

experimental methods

Feminist and Experimental Philosophy Workshop

Author: Jordan Wylie

Date: May 7th, 2021

****Slides adapted from slides by Natasha Karp & Maureen Coyle****

workshop contents

Part 1:

Introduction

General concepts & definitions

Part 3:

Survey Research

Experiment creation + rules /
steps to follow

Part 2:

Experimental Design

Part 4:

Open science + collaborations

Ethics and the IRB

Part 1: Introduction

Who are you?

What is experimental philosophy?

1. In the late 17th and 18th centuries, a name for the new discipline of experimental science then emerging. Use of the term often went with an optimism about the ability of experimental science to answer the questions that had been posed but unsolved by “natural philosophy.” The systematic work of Isaac Newton is often given as a defining example of experimental philosophy.
2. A late 20th-century movement holding that modern experimental science, particularly neuroscience, will ultimately uncover the biological foundations of thought and thereby provide a material answer to the questions of epistemology. In other words, experimental philosophy holds that answers to philosophical questions regarding the mind and its activities can, and likely will, be reduced to questions of how the brain functions. See reductionism.

–APA Dictionary of Psychology

“The empirical investigation of philosophical intuitions, the factors that affect them, and the psychological and neurological mechanisms that underlie them.”

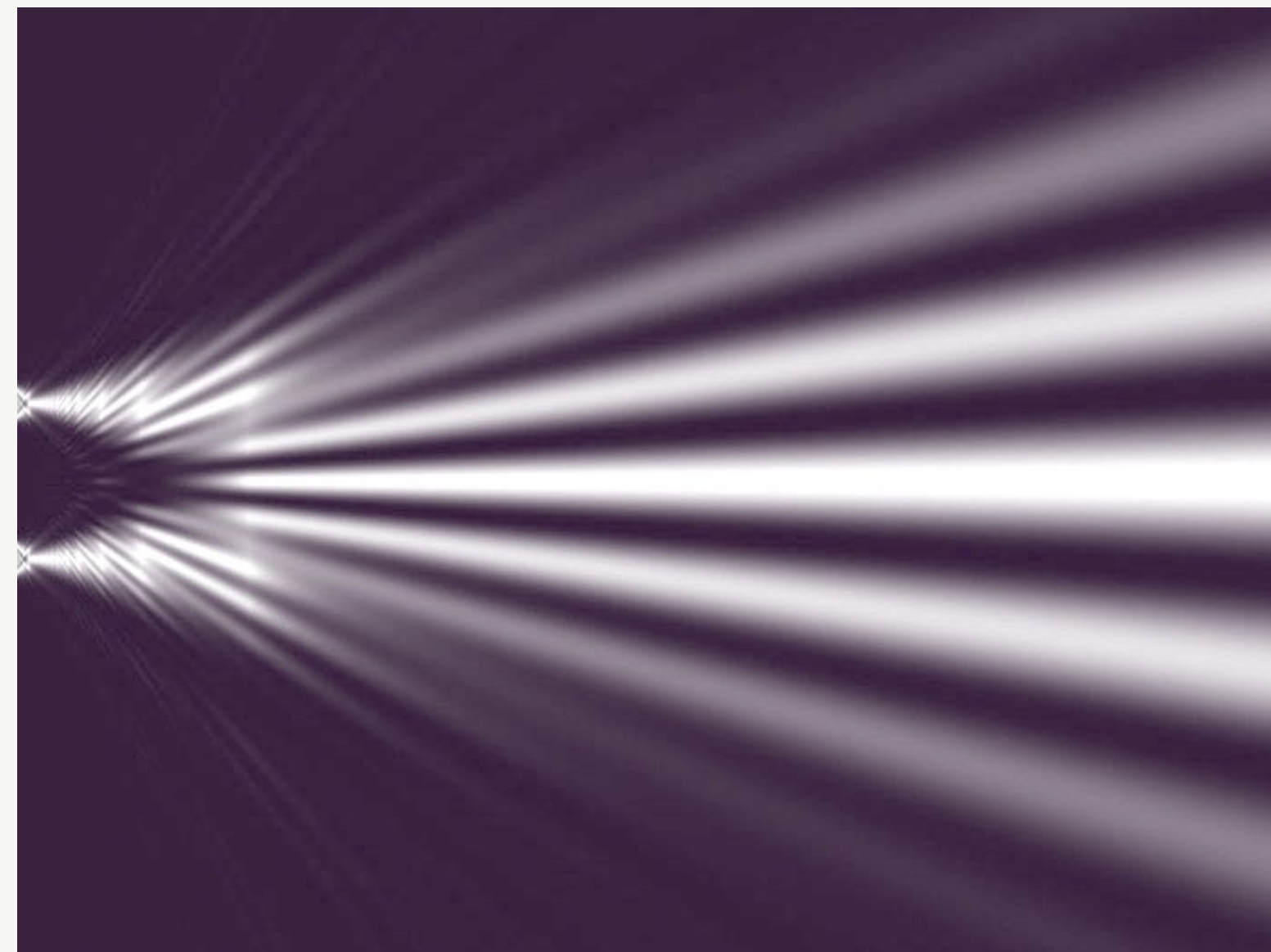
–Stephen Stich

“I claim that as a result of a misinterpretation of the approach of the basic natural sciences and a focus on design, experiment, and certainty over relevance, reality, and durability, much of the current field of modern social psychology has an unnecessarily narrow focus that, among other things, (a) pays little attention to powerful cultural influences (though this has been changing in the last decade), (b) discourages the discovery of new phenomena and creativity (Wegner, 1992), (c) discourages the description of basic regularities in the social world, and (d) presents a rather narrow model of what is acceptable science to graduate students in the area.”

–Paul Rozin (2001)

when experiments tell us things

- An over reliance on experimentation is a valid criticism of a lot of psychology (“physics envy”)
- Nonetheless, there are times when experiments really can tell us more than we’d otherwise glean
- E.g., the side-effect effect (Knobe, 2007)



Scientists are often preoccupied with *doing* research rather than focusing on how to do it well


Part 1: General Concepts + Definitions

language of experiments

Descriptive: Describe what is going on, what exists, what intuitions people hold, etc

- Example: Moral foundations theory 

Relational: Associations between two or more variables

- Example: On average, conservatives have greater amygdala activity 

Causal: Designed to determine whether one or more variables causes or affects one or more outcome variables.

- Example: State emotion manipulations affect moral condemnation 

language of experiments

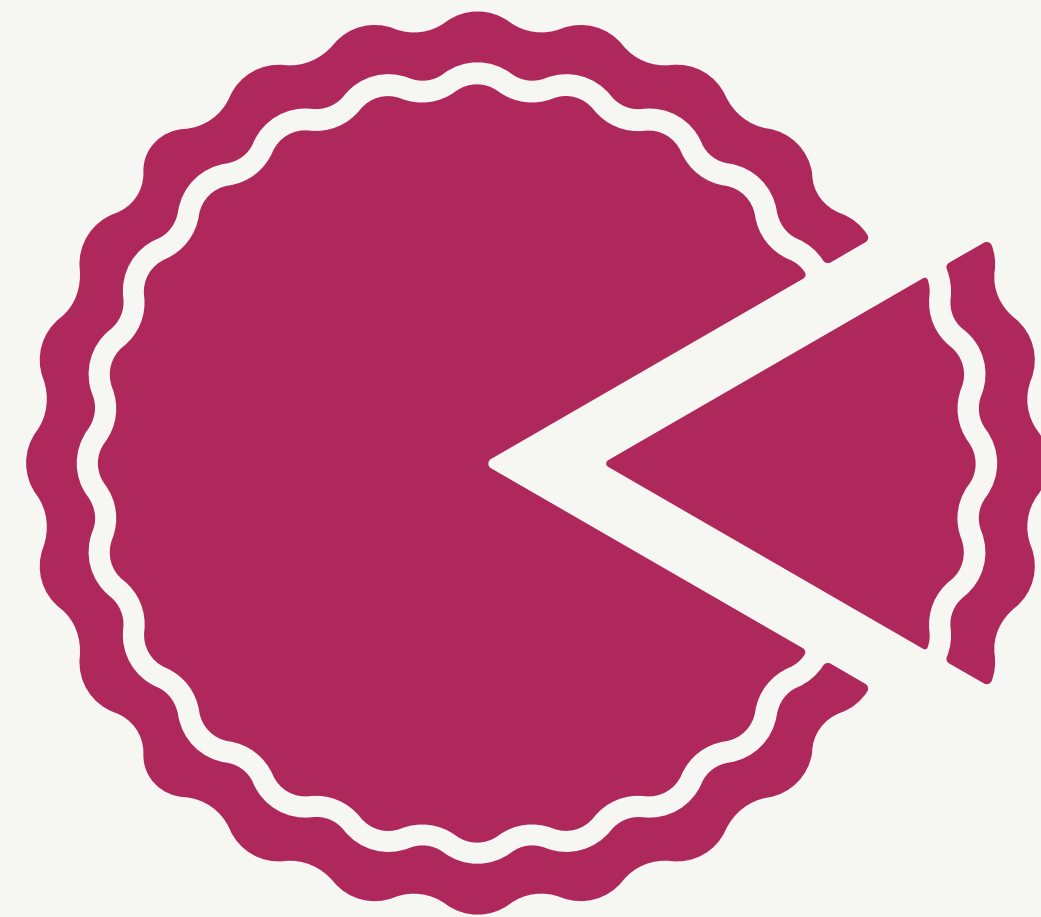
Cross-sectional: Study that takes place at a single point in time (taking a slice out of the cake of whatever we are studying)

Longitudinal: Study that takes place over time—at least two waves of measurement

- Repeated measures: two or more measurements in the same time point
- Time series: two or more measurements at different time points (or different waves)

language of experiments

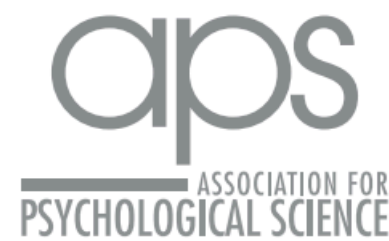
How would one investigate rates of cardiovascular disease in the population?



**Draw a sample of 1000 people,
administer some cardio measure,
and then analyze the rate**

language of experiments

Research Article



The Effect of a Supreme Court Decision Regarding Gay Marriage on Social Norms and Personal Attitudes



Margaret E. Tankard¹ and Elizabeth Levy Paluck^{2,3}

¹Behavioral and Policy Sciences Department, RAND Corporation, Santa Monica, California;

²Department of Psychology, Princeton University; and ³Woodrow Wilson School of Public and International Affairs, Princeton University

Abstract

We propose that institutions such as the U.S. Supreme Court can lead individuals to update their perceptions of social norms, in contrast to the mixed evidence on whether institutions shape individuals' personal opinions. We studied reactions to the June 2015 U.S. Supreme Court ruling in favor of same-sex marriage. In a controlled experimental setting, we found that a favorable ruling, when presented as likely, shifted perceived norms and personal attitudes toward increased support for gay marriage and gay people. Next, a five-wave longitudinal time-series study using a sample of 1,063 people found an increase in perceived social norms supporting gay marriage after the ruling but no change in personal attitudes. This pattern was replicated in a separate between-subjects data set. These findings provide the first experimental evidence that an institutional decision can change perceptions of social norms, which have been shown to guide behavior, even when individual opinions are unchanged.

Psychological Science
1–11
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797617709594
www.psychologicalscience.org/PS
SAGE

Take multiple measures from the same people. Here is the most commonly thought of type of longitudinal experiment



language of experiments

Population: the full unit of interest

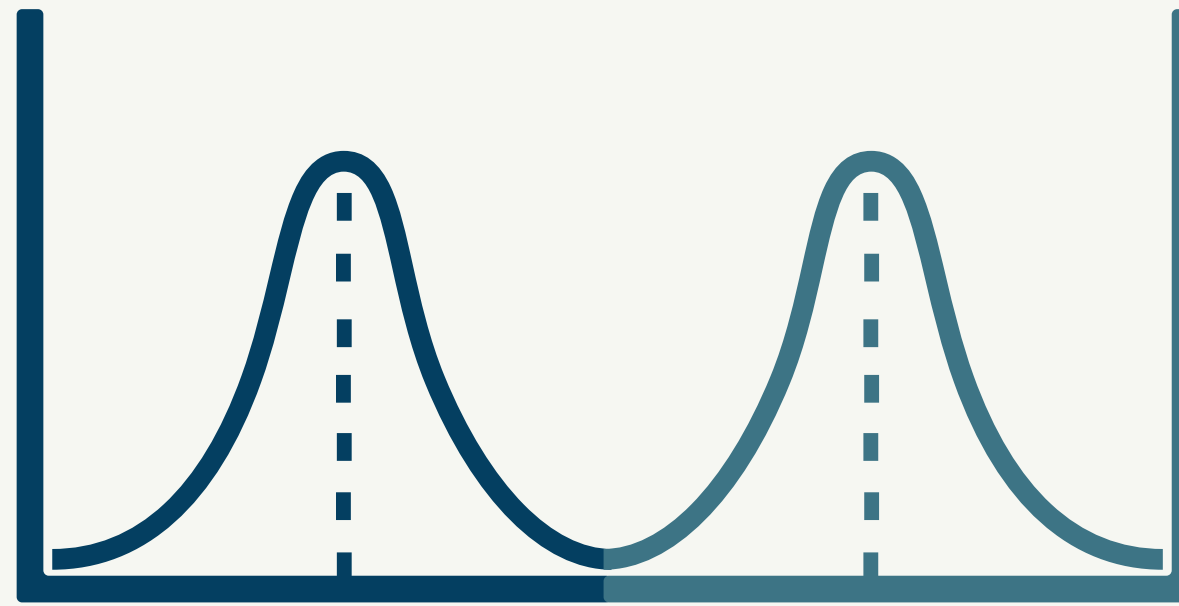
Sample: selected units from a population of interest (how generalizable to the population depends on sampling)

Example: I'm interested in the IQ for major league baseball players, so I sample Mets and Yankees players.

Example: I'm interested in the IQ for major league baseball players based in the NYC/NJ region, so I sample Mets and Yankees players.

language of experiments

Effect Size: A quantitative measure of the magnitude of an experimental effect



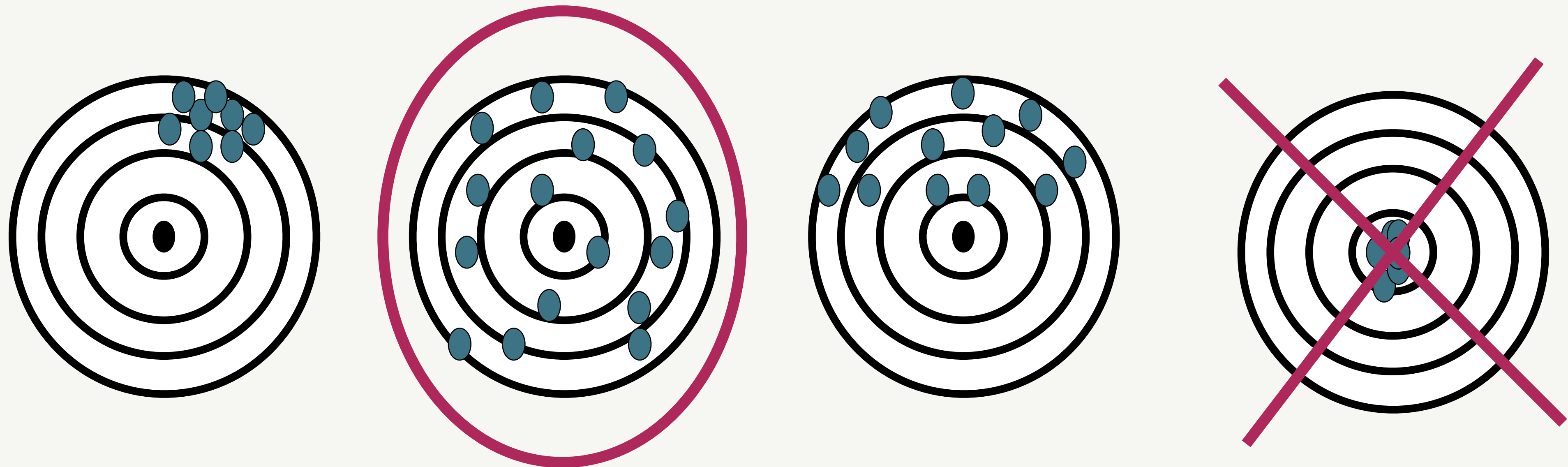
It is important to assess statistical versus practical significance (SESOI)

Even very tiny effects can translate into fairly important treatment effects when you consider the real life odds of experiencing a given outcome

language of experiments

Validity: The extent to which tools/instruments measure what they're intended to measure

