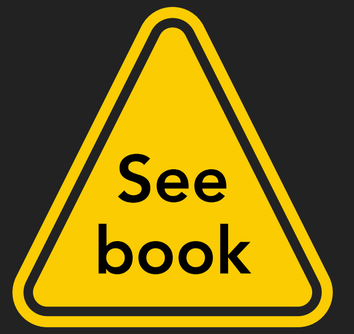
A Few Threats to Internal Validity

- Ambiguous Temporal Precedence:
  - Which variable occurred first?
- Selection:
  - Systematic differences over conditions in respondent characteristics.
- History:
  - Events occurring concurrently with treatment.
- Maturation:
  - Naturally occurring changes over time confused with a treatment effect.
- Regression:
  - When units are selected for their extreme scores, they will often have less extreme scores on other variables.
- Attrition:
  - Loss of respondents to treatment or to measurement.
- Testing:
  - Exposure to a test can affect scores on subsequent exposures to that test.
- Instrumentation:
  - The nature of a measure may change over time or conditions.

Statistical Conclusion Validity

- Two related statistical inferences that affect the covariation component of causal inferences:
  - whether the presumed cause and effect covary.
  - how strongly they covary.
- Type I error:
  - incorrectly conclude that cause and effect covary when they do not.
- Type II error:
  - incorrectly conclude that they do not covary when they do.

A Few Threats to Statistical Conclusion Validity



- Low Statistical Power:
  - → Type II errors
- Violated assumptions of statistical tests:
  - Either over- or underestimate the size and significance of an effect.
- Fishing:
  - Repeated tests can inflate statistical significance.
- Unreliability of measures
- Restriction of range on variable:
  - Typically weakens the relationship between it and another variable.
  - E.g., don't dichotomize.

Hypothesis Tests

- Aka “significance tests”
- Purpose:
  - Could random chance be responsible for an observed effect?
- Null hypothesis (H_0):
  - The hypothesis that chance is to blame.
  - e.g., “There is no difference in the mean time to complete a task using NL2Code vs. writing code from scratch.”
- Alternative hypothesis (H_a):
  - Counterpoint to the null (what you hope to prove).
  - e.g., “It takes less time on average to complete a task using NL2Code rather than by writing code from scratch.”

Aside: Why Do We Need a Hypothesis? Why Not Just Look at the Outcome of the Experiment and Go With Whichever Treatment Does Better?

- Experiment: invent a series of 50 coin flips.
  - Write down a series of random 1s and 0s: [1, 0, 1, 0, 1, 0, …]

Aside: Why Do We Need a Hypothesis? Why Not Just Look at the Outcome of the Experiment and Go With Whichever Treatment Does Better?

- Experiment: invent a series of 50 coin flips.
  - Write down a series of random 1s and 0s: [1, 0, 1, 0, 1, 0, …]
- Humans have a tendency to underestimate randomness.
- Computer-generated "real" coin flip results vs made-up human results:
  - the real ones will have longer runs of 1s or 0s.
  - median length of subsequences of 1s in a row:
    - 5 for the computer-generated sequences
    - only 4 for the human-generated set
- When most of us are inventing random coin flips and we have gotten three or four 1s in a row, we tell ourselves that, for the series to look random, we had better switch to 0.

Aside: How Do You Interpret the P-Value?

- H_0: "There is no difference in the mean time to complete a task using NL2Code vs. writing code from scratch."
- H_a: "It takes less time on average to complete a task using NL2Code rather than writing code from scratch."

- You run some statistical test (e.g., t-test) and obtain a P-value.

Aside: P-Value Controversy

- What we would like the p-value to convey:
  - (We hope for a low value, so we can conclude that we've proved something.)

