Threats to measurement validity

Why statistics

- An essential aid to “signal detection”
- A universal language for communicating what we find.
- Required for competent evaluation of others’ work.

Goal of today

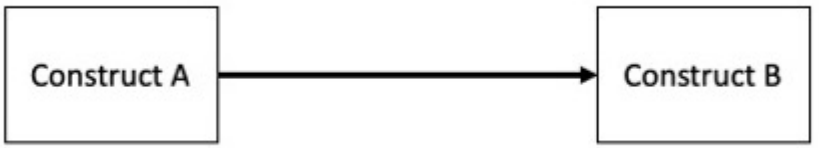
Advanced skill in quantitative methods carries with it the responsibility to use those skills carefully and ethically.

Today, we'll discuss methodological issues present in statistics.

- It can be tempting to use statistics to fix poor research design.
- *These issues cannot be fixed quantitatively* (even when it looks like they can).
- *Do you believe any of the statistics you ran all semester?*

Constructs

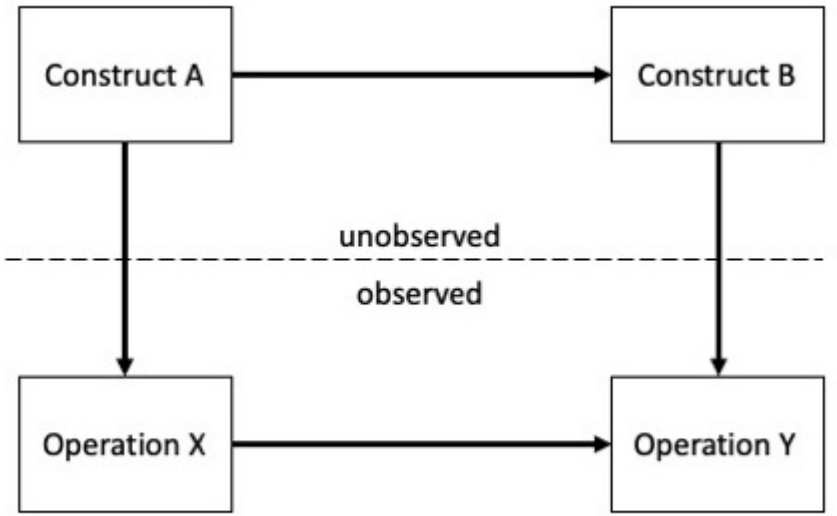
- Our basic goal in science is to make inferences about the causal relations between constructs.



- We can’t do that directly, so we rely on proxies for those constructs

> We are measuring the invisible

Measuring the Invisible



In order to infer that A --> B, we have to make three assumptions:

- X is a good proxy for A
- Y is a good proxy for B
- X and Y are causally related

Measuring the Invisible

- When the first two assumptions are true, the relation between X and Y will provide a good estimate of the relation between A and B.
- What threatens our ability to carry out this seemingly simple task?
    - How do quantitative methods help us solve these problems?

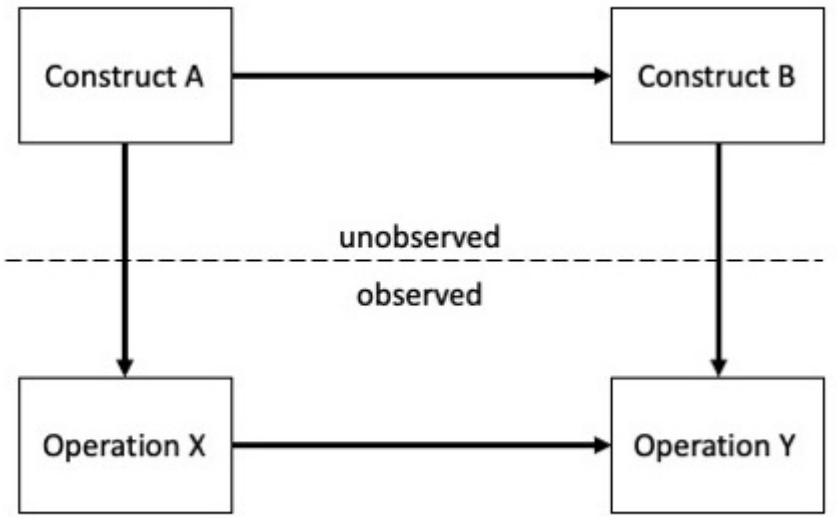
Validity

Four kinds of validity in research threaten our ability to make valid causal inferences. Solving each problem either directly requires quantitative methods or makes use of principles that are central to quantitative methods.