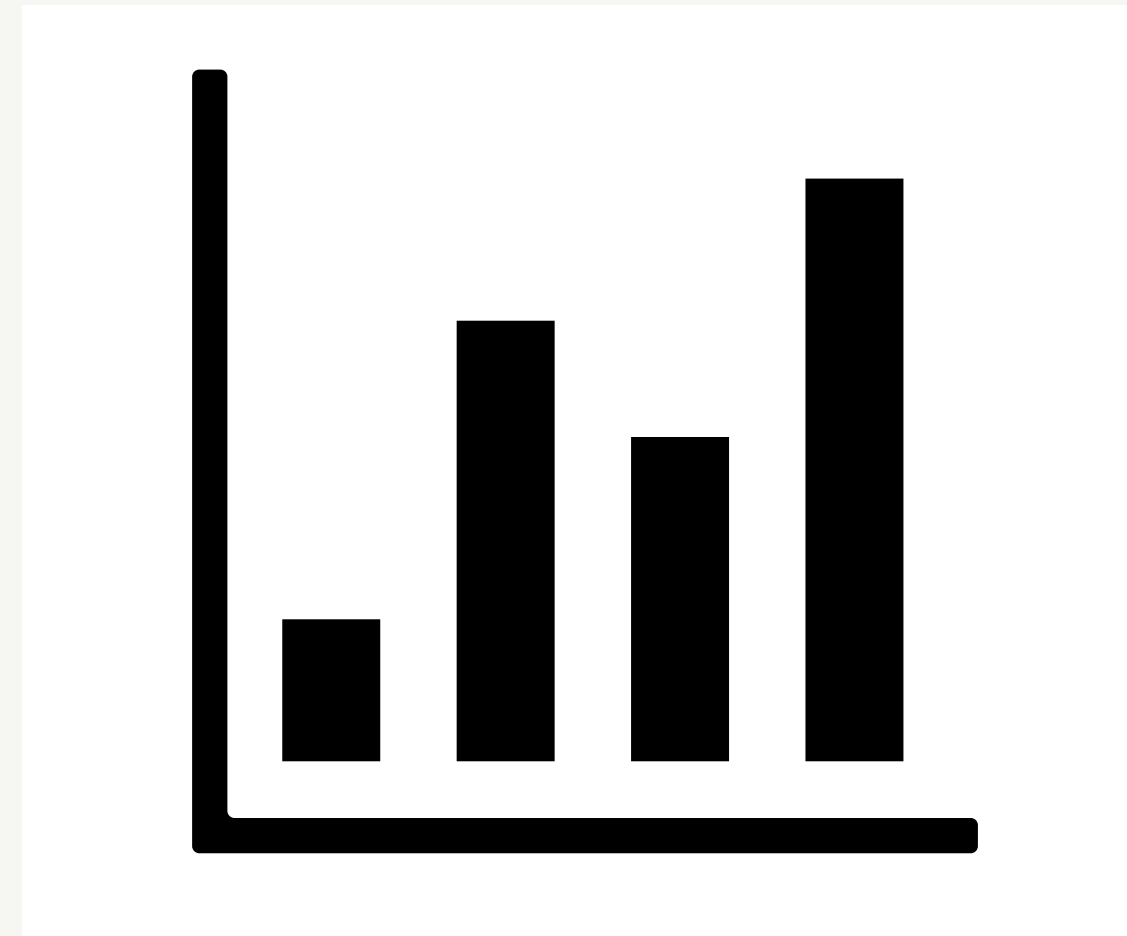
Reliability: The extent to which measurements/outcomes are consistent over time



Experimental reliability



Time 1



Time 2

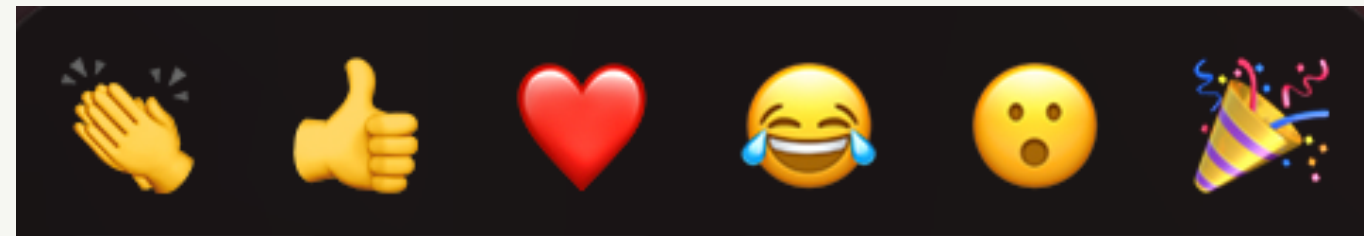


Time 3

Replicable: Effect stands in another experimental settings

Zoom Poll!

Using the emoji reactions on zoom, please answer the following question:



What is the difference between validity and reliability?

- A. Reliability is about precision of measurement, validity is about accuracy ❤️
- B. Reliability is about accuracy of measurement, validity is about precision 🙌

language of experiments

Error: Fluctuations in results that effects validity and/or reliability

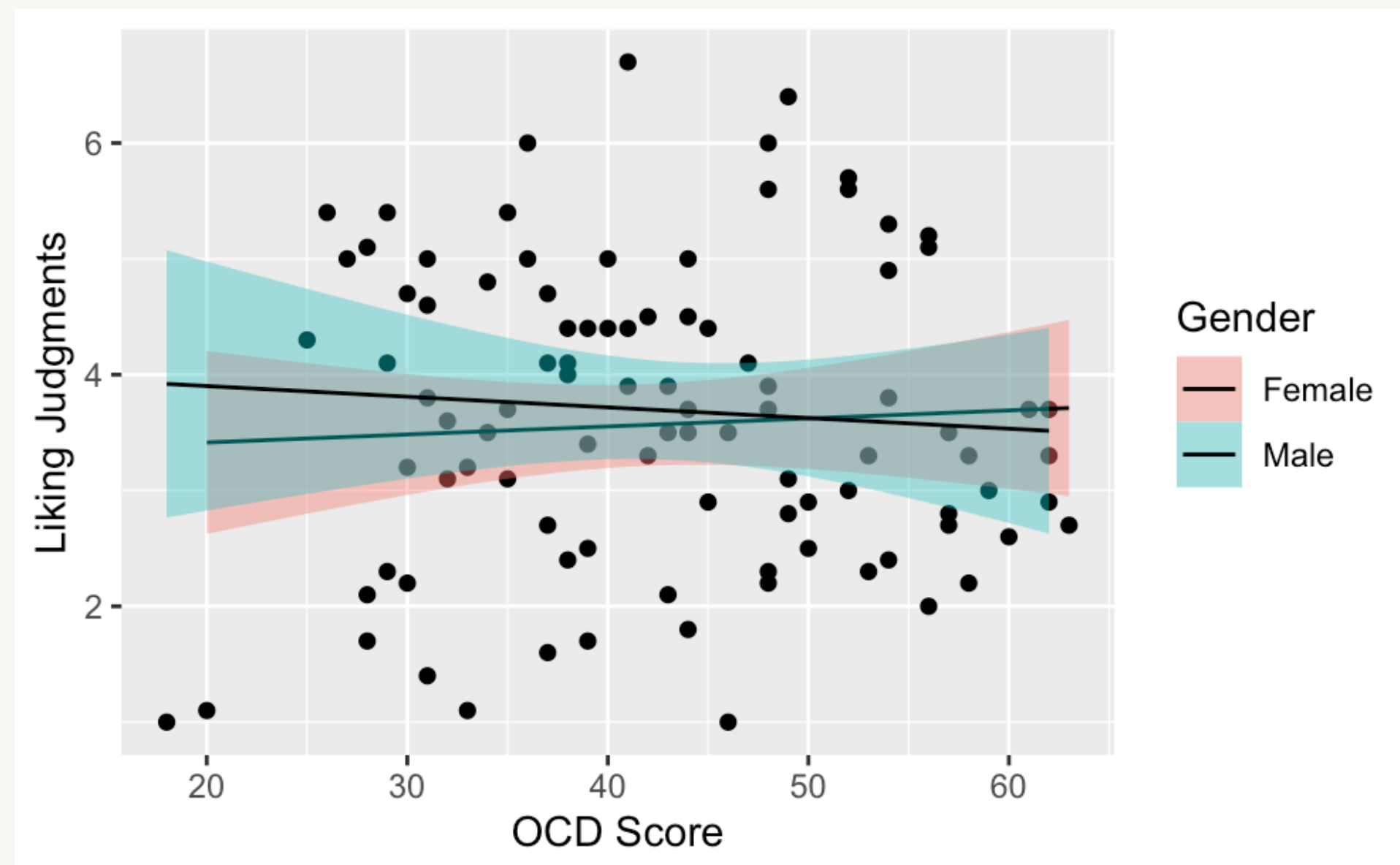
[Observed value = actual value - chance/random error]

Chance/random error (also called noise): Uncorrelated with actual value, random fluctuations around actual value, over time will cancel each other out and thus represent or be very close to actual value.

Systematic error (also called bias): Will also push value in a certain direction - resulting in a mean that's too big or too small. This does not cancel out due to the errors all being in the same, systematic direction. This is a main concern for internal validity!

language of experiments

Chance/random error



Systematic error

- 60Hz signal
- One experimenter behaves different toward participants
- Some participants have poorly fitted caps

validity

construct validity

Is the extent to which a test measures the concept or construct that it is intended to measure.

Example: To what extent is an IQ questionnaire actually measuring "intelligence"?

example

19th century 1879 French Neurologist Paul Broca reported “Evidence for the intellectual inferiority of women”. He found that the brain weight of men > brain weight of women

Is brain weight a good measure for intelligence?

What was the source of the difference?

external validity

Relates to whether the findings of a study be generalized to other people and environments.

Your inference space

ecological validity

A subset of external validity—it relates to the breadth of the population sampled and how well the experimenter can justify extending the results to broader population.

Convenience samples are not very representative!

example

What about hypothetical moral decisions?



Cognition
Volume 123, Issue 3, June 2012, Pages 434-441



What we say and what we do: The relationship between real and hypothetical moral choices

Oriel FeldmanHall ^{a, b}  , Dean Mobbs ^b, Davy Evans ^{a, b}, Lucy Hiscox ^b, Lauren Navrady ^b, Tim Dalgleish ^b

[Show more](#) 

internal validity

Is your independent variable (and not a lurking or confounding variable) causing the effect you are measuring?

Requirements:

1. the "cause" precedes the "effect" in time
2. the "cause" and the "effect" are related
3. there are no plausible alternative explanations for the observed observation

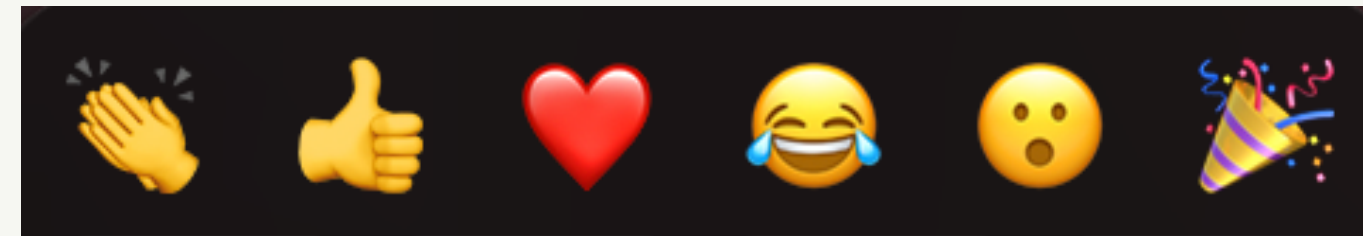
internal validity

3. there are no plausible alternative explanations for the observed observation

- Subject-experimenter artifacts
- Demand characteristics
- Experimenter expectancy effects
- Ceiling & floor effects
- Bad questionnaire items

Zoom Poll!

Using the emoji reactions on zoom, please answer the following question:



What is the difference between internal and external validity?

- A. Internal validity is about generalizability and external is about the measurements 👏
- B. Internal validity is about the measurements and external is about the generalizability 👍

Internal Validity

Did we isolate/rule out
an effect in our
experiment?

External Validity

Would the same thing happen in
other settings?

Other
labs

Everyday
settings

Construct Validity

Does the measured
variable represent
the construct of
interest?

**Valid
Result**

problems that pop up!

1. Challenges determining causality (!!!!)
2. Lack of a control
3. Confounds
4. No blinding
5. Lack of randomization
6. Type two errors
7. Type one errors



1: causality

General thought: If you manipulate environment, and the only difference is control versus treatment => assign causality

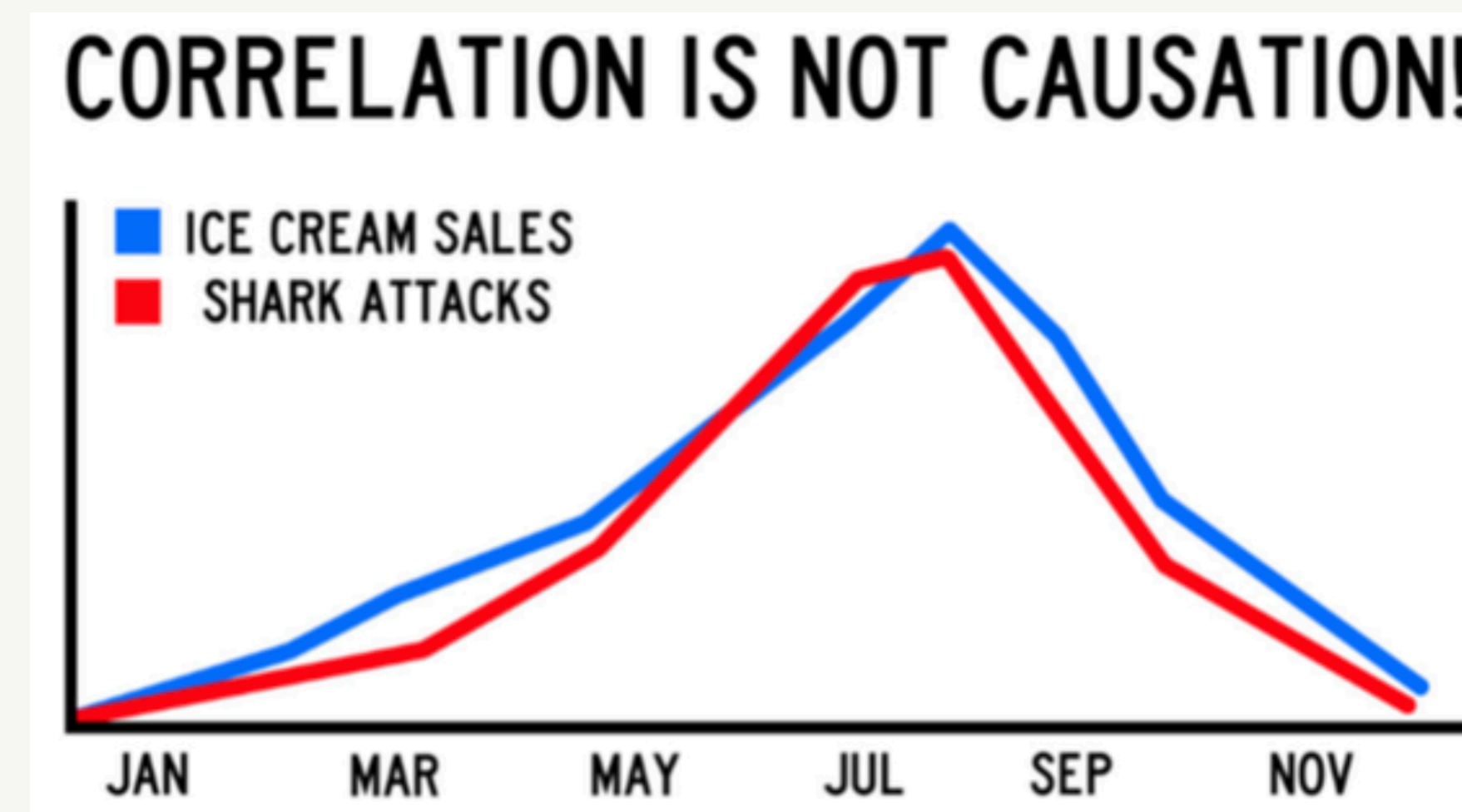
- However, there are often other factors that could interfere
 - These are called confounding or lurking factors

To understand confounders, need:

- knowledge of the system being studied
- technical knowledge of methods by which the data is collected

1: causality versus correlation

Correlation does not equal causation



1: causal reasoning

Cross Sectional Mediation Bot
@MediationBot

All models are wrong, especially cross sectional mediation models.

