Research Methodology for Public Policy

Lecture 5 — Measurement 2: Validity and Reliability

Dr. Colin Kuehl

2026-07-13

Northern Illinois University

SCHOOL OF PUBLIC POLICY
CHIANG MAI UNIVERSITY

Welcome Back & Agenda

Warm - Up 1

Thinking about your preferred question for your final project

On a piece of scratch paper . . .

- What is your dependent variable? How will you operationalize it?
- What is your independent variable? How will you operationalize it?
- What is your hypothesis?

Warm - Up 2

Share your response with a partner (preferably someone new to the project)

Warm - Up 2

Share your proposal with a partner

Partner,

1. Write down their project in terms of the core model.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

2. Quickly sketch a plot of their hypothesized relationship.

Problem Set Review

Problem Set Review

Problem Set Review

- Excellent job on the R coding. Very few errors.
- Those that struggled indicated so on their final question.
- Make sure to fully answer each question

Problem Set Review

Problem Set Review

- In ways your parents would understand,under what conditions can we claim causality.

Problem Set Review

- In ways your parents would understand,under what conditions can we claim causality.
- Imagine you are running a research project on the effect of education on youth political participation. List some of the factors (at least 3) that could lead to an endogeneity problem.

Problem Set Review

- In ways your parents would understand,under what conditions can we claim causality.
- Imagine you are running a research project on the effect of education on youth political participation. List some of the factors (at least 3) that could lead to an endogeneity problem.
- Anything else Dr. Colin should know about you, this class, CMU, life, etc? Feel free to vent

Problem Set #2

Your next problem set is posted and due tomorrow at 8 PM(You need rest for Wednesdays)

Your next problem set is posted and due tomorrow at 8 PM(You need rest for Wednesdays)

- Shorter than problem set #1
- Mostly review, new R skills and concepts from Wednesday and Thursday
- You will have time to work on it this afternoon and Tuesday during practicaum

Course logistics

Course logistics

1. Research Questions due tonight at midnight

- Submit on webpage (just like problem sets, but on final projects page)
- Do for two questions, but it is OK if one is much stronger than the other

1. Problem set due tomorrow at 8 PM

2. Midterm on Thursday

- We will have a review session on Thursday morning
- Midterm will be during afternoon session
- 45 minutes paper exam/45 minutes laptop R
- Will cover material covered through Wednesday

1. Problem Set #3 due Friday night

Quick Recap: Lecture 4

- A good concept must be **concrete** (clearly defined) and must **vary** across the units we study
- Operationalization turns an abstract concept into a concrete, measurable indicator
- Today: even a well-defined, varying concept can still be measured *badly* - today we learn the researcher vocabularly to describe *bad* measures
- Two distinct questions to ask of any measure, every time

Today's Agenda

- Warm-up
- Problem Set Review
- Course Logistics
- Recap & questions from Lecture 4
- Two questions every measure must answer: reliability and validity
- Reliability: is it consistent?
- Types of validity: face, content, criterion, construct
- Convergent and discriminant validity
- Internal vs. external validity, revisited from Lecture 2
- The dartboard: seeing both at once
- Strategies for improving reliability
- A worked example: polling
- Measures of central tendency (from lecture 4)

Two Questions Every Measure Must Answer

Beyond "Does It Vary?"

- Lecture 4 set the bar at *concrete and varying* — necessary, but not sufficient
- A measure can satisfy both of those conditions and still be a bad measure
- Two further questions to ask of any measure, every time:
  1. **Reliability** — if I measured this again, would I get a similar answer?
  2. **Validity** — am I actually measuring what I think I'm measuring?
- A measure can be reliable without being valid. It cannot be called "good" unless it is both.

