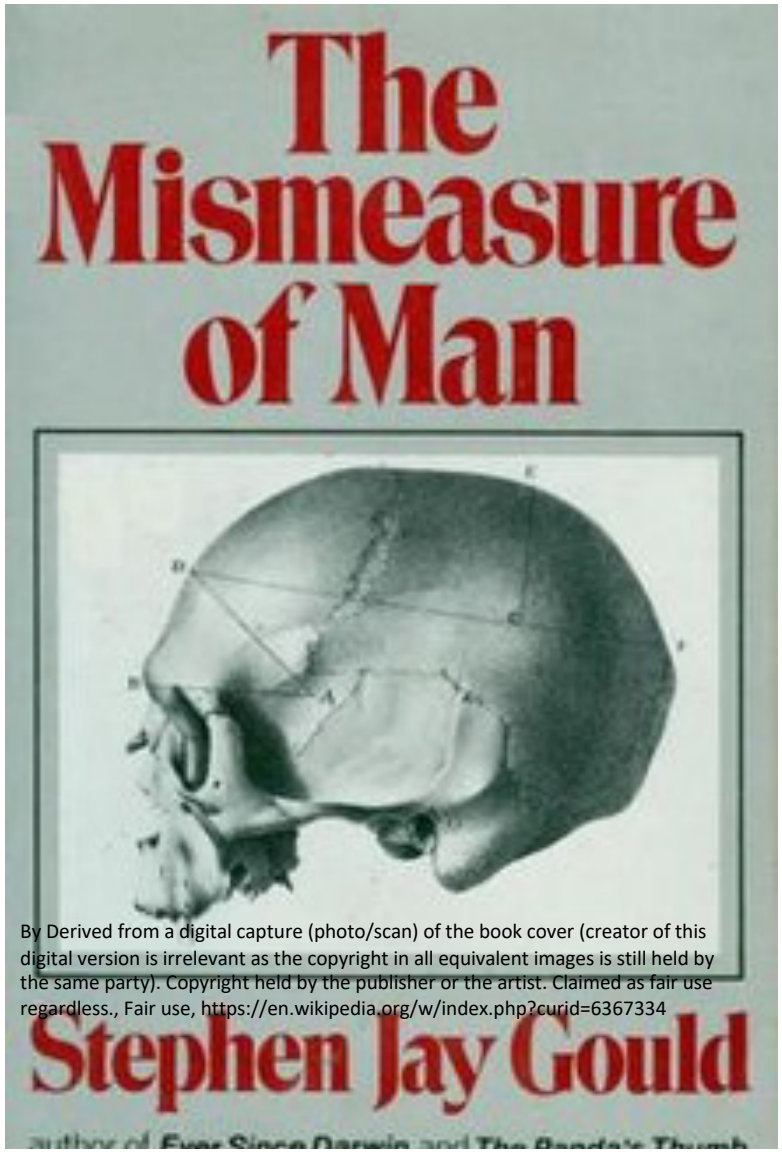
Lecture IV: Measurement

EMSE 6577: Data-Driven Policy for Analytics

David A. Broniatowski

Agenda

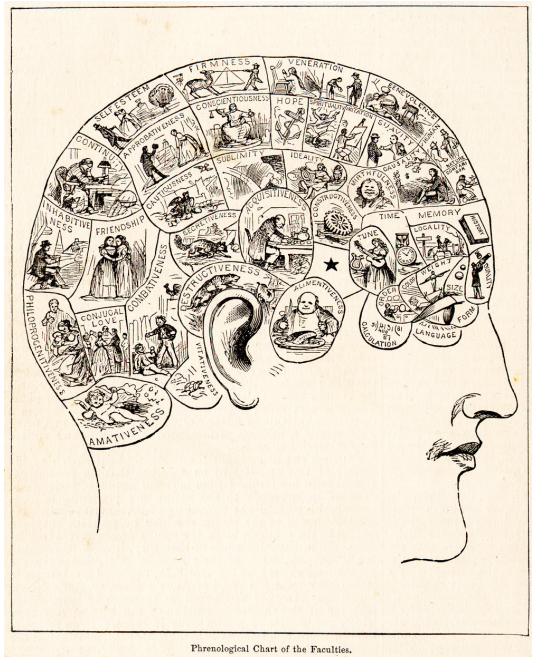
- Construct Validity: What are you measuring really?
  - Definition of constructs and construct validity
- Convergent and Discriminant Validity
- In-class exercise: Measuring “comprehensibility” of Tweets
- Threats to construct validity
- The Multi-Trait Multi Method Matrix
- Reliability
  - Relation to bias-variance tradeoff
- Levels of Measurement and Scaling
- Unobtrusive Measures and Ethics



By Derived from a digital capture (photo/scan) of the book cover (creator of this digital version is irrelevant as the copyright in all equivalent images is still held by the same party). Copyright held by the publisher or the artist. Claimed as fair use regardless., Fair use, https://en.wikipedia.org/w/index.php?curid=6367334

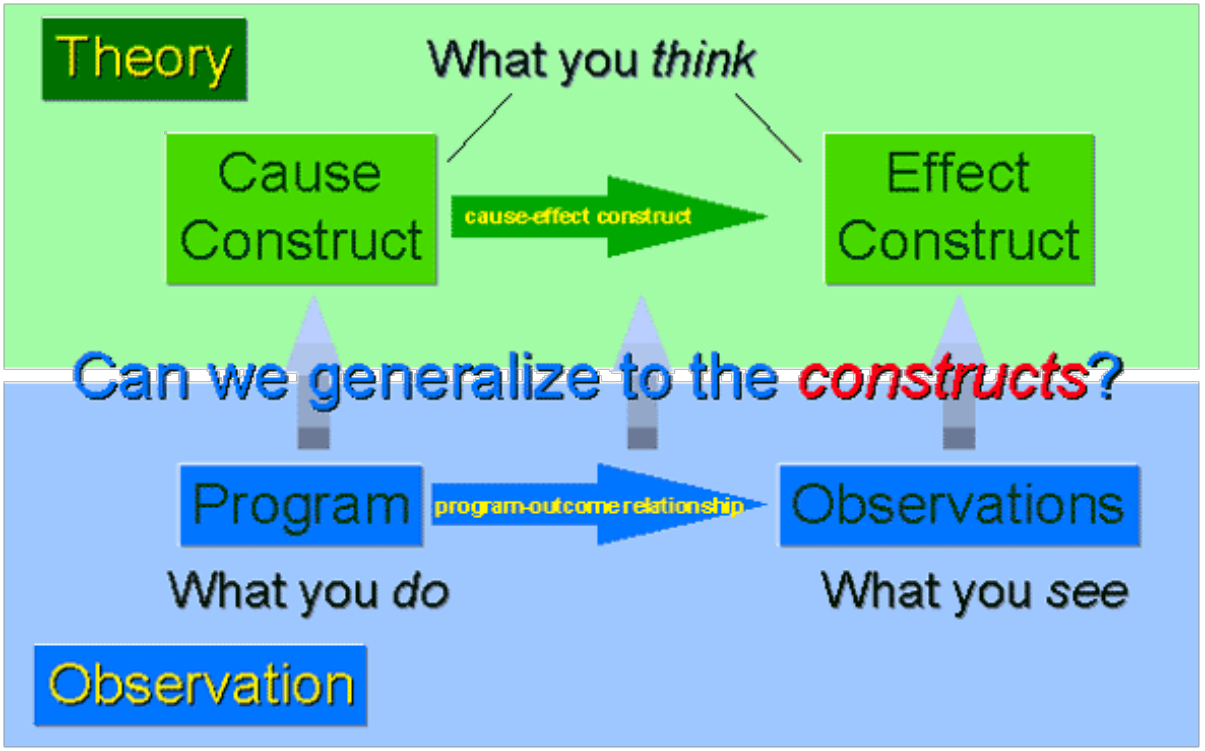
Construct Validity

- How well does your measure represent your concept/theory?
- The “myth of metrics”: Data are objective because they are based on math
- Math is not biased, and numbers don’t lie
- But are we measuring the right things?



By Not credited - From People's Cyclopedia of Universal Knowledge (1883)Transferred from en.wikipedia Original uploader was Whbonney at en.wikipedia, Public Domain, https://commons.wikimedia.org/w/index.php?curid=6693422

What is a construct?



http://www.socialresearchmethods.net/kb/Assets/images/conidea1.gif

What is a construct?

- A “construct” is a theoretical abstraction. You can never directly observe constructs. Instead, constructs have to be “operationalized” into something we can measure.
- The extent to which our measures correspond to our constructs is construct validity – it’s a match between observation and theory

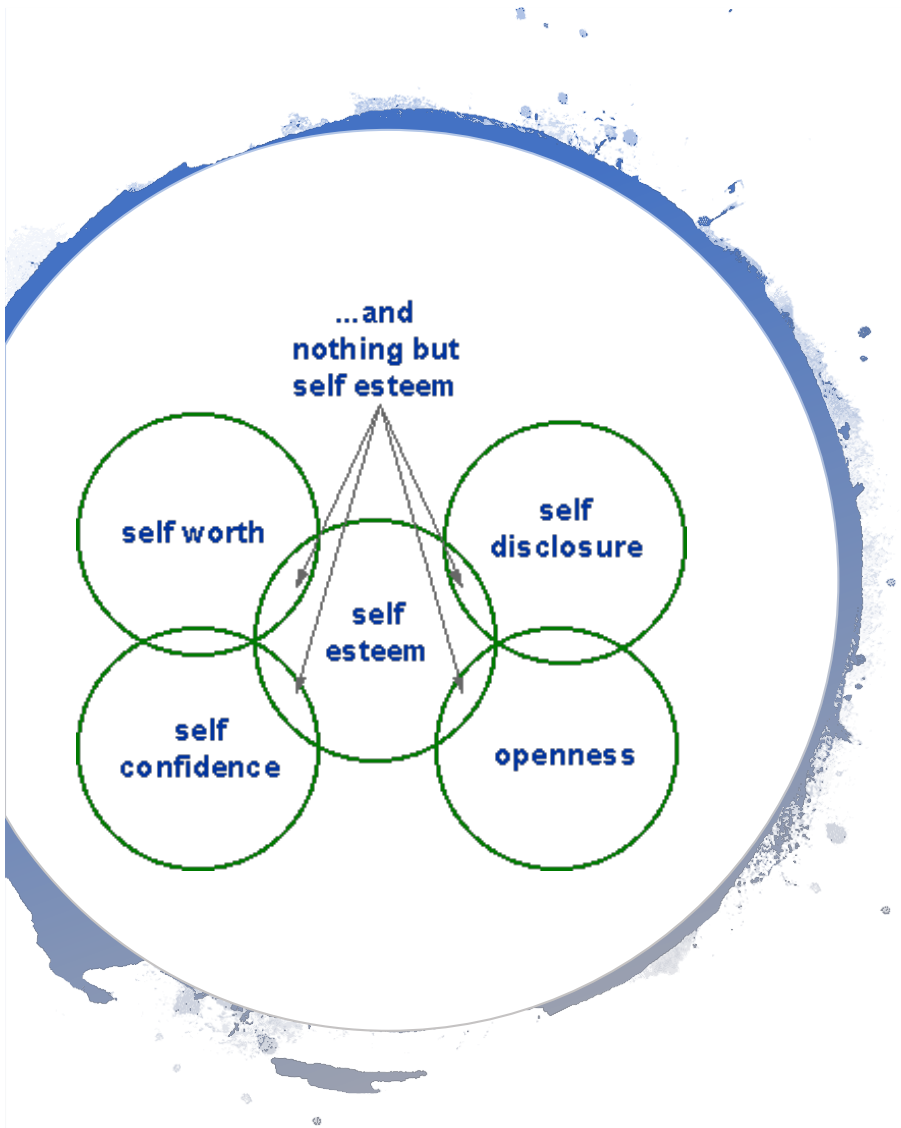
Constructs	Operationalizations
Opioid Use	**Concentration** of opioids in wastewater
Experts’ meaningfulness for system	**Average rating** of Likert scale questions
Quality of restaurant	**Average number** of stars in online review
Team quality	**Average** number of RBI
Suitability for tidal power	**Average** wave height?

Some ways of defining construct validity

- There's a big problem: constructs are in our heads, so, by definition, there is no "ground truth"
- How do we know if we're measuring the right thing?
- Not ideal
  - Face Validity: Does it pass the "giggle test"?
  - Content Validity: Define a checklist for what a good measure looks like, and adopt a measure that does
  - What are some pros/cons of these?

Some better ways of defining construct validity

- There's a big problem: constructs are in our heads, so, by definition, there is no "ground truth"
- How do we know if we're measuring the right thing?
- Better
  - Predictive Validity: Does it predict what it should be able to predict (Twitter predict the flu)?
  - Concurrent Validity: Does it help us to differentiate between two things that are different (influenza awareness vs. influenza infection)?
  - Convergent Validity: Is it similar to other things that measure the same thing? (ILI vs. Twitter mentions)
  - Discriminant Validity: Is it different from other things that measure different things? (Twitter mentions vs. Zika)



Two approaches

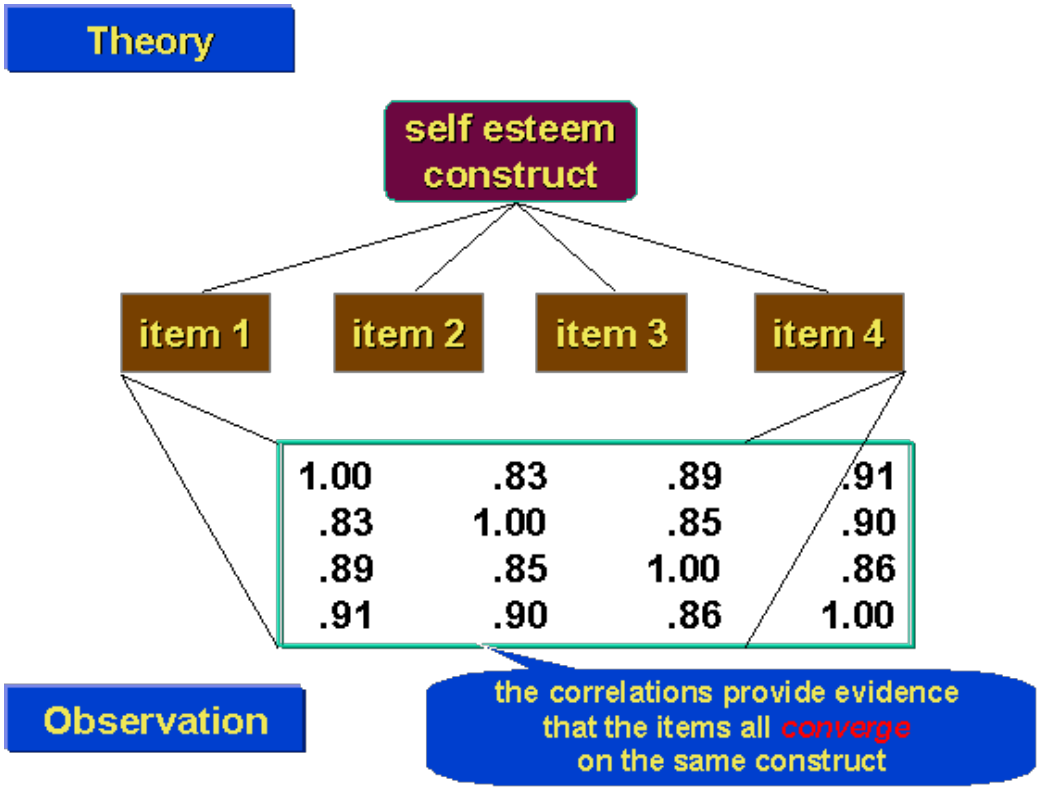
- "Definitionalist": Define the construct so precisely that it is essentially the measure. Using this approach, we can never talk about abstract concepts like "stock value" or "athleticism"– we can only talk about "profit" or "racing speed".
  - Remember positivism?
- "Relationalist": Define similar concepts to the one you want to measure, and see how they converge or diverge

Convergent and Discriminant Validity

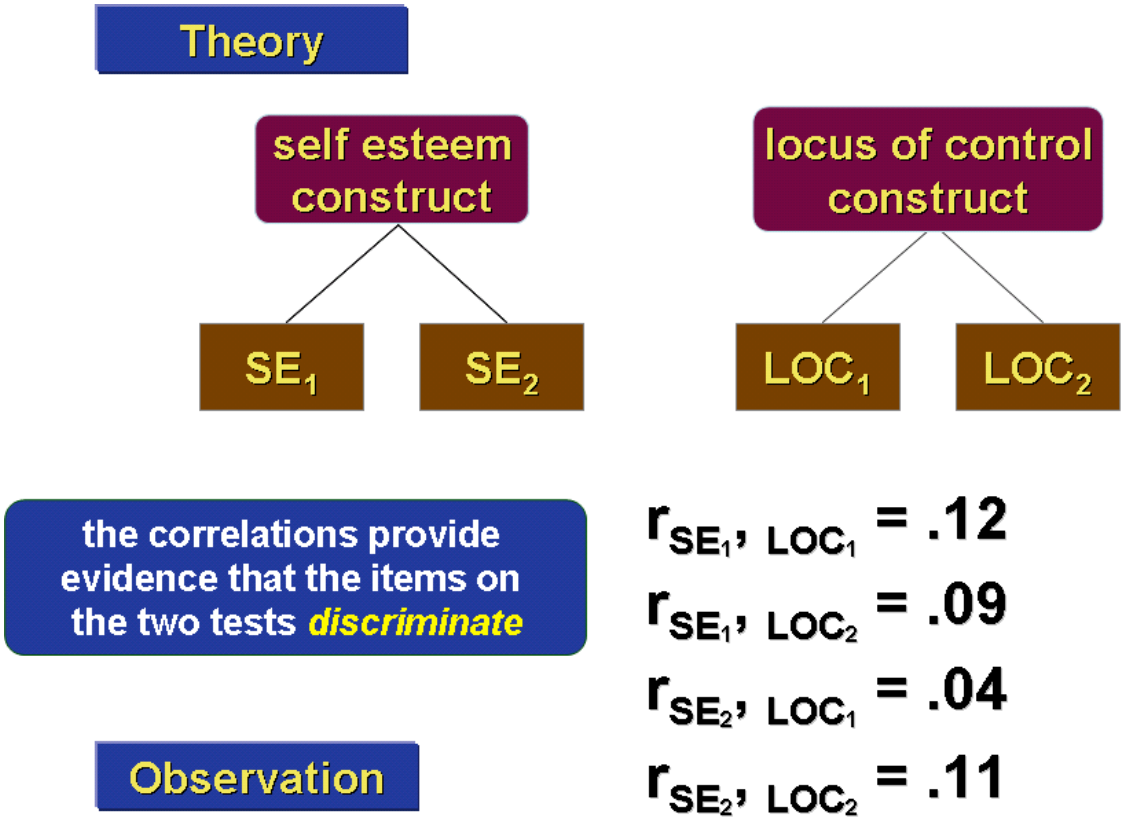
- Example: Can we directly measure the mass of an object?
  - What does a scale measure?
  - Massmeter, a device for measuring the inertial mass of an astronaut in weightlessness. The mass is calculated via the oscillation period for a spring with the astronaut attached (Tsiolkovsky State Museum of the History of Cosmonautics)

Convergent Validity

- "Measures of constructs that theoretically should be related to each other are, in fact, observed to be related to each other (that is, you should be able to show a correspondence or convergence between similar constructs)"
- Things that are similar ARE correlated



http://www.socialresearchmethods.net/kb/convdisc.php