⋮ Follow

2: no control group

- Negative controls
 - Check for unrelated effects
- Placebo
- Positive controls
 - Check the procedure is observing the effect

3: confounds

- A variable which is related to one or more of the variables defined in a study.
- May mask an actual association or falsely demonstrate an apparent association.
- *Example: Slightly overweight people live longer than thin people*

3: confounds

Comparing treated versus control but they're measured :

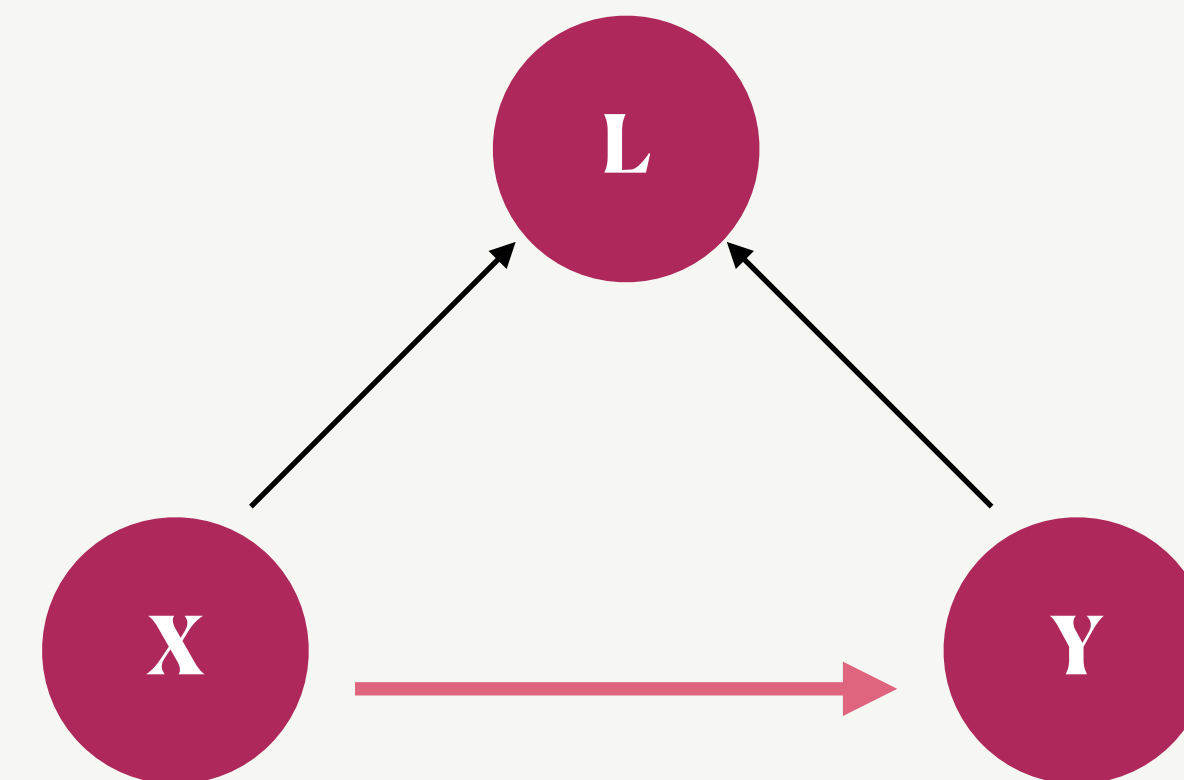
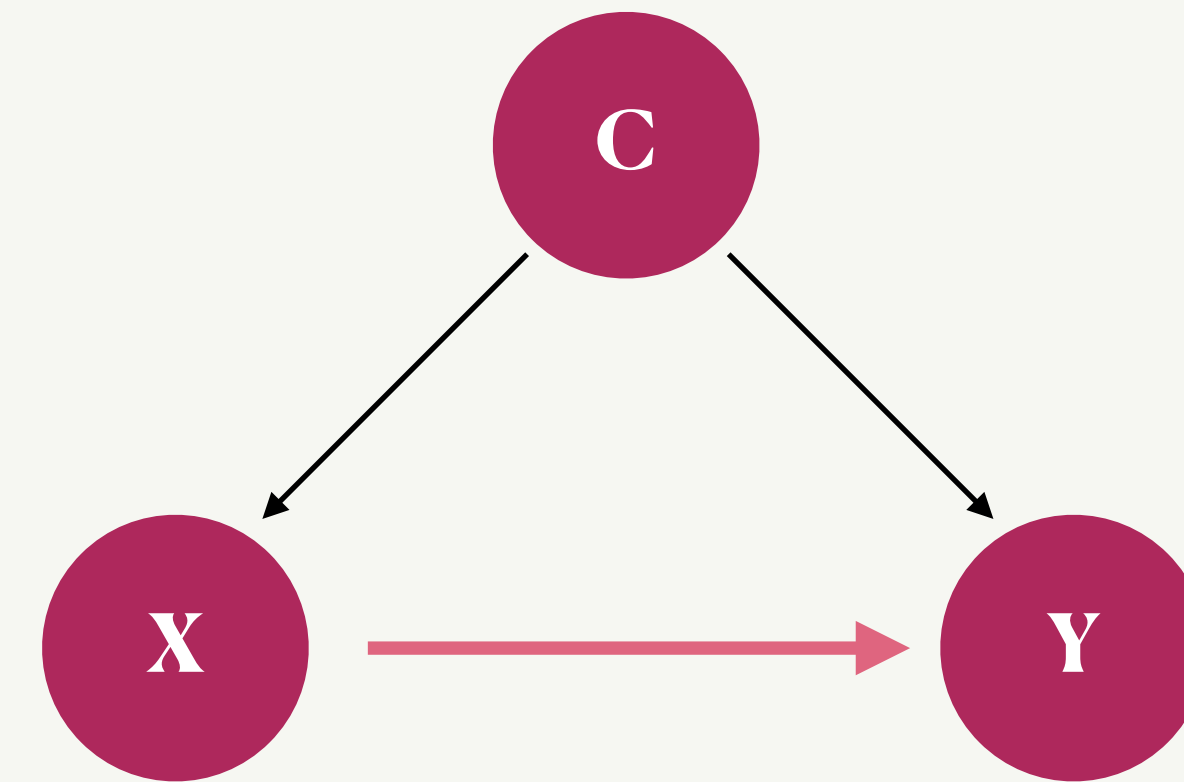
- In different rooms
- On different days
- By different operator
- On different shelves/machine/plate
- One group is measured first

= bias

3: confounds vs. lurking variables

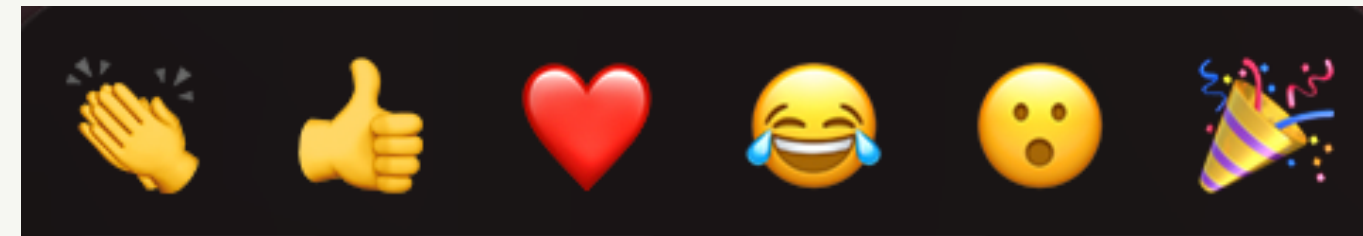
Confound vs. Lurking

- Confounds change an existing relationship
 - E.g., video games and aggression
- Lurking variables connect otherwise unconnected variables
 - E.g., Number of firefighters at a scene vs damage done by a fire



Zoom Poll!

Using the emoji reactions on zoom, please answer the following question:



What is a lurking variable?

A. A variable creates spurious relationships 🎉

3: managing confounds

- Known sources:
 - Fix the factor of interest
 - Source of variance of interest
 - E.g. sex, concentration, dosage time, age
 - Block the source of noise you want to account for to increase generalisability but maintain sensitivity
 - E.g. batch, operator, plate, time of day
- Unknown sources:
 - **Randomization**
 - Usually: plate location or processing order

4: no blinding

Blind to hypotheses/Condition: Person interacting with participants should not be privy to experimental expectations

- Not as important in online settings
- Important anytime you are interacting with another person

5: lack of randomization

- Spread any unknown, inescapable variation amongst all subjects with equal probability
- Avoids systematic bias
- When?
 - Assigning to treatment
 - Measurement/sample processing

quick review

- Treatment = planned, systematic variability 😊
- Noise = chance-like variability 😐
- Confounding = unplanned, systematic variability 😞

- **Controlling:** removes variation from potential confounders
- **Randomizing:** converts into chance-like variability.
- **Factor/Block:** converts into planned, systematic variability.

6: type II errors

False negatives!

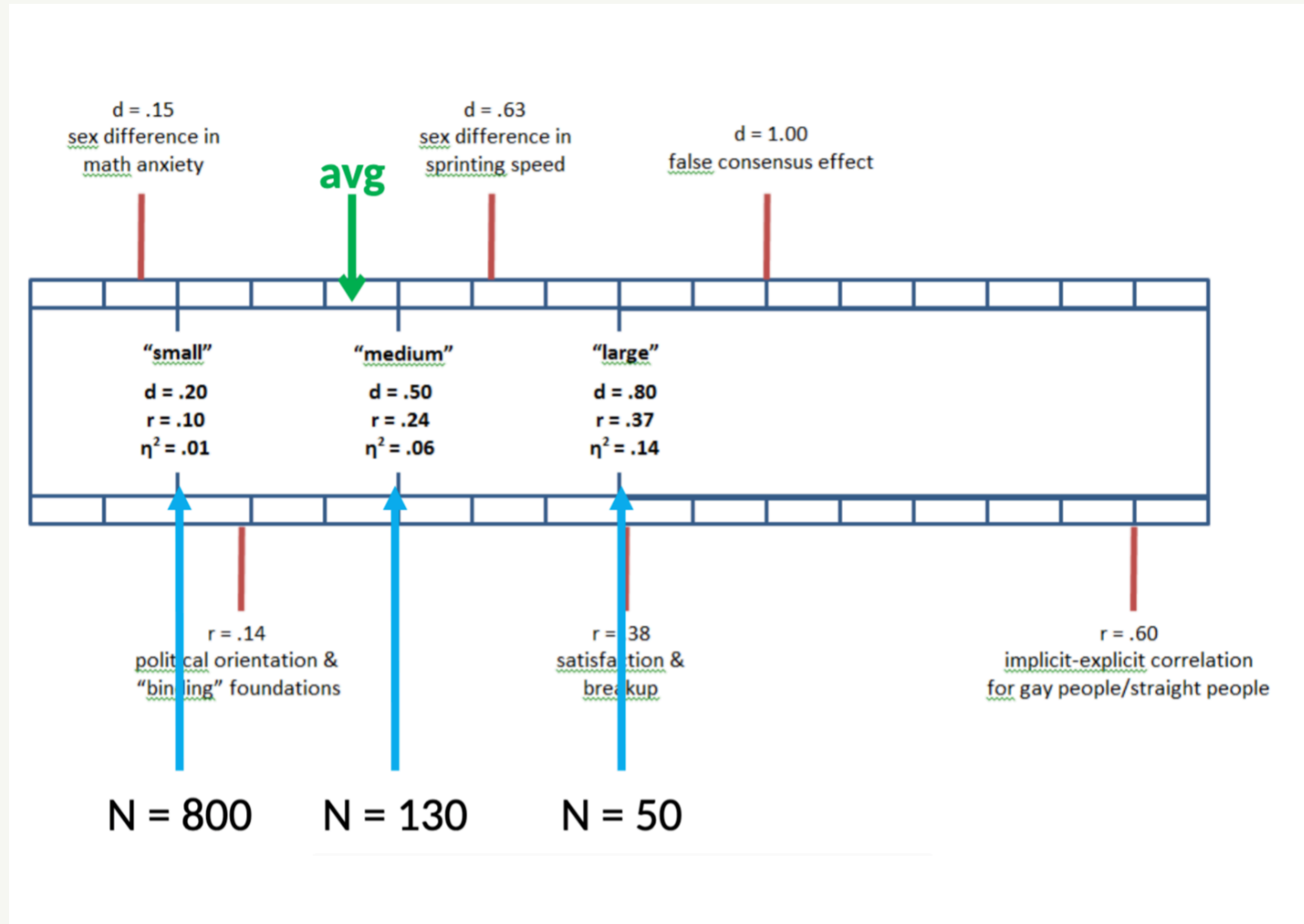
$$1 - P(R \mid H_a \text{ is true})$$

- The ability of the test to detect an effect when it exists [$1 - \beta$ (Type II error)]
- Increases with the size of the sample, the size of the effect, and the significance criterion (p-value)



6: type II errors

- The power of a statistical test is the probability that it will correctly lead to the rejection of a false null hypothesis (*1 - type 2 error rate or β*)
- Varies between 0 and 1
- Target power: $\beta = .80$ or 80%



7: type I errors

- False positive!

$$P(R \mid H_0 \text{ is true})$$

- The error of rejecting a null hypothesis when it is actually true.



7: type I errors

- We typically want a type I error rate of 5% ($\alpha = .05$).
- However, sometimes the real rate with 0.05 threshold is 40%
- Why?
 - Assumptions of the statistical test are not met
 - Normality
 - Independent readings
 - Multiple testing (!!!!!!!!!!!!!)

7: type I errors

- Say you have a set of hypotheses that you wish to test simultaneously. The first idea that might come to mind is to test each hypothesis separately, using some level of significance α . At first blush, this doesn't seem like a bad idea. However, consider a case where you have 20 hypotheses to test, and a significance level of 0.05.
- What's the probability of observing at least one significant result just due to chance?

7: type I errors

- What's the probability of observing at least one significant result just due to chance?

$$\mathbf{P(\text{at least one significant result}) = 1 - P(\text{no significant results}) = 1 - (1 - 0.05)^{20} \approx 0.64}$$

6: type I & II errors

	$H_0 = \text{True}$	$H_0 = \text{False}$
Reject H_0	Type I error (false pos)	Correct! (true pos)
Fail to reject H_0	Correct! (true neg)	Type II error (false neg)

questions?

workshop contents

Part 1:

Introduction

General concepts & definitions

Part 3:

Survey Research

Experiment creation + rules /
steps to follow

Part 2:

Experimental Design

Part 4:

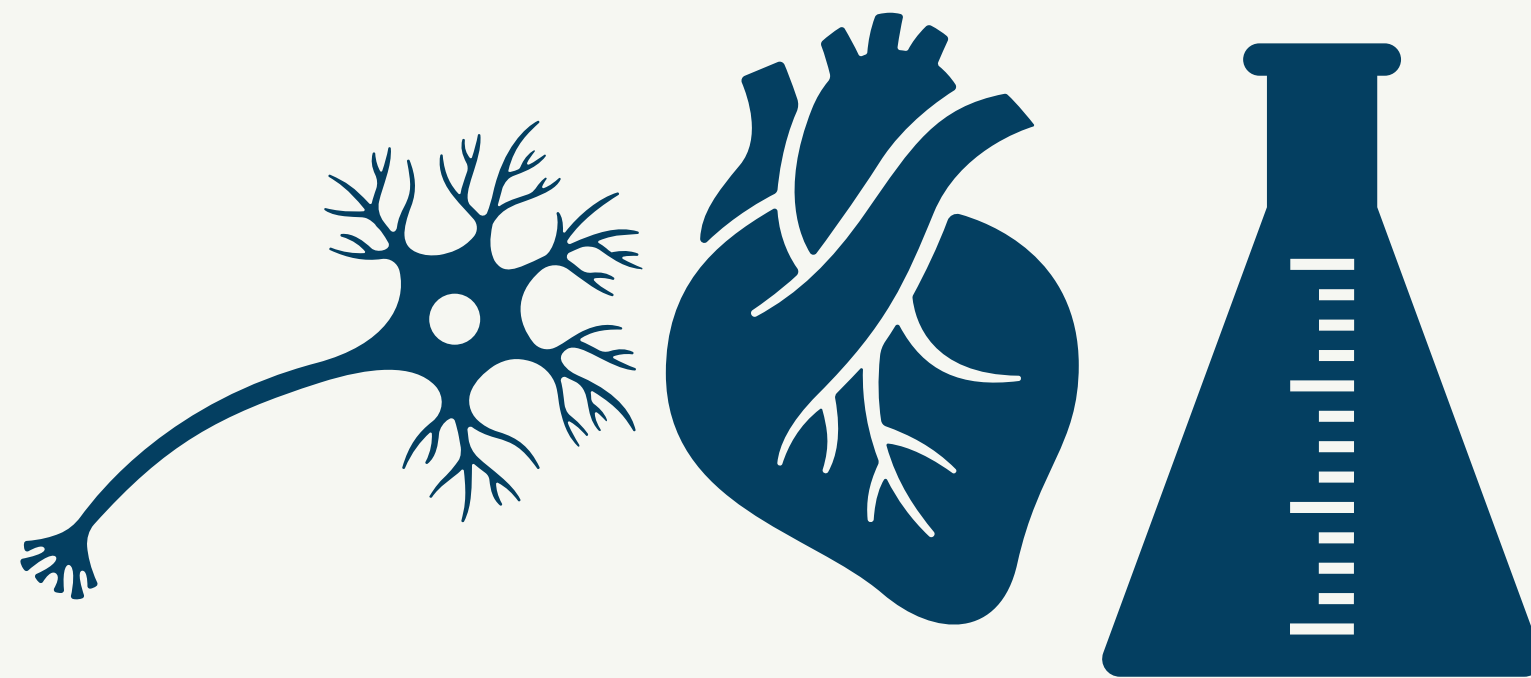
Open science + collaborations

Ethics and the IRB

Part 2: Design

variables

Two primary variable types: dependent and independent



variables

What is a dependent variable?

- Variable whose changes are viewed as dependent on changes in one or more (independent) variables
- Variable that is measured, or being affected (the result)
- Often referred to as y , criterion, or outcome variable

variables

What is an independent variable?

