Reliability and Validity

Week 7

- Extra credit opportunity: Take the survey (link in Announcement on Canvas)
- Reliability
  - Statistical ways to examine reliability
  - Improving reliability
- Validity
  - Types/Approaches to validity
- Measurement problems related to reliability and validity

Criteria of Measurement Quality

How do we judge success / failure for measuring concepts?

- reliability
- validity

- consistency of measurement
- confidence in measures

Reliability

- refers to **whether a measure produces stable, consistent measurement**
- analogous to precision
- if you measure multiple times, values should not change

Scalar reliability

- reliability of research scales
- most commonly assessed

Scalar Reliability: An Example

Instructions: On the scales below, please indicate your feelings about "**Higher Education**." Numbers "1" and "7" indicate a very strong feeling. Numbers "2" and "6" indicate a strong feeling. Numbers "3" and "5" indicate a fairly week feeling. Number "4" indicates you are undecided or do not understand the adjective pairs themselves. There are no right or wrong answers. *Only circle one number per line.*

1.	Good	1	2	3	4	5	6	7	Bad
2.	Wrong	1	2	3	4	5	6	7	Right
3.	Harmful	1	2	3	4	5	6	7	Beneficia
4.	Fair	1	2	3	4	5	6	7	Unfair
5.	Wise	1	2	3	4	5	6	7	Foolish
6.	Negative	1	2	3	4	5	6	7	Positive

We have to examine how people are answering our measure of attitudes toward higher education. We will talk about how individual respondents can answer scales such as this, but when assessing reliability of a measure, we need to look collectively at the respondents' scores to a scale, not just individual scores. Some people will respond inconsistently, but, hopefully, most will respond consistently. We need some way to evaluate whether these responses are consistent across the sample—we need statistical ways of doing this.

This is what responses should look like if participants have positive (top) or negative (botton) feelings about higher education.

1.	Good	(1)	2	3	4	5	6	7	Bad
2.	Wrong	1	2	3	4	5	6	(7)	Right
3.	Harmful	1	2	3	4	5	6	(7)	Beneficial
4.	Fair	(1)	2	3	4	5	6	7	Unfair
5.	Wise	(1)	2	3	4	5	6	7	Foolish
6.	Negative	1	2	3	4	5	6	(7)	Positive

OR

1.	Good	1	2	3	4	5	(6)	7	Bad
2.	Wrong	1	(2)	3	4	5	6	7	Right
3.	Harmful	1	(2)	3	4	5	6	7	Beneficial
4.	Fair	1	2	3	4	5	(6)	7	Unfair
5.	Wise	1	2	3	4	5	(6)	7	Foolish
6.	Negative	1	(2)	3	4	5	6	7	Positive

On the scale below, please indicate your feelings about **"Higher Education."**

1.	Good	1	2	3	4	(5)	6	7	Bad
2.	Wrong	1	2	3	4	5	6	(7)	Right
3.	Harmful	1	2	3	4	5	6	(7)	Beneficial
4.	Fair	1	2	3	4	5	6	(7)	Unfair
5.	Wise	1	2	3	4	5	6	(7)	Foolish
6.	Negative	1	2	3	4	5	6	(7)	Positive

But sometimes we get responses that look like this. What's the problem here? … However, we don't know whether these responses are "real" or whether they are due to lack of attention, misunderstanding of the questions, etc. We have to have statistical ways to collectively look at how reliably respondents' have answered this question using this scale to find out.

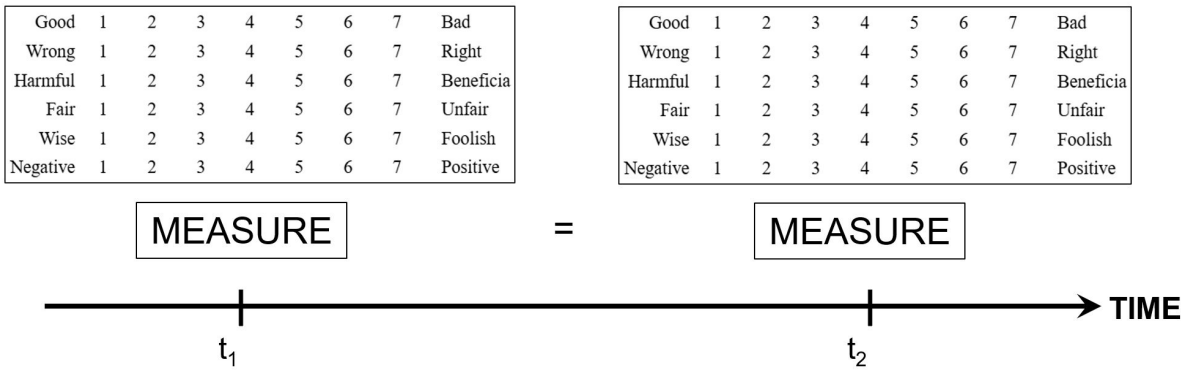
Statistical Ways to Examine Reliability

1) Test-retest reliability
2) Alternate forms reliability
3) Average inter-item reliability
4) Split-half reliability
5) Cronbach's alpha (α)

Test-Retest Reliability

Assessing consistency over time

- use the test-retest method with the same respondents
- compare responses to the same measure at t_1 and t_2