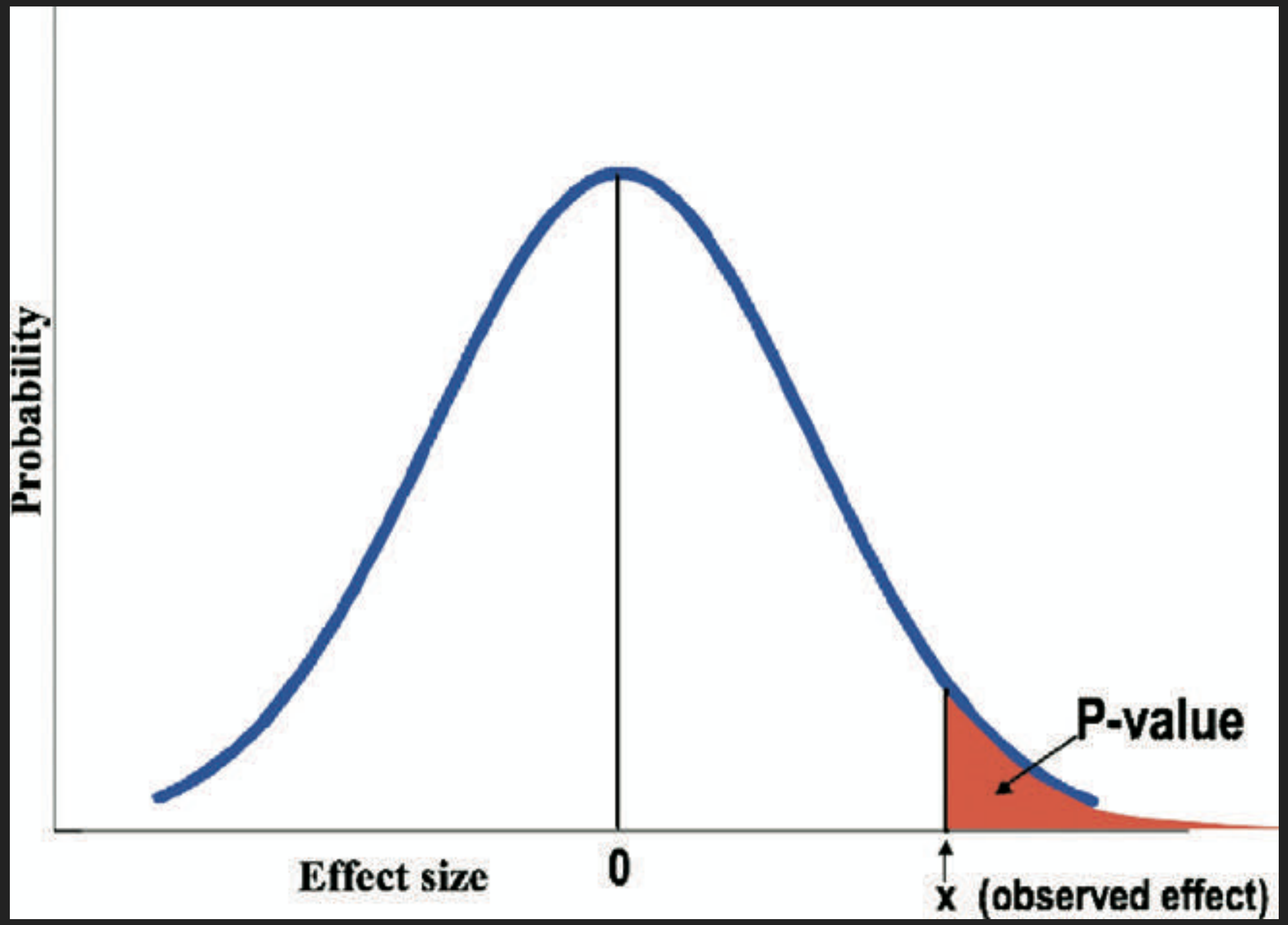
The probability that the result is due to chance: $P(H_0|D)$

- What the p-value actually represents:

The probability that, given a chance model, results as extreme as the observed results could occur: $P(D|H_0)$

Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI.
In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

The P Value Is the Probability of the Observed Outcome (X) Plus all "More Extreme" Outcomes



Graphical depiction of the definition of a (one-sided) P value. The curve represents the probability of every observed outcome under the null hypothesis.