- In experiments, the variable that the experimenter manipulates or varies to determine whether there are effects on another variable (the dependent variable)
- Variable that you expect causes the result
- Often referred to as x , predictor, or experimental variable

variables

Conceptual variables: are about abstract constructs (e.g., depression)

Operational variables: are the concrete operations, measures, or procedures used to measure the concept in practice (e.g., BDI)

Operationalization: researcher defined measurement of a construct

- Cognition
- Attention
- Memory
- Aggression

variables

What is a covariate?

- A variable that covaries with the dependent variable, so we want to control for it (or else it may be a confound!)
- Often demographics variables like age, gender, political ideology

variables

Nominal: No inherent value or mathematical value

- All dichotomous variables are like this (male or female, race, political side)
- Often called categorical variables
- You typically analyze these with non parametric techniques
- If you are measuring them as outcome then looking at things like chi squares or logistic regression (true and not arbitrary binary), if its survey then you often dummy code and will run a correlation or descriptive

variables

Interval: The data can be quantified with a constant distance

- E.g., things like intelligence tests, anxiety scales, personality scales, temperature, etc... all are interval.
- But in any interval there is no natural zero (you wouldn't talk about the absence of intelligence or temperature) and thus the ratio between are not meaningful
- Agreement and satisfaction are often used as interval in social psych

variables

Ordinal: Responses presented are rank ordered

- The distance between the response options is not consistent
- E.g., army rank
- E.g., educational attainment (less than high school, some college, post grad, etc..)
- Can use non parametrical stats too (spearman's rho; Mann Whitney t-test)

variables

Ratio: Interval but have natural zero

- Typically more in the other sciences than Xphi & psych
- E.g., weight, volume, amount of time past; ratios between are meaningful
120lbs is 2 times 60

hypotheses

What is a hypothesis?

- Testable prediction of what will happen given a certain set of conditions
- Tentative guess about a behavior that usually is related to some other behavior or influence
- Two parts: null & alternative

hypotheses

Do baseball fans in NYC spend more money at the concession stand than baseball fans in Boston?

null (H_0): NYC \$\$ = Boston \$\$

alt (H_a): NYC \$\$ > Boston \$\$ (or < or !=)

hypotheses

A good hypothesis is:

- Replicable, falsifiable, parsimonious, precise
- A specific prediction about the relationship between two or more variables (experimental hypotheses usually include control or comparison groups)
- Theoretically based (based on previous knowledge)
- Testable (has obtainable answer)
- Novel yet consistent with previous research*

hypotheses

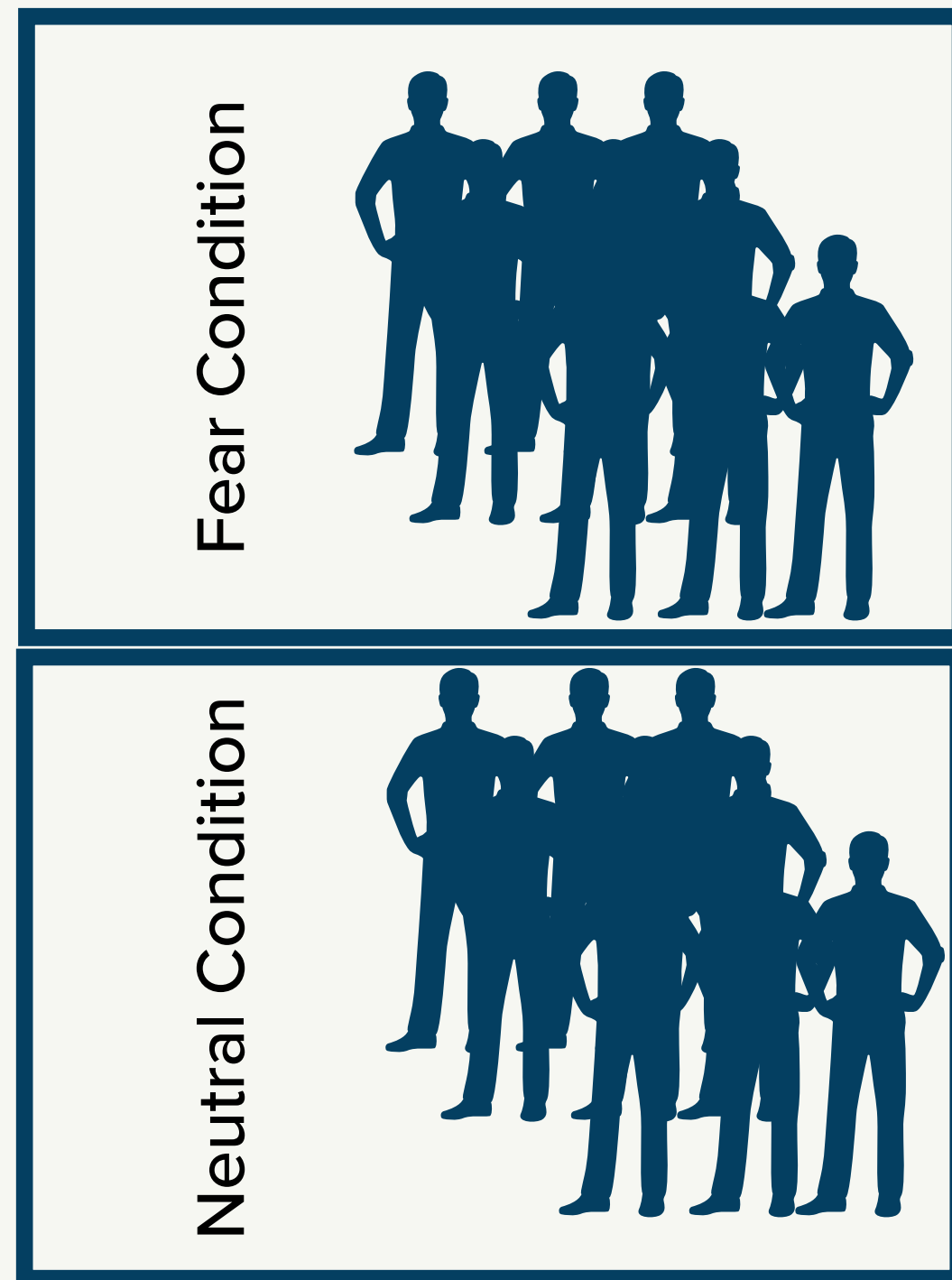
- Typically with an experimental hypothesis, there is a comparison between 2+ levels/groups of an independent variable and a specified direction of the pattern of results for 1+ dependent variable
- Experimental studies have at least one experimental condition and one control condition

conditions

- Experimental condition- group that experiences the independent variable or manipulation
- Control condition- group that does not experience the independent variable or not anticipated to experience effect of independent variable; comparison group

condition & factor types

Between vs. Within Subjects



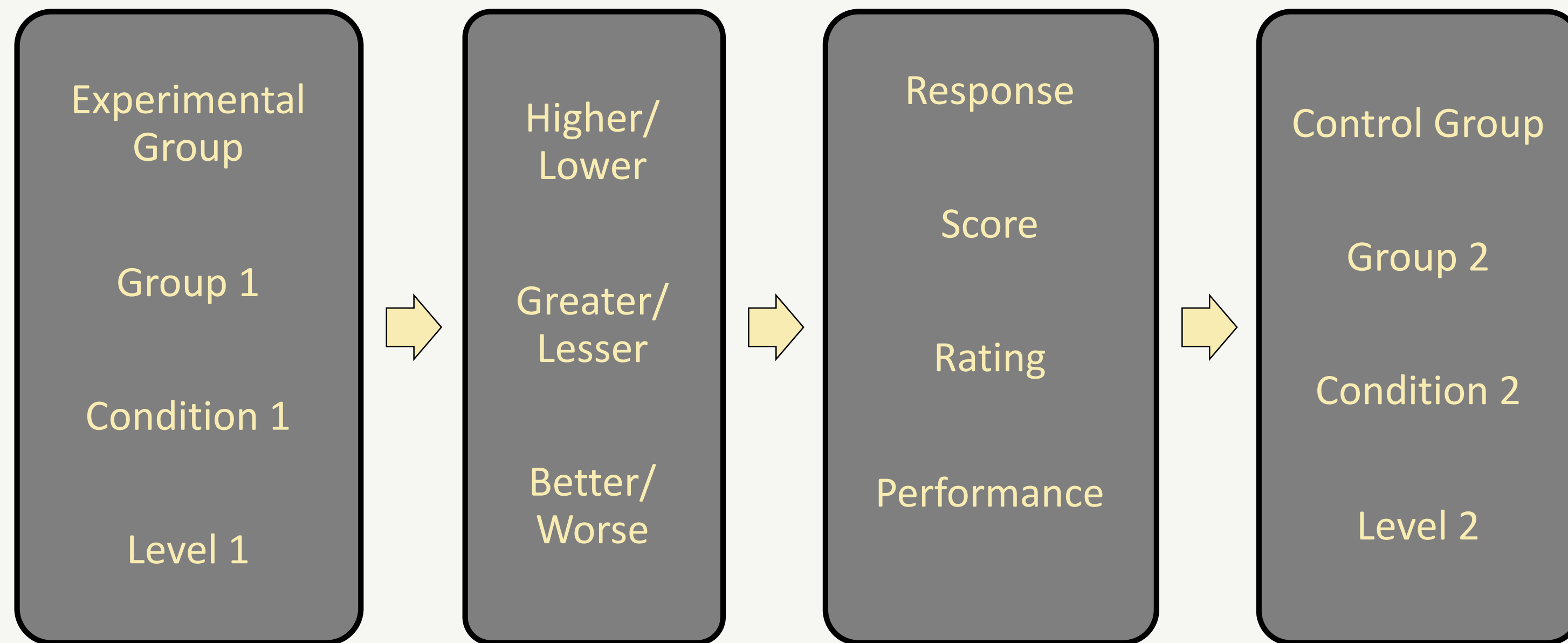
Trustworthiness

Dominance

Neutral



conditions + predictions



effect types

Main Effect: effect of IV on DV ignoring other IVs

Interaction Effect: effect of IV on DV depending on the level of another IV

Simple Effect: the effect of IV on DV for one level of another IV

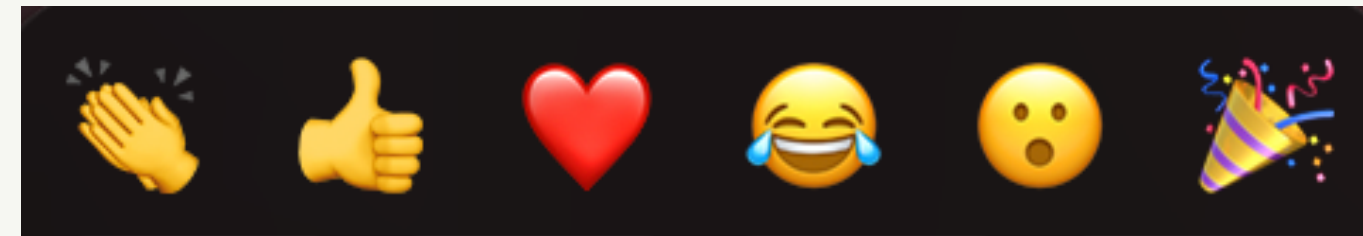
Moderation: interaction between IV and DV where the effect of IV on DV is moderated by a third variable

Mediation: explains the relationship between IV and DV by showing how they are related**

**Political Orientation significantly
moderates the relationship between
sleep and violence. The interaction
between sleep and political
orientation was significant in the
no training (vs. training) condition.**

Zoom Poll!

Using the emoji reactions on zoom, please answer the following question:



What is an interaction?

- A. A variable that mediates relationships 🎉
- B. A variable that moderates relationships 😮