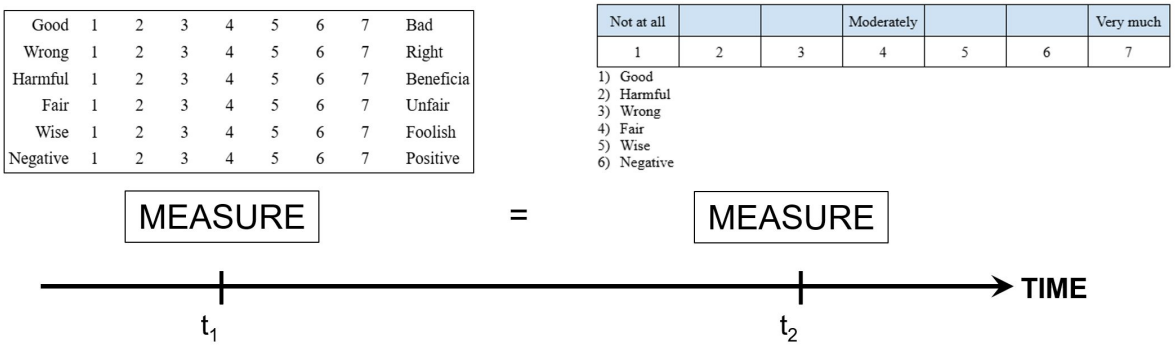
- good reliability coefficient is between ~0.7 and 0.9

Test–retest reliability for attitude measures is estimated by administering the instrument to the same group of people on two occasions, separated by some given amount of time. Good attitude-measuring instruments normally produce test–retest reliability coefficients of 0.70 or above, and many coefficients are 0.90 or above. If a researcher finds a test–retest reliability below 0.70, then her or his participants are not responding to the scale consistently. Conversely, test–retest reliability coefficients of about 0.90 indicate that people are filling out the scale at Time 1 and Time 2 in almost identical fashion.

Alternate Forms Reliability

Also assessing consistency over time

- use the test-retest method with the same respondents
- compare responses to the different measure at t_1 and t_2
- good reliability coefficient will be slightly lower than test-retest method



Assessing inter-item reliability

- Average inter-item reliability
- Split-half reliability
- Cronbach's alpha (α)

I'm going to describe a few ways in which researchers assess reliability. This is to give you an idea of how reliability is assessed.

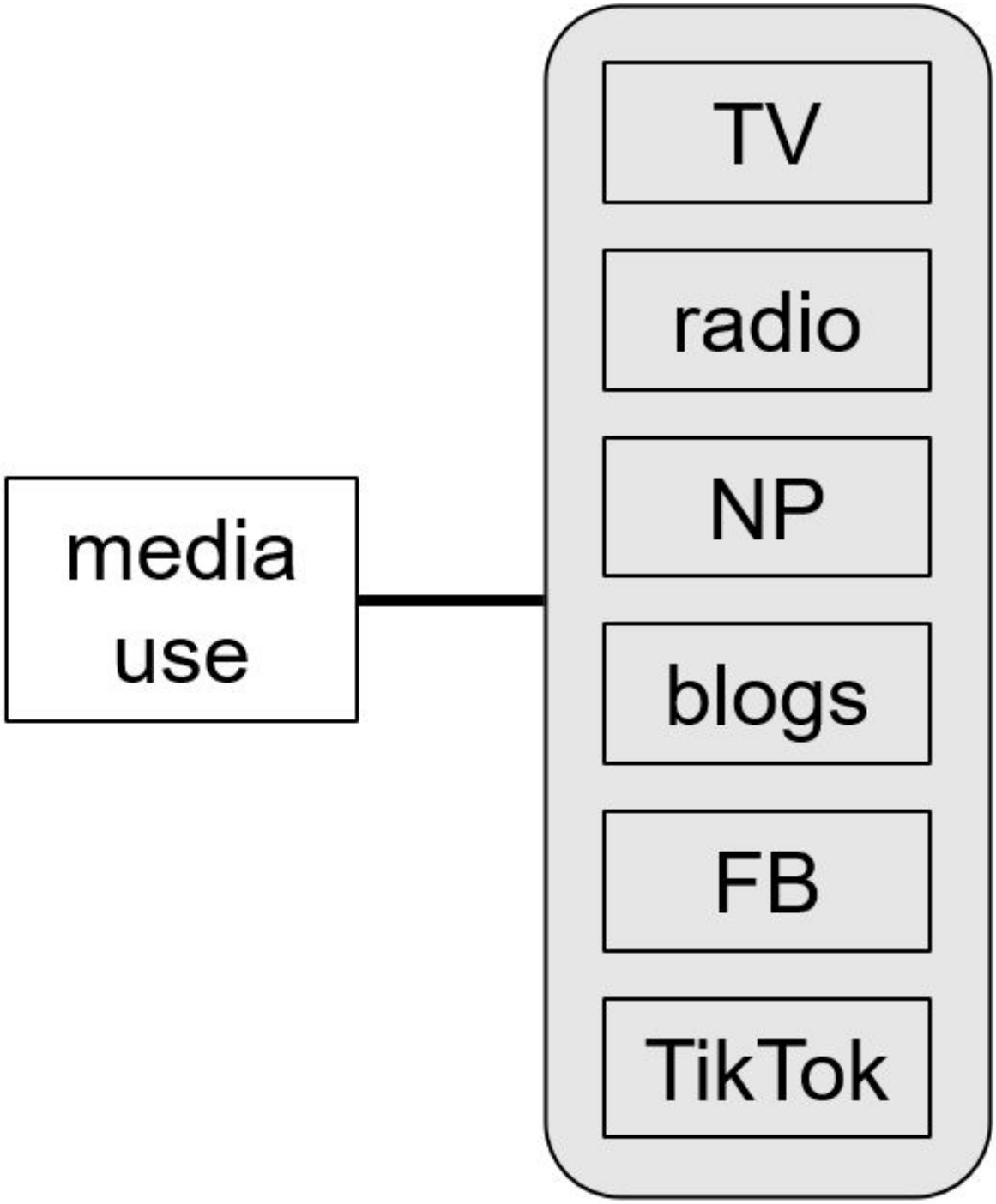
You will not need to know how to do the math. I just want you to understand the logic behind the different ways of calculating reliability.

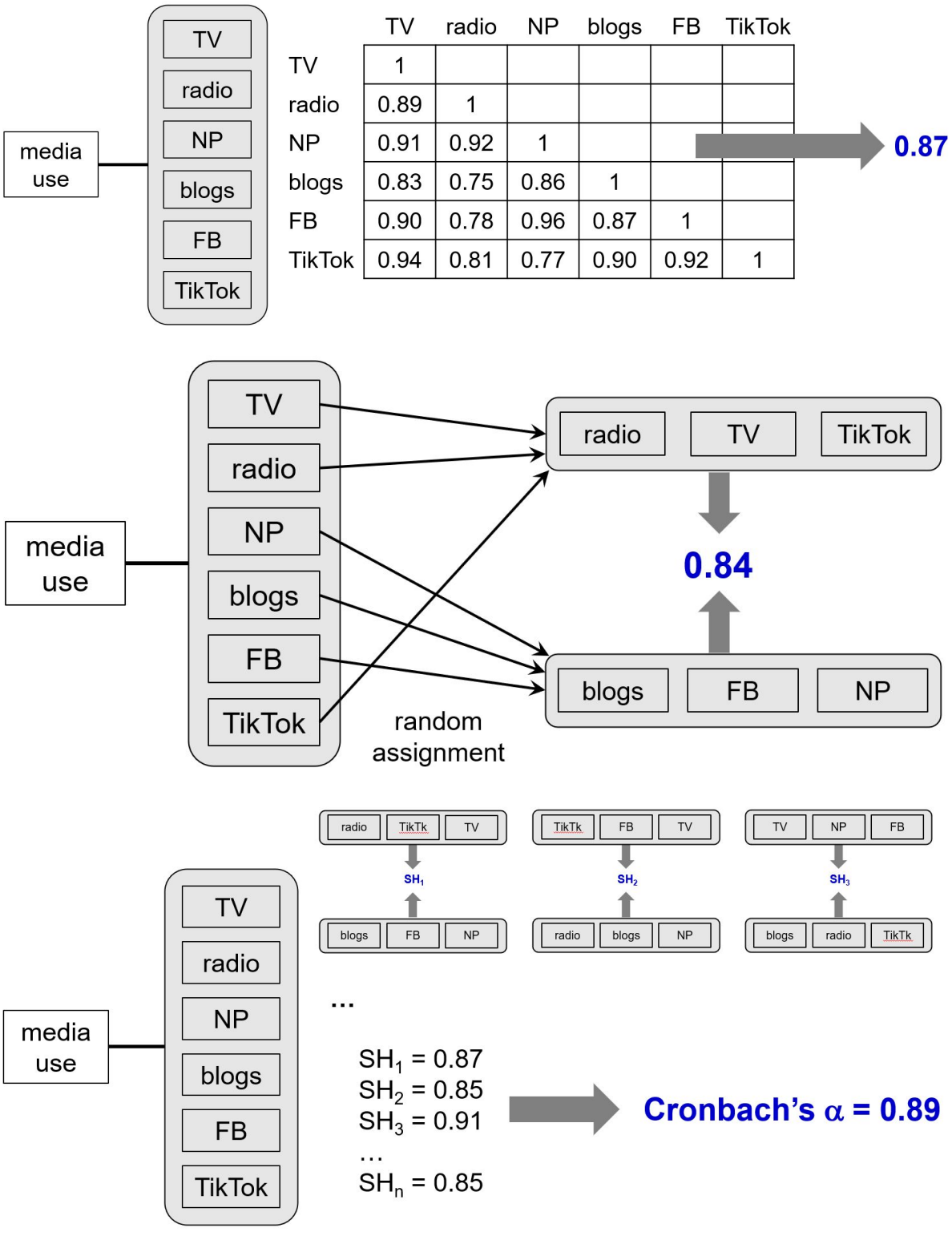
Average Inter-item Reliability

To determine average inter-item correlation, we first assess the correlation between each pair of items. Then we average correlations across those pairs of items. If there is low reliability (i.e., low number), then you know that some of the indicators are not a consistent measure of media use.

Split-half Reliability

Involves splitting half of the items on a test to form two sets of items—randomly assign items to sets/groups. Calculate average inter-item correlation for each set, then the correlation between sets, which is the split-half reliability estimate. If measures are reliable (indicators are consistent and homogeneous), both sets of items should measure concept in the same way, i.e., there should be high correlation between sets.


media use
TV
radio
NP
blogs
FB
TikTok


media use
TV
radio
NP
blogs
FB
TikTok
TV radio NP blogs FB TikTok
TV 1
radio 0.89 1
NP 0.91 0.92 1
blogs 0.83 0.75 0.86 1
FB 0.90 0.78 0.96 0.87 1
TikTok 0.94 0.81 0.77 0.90 0.92 1
0.87
media use
TV
radio
NP
blogs
FB
TikTok
random assignment
radio TV TikTok
0.84
blogs FB NP
media use
TV
radio
NP
blogs
FB
TikTok
radio TikTk TV
SH1
blogs FB NP
TikTk FB TV
SH2
radio blogs NP
TV NP FB
SH3
blogs radio TikTk
...
SH1 = 0.87
SH2 = 0.85
SH3 = 0.91
…
SHn = 0.85
Cronbach's α = 0.89

0.90	Excellent
0.80 - 0.90	Good
0.70 - 0.80	Respectable
0.65 - 0.70	Minimally acceptable
0.60 - 0.65	Undesirable
0.60	Unacceptable

Cronbach's alpha (α)

Mathematically equivalent to all possible split-half estimates of reliability. Cronbach's alpha is the most often used measure of reliability among indicators. To calculate:

- Compute one split-half reliability
- Randomly divide items into another two sets and recompute
- Repeat until all possible split half estimates have been computed

Problem that we must be aware of: More indicators = higher Cronbach's alpha. We want a high Cronbach's alpha (indicates high reliability), but we do not want an artificially inflated one (i.e., one that is higher just because there are more items vs. higher because of more reliable indicators).

Interpreting Reliability

Increasing Reliability

1) *Item construction*

    - concept explication!

2) *Length of instrument*

    - reliability is associated with the number of items you have measuring a specific concept in the measurement instrument
    - more is generally better

3) *Administration of test*

- standard conditions
- clear, consistent instructions

How to Achieve High Reliability and Validity

Concept explication!

- thorough meaning analysis
- good conceptual and operational definitions

Remember that…

- conceptual definitions = meaning of concepts
- operational definitions = how concepts are measured

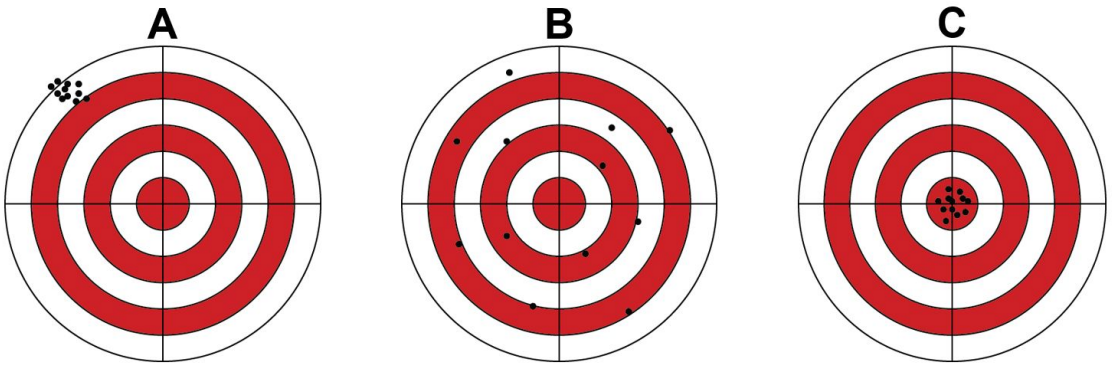
Week 7

- Extra credit opportunity: Take the survey (link in Announcement on Canvas)
- Reliability
  - Statistical ways to examine reliability
  - Improving reliability
- Validity
  - Types/Approaches to validity
- Measurement problems related to reliability and validity

Validity

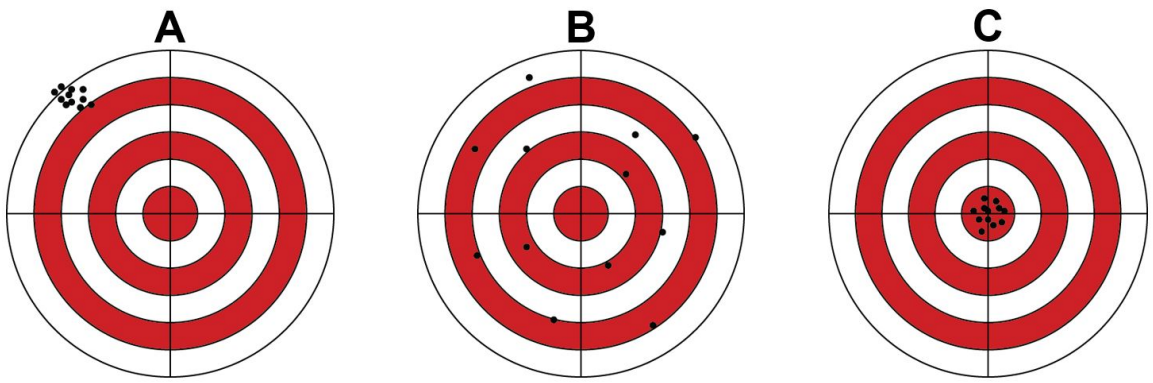
- degree to which the instrument measures what it is intended to measure
- is the instrument measuring what we want it to?
- analogous to accuracy
- also extends to...
  - precision in design (**internal validity**)
  - ability to generalize (**external validity**)

An Analogy: Reliable? Valid?

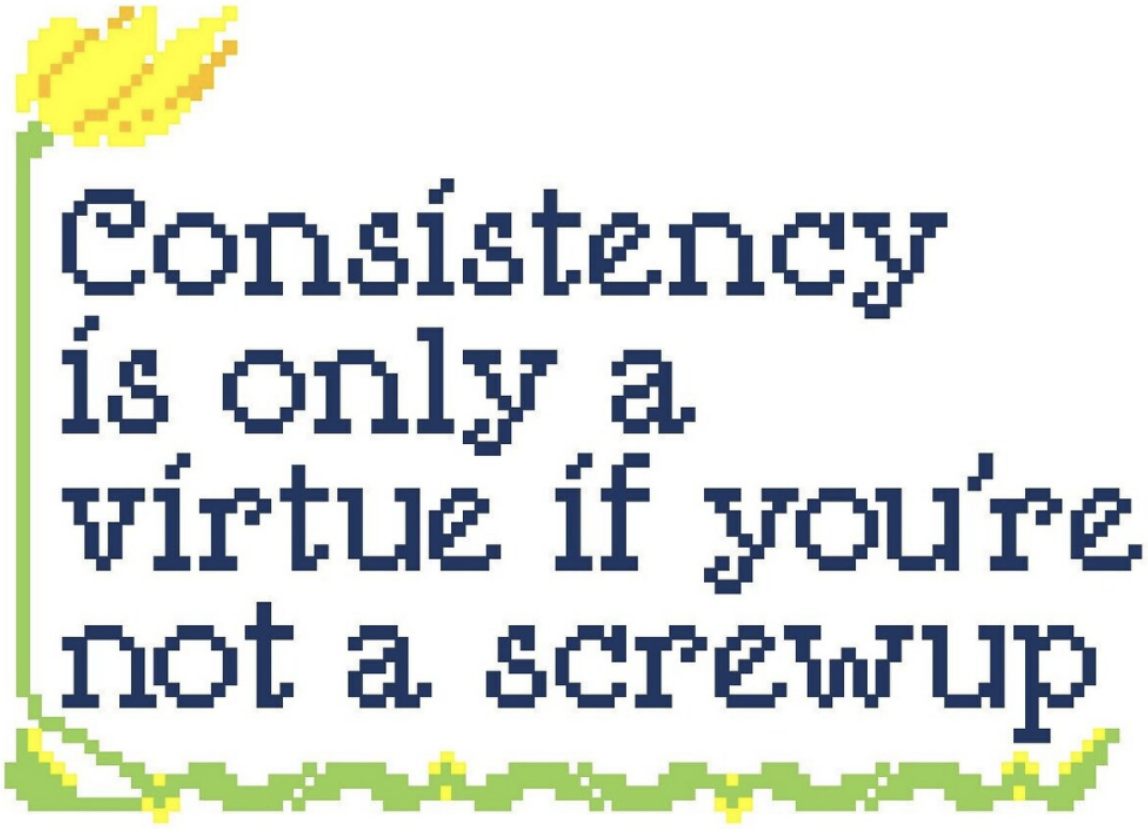


An Analogy: Reliable? Valid?

Bathroom scale example: If the scale is reliable it tells you the same weight every time you step on it as long as your weight has not actually changed. However, if the scale is not working properly, this number may not be your actual weight. If so, this is an example of a scale that is reliable, or consistent, but not valid. For the scale to be valid and reliable, not only does it need to tell you the same weight every time you step on the scale, but it also has to measure your actual weight.

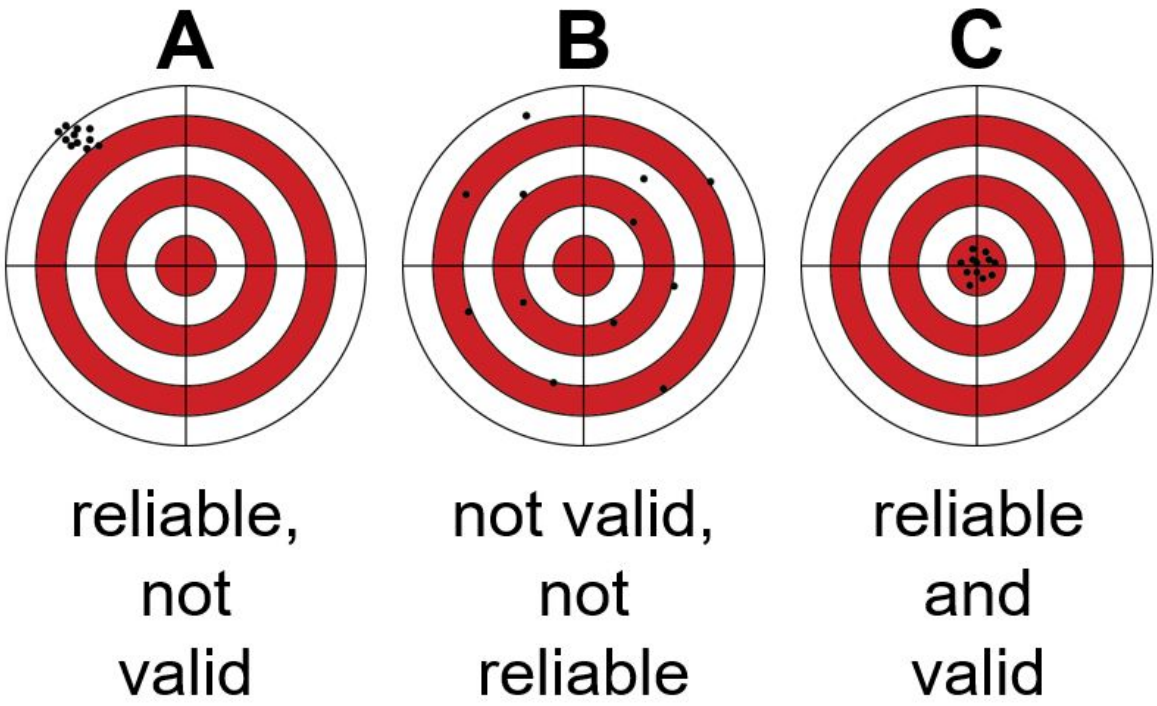


reliable, not valid	not valid, not reliable	reliable and valid
precise but not accurate	not accurate, not precise	precise and accurate



Does One Rely on the Other?

Necessary or Sufficient Condition?



Reliability is a **necessary condition** for validity.

Reliability is not a **sufficient condition** for validity.

Necessary condition –> If it is not reliable, it cannot be valid.

That being said… Sufficient condition –> Even though a measure is reliable, it does not mean it is valid.

Just because we have reliability, doesn't mean we will get an adequate validity.

Validity

- degree to which the instrument measures what it is intended to measure
- is the instrument measuring what we want it to?
- analogous to accuracy
- also extends to…
  - precision in design (**internal validity**)
  - ability to generalize (**external validity**)

Types of External Validity

Before data collection

1) Face or content validity

After data collection

1) Criterion validity
2) Construct validity

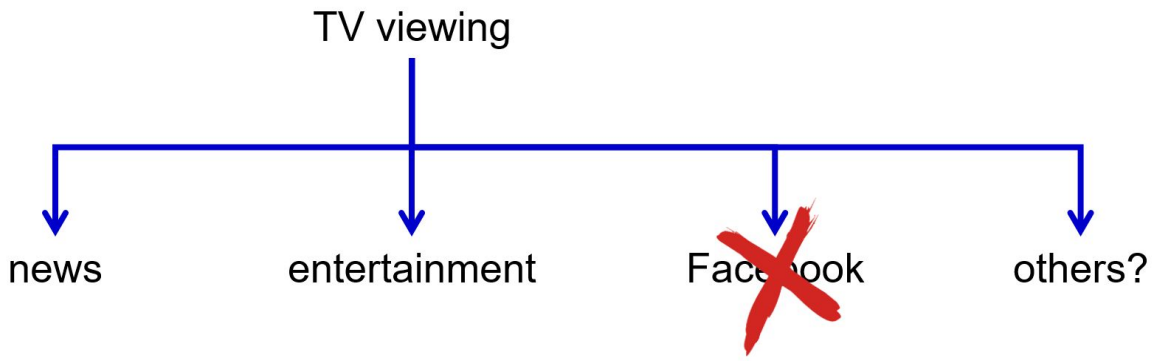
---

1. Face or Content Validity

"On its face"

- subjective validity judgment
- reasonable measure of the concept?
- exhaustive?

Face validity is a type of validity that is logical and intuitive. It is our subjective judgment of whether a measure is valid "on its face." The question we are asking ourselves here is: Do our measures make sense for the variable we wish to measure? Face validity is the quality of an indicator that makes it seem a reasonable measure of some variable/concept. This quote, "If it looks like a duck…" is probably one of the simplest manifestations of face validity. Let's say, for example, that we wish to measure the concept TV viewing. After concept explication, we come up with 3 indicators—exposure to news, entertainment, and Twitter. Now, we should ask ourselves whether these indicators are reasonable measures of the concept of TV viewing. Obviously, exposure to Twitter has little to do with this concept and does not pass our test of face validity.



---

2. Criterion Validity

Extent to which a new measure can accurately predict a well-accepted (external) criterion

Does measure relate to some external criterion?

- empirical evidence to test validity
- e.g., student success in college and SAT/ACT scores

3 sub-types

A) predictive
B) concurrent
C) retrospective

Few attitude-measuring instruments are actually submitted to the test of predictive validity. This is because, in most cases, we are not interested in predicting a specific behavior. If the scores are highly correlated, we have an indication that one is, to the degree of the correlation, as valid as the other.

Read about 3 sub-types in your textbook.

---

3. Construct Validity

All constructs are a result of theoretical development

- impacts inferences
- validity test of the theoretical construct itself

3 common ways of measuring construct validity

A) relying on theory
B) measuring known groups
C) factorial validity

The last type of external validity we will learn about is construct validity. Construct validity is the degree to which a measure relates to other measures as expected based on theory. Important bit: Based on theoretical relationships among variables. We often assess construct validity when we think about how valid survey questions are, i.e., does this question capture what I am trying to measure? To assess construct validity, we use theoretical links, convergent validity, and discriminant validity. Note that convergent and discriminant validity are also considered sub-types of construct validity and these work together. Neither one alone is sufficient to say that you have construct validity. If you can show that you have evidence for both convergent and discriminant validity, then you can say you have evidence for construct validity.

3a. Relying on theory

Theory: People who prefer sweet over tart flavors should show attitudinal preference for sweet fruit.

Hypothesis: People who prefer honey crisp over Granny Smith apples should score higher on an instrument that measures preference for sweet fruit.

Experiment: Measure preference for sweet fruit among 2 groups of people.

- Group A = people who prefer honey crisp apples
- Group B = people who prefer Granny Smith apples
- Instrument: "I prefer sweet apples."
  - "Strongly disagree" (1) to "Strongly agree" (7)

Scores should be higher among respondents in Group A.

3b. Measuring known groups

Less confounding with theory

Example: Developing a measure of `racial tolerance`

- National Association for the Advancement of Colored People (NAACP)

- Ku Klux Klan (KKK)

If measure of `racial intolerance` is reasonably valid, what should the scores look like?

Should observe large difference in scores between respondents associated with NAACP vs. those associated with KKK

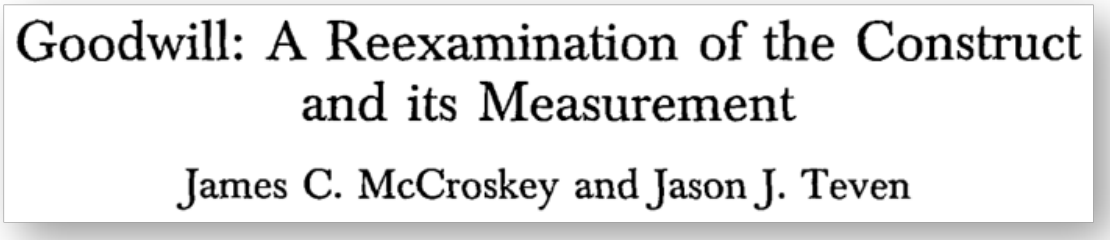
Another approach to construct validity that is less susceptible to the confounding with theory is the measurement of known groups.

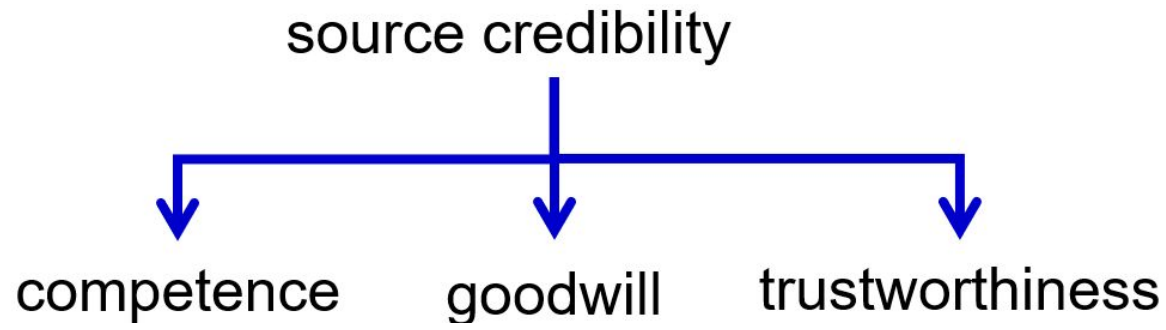
If our measure is reasonably valid, we should observe large differences in the scores between these two groups, unless the former group was much more tolerant or the latter group much less tolerant than is customarily presumed.

3c. Factorial validity

This approach is based on a statistical technique called factor analysis

- tells us how many groups of items (factors) there are in an instrument





In Ch. 7

A third approach to construct validity that deserves a category to itself is factorial validity, which is based on the statistical technique known as factor analysis. Factor analysis is a highly sophisticated statistical technique for examining a series of items to see which items correlate well with one another but are not highly correlated with other items or groups of items—it tells us how many groups of items, or factors, there are in our instrument.

This approach to validity was employed by McCroskey (1966) in the development of measuring instruments to tap source credibility (as discussed in Chapter 7 in the section called "One Measure, Multiple Factors."). At the outset, McCroskey presumed that this attitude was like other attitudes—that is, unidimensional. He therefore developed a set of items to measure this attitude. However, when he subjected these items to factor analysis, he found that there were two dimensions to this attitude rather than one. More simply put, he thought he was measuring "apples," but he found he was measuring "Granny Smith apples" and "Red Delicious apples." In essence, there were two parts (competence and trustworthiness) that created participants' perceptions of credibility, just like Granny Smith and Red Delicious are two different types of apples. As a result, he was forced to develop two separate measures in order to measure these two new constructs.

The only serious problem with the approach to validity through factor analysis is that it is a strictly negative approach. That is, factor analysis can tell someone that she is not measuring her construct, but cannot tell her that she is at least independent of other considerations.

Below are several oppositely worded adjective pairs that represent how you may feel about the U.S. President. Circle the number between the adjectives which best represents your feelings (Fig 7.3 in textbook).

Use these items to measure their hypothesized three dimensions of source credibility (goodwill, trustworthiness, and competence), then use factor analysis to test this hypothesis.

1.	Unintelligent	1	2	3	4	5	6	7	Intelligent
2.	Untrained	1	2	3	4	5	6	7	Trained
3.	Doesn't care about me	1	2	3	4	5	6	7	Cares about me
4.	Dishonest	1	2	3	4	5	6	7	Honest
5.	Doesn't have my interests at heart	1	2	3	4	5	6	7	Has my interests at heart
6.	Untrustworthy	1	2	3	4	5	6	7	Trustworthy
7.	Inexpert	1	2	3	4	5	6	7	Expert
8.	Self-centered	1	2	3	4	5	6	7	Not self-centered
9.	Not concerned with me	1	2	3	4	5	6	7	Concerned with me
10.	Dishonorable	1	2	3	4	5	6	7	Honorable
11.	Uninformed	1	2	3	4	5	6	7	Informed
12.	Immoral	1	2	3	4	5	6	7	Moral
13.	Incompetent	1	2	3	4	5	6	7	Competent
14.	Unethical	1	2	3	4	5	6	7	Ethical
15.	Insensitive	1	2	3	4	5	6	7	Sensitive
16.	Stupid	1	2	3	4	5	6	7	Bright
17.	Phony	1	2	3	4	5	6	7	Genuine
18.	Not understanding	1	2	3	4	5	6	7	Understanding

Instructions: Please indicate your impression of the person noted below by circling the appropriate number between the pairs of adjectives below. The closer the number is to an adjective, the more certain you are of your evaluation.

Competence

Intelligent 1 2 3 4 5 6 7 Unintelligent
Untrained 1 2 3 4 5 6 7 Trained
Inexpert 1 2 3 4 5 6 7 Expert
Informed 1 2 3 4 5 6 7 Uninformed
Incompetent 1 2 3 4 5 6 7 Competent
Bright 1 2 3 4 5 6 7 Stupid

Goodwill

Cares about me 1 2 3 4 5 6 7 Doesn't care about me
Has my interests at heart 1 2 3 4 5 6 7 Doesn't have my interests at heart
Self-centered 1 2 3 4 5 6 7 Not self-centered
Concerned with me 1 2 3 4 5 6 7 Unconcerned with me
Insensitive 1 2 3 4 5 6 7 Sensitive
Not understanding 1 2 3 4 5 6 7 Understanding

Trustworthiness

Honest 1 2 3 4 5 6 7 Dishonest
Untrustworthy 1 2 3 4 5 6 7 Trustworthy
Honorable 1 2 3 4 5 6 7 Dishonorable
Moral 1 2 3 4 5 6 7 Immoral
Unethical 1 2 3 4 5 6 7 Ethical
Phoney 1 2 3 4 5 6 7 Genuine

Threats to Validity

Poor concept explication

Social threats to validity

Social Threats to Validity

Hypothesis guessing

- R tries to guess what you are measuring and responds accordingly

Evaluation apprehension

- R has anxiety at being evaluated

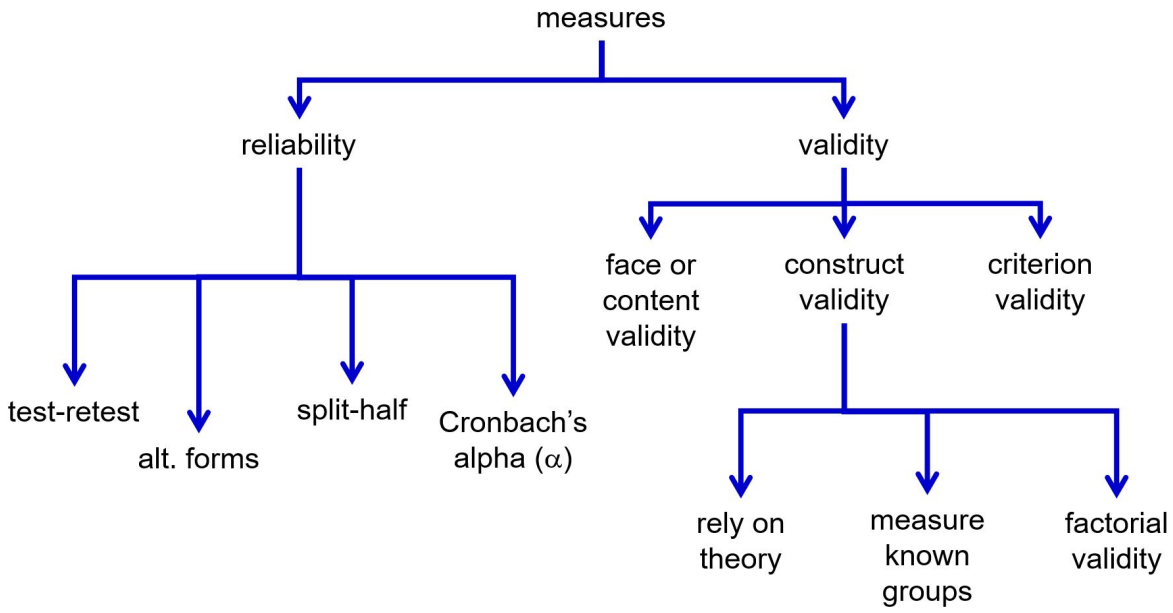
Experimenter expectancies

- researcher unknowingly encourages R to respond in certain way

Social desirability bias

- R responds in a way they think they should
- project favorable image of themselves
- "How often do you drink alcohol?"

Reliability & Validity: Summary



Problems with Measurement

Faking responses

- deliberate attempt to alter results
- acquiescence: cooperating with the researcher
- social desirability: responses that are acceptable
- screw-you effect: opposite of acquiescence

Poor measurement items

- **solution:** identify them, throw them out, develop better items