- Statistical conclusion validity
- Internal validity
- Construct validity
- External validity

Statistical conclusion validity

- **Definition:** the validity of the inference that X and Y are related (aka that you are making appropriate and reasonable conclusions). How well do the numbers support the claim? Do you believe the numbers or do you think they are lying to you?



(Some) threats to statistical conclusion validity

- low statistical power
- violations of assumptions
- fishing and the error rate problem
- unreliable measures
- restricted range
- unreliable treatment implementation

(Some) threats to statistical conclusion validity

- **low statistical power**
- violations of assumptions
- fishing and the error rate problem
- unreliable measures
- restricted range
- unreliable treatment implementation

Power = ability to detect an effect, if one is there

(Some) threats to statistical conclusion validity

- low statistical power
- **violations of assumptions**
- fishing and the error rate problem
- unreliable measures
- restricted range
- unreliable treatment implementation

Statistical tests have assumptions

- If we violate those assumptions, is that a useful test?

(Some) threats to statistical conclusion validity

- low statistical power
- violations of assumptions
- **fishing and the error rate problem**
- unreliable measures
- restricted range
- unreliable treatment implementation

Independence Assumption

- We're OK with 5/100 or 1/20
- But if you do 20 tests on the same data...

(Some) threats to statistical conclusion validity

- low statistical power
- violations of assumptions
- fishing and the error rate problem
- **unreliable measures**
- restricted range
- unreliable treatment implementation

Reliability = consistency

(Some) threats to statistical conclusion validity

- low statistical power
- violations of assumptions
- fishing and the error rate problem
- unreliable measures
- **restricted range**
- unreliable treatment implementation

How do you know if there's actually a relationship, when you neglect real values?

(Some) threats to statistical conclusion validity

- low statistical power
- violations of assumptions
- fishing and the error rate problem
- unreliable measures
- restricted range
- unreliable treatment implementation/group assignment

If you say you randomly assigned participants into groups, that better be the case...

Internal validity

- **Definition:** the validity of the inference that X and Y are ***causally*** related. There are no other possible explanations
    - Given that X and Y are correlated, can we validly infer that the relation is causal?
- Requirements:
    - Temporal precedence
    - No confounds

Threats to internal validity

- ambiguous temporal precedence
- selection
- attrition
- history
- maturation
- regression
- testing
- instrumentation

Threats to internal validity

- **ambiguous temporal precedence**
- selection
- attrition
- history
- maturation
- regression
- testing
- instrumentation

Temporal precedence can be established in an experiment because treatment precedes outcome.

But, when treatment is not possible, then logic and common sense can sometimes dictate temporal precedence.

- prenatal nutrition and cognitive development
- depression and cancer

Threats to internal validity

- ambiguous temporal precedence
- **selection**
- attrition
- history
- maturation
- regression
- testing
- instrumentation

Any systematic differences between groups that might account for an observed effect.

- Test scores of students who visit the Psychology tutoring center vs students who do not visit tutoring center.

How to combat this?

Threats to internal validity

- ambiguous temporal precedence
- selection
- **attrition**
- history
- maturation
- regression
- testing
- instrumentation

Even if random assignment is used, participants may drop out of the study, producing unequal groups, a situation that has the same inferential problems as selection.

Threats to internal validity

- ambiguous temporal precedence
- selection
- attrition
- **history**
- maturation
- regression
- testing
- instrumentation

History refers to any event that occurs between the beginning of treatment and the measurement of outcome that might have produced the observed effect.

- A marketing campaign intended to increase beer sales happens to coincide with other events that might have the same effect: a particularly hot period of weather, a long losing streak by the St. Louis Cardinals, etc.

Threats to internal validity

- ambiguous temporal precedence
- selection
- attrition
- history
- **maturation**
- regression
- testing
- instrumentation

Maturation refers to changes in the organism that occur regardless of treatment and that may masquerade as a treatment effect.

- A school-wide educational intervention to increase achievement test scores. The entire school must get the same curriculum, so a control group in the school is not possible.

Threats to internal validity

- ambiguous temporal precedence
- selection
- attrition
- history
- maturation
- **regression**
- testing
- instrumentation

Regression (to the mean) occurs when participants are selected because of their extreme scores and those scores are unreliable. The scores will regress toward the mean at the second assessment

- *Sports Illustrated* cover jinx
- Tall men father not-so-tall sons (Galton)

Threats to internal validity

- ambiguous temporal precedence
- selection
- attrition
- history
- maturation
- regression
- **testing**
- instrumentation

Testing refers to the possible change that may occur just because participants have been previously measured. These are often called practice or fatigue effects.