effect types

	IV2	
IV1	5	5
	10	10

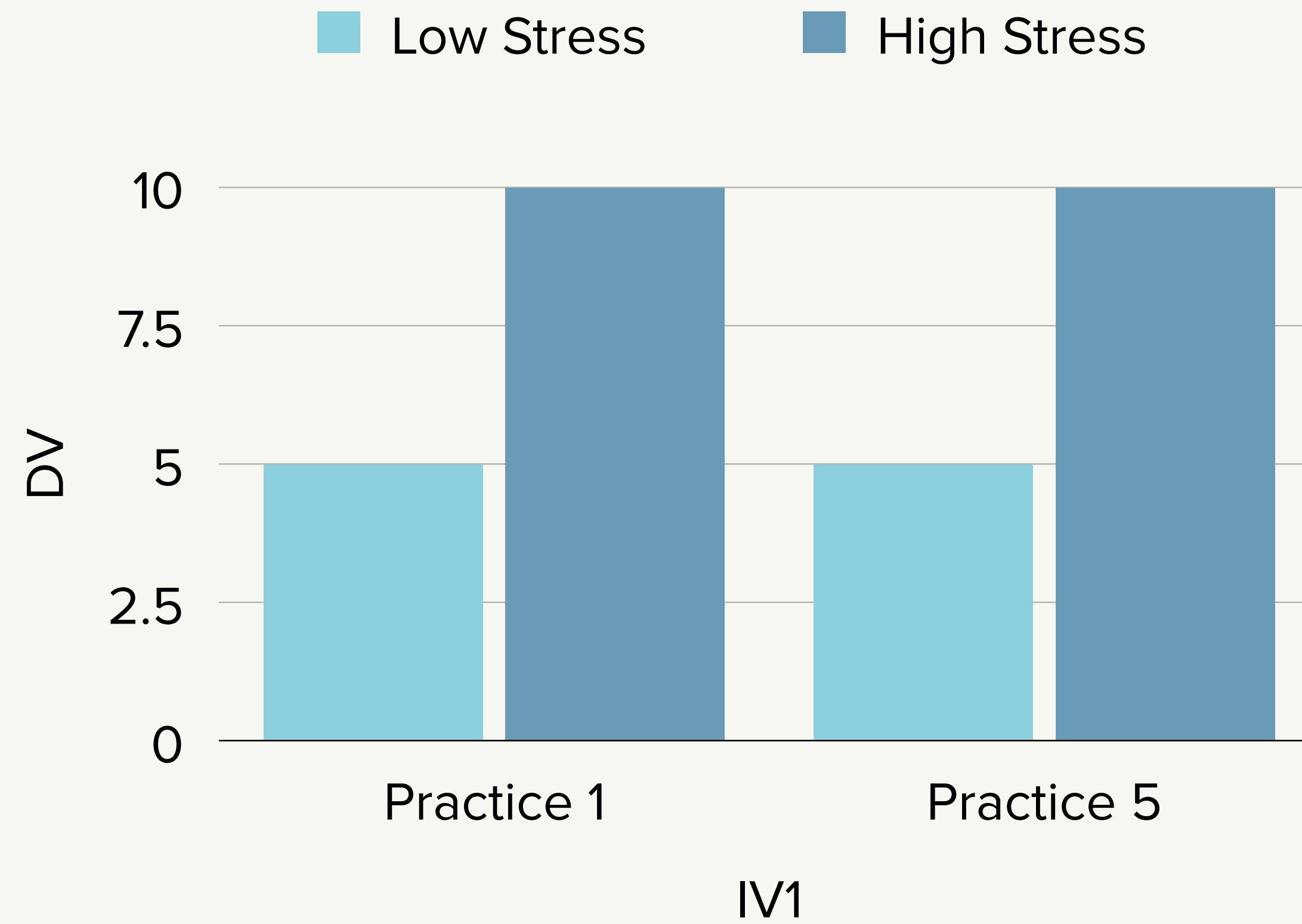
	IV2	
IV1	10	5
	5	5

	IV2	
IV1	10	10
	5	5

	IV2	
IV1	10	5
	5	10

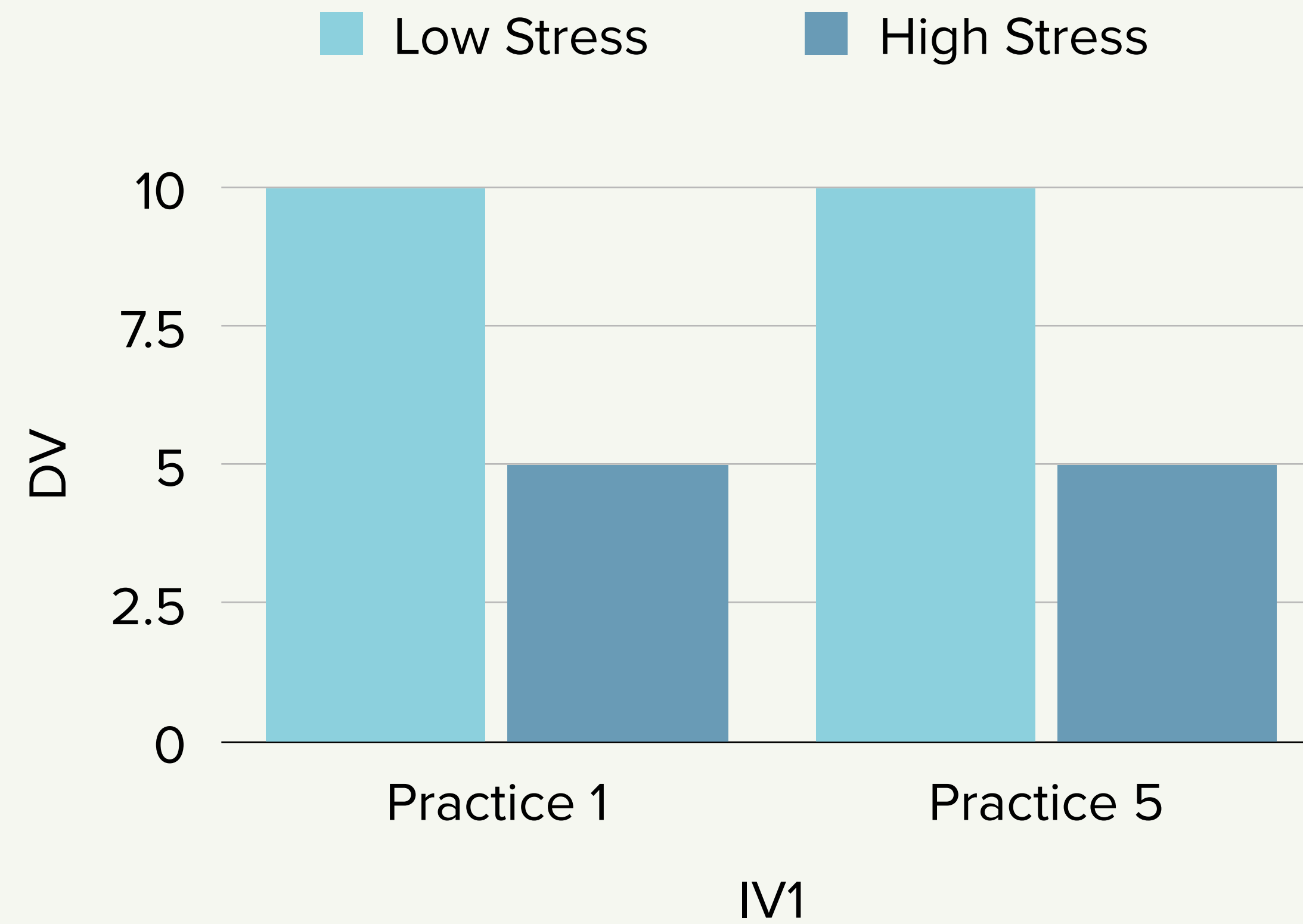
effect types

	IV2	
IV1	5	5
	10	10



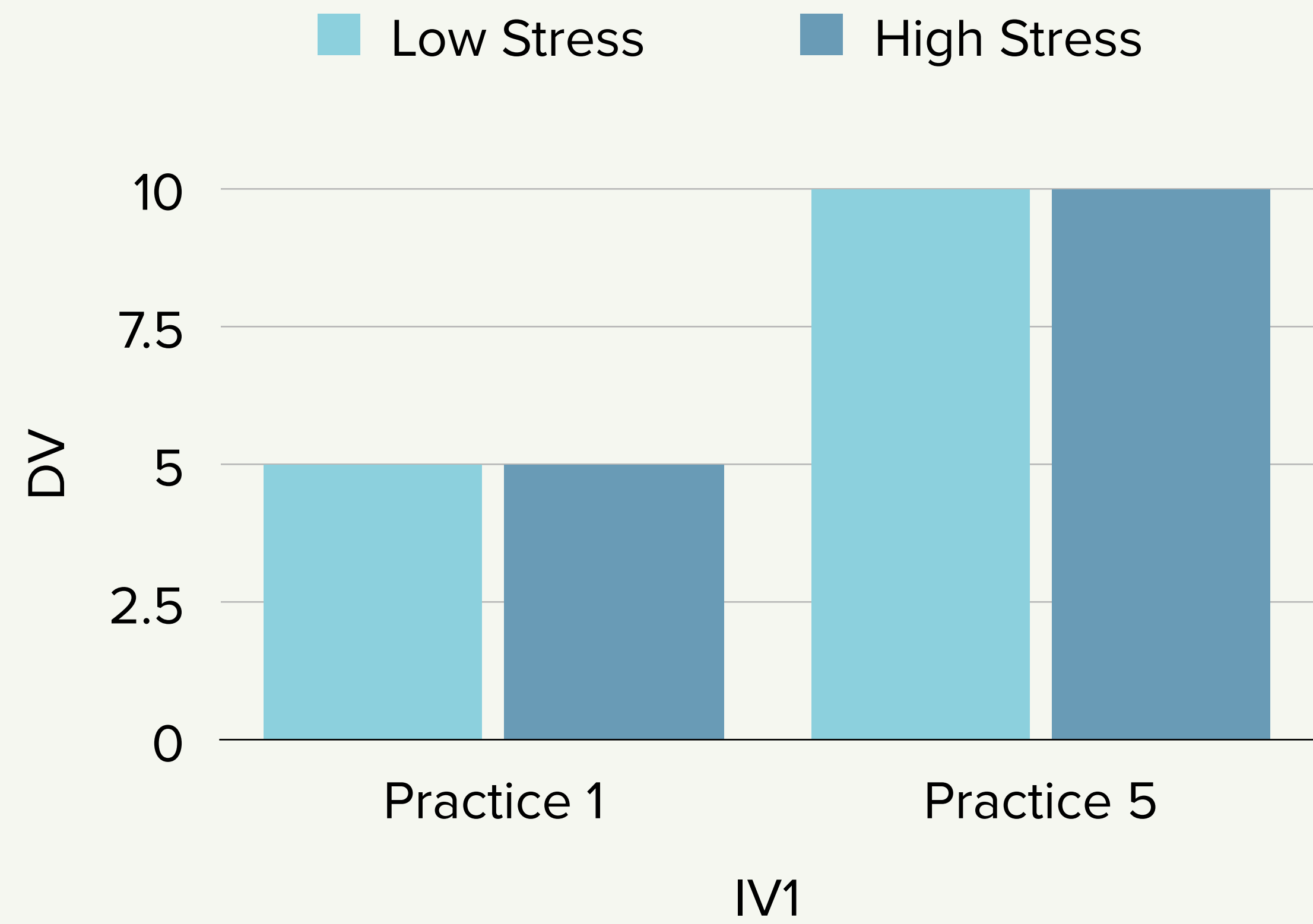
effect types

	IV2	
IV1	10	10
	5	5



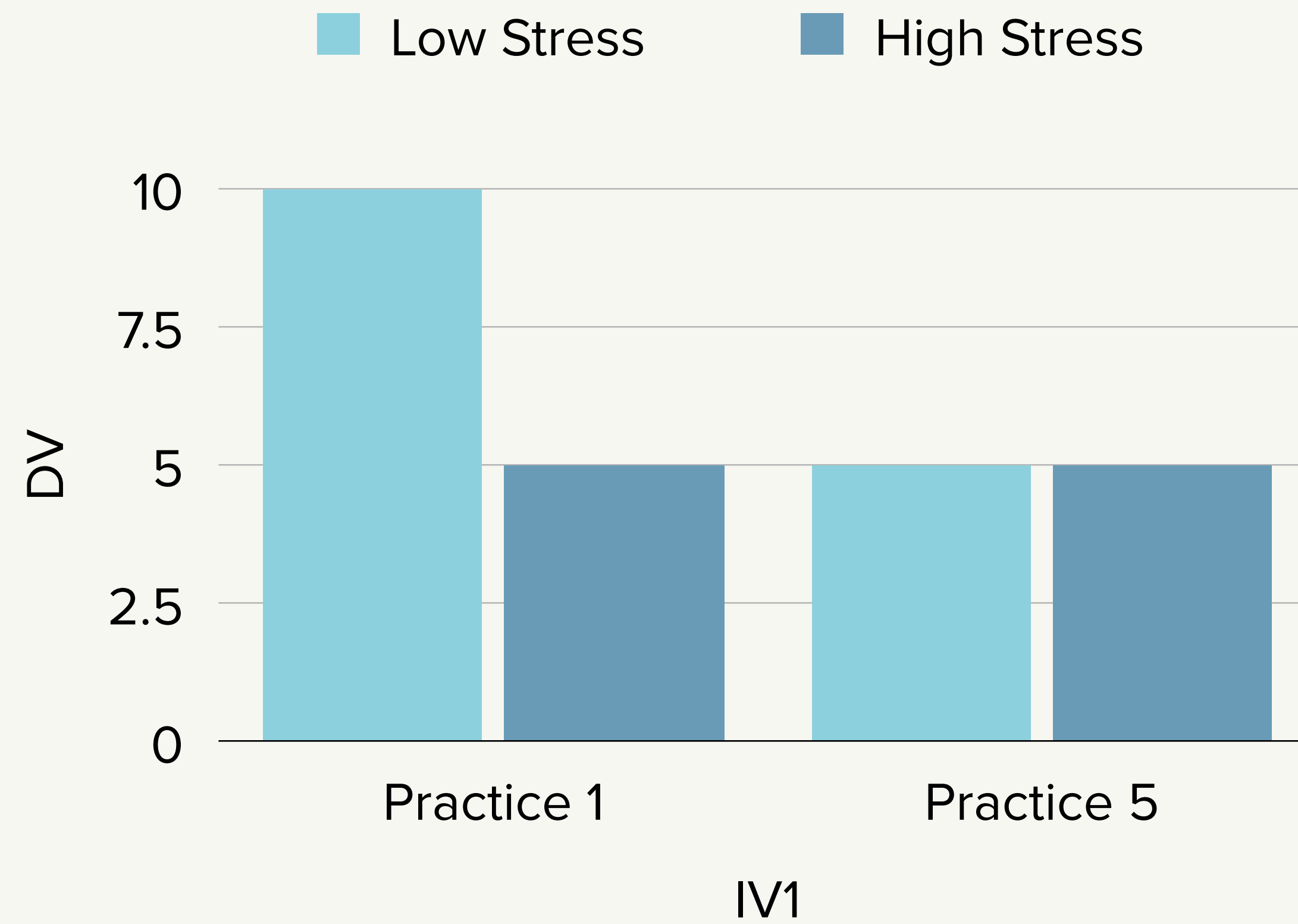
effect types

	IV2	
IV1	5	10
	5	10



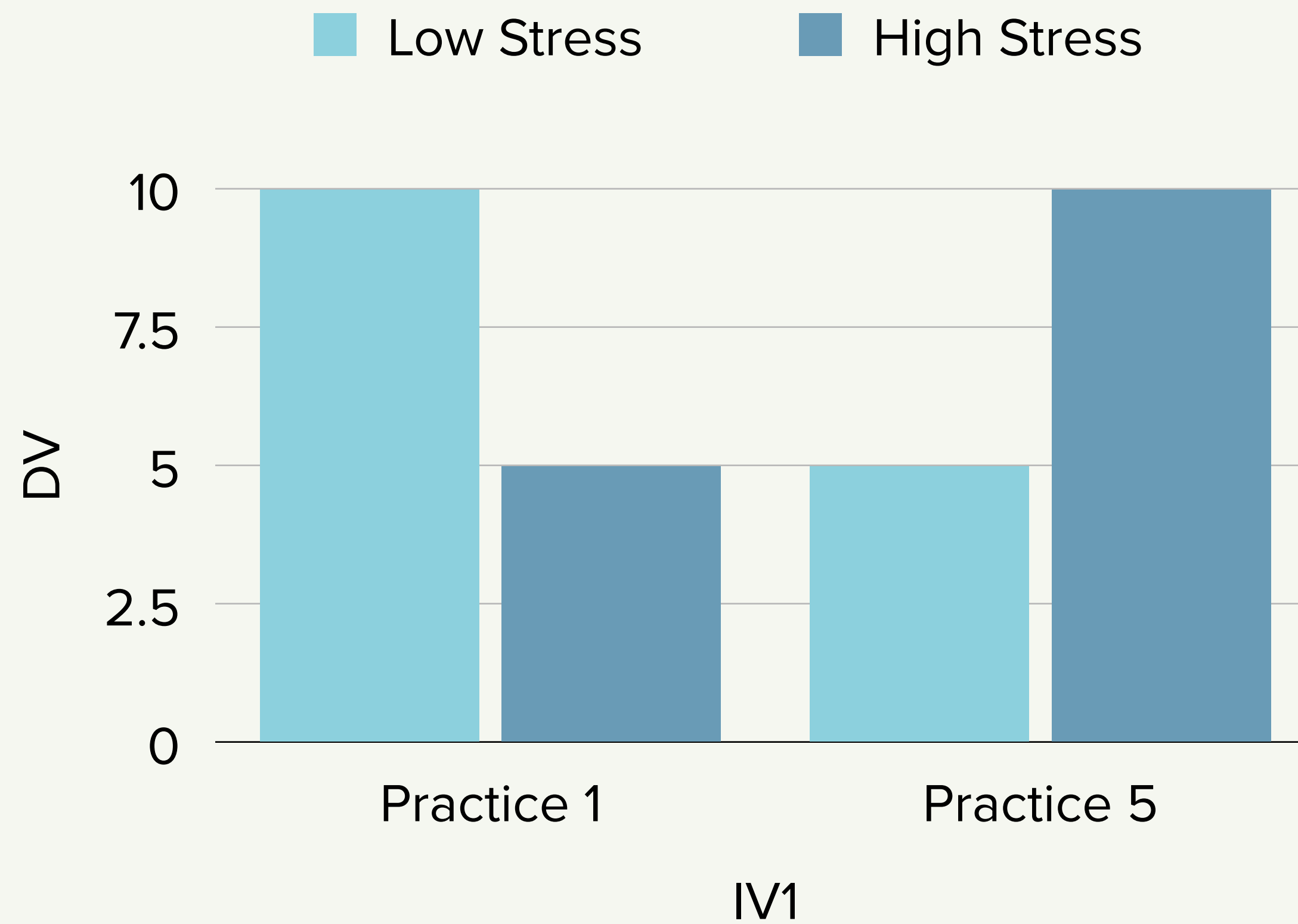
effect types

	IV2	
IV1	10	5
	5	5



effect types

	IV2	
IV1	10	5
	5	10



designs

One-factor (IV) experiments: One manipulation, regardless of number of DVs

- E.g., Does framing type affect message persuasiveness?

Matched groups design: B/WS design in which one creates two groups that are matched on a variable that is likely to be highly correlated with the outcome variable of interest (typically pretest on the variable)

- E.g., Efficacy of Alzheimer's drug depends on age - match on this

Factorial designs (two or more IVs): Experiments with two (or more) independent variables, or factors (2nd is often a moderator).

- E.g., Does framing and emotional content affect message persuasiveness?

factorial designs

Design Statements reflect the number and type of IVs in an experiment:

- Two between-subjects variables with two levels each: 2 X 2 between subjects design.
- Two within-subjects variables, one has two levels and the other has three levels: 2 X 3 within-subjects design, or 2 X 3 repeated measures design.
- One within-subjects variable and one between-subjects variable, each with two levels: 2 X 2 mixed-model design.

factorial designs

Three factors or more: Interaction effects can be difficult to predict and explain (and power).

As a general rule, stay away from four way interactions!

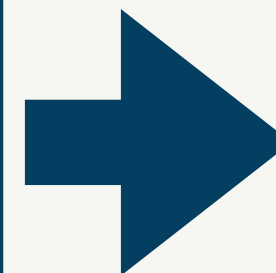
design + stats (parametric)

Two levels

Two groups with 2 levels
each

Variable relationships

Multiple dependent
variables



T-test

ANOVA, ANCOVA,
Regression, Mixed-
models

Regression + Correlation

MANOVA

practice

- Students watched a cartoon either alone or with others and then rated how funny they found the cartoon to be.
 - Independent Variable:
 - Dependent Variable:
 - Design:

practice

- A comprehension test was given to students after they had studied textbook material either in silence or with music turned on and either in their room or in the library.
 - Independent Variable:
 - Dependent Variable:
 - Design:

practice

- Workers at a company were assigned to either complete a stress management training program or not complete the program. Then, they were asked to complete a mindfulness meditation and a meditation exercise on Headspace. The number of sick days taken by the workers was examined for the next two months.
 - Independent Variable:
 - Dependent Variable:
 - Design:

kiss

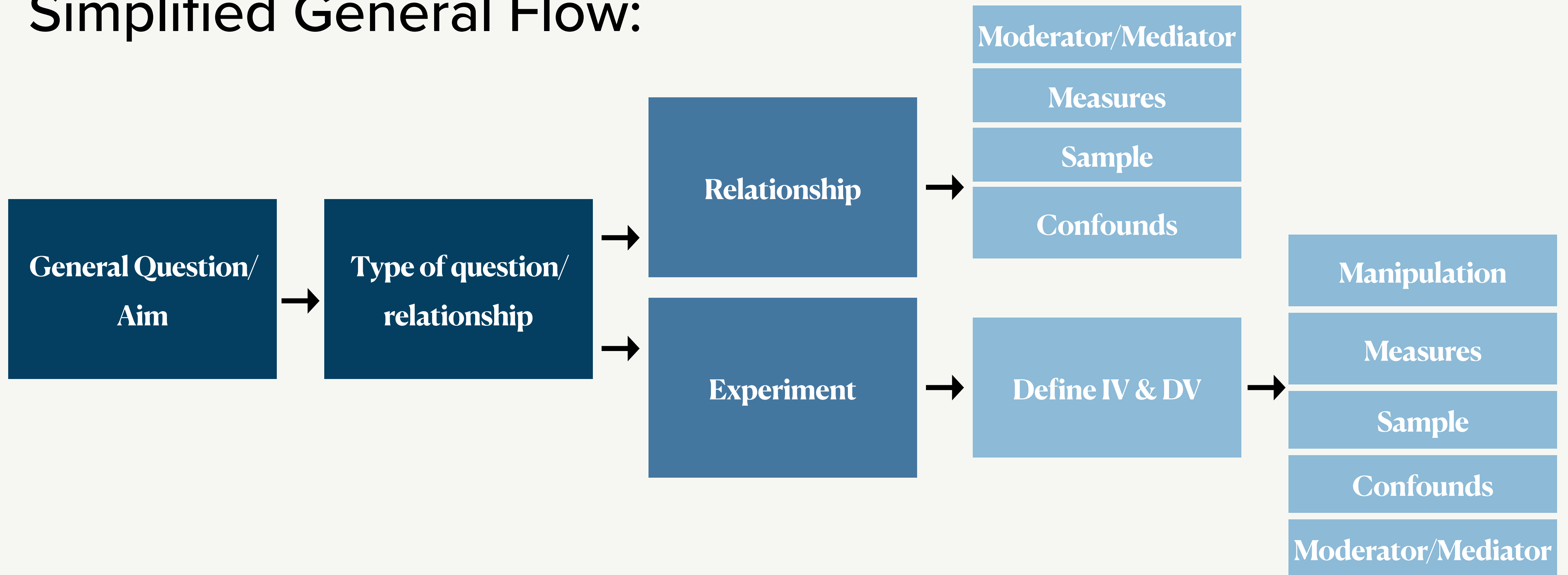
“Keep It Simple, Stupid”

designing experiments

- It is tough!
- Lots of research needed to find the paradigms and operationalizations that best suit your question
- Takes time (especially wrt confounds/colliders)
- Isn't always revealing

designing experiments

Simplified General Flow:



designing experiments

Toolkit:

**Theoretically sound
concepts +
operationalization of
variables**

**Validity of
experimental
paradigm/ Survey
questions**

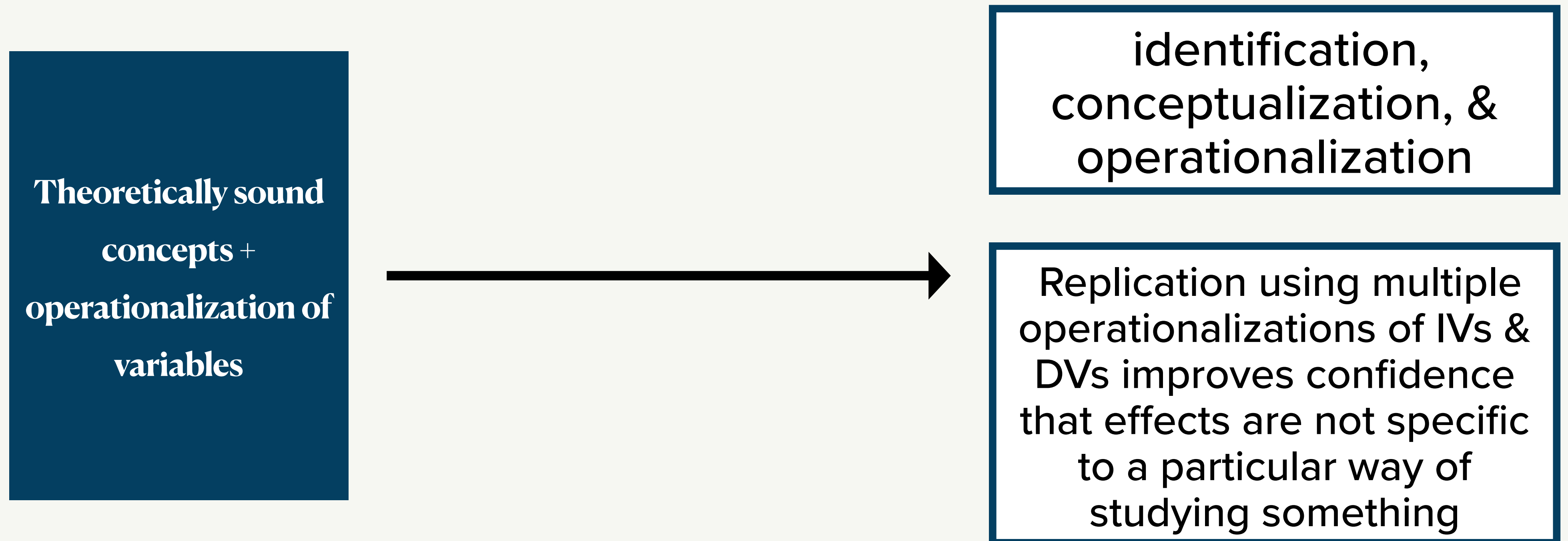
**Does experimental
design capture what
you intend?**

**Enough power to
detect effects/
Conduct the correct
analyses**

**Manipulation Checks/
Cronbach's Alpha/
Correlations among
variables of interest/
ICCs**

designing experiments

Toolkit:



designing experiments

Toolkit:

Validity of
experimental
paradigm/ Survey
questions



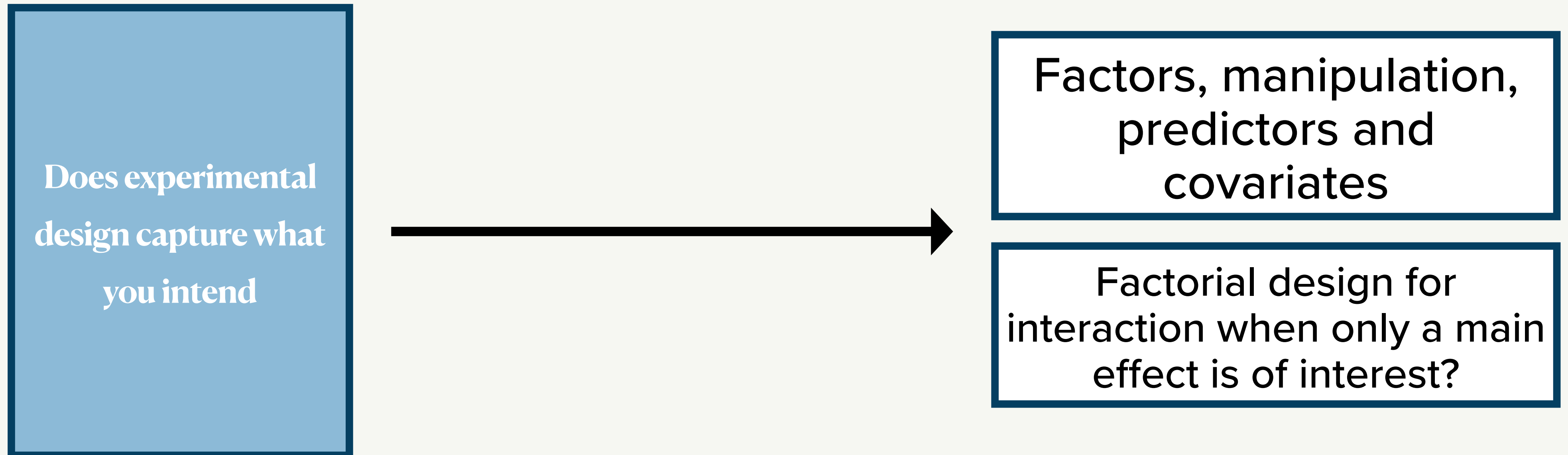
Is this the best way of testing your question? Is it the right level of analysis?

Are there other ways to phrase questions or approach the problem that you didn't yet consider?

E.g., Dictator game vs. Trust game vs. 3PP

designing experiments

Toolkit:



designing experiments

Toolkit:

Enough power to
detect effects/
Conduct the correct
analyses



A priori power analysis using
a sample size estimated
from either SESOI or
previous research (but
publication bias)

Simulation is another option

ANOVA, ANCOVA,
MANOVA, Regression, T-test

designing experiments

Toolkit:

**Manipulation Checks/
Cronbach's Alpha/
Correlations among
variables of interest/
ICCs**



Make sure your
manipulation did what
you intended it to do

Make sure your measures
are measuring what you
want them to

Make sure your mixed-
effects structure holds

designing experiments in philosophy

Experiments in philosophy

- In Western philosophy, typically a philosopher describes a situation (usually imaginary), and asks whether some of the people or objects or events in the situation described have some philosophically interesting property or relation
 - E.g., Is the action describe morally wrong?
 - E.g., When the speaker in the story uses the word `water' does the word refer to H₂O?

designing experiments in philosophy

- These kinds of questions make good experiments!
 - E.g., Phillips's work on semantic theories of modal terms—words like 'might' or 'could'.
 - E.g., Prinz's work on emotion and amplification of moral condemnation.

designing experiments in philosophy

- But sometimes philosophers may be interested in *the nature* of knowledge—what knowledge is and is not
 - E.g., Political philosophers may be concerned with *the nature* of justice

questions?

workshop contents

Part 1:

Introduction

General concepts & definitions

Part 3:

Survey Research

Experiment creation + rules /
steps to follow

Part 2:

Experimental Design

Part 4:

Open science + collaborations

Ethics and the IRB

Part 3: Survey Research

survey research

- A powerful tool for research with human participants
- Used to investigate descriptive, relational, or experimental questions
- Used to investigate quantitative (e.g., moral judgments) or qualitative (e.g., radical imagination) relationships
- Collected online!



survey research

Can include manipulations:

These are usually in the written form for survey methods, but can also be with images, time restraint, audio, video, etc.

survey research

Different types of sampling/designs:

- Cross-sectional surveys: data collection at a single point from sample
- Repeated cross sectional: multiple independent surveys conducted over time
- Panel survey: data collected from same Ps at two or more points of time

qual research

Many types, including observational, focus groups, archival data & content analysis. Important to:

- Sort the data consistent among judges (and calculate IRR)
- Decide what categories & units and define them (or count them)
- If not all material is available to be analyzed, an appropriate sampling procedure be used

qual research

When to use qual research:

- If context is central to RQ (e.g. learning about worker satisfaction at specific company)
- If participant interpretation is central to RQ (Ps can explain why they feel the way they do)
- If depth is important (can provide more detail)
- If research is exploratory (help develop specific RQs for other methods)
- If topic may cause discomfort (may get findings otherwise missed)

qual research

Content Analysis:

- For any research approach that yields textual data
- Meaning condensation and categorization using frequencies/co-occurrence

survey research

One of the most important aspects of survey research is measurement



survey research

Measurement: a system for conceptualizing, observing, and describing the quality and quantity of a phenomena.

1. measurements must have a purpose
2. measurement is usually concerned with properties and attributes of something, rather than with that something specifically
3. entails a system of rules for consistency
4. typically (but not always) based on numerical descriptions of representations
5. must represent observable and meaningful properties

survey research

Measurement techniques:

- Self-report measures
 - Categorical judgements
 - Response Alternatives
 - Dimensional judgments
 - Questionnaires
 - Open-ended responses (qualitative)

survey research

Measurement techniques:

Categorical judgments:

- E.g., forced choice judgments (punish vs. no punish in a moral decision task)
 - “What do you like most about Apple products?”
 - “Good customer service”, “Good design”, “Easy to use”, “Quick to learn”
- Categories should not overlap
- Category options should be complete
- Category options should pertain to the question asked

survey research

Measurement techniques:

Categorical judgments:

- Pros: can reduce halo effect
- Cons: validity concerns

survey research

Measurement techniques:

Dimensional judgments:

- Responses based on a continuous (or ordinal) scale
 - numerical: Likert-type scales
 - graphic: slider scale where p's respond via slider or marking the line (can have numbers or only anchors)
- No formal guidelines on how long a scale should be, but anchors should simple, clear, and unambiguous
 - unipolar scale: levels of the same dimension (e.g., somewhat warm, moderately warm, very warm)
 - bipolar scale: negative and positive extremes of a dimension (e.g., very cold to very warm)
 - magnitude scale: p's rate intensity of a list of variable based on an anchored example (i.e., how severe is murder compared to stealing a bike?)



survey research

Measurement techniques:

Dimensional judgments:

- Pros: More variability than with categorical judgments, typically able to capture a wider range of options
- Cons: Issues can arise when the scales are not created in a valid way

What kind of scales to use?

- can make midpoint 0
- Introversion and extraversion are like this -> 0 as midpoint - everything below is intra and everything above is extra
- pos and neg for emotions
- regardless, mid point is usually average - then things below and above as lower frequency

survey research

Measurement techniques:

- Regardless of the method or measures used, intrinsic measurement factors are very important
 - They include length of stimulus exposure, quality of stimulus materials, order of presentation.
 - Usually in surveys, you want it to be self-paced (unless time is a factor)
 - Sometimes with a one-shot judgment, you might constrain time

survey research

Best to use previously validated measures where possible, but when creating your own items, consider:

- Biases
- Loaded questions
- Leading questions
- Double Barreled questions
- Double negatives

survey research

Biases:

- Halo bias: all ratings inflated towards top of scale
- Leniency bias - ppl give the benefit of the doubt and are more lenient in their responses
- Error of leniency: Participants rate someone whom they're familiar more positively
 - participants made aware of this bias may overcompensate and commit severity error, where they rate them more negatively.
- Error of central tendency: Participants hesitate to give extreme rating, and answer toward the middle of the scale.
- Logical error in rating: Participants rate variables that they few as related similarly, similar to halo effect in that it results in high correlations and low variability.

survey research

Loaded questions:

- Questions that contain a false, presumptive, or that contain emotionally provocative terms
- There is a lot of content that assumes something about the participant (e.g., have you stopped beating your wife?)
 - That assumes that you were or have
- Questions that contain emotionally provocative (typically neg)
 - How do you feel about the liberal agenda?



survey research

Leading questions:

- Pull you to answer in a certain way
 - E.g., How fast were the cars going before they crashed into each other yields a different response than using collided instead
 - E.g., How tall some is versus how short - leads to higher estimates than the other
 - Something like “What is the estimated height of this person?”



survey research

Double-barreled questions:

- Do you think the gov. should increase taxes and provide more support for single parent families?
- Someone who disagrees with the first part and agrees with second wouldn't know how to respond - and may be forced to choose which to respond to leading to an increase in noise



survey research

Double negatives:

- E.g., Organization should not be required to monitor overtime over employees
- If someone answers disagree then that means that they should
- Especially if someone is fatigued
- Also beware of questions that include jargon

survey research

Issues with self-report measures:

- Ceiling & Floor
- Blanks
- Memory biases
- How participants approach questions

survey research

Ceiling Effect:

- When you are asking a questions, the idea is that you want to be able to distinguish among people - hopefully distinguish between low, mid, and high - if everyone is high then you have restricted range problem and no variability and cannot thus look at correlation of that item
 - One way to avoid this is to elongate the scale or add descriptors
 - If you had a question about satisfaction at a job where you think everyone is a little satisfied, you can have the scale say dissatisfied, neutral, and then lots of words for the next 7 or so
 - In this case, this would assume the middle of the scale as moderate satisfaction



survey research

Floor Effect:

- Same as ceiling but the other direction
 - Example is if you asked participants who are generally satisfied at work "I intend to leave in 60 months", they will probably all cluster around no or very low ratings



survey research

- Never put "leave this blank"
- It is impossible to tell nonresponse from intentional blanks
- **ESPECIALLY** for an attention check item!

survey research

Memory-related biases

- Repeated administrations of a survey may result in responses that are anchored by previous responses.
- Correlations among similarly or identically worded questions will be artificially inflated.
 - **Solution:** Create sufficiently long lags between administrations so that memory for previous responses fades

survey research

How participants approach questions :