The P Value Is the Probability of the Observed Outcome (X) Plus all "More Extreme" Outcomes

- Not the probability that the null hypothesis is true!
- Example: Is a coin fair or not?
  - H_0: The coin is fair: P(Heads) = P(Tails) = 1/2
  - H_a: The coin is biased: P(Heads) ≠ 1/2

Consider Four Consecutive Coin Flips:

Probability

- First toss:

?

Consider Four Consecutive Coin Flips:

Probability

- First toss:  0.5
- Second toss:   ?

Consider Four Consecutive Coin Flips:

Probability

- First toss:  0.5
- Second toss:  0.25
- Third toss:  0.125
- Fourth toss:  0.0625

Is Coin Fair?

- Two-sided P = 0.125.

0.0625

0.0625

- This does not mean that the probability of the coin being fair is only 12.5%!

Is Coin Fair?

- Two-sided P = 0.125.

0.0625

0.0625

- This does not mean that the probability of the coin being fair is only 12.5%!

$$P(H_0|D) = \frac{P(D|H_0)\,P(H_0)}{P(D)}$$

Common false belief that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to external evidence or the plausibility of the underlying mechanism.

Twelve P-Value Misconceptions

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

Goodman, S. (2008, July). A dirty dozen: twelve p-value misconceptions. In Seminars in hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.

Type I and Type II Errors

		Study conclusion	
		No difference	Using NL2Code is faster
Reality	No difference	✓	Type I error
	Using NL2Code is faster	Type II error	✓

Type I and Type II Errors

- In assessing statistical significance, two types of error are possible:
  - Type I: you mistakenly conclude an effect is real, when it is really just due to chance
    - False positives
  - Type II: you mistakenly conclude that an effect is due to chance, when it actually is real
    - False negatives
- The basic function of hypothesis tests is to protect against being fooled by random chance; thus they are typically structured to minimize Type I errors.

Controlling the Risks of Type I and Type II Errors

- The probability of making a Type I error is called alpha.
  - (or “significance level”, “P-value”)
- The probability of making a Type II error is called beta.
- The statistical power of a test, defined as $1 - \beta$, refers to the probability of successfully rejecting a null hypothesis when it is false and should be rejected.
- To reduce errors:
  - Type I: $P < 0.05$
  - Type II: large sample size

Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the probability of Type I errors (false positives)?

Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the probability of Type I errors (false positives)?

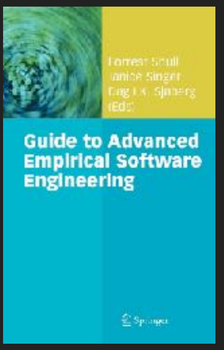
- The probability that one will correctly test nonsignificant is 0.95
- The probability that all 20 will correctly test nonsignificant is:
  - $0.95 \times 0.95 \times 0.95...$, or $0.95^{20} = 0.36$
- The probability that at least one predictor will (falsely) test significant:
  - 1 – (probability that all will be nonsignificant) = 0.64 