- Students do better on the first half of test compared to the second.
- Students do better in the second half of the term compared to the first.

Threats to internal validity

- ambiguous temporal precedence
- selection
- attrition
- history
- maturation
- regression
- testing
- **instrumentation**

Change may occur because the measurement changes over time, perhaps becoming more or less reliable.

Instrumentation reflects changes in the measurement; testing reflects changes in the object of measurement.

"When a measure becomes a target, it ceases to be a good measure." (Goodhart)

GOODHART'S LAW

WHEN A MEASURE BECOMES A TARGET,
IT CEASES TO BE A GOOD MEASURE

IF YOU MEASURE PEOPLE ON...	NUMBER OF NAILS MADE	WEIGHT OF NAILS MADE
THEN YOU MIGHT GET	1000'S OF TINY NAILS	A FEW GIANT, HEAVY NAILS

sketchplanations

Limitations

OpenAI acknowledges that ChatGPT "sometimes writes plausible-sounding but incorrect or nonsensical answers".[10] This behavior is common for large language models, and is called "hallucination".[37] The reward model of ChatGPT, designed around human oversight, can be over-optimized and thus hinder performance, in an example of an optimization pathology known as Goodhart's law.[38]

The key point with internal validity is that something else besides the treatment is a plausible alternative explanation for any apparent treatment effect.

Solving threats to internal validity is a **research design problem**, not a statistics problem. Nonetheless, quantitative methods play a key role in making the case for internal validity.

Causal is the "C" word. You better be reallllly sure you mean it before you use it.

Removing the influence of other variables

If the "other variables" can be measured, their influence can be statistically controlled so that the hypothesized relation can be detected more accurately.

However:

Statistical control should best be thought of as a method of last resort, to be used when design controls are not available or have failed.

Construct validity

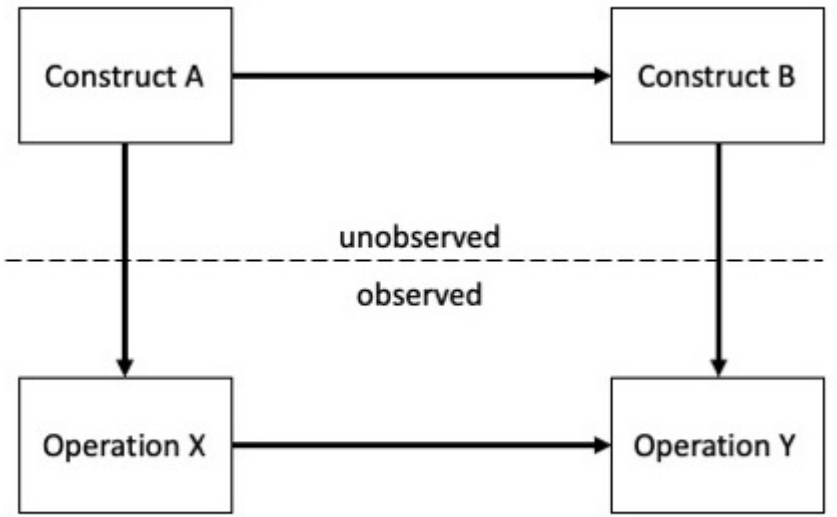
CONSTRUCT VALIDITY IN PSYCHOLOGICAL TESTS[1]

Lee J. Cronbach and Paul E. Meehl

Validation of psychological tests has not yet been adequately conceptualized, as the APA Committee on Test Standards learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term construct validity.[2] This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations

Construct validity

- The validity of the inference that a given operationalization of of a construct does a good job representing the construct.



- Construct validity refers to the correctness of the label that is applied to the operation. It depends on first demonstrating adequate reliability and then is bolstered by demonstrating relations of the target operation to other operations.

How do you establish construct validity?

Show that the variable correlates with all of the things it should correlate with and none of the things it shouldn't

- Convergent validity
- Divergent/discriminant validity

How do we figure out what it should correlate with?