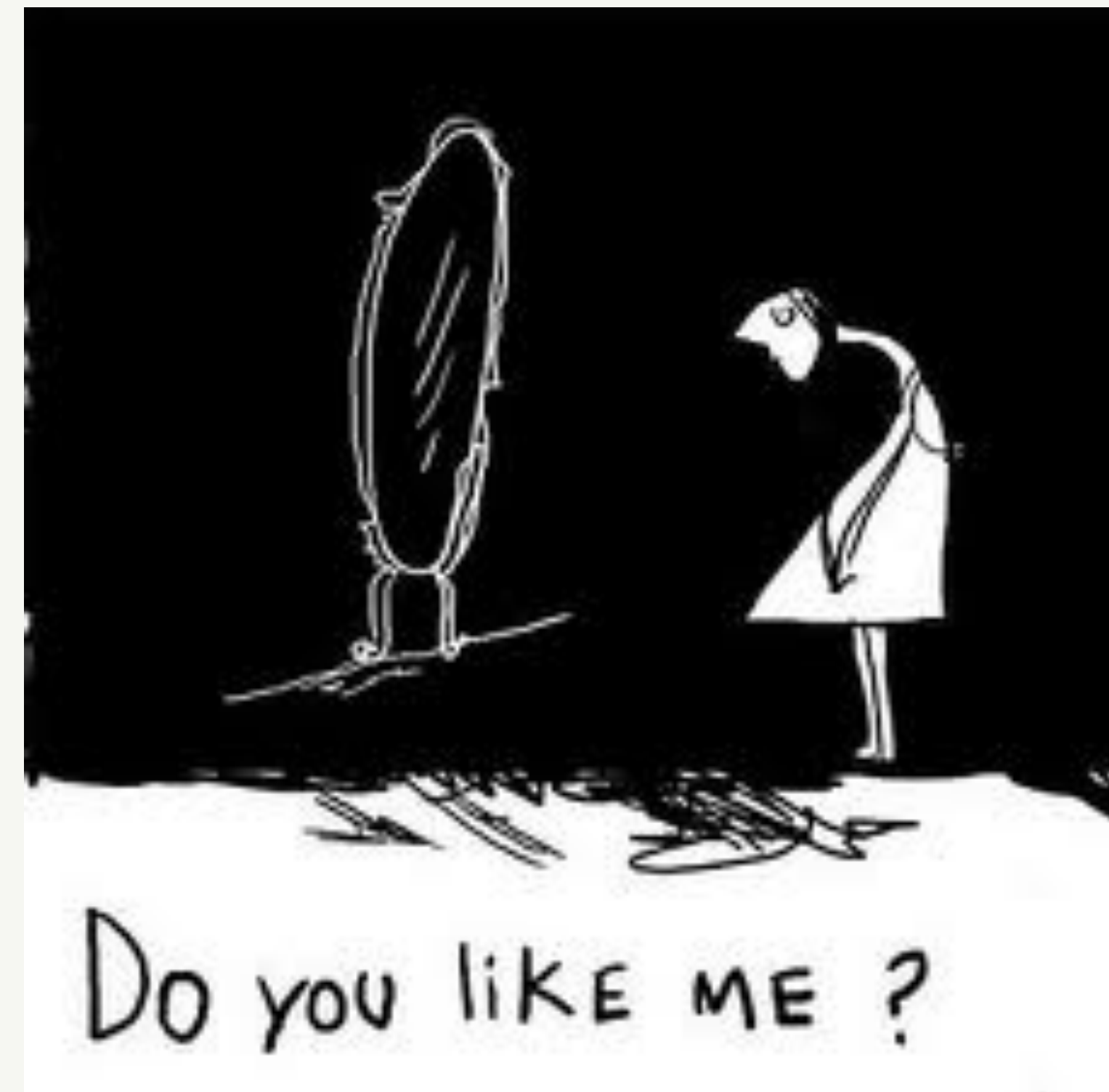
- Sometimes participants may not know answers, and guess.
- Participants may not be open and forthcoming when answering
- Participants **often** do not and cannot realistically evaluate their behaviors
- Subjectivity of answers across different participants
 - Does your (1) NOT AT ALL HUNGRY represent the same for me?
- 1-10 scales elicit different responses than -5 to +5 scales

survey research

- Participants often have self-presentation concerns
- Guilt, shame, embarrassment (among other neg emotions) can be tough to elicit honest responses on
- You have to be extremely careful about how you ask - questions about sexual behavior, criminal history, unethical behaviors at work, etc
- Phrase so that participants are less likely to get offended
- **ALWAYS** preface items with a request for honest responding - **AND ASSURING ANONYMITY**

survey research

- Marlow crowne: measure of social desirability
- Often times, asking at the end “Where you honest” can elicit similar responses



survey research

Questionnaires:

- Already validated questionnaires are a great way to measure things
- However, if you edit the phrasing, scale anchors, etc, then you are no longer really working with the validated measure
- Shortening scales also means you are not using a validated measure (unless the shortened measure is validated; e.g., RWA vs. RWA-short)
- Important to follow the validated measures closely, including the wording of the instructions

survey research

Developing questionnaires:

- Piloting initial questions to test wording and others conception of the question
- Avoid questions that require narrow answers
- Avoid overly complex questions
- Funnel sequence of questions: strategy that moves from broad questions and ends with specific questions
- Acquiescent response set (yea-saying): Ps always respond in agreement or positively with the item (reverse of this is nay-saying).
- For cross-cultural surveys use “back-translation”: have bilinguals translate survey from source language to target language and compensate/replace gaps in translation with analogous terms.

survey research

Things to keep in mind when developing questionnaire items

- How will the respondent interpret the question?
- Is the question ambiguous? Will it be interpreted differently by different people?
- Is the question written in plain language? Does it include jargon that is difficult for laypeople to understand?
- Is there a social desirability component to the question? Will people lie to look good or to protect a positive self-image?
- Is the question biased? Does it “pull” for a particular type of response?
- Does the operational definition fit the conceptual definition? Are you really tapping into the construct that you are interested in measuring?
- Is there a potential restriction of range problem? Will you have a ceiling or floor effect?
- Is the questionnaire so long that people will become fatigued and fail to attend carefully to the items?

survey research

- Anything you intend to collapse into a single measure needs some reliability statistics first
 - Correlations among items
 - Cronbach's alpha: measures the extent to which items measuring the same construct relate to one another
 - Minimum of 3 items needed before you can collapse
- This includes questionnaire items that are to be collapsed and judgments that center around a central concept (e.g., multiple ways to ask about punishment or blame)

survey research

- How much tv do you watch?
 - Scale from 1 hour to 3 hours a day
 - 16.2% indicated watching tv for more than 2.5 hours a day

survey research

- How much tv do you watch?
 - Scale from 1 hour to 6 hours a day
 - 37.5% say they watch more than 2.5 hours a day

survey research

- If you really wanted to ask this type of question, you may want to have them write it in.
- Anchors have a huge effect on responding (and the mind in general)
- You can also use qualifiers like "frequently, not frequently, etc"
 - But the problem here would be that people have varied understandings of what frequency means, so if ill-defined, you'll get noise in your data

online data collection

online data collection

- When, what, and how matter
 - What you are asking about and when — E.g., asking about purity during the pandemic is not likely to lead to replicable effects
 - How you ask the questions - using what measure is also important
- Length & type matter
 - Long surveys lead to attrition and poor data quality
 - As short as humanly possible is ideal (think about how we interact with social media)
 - If all of the questions are the same, this can lead to a boring survey that has low engagement and high errors (leading to more noise)

online data collection

- How you ask questions about attention and engagement also matters
 - Attention checks are **CRITICAL** to ensure you have good data quality
 - To check for participants who are paying attention
 - To check for participants who may display non-variance in responding
 - To check for bots
 - If you phrase your attention checks as “Gotchas!”, you might change how the participant interacts with the rest of the survey
 - Also important to see how long it took participants to complete the survey

online data collection

- Attention checks and bot checks should be administered throughout the survey
- Both self-attention and demand questions should come at the end of the survey
 - Examples on Qualtrics!

online data collection

- Prolific vs. MTurk vs. CloudResearch
- These are the most common sources of online participants - they take surveys in exchange for payment and, thus, may be different than an average person
- These services allow you to access specific populations (e.g., democrats, people who live in Australia, etc)
- Data may differ from data collected at CUNY

questions?

Part 4: Other Important Things

reproducibility

Robust experimental design is crucial to reproducible and generalizable findings

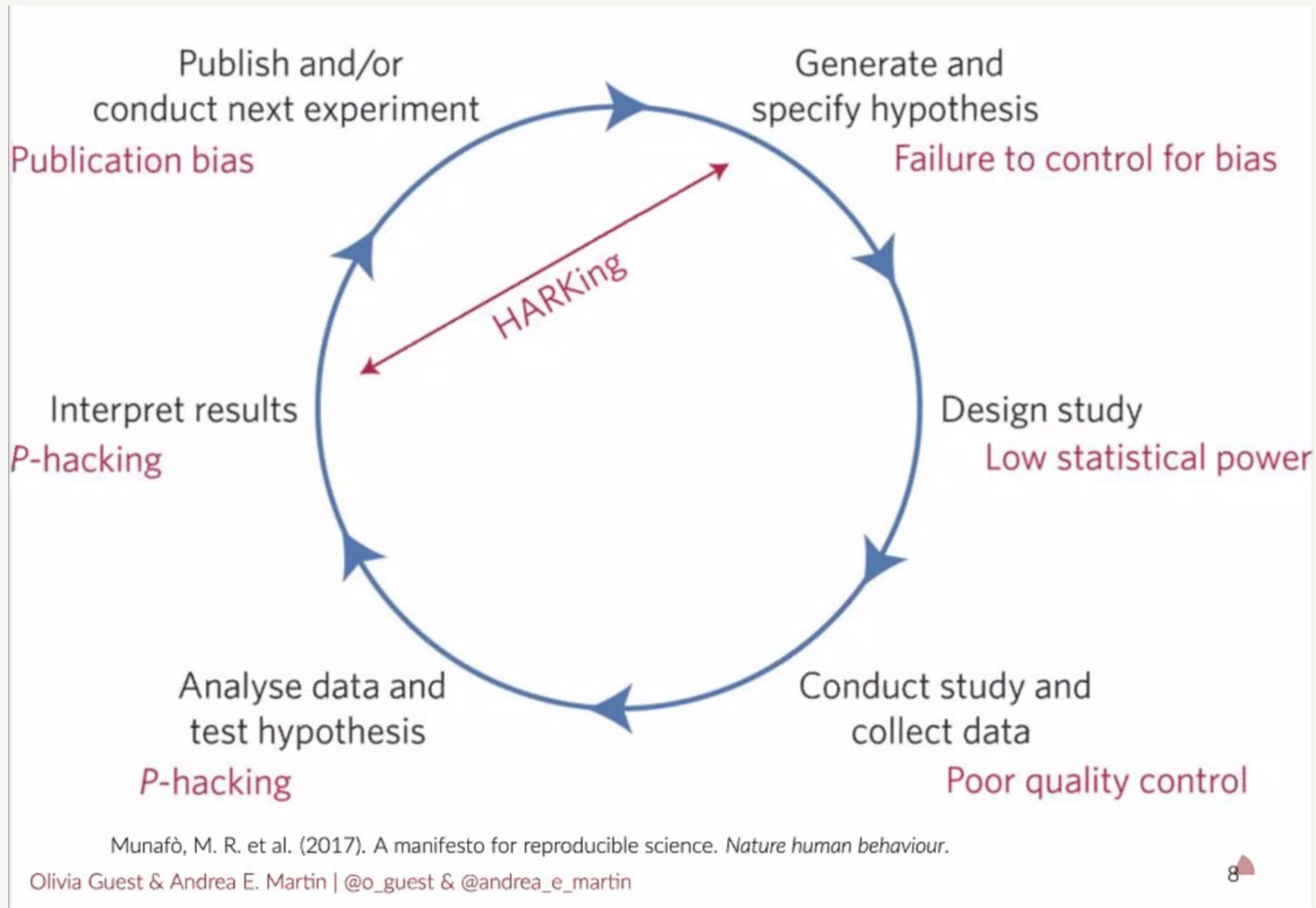
The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are precognition (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

reproducibility

- In an attempt to replicate 100 studies published in 3 top psychology journals, the following results were found:
 - Average effect size in replications was about half the magnitude of average effect size of original studies
 - Only 36% of replications had significant results ($p < .05$) while 97% of original studies had significant results
- Labeled by some as a replication or reproducibility crisis because only around 1/3 of studies were successfully replicated

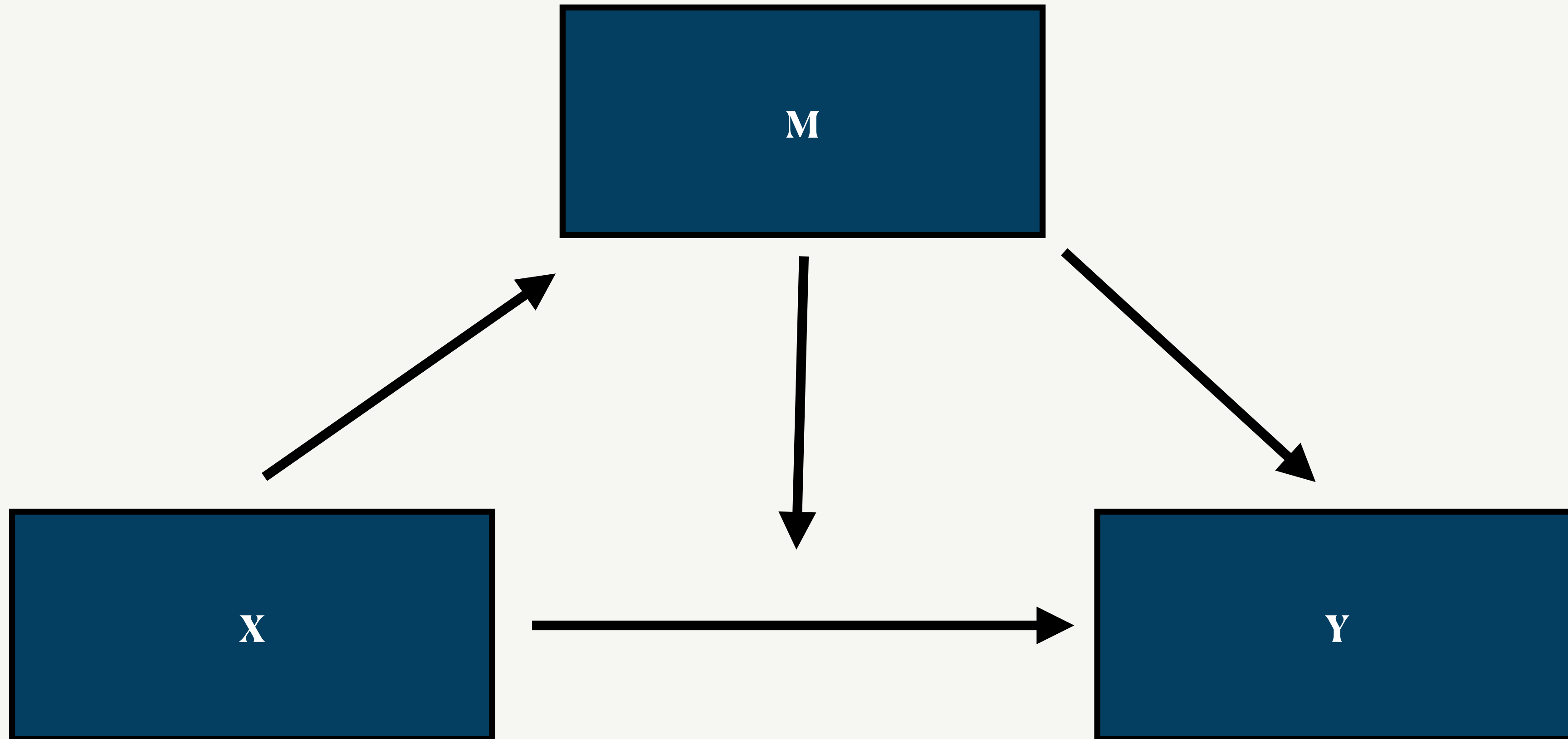
reproducibility


- Issues with measurement, causal claims, and analyses are also major causes of non-replicable research
 - Measurement is hard!
 - Overstated causal claims (or claiming causation where none exists) is dangerous
 - Analytic methods are also important to tackling some of these issues (but not all!)



QMPs

- Aside from questionable research practices (QRPs) that have really dominated critiques of psychological research, there are also issues with measurement strategies and norms
- Cronbach's alpha (α) is used to establish validity of a measure
 - But Cronbach's alpha is a measure of reliability. Good internal reliability doesn't tell you much about the validity of a measure!
- This is unfolding now and the consequences are not clear. Best strategy is to keep note of and report ALL measurement decisions
 - Where did the measure come from?
 - Did you make it up? How do you know it measures what you think it does?
 - What items did you use? Did you use all of the items in the validated measure?
 - What scale anchors did you use? What wording did you use or tweak?



 **Brenton Wiernik** 🏳️‍🌈 @bmwiernik · Nov 16, 2020


(1) interaction effects tend to be maybe half of the size of main effects; given that, the larger standard errors associated with multiplicative terms makes them super noisy (see statmodeling.stat.columbia.edu/2018/03/15/nee... for a demo, the Cohen et al regression book for formal treatment of int SEs).

4


 **Brenton Wiernik** 🏳️‍🌈 @bmwiernik · Nov 16, 2020

(2) if you are talking about interactions between two measured variables, not assigned conditions, then reliability will likely be terrible. The reliability of the int term is roughly the product of the two reliabilities (somewhat higher if measures correlated), so >>

1

 **Brenton Wiernik** 🏳️‍🌈 @bmwiernik · Nov 16, 2020

if you have two measures with reliability = .70, then the interaction likely has a reliability \approx .50. See

 **Multiple Regression: Testing and Interpreting Intera...**
Multiple Regression: Testing and Interpreting Interactions
amazon.com

other important things

Missing data




- Is the data missing from the same condition? This would be an issue.
- Is the data missing at random? Probably fine!
- Is there a lot of missing data (or high attrition), then you probably have a larger issue validity issue (potential fatal for the experiment)

other important things

Volunteer bias

- Certain types of people may volunteer to participate, thereby biasing the results
- Volunteer subjects (who tend to be educated and smart) may skew norms in test standardization procedures.
- Volunteers (who tend to be higher in need for approval) may affect the internal validity of an experiment.
- Volunteerism may also affect the external validity of results such that results cannot be generalized to the rest of the (nonvolunteer) population.

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		—	+00	-0-	--0	-10	—	—	0	0	—
Discounting (δ)	—		0	0	0	+	0	0	0	+00	0
Dictator	+00	0		—	—	—	—	—	00+	0-0	-00
Prisoner's Dilemma	-0-	0	—		+	+++	+	+00	0	0+0	+
Lying	--0	0	—	+		+00	+00	+	0	0++	+00
Cognitive	--0	+	—	+++	+00		+	+	-00	0	+
Confidence	—	0	—	+	+00	+		+	0	+00	+
Compete	—	0	—	+00	+	+	+		0	+00	+
IAT Race	0	0	00+	0	0	-00	0	0		+	0
IAT Gender	0	+00	0-0	0+0	0++	0	+00	+00	+		+
Male	—	0	-00	+	+00	+	+	+	0	+	

Notes:  indicates complete agreement,  complete disagreement, and  two out of three samples agreeing.

other important things

Participant agency!