Credits

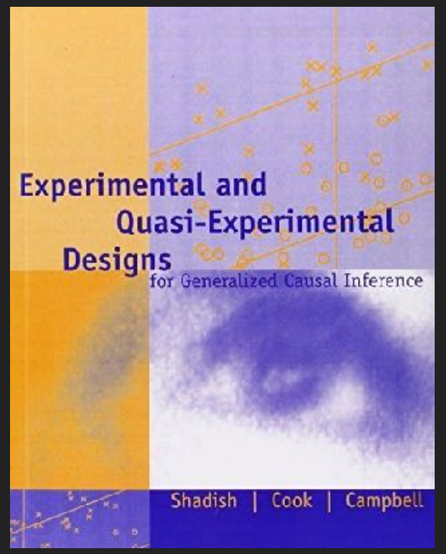
- Graphics: Dave DiCello photography (cover)
- Chapters from Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Publishing
  - Ch1: Experiments and generalized causal inference
  - Ch2: Statistical conclusion validity and internal validity
  - Ch3: Construct validity and external validity
  - Ch8: Randomized experiments
- Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.
- Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (2007). Statistics.
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. In Seminars in Hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.
- Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann.
  - Ch 3: Experimental design
  - Ch 4: Statistical analysis
- MacKenzie, I. S. (2012). Human-computer interaction: An empirical research perspective.
  - Ch 6: Hypothesis testing
- Robertson, J., & Kaptein, M. (Eds.). (2016). Modern statistical methods for HCI. Cham: Springer.
  - Ch 5: Effect sizes and power analysis
  - Ch 13: Fair statistical communication
  - Ch 14: Improving statistical practice
- Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

Read

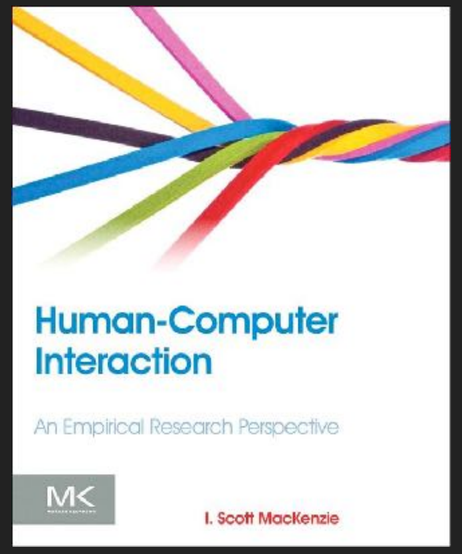
Ch 10 (Analysis and interpretation)



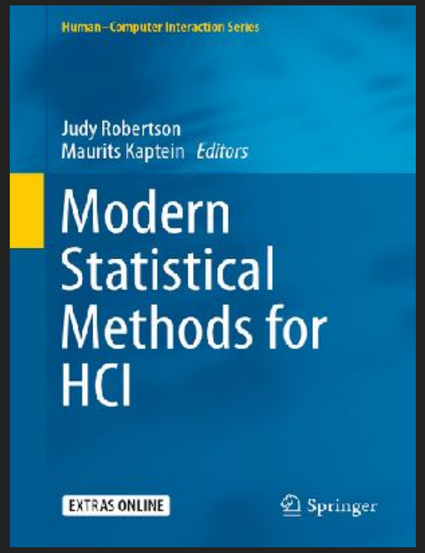
Ch 6 (Statistical methods and measurement)



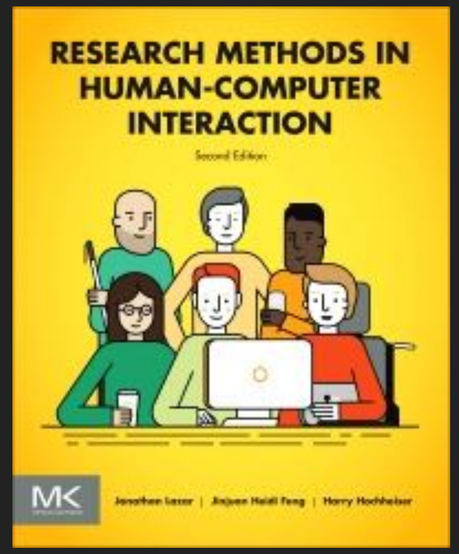
Ch 1 (Experiments and causality)
Ch 2 & 3 (Validity)
Ch 8 (Randomized experiments)



Ch 6 (Hypothesis testing)



Ch 5 (Effect sizes and power analysis)
Ch 13 (Fair statistical communication)
Ch 14 (Improving statistical practice)



Ch 3 (Experimental design)
Ch 4 (Statistical analysis)