- **theory**

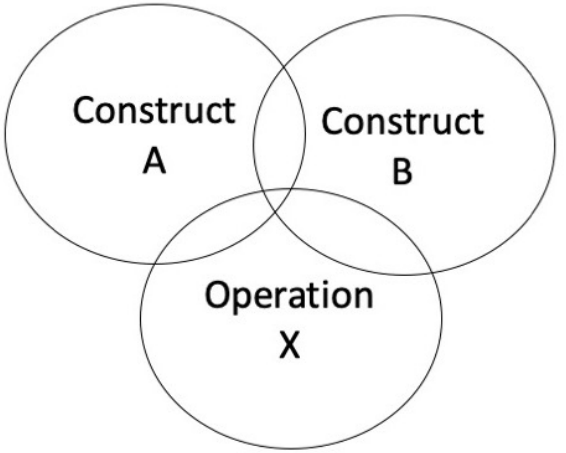
Threats to construct validity

- **inadequate explication of constructs**
- **construct confounding**
- confounding constructs with levels of constructs
- reactive self-report changes
- reactivity to the experimental situation
- experimenter expectancy
- novelty and disruption effects

Construct confounding

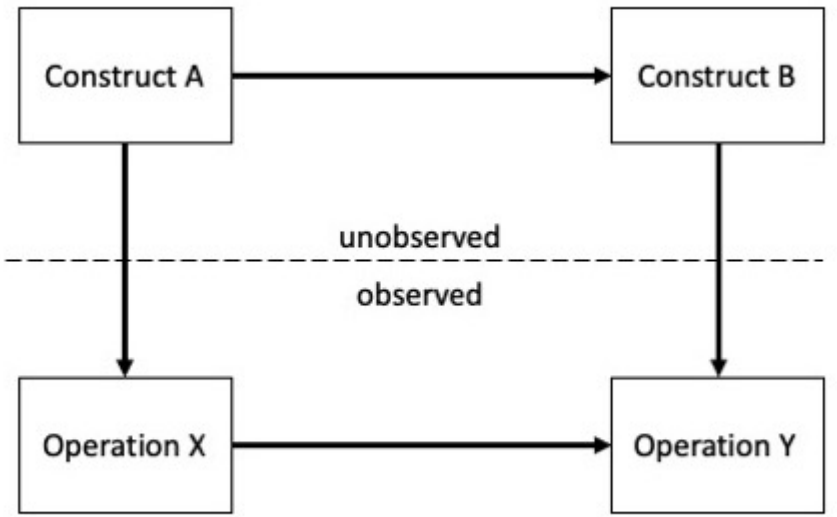
Operations usually tap more than one construct. Failure to recognize the full set of constructs embedded in the operation can lead to incorrect inferences about the constructs that are active.

A self-report of optimism might also reflect self-esteem or positive affect. How can quantitative methods help?



External validity

- **Definition:** The validity of the inference that a causal relation between operations *generalizes* to other units, treatments, observations, or settings.



Threats to external validity

- sampling bias
- experimenter effects
- Hawthorne effect
- testing effects
- situation effects

Threats to external validity

- **sampling bias**
- experimenter effects
- Hawthorne effect
- testing effects
- situation effects

We can't study an entire population, so instead we take samples. If your sample is not representative of the population however, then how on Earth can you generalize to that population?

Threats to external validity

- sampling bias
- **experimenter effects**
- Hawthorne effect
- testing effects
- situation effects

What if I told all my research participants that my job depended on the outcome of the study they are in? The participants might change their behaviors. And their behavior is what we are studying. So anything I find might not be generalizable (even if unintentional). Need to remain neutral!

Threats to external validity

- sampling bias
- experimenter effects
- **Hawthorne effect**
- testing effects
- situation effects

The fact that people know they are being observed is enough to change behavior.

Ex: if someone is in a study about stress, and they know they are in the study, they may make themselves seem more stressed than they actually are. Surveys with scores reflecting higher levels of stress than what they might fill out if they didn't think they were being watched.

Threats to external validity

- sampling bias
- experimenter effects
- Hawthorne effect
- **testing effects**
- situation effects

Testing effects are especially critical in pre/post designs. At the pre-test, they are nervous. But at the post-test, they know what to expect and are less anxious etc. Really problematic when studying something like, say, anxiety.

Threats to external validity

- sampling bias
- experimenter effects
- Hawthorne effect
- testing effects
- **situation effects**

Situation effects can be things like time of day, the setting of the experiment/location etc.

What if you study recall memory. You have all your participants come in before 10am. You find an effect.

Now you repeat the study, but you have all your participants come in after 8pm. You don't find an effect.

Your Ethical Duty as a Scientist

Advanced skill in quantitative methods carries with it the responsibility to use those skills **carefully and ethically**.

- Know the shortcomings of your study
- REPORT the shortcomings of your study
- Let your readers understand the limitations
- Do NOT overstate your findings (or let the press overstate them)

Next time

Exam