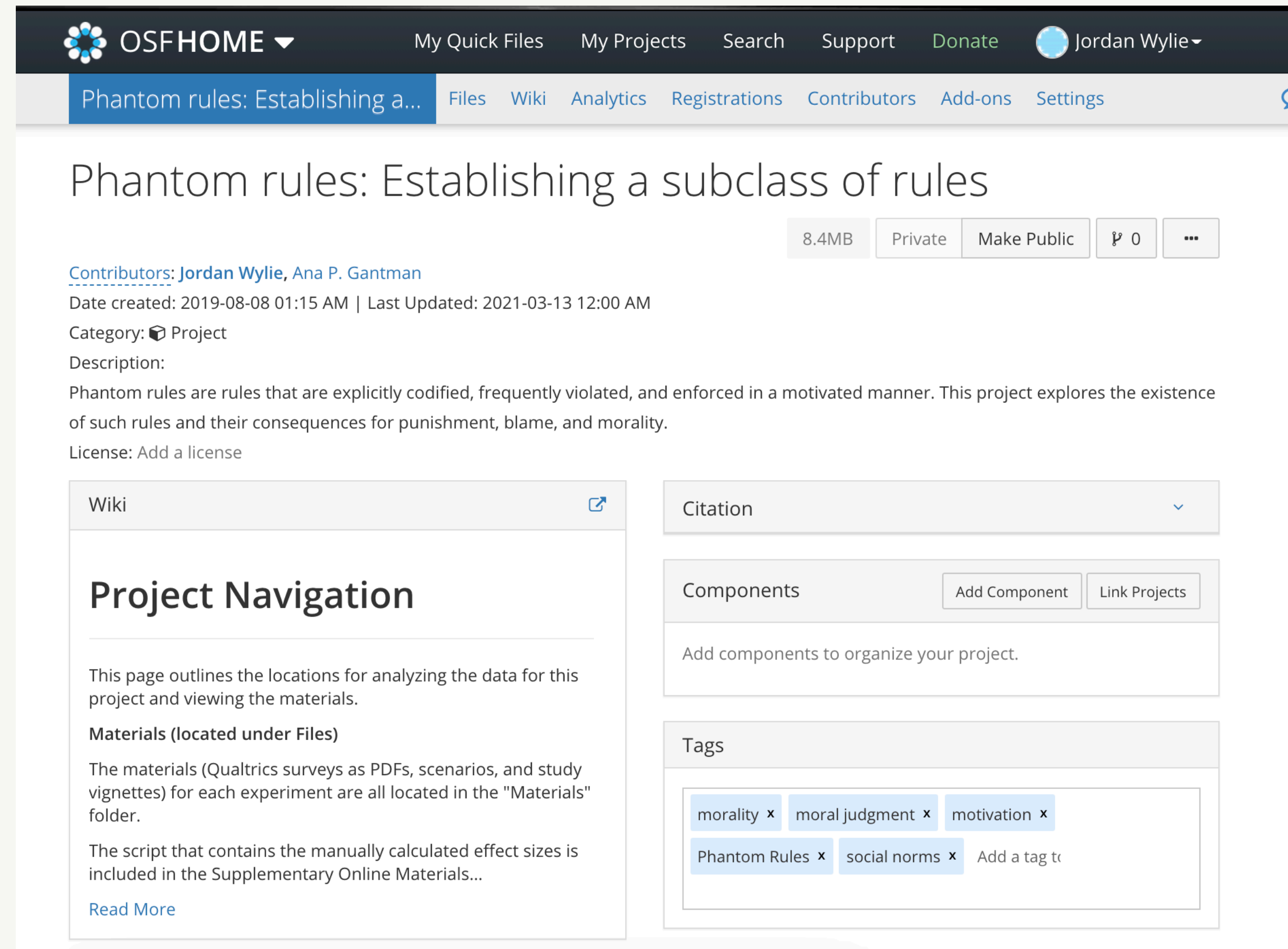
- Often times when setting up experiments, we forget that there are people on the other side who we do not want to leave feeling worse in any way (especially when our questions don't tend to be matters of great importance)
 - This will come up with the IRB, but often we as researchers can avoid putting participants in bad situations with a little thought
 - E.g., email replies experiment
- For some questions, you may want to consider PARE research & design
 - PARE = Participatory Action Research
 - <https://publicscienceproject.org/>

other important things

- When experiments CANT tell us things
 - Meta-ethical research suggests that moral judgments might not mean to lay participants what we think they do
 - Top down philosopher and researcher distinctions might not map well to
- Descriptive work is **really** important

random tips

- OSF & Bro-pen science
 - Pre-registration is the process of time stamping your hypotheses. It should be done **BEFORE** data collection
 - You can also pre-register your analyses (which is recommended). Deciding which tests you will run before you collect data is often really useful to thinking about the project more broadly
 - On OSF, I use the “As predicted” template to pre-register my hypotheses and analyses, but you can use whatever template suits you best.
 - You can also upload your materials onto OSF (or GitHub)
 - You can also use aspredicted.com to pre-register your hypotheses (not my personal preference).

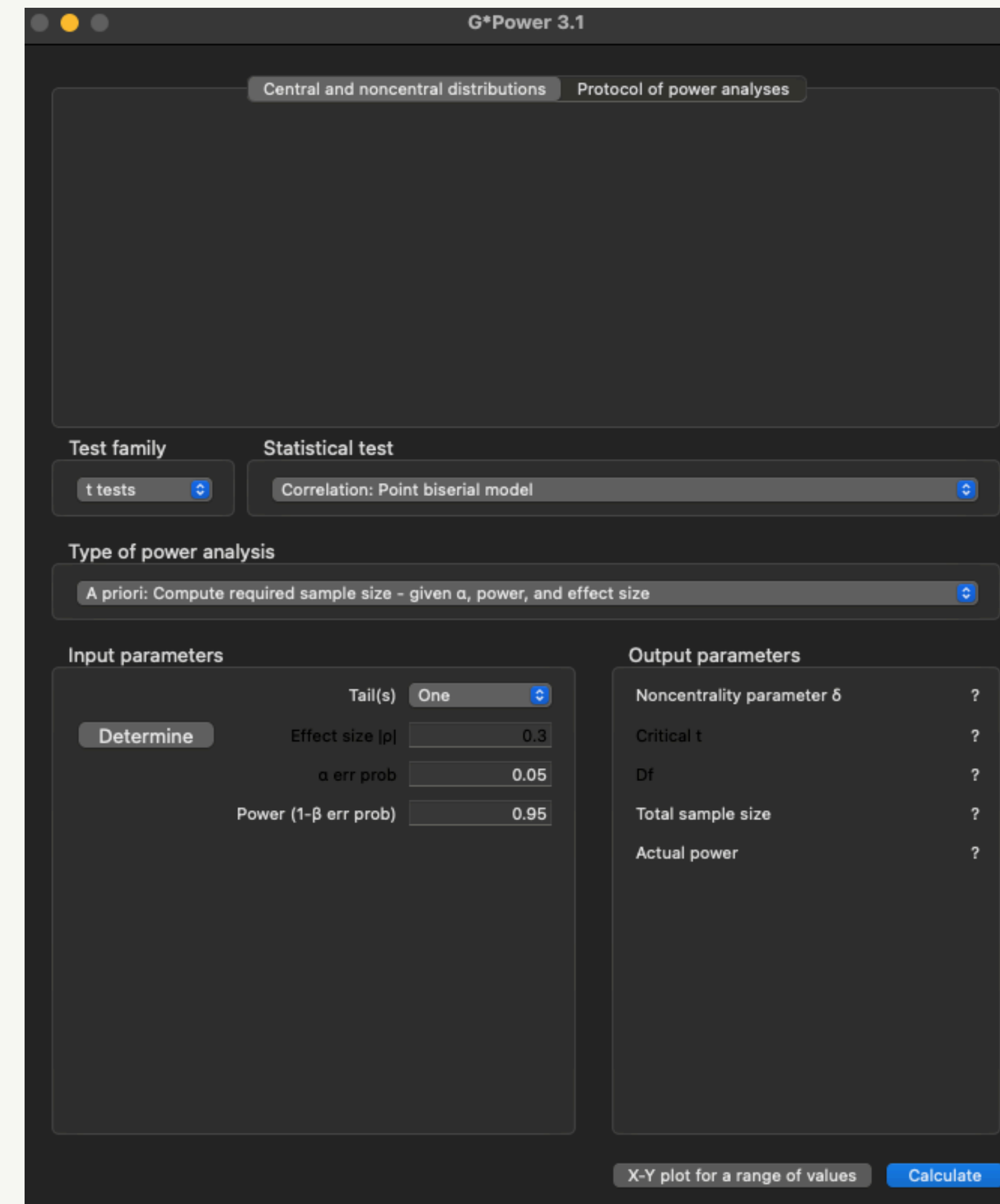


The screenshot shows the OSFHOME interface for a project titled "Phantom rules: Establishing a subclass of rules". The page includes a navigation bar with "OSFHOM" and "My Quick Files", "My Projects", "Search", "Support", "Donate", and a user profile for "Jordan Wylie". Below the navigation bar, there are tabs for "Files", "Wiki", "Analytics", "Registrations", "Contributors", "Add-ons", and "Settings". The main content area displays the project title, a file size of 8.4MB, and options to "Private", "Make Public", and "0" (likely a version or count). It also lists contributors "Jordan Wylie, Ana P. Gantman", the creation date "2019-08-08 01:15 AM", and the last update date "2021-03-13 12:00 AM". The category is "Project" and the description states: "Phantom rules are rules that are explicitly codified, frequently violated, and enforced in a motivated manner. This project explores the existence of such rules and their consequences for punishment, blame, and morality." There is a "License: Add a license" option. The page is divided into sections: "Wiki" (with a "Project Navigation" section), "Citation", "Components" (with "Add Component" and "Link Projects" buttons), and "Tags" (with tags for "morality", "moral judgment", "motivation", "Phantom Rules", and "social norms").

random tips

Power analysis:

G*Power (free)



random tips

Statistical test

✓ Correlation: Point biserial model

Linear bivariate regression: One group, size of slope

Linear bivariate regression: Two groups, difference between intercepts

Linear bivariate regression: Two groups, difference between slopes

Linear multiple regression: Fixed model, single regression coefficient

Means: Difference between two dependent means (matched pairs)

Means: Difference between two independent means (two groups)

Means: Difference from constant (one sample case)

Means: Wilcoxon signed-rank test (matched pairs)

Means: Wilcoxon signed-rank test (one sample case)

Means: Wilcoxon-Mann-Whitney test (two groups)

• Generic t test

random tips

Statistical test

✓ ANCOVA: Fixed effects, main effects and interactions

ANOVA: Fixed effects, omnibus, one-way

ANOVA: Fixed effects, special, main effects and interactions

ANOVA: Repeated measures, between factors

ANOVA: Repeated measures, within factors

ANOVA: Repeated measures, within-between interaction

Hotellings T^2 : One group mean vector

Hotellings T^2 : Two group mean vectors

MANOVA: Global effects

MANOVA: Special effects and interactions

MANOVA: Repeated measures, between factors

MANOVA: Repeated measures, within factors

MANOVA: Repeated measures, within-between interaction

Linear multiple regression: Fixed model, R^2 deviation from zero

Linear multiple regression: Fixed model, R^2 increase

Variance: Test of equality (two sample case)

• Generic F test

random tips

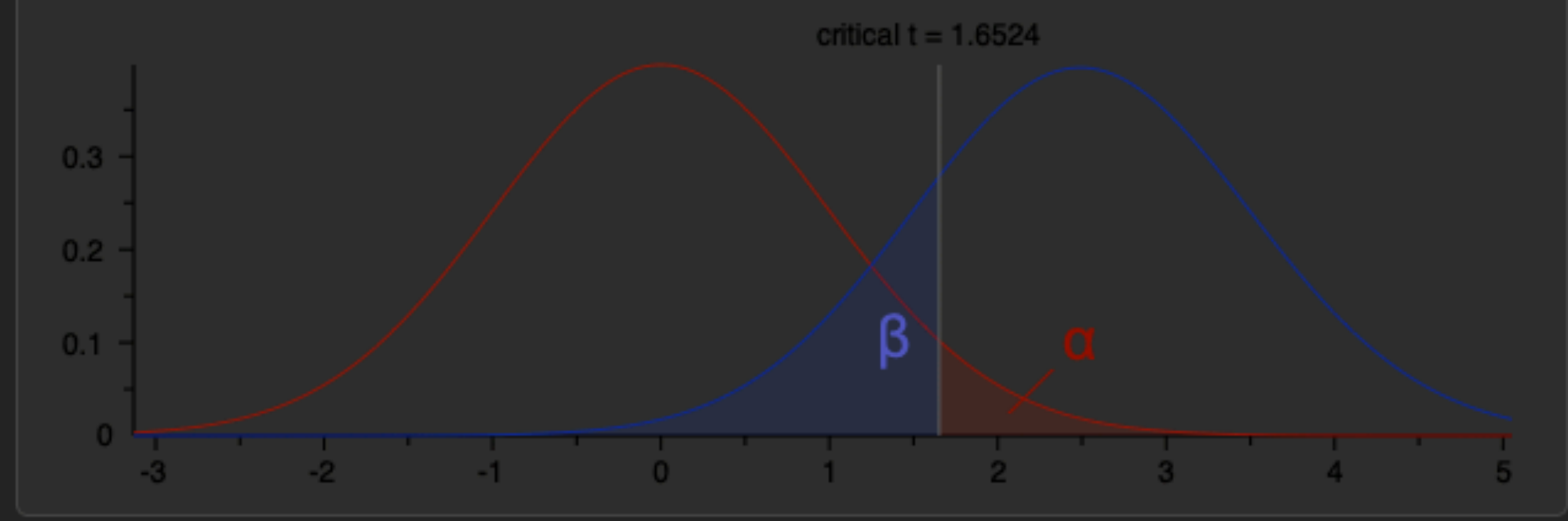
✓ A priori: Compute required sample size - given α , power, and effect size

Compromise: Compute implied α & power - given β/α ratio, sample size, and effect size

Criterion: Compute required α - given power, effect size, and sample size

Post hoc: Compute achieved power - given α , sample size, and effect size

Sensitivity: Compute required effect size - given α , power, and sample size



Test family:

Statistical test:

Type of power analysis:

Input parameters

Tail(s)

Effect size d

α err prob

Power ($1-\beta$ err prob)

Allocation ratio N2/N1

Output parameters

Noncentrality parameter δ	2.4994999
Critical t	1.6524320
Df	202
Sample size group 1	102
Sample size group 2	102
Total sample size	204
Actual power	0.8012966

random tips

- 1. Determine a priori the smallest effect size of interest (SESOI)
- 2. Power your study to reliably detect the SESOI
- 3. Perform an equivalence test using the SESOI as the equivalence bounds
- <https://journals.sagepub.com/doi/10.1177/2515245918770963>

random tips

- My intro to R
- <https://github.com/jwylie21/BeginninginR>

Basics Tutorial

Jordan Wylie 10/7/2020

Basics of R Tutorial

This tutorial is to help new users get started with R with the goal of analyzing data. This tutorial was specifically made for undergraduate research assistants, but should be helpful for anyone looking to get started with data analysis in R.

If your goal is to become literate in R programming language, there are many other resources better suited to that goal.

For example:

- <https://swirlstats.com/students.html>
- <https://www.datacamp.com/courses/introduction-to-the-tidyverse>
- <https://kiirstio.wixsite.com/kowen/post/the-25-days-of-christmas-an-r-advent-calendar>

This was created in RStudio (using RMarkdown).

R studio is an Integrated Development Environment (IDE) for R, you will need to download both R and RStudio

Here is the link to R: <https://cran.rstudio.com/>

random tips



JASP

jamovi Stats.
Open.
Now.

random tips

- Sci-hub



random tips

Some potentially useful resources

- <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190954>

The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes

Dimension	Label	Lower Anchor	Upper Anchor
Valence	This image is . . .	UNPLEASANT or NEGATIVE	PLEASANT or POSITIVE
Arousal	This image is . . .	CALMING	EXCITING
Morality	This image portrays something . . .	IMMORAL / BLAMEWORTHY	MORAL / PRAISEWORTHY
Harm	This image makes me think about the concept of CARE / HARM	NOT AT ALL	VERY MUCH
Fairness	This image makes me think about the concept of FAIRNESS / CHEATING	NOT AT ALL	VERY MUCH
Ingroup	This image makes me think about the concept of LOYALTY / BETRAYAL	NOT AT ALL	VERY MUCH
Authority	This image makes me think about the concept of RESPECT / SUBVERSION	NOT AT ALL	VERY MUCH
Purity	This image makes me think about the concept of SANCTITY / DEGRADATION	NOT AT ALL	VERY MUCH

<https://doi.org/10.1371/journal.pone.0190954.t003>

**WHEN DID YOU BECOME AN EXPERT IN
EXPERIMENTAL PHILOSOPHY**



LAST NIGHT

specific questions?