A Quick Example

- Suppose my measure repeatedly classifies a well-known conservative politician as a liberal
- It might be perfectly **reliable** — it gives the same wrong answer every time
- But it is clearly **not valid**
- Reliability is necessary, but not sufficient, for validity
- Keep this asymmetry in mind: it is the single most-tested idea from today's material

Reliability

What Reliability Means

- **Reliability**: the degree to which a measure is *consistent* in capturing the concept
- Two re-measurements of the same stable thing should give similar results
- Unreliable measures inject **random** error into your analysis — noisy, but not necessarily biased in any one direction
- Random error tends to attenuate (weaken) estimated relationships rather than create fake ones

Evaluating Reliability

- **Test-retest**: measure the same units twice; do the scores correlate?
- **Alternative-form**: use two different versions of an instrument; do they agree?
- **Split-half**: split a multi-item scale in half; do the two halves correlate?
- **Cronbach's alpha**: a single statistic summarizing internal consistency across all items in a scale
- None of these tells you anything about validity — they are purely about consistency

Where Reliability Breaks Down

- Ambiguous question wording — respondents interpret it differently each time
- Coder/rater inconsistency in content analysis or qualitative coding
  - My own work on categorizing tweets has big issues here
- Instrument or survey-mode changes between waves (phone vs. web, different interviewers)
- Mood, fatigue, or context effects that shift answers without the underlying concept changing

Validity: The Big Picture

What Validity Means

- **Validity**: the degree to which a measure captures the *true* value of the underlying concept
  - *Are you measuring what you are trying to measure?*
- There is no single statistical test that "proves" validity — it's a cumulative, theory-driven judgment
- Researchers typically build a *validity argument*: multiple, converging pieces of evidence
  - Use literature review
- We'll walk through four major types: face, content, criterion, and construct validity

A Map of Validity Types

Asking "does this look right?"

- Face validity
- Content validity

Asking "does this behave right?"

- Criterion validity
- Construct validity
  (incl. convergent/discriminant)

- Moving down this list generally means *more* evidence, but *more* work to collect it

Face and Content Validity

Face Validity

- **Face validity**: on its face, does this measure plausibly capture the concept?
- The weakest form of validity evidence — essentially "it looks right to me"
- Useful as a first, cheap sanity check; never sufficient on its own
- Example: a survey item asking "How many times did you vote in the last 5 elections?" has face validity as a measure of political participation

Content Validity

- **Content validity**: does the measure cover the *full conceptual space* of the concept, not just one slice of it?
- Ask: if my concept has multiple dimensions, does my measure (or battery of items) touch all of them?
- Example: "political knowledge" includes knowing current officeholders, basic institutions, *and* current events — a quiz that only asks about officeholders has weak content validity
- Content validity is partly a logical/theoretical judgment, not a statistical one

Criterion Validity

Criterion Validity: Comparing to a Known Standard

- **Criterion validity**: does the measure correlate with an outcome or standard it should predict or match?
- Two common sub-types:
  - **Concurrent validity**: correlates with a criterion measured *at the same time*
  - **Predictive validity**: correlates with a criterion measured *in the future*
- Requires having some trusted external benchmark — not always available

A Worked Example

- New survey-based "civic engagement index" vs. actual voter-file turnout records
- If people who score high on the index also show up in the voter file at high rates, that's **concurrent criterion validity**
- If high scorers today are more likely to vote in *next year's* election, that's **predictive criterion validity**
- The catch: you need an independent, trusted record (the voter file) to validate against in the first place

Construct Validity

What Construct Validity Means

- **Construct validity**: does the measure relate to other variables the way the underlying theory says it should?
- The broadest and most theory-dependent form of validity
- Evaluated by checking a whole *network* of expected relationships, not just one correlation
- Two especially useful sub-checks: convergent and discriminant validity

Convergent Validity

- **Convergent validity**: does this measure correlate with *other* measures of the *same* (or a closely related) concept?
- If you have two different operationalizations of "institutional trust," they should move together
- High convergence is reassuring — it suggests you're not just picking up noise or an idiosyncratic artifact of one instrument

Discriminant Validity

- **Discriminant validity**: does this measure *fail* to correlate with measures of *conceptually distinct* concepts?
- If your "institutional trust" measure correlates just as strongly with "general extroversion," that's a red flag
- A good measure should be distinguishable from things it is *not* supposed to be
- Convergent + discriminant validity together are sometimes called a "multitrait-multimethod" check

Putting Convergent and Discriminant Together

Convergent (want high)

- Trust-in-government Index A
- Trust-in-government Index B
- These *should* correlate strongly

Discriminant (want low)

- Trust-in-government Index A
- General personality/extroversion scale
- These should *not* correlate strongly

- A construct-valid measure looks like its conceptual siblings and unlike its conceptual strangers

Internal and External Validity, Revisited

- Lecture 2 introduced internal and external validity in the context of *causal claims*
- The same pair of ideas applies directly to *measurement*
- **Internal validity** (here): is the causal story about this measure's role, within this study, trustworthy?
- **External validity** (here): does this measure — and the relationship it's embedded in — generalize beyond this study?

In the Words of QRM

"Internal validity basically means we can make a rigorous statement within the context of our study." (QUANT Ch. 2)
"External validity means that we can generalize the results of our study. It asks whether our findings are applicable in other settings." (QUANT Ch. 2)

A Measure Can Have Internal Validity and Still Disappoint You

- Book example: suppose "being well-fed" is operationalized as eating 20 Hostess Twinkies in an hour
- If Twinkie-eaters are more productive than non-eaters, you may have shown a causal effect *of Twinkie-eating* — internal validity is fine
- But if your *theory* is about a balanced, healthy diet, this measure tells you little about that theory
- Lesson: internal validity is about your design; it does not rescue a poorly chosen operationalization

Threats to Internal Validity

QUANT Ch. 2 catalogs the classic threats researchers must rule out:

- History — events between measurements, other than the treatment, that could explain change
- Maturation — units change naturally over time regardless of treatment
- Testing — the act of measuring once changes how units respond the second time
- Instrumentation — the measuring instrument itself changes between waves
- Selection — groups differ systematically before treatment ever happens
- Mortality / attrition — who drops out of the study is not random

Threats to External Validity

- Unrepresentative samples — your sample doesn't look like the population you want to generalize to
- Artificiality of setting — lab conditions don't resemble the real world (QUANT's news-story-and-kitten-video example)
- Reactivity — subjects behave differently *because* they know they're being studied
- Interaction of testing and treatment — a pre-test sensitizes subjects in a way that real-world exposure wouldn't

The Classic Trade-off

- Tightly controlled designs (lab experiments) tend to maximize **internal** validity
- Messier, real-world designs (field studies, large surveys) tend to maximize **external** validity
- "Design includes trade-offs, e.g., internal vs. external validity, and compromises based on time, resources, and opportunities." *(QUANT Ch. 2)*
- There is rarely a design that maximizes both at once — know which one your project needs more

Putting Reliability and Validity Together: The Dartboard

Seeing Both at Once

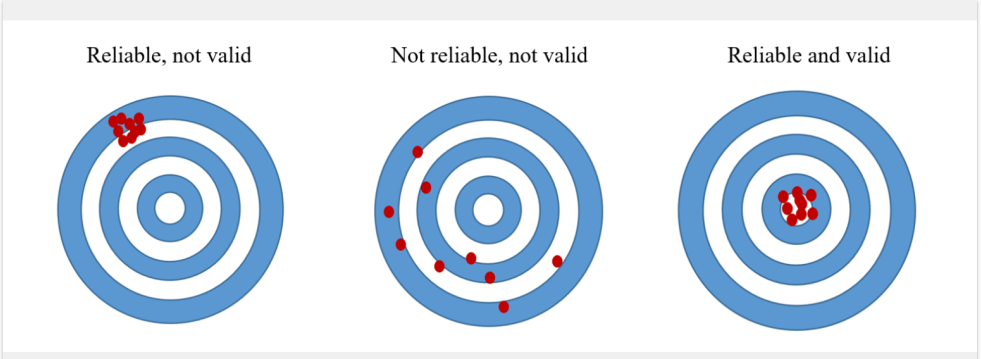


Figure 4.2: Validity and reliability

Reading the Dartboard

- **Reliable, not valid**: tightly clustered throws — but off-target (consistent bias)
- **Not reliable, not valid**: scattered *and* off-target
- **Reliable and valid**: tightly clustered *on* the bullseye, every time
- A good measure is a bullseye, every time — not just close on average
- Notice there is no "valid but not reliable, tightly grouped" panel: if it's valid *and* scattered, the average might still be on target, but no single throw is trustworthy

Strategies for Improving Reliability

Multiple Indicators

- Relying on a single survey item or single coder invites random error
- Use **multiple indicators** of the same concept and combine them
- Random error in one item is more likely to cancel out across several items than to compound
- This is the logic behind nearly every multi-item battery you'll see in survey research

Building an Index

- Combine multiple indicators into a single composite **index** (e.g., simple sum, weighted average, factor score)
- Benefits: smooths out item-level noise, often improves both reliability *and* content validity
- Costs: can obscure which specific item is driving a result; requires the items to plausibly tap the same concept; forces arbitrary decision on weighting of each component
- Always report internal consistency (e.g., Cronbach's alpha) for any index you build

Other Practical Strategies

- Pilot test instruments before fielding them at scale
- Standardize administration — same wording, same mode, same training for coders/interviewers
- Triangulate across data sources (survey + administrative records + observational data)
- None of these *guarantees* validity — but they all reduce avoidable noise and bias

Evaluating Reliability and Validity in Practice

In Your Own Research

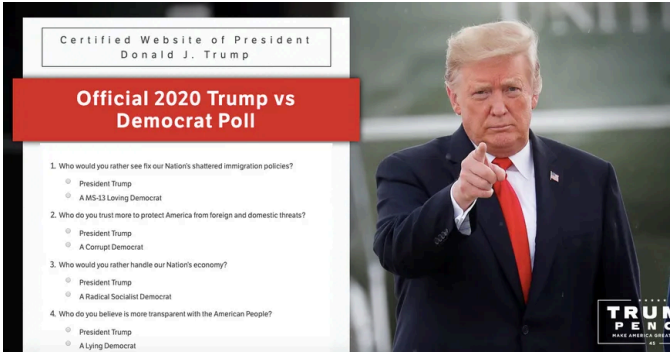
- Before you trust a measure, ask both questions explicitly — don't assume either
- Borrowing a measure from a published, widely-used study buys you some evidence, but doesn't guarantee it transfers to *your* context or population
- Pilot testing, multiple items per concept, and triangulating across data sources all help with both
- No measure is perfect

Two Failure Modes to Watch For

- Treating face validity as if it were sufficient — "it looks right to me" is not evidence
- Reporting reliability statistics (e.g., Cronbach's alpha) as if they were validity evidence — they are not the same thing
- A scale can have excellent internal consistency and still measure the wrong construct entirely

An Example: Polling

Are Election Polls Reliable? Valid?

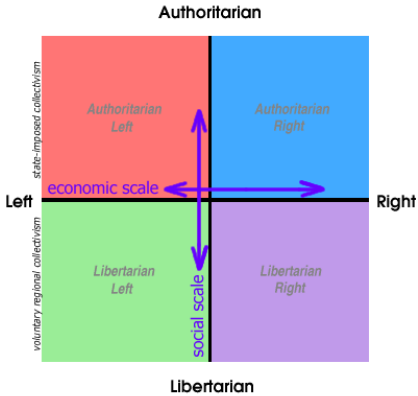


- Reliability question: would a similarly-worded poll, fielded the same week, get a similar topline number?
- Validity question: does "who would you vote for if the election were today" actually predict turnout-weighted behavior on election day?
- Polling misses are often a *validity* problem — likely-voter models, non-response

Diagnosing the Miss

- If five different pollsters, using five different methodologies, all land within a point of each other: that's evidence of **reliability**, not necessarily validity
- If all five then miss the actual result in the *same* direction: that's a shared **validity** problem — likely a flawed likely-voter model or systematic non-response
- This is exactly the "reliable but not valid" panel of the dartboard, applied to a real and consequential measurement problem

Two Axes, One Lesson



- Any time you collapse a multidimensional concept onto a single axis or score, ask both questions again
- Reliability: would this placement be stable if remeasured?

Questions

Questions

- Any questions about relaibility and/or validity?
- Do we need more examples?

Sampling from a Population

Samples and Populations

Population	The overall collection of individuals, beyond just the sample
Sample	The collection of individuals on which statistical analyses are performed, and from which general trends for the *population* are inferred
Individual	Also called an 'object' or 'unit,' a single data point contributing to the sample

- **Population**: the full set of units you ultimately want to learn about
- **Sample**: the subset you actually observe and measure
- **Individual / unit**: a single data point within the sample
- Good sampling — covered in detail later this course — is what lets us generalize from sample back to population

Describing a Single Variable: Central Tendency

Description vs. Inference

- Our usual long-run goal is **inference**: a generalizable relationship, $X \rightarrow Y$
- But **description** matters in its own right: introducing a new dataset, polling, characterizing the state of the world
- "Descriptive statistics only allow researchers to speak about the data at hand" — generalizing beyond it is the separate job of inferential statistics
- Even when inference is the goal, you must first describe your variables accurately — this is where measurement and analysis meet

Measures of Central Tendency

Describing the "middle" of a variable:

- **Mean**: sum of all values divided by the number of values — "the most widely used summary statistic," but sensitive to outliers *(EMPS Ch. 4, paraphrased)*
- **Median**: the value that splits an ordered distribution exactly in half — far less sensitive to outliers
- **Mode**: the most frequently occurring value — usable with any variable type, especially useful for categorical data

```
# we know what this looks like in R
mean(data$income); median(data$income); table(data$region)
```

Why the Mean Can Mislead

- Ten seminar students earn between \$1,000–\$2,600/month; a guest speaker earns \$60,000/month
- The room's **mean** income: \$7,054 — yet all ten students earn under \$3,000
- The room's **median** income: \$2,300 — a far better description of the "typical" person in the room
- One extreme value (an outlier) can drag the mean far from where most of the data actually sits (aka the Elon Musk walks into a bar problem)

Choosing the Right Measure

- Nominal data: mode only — mean and median aren't meaningful
  - Count by group
- Ordinal data: median or mode
- Interval / ratio data: mean, median, *and* mode are all available — choose based on skew and outliers
  - This is why look at histogram, not just summary statistics

Key Terms & Looking Ahead

Key Terms from Today

- Reliability
- Test-retest / split-half
- Cronbach's alpha
- Face validity
- Content validity
- Criterion validity (concurrent / predictive)
- Construct validity
- Convergent validity
- Discriminant validity
- Internal validity (measurement context)
- External validity (measurement context)
- Index / multiple indicators

Before Next Session

Lecture 6: Probability and Confidence Intervals

- Read: QUANT, Ch. 4 (*Probability*)
- Read: QUANT, Ch. 5, secs. 5.1–5.2 (*Inference*: populations, samples, the normal distribution)

Discussion

Take the measure you sketched in Lecture 4. Is it more likely to have a reliability problem, a validity problem, or both? What evidence — face, content, criterion, or construct — would you need to find out?

Questions?

Thank you — see you in the Practicum this afternoon.

See You in the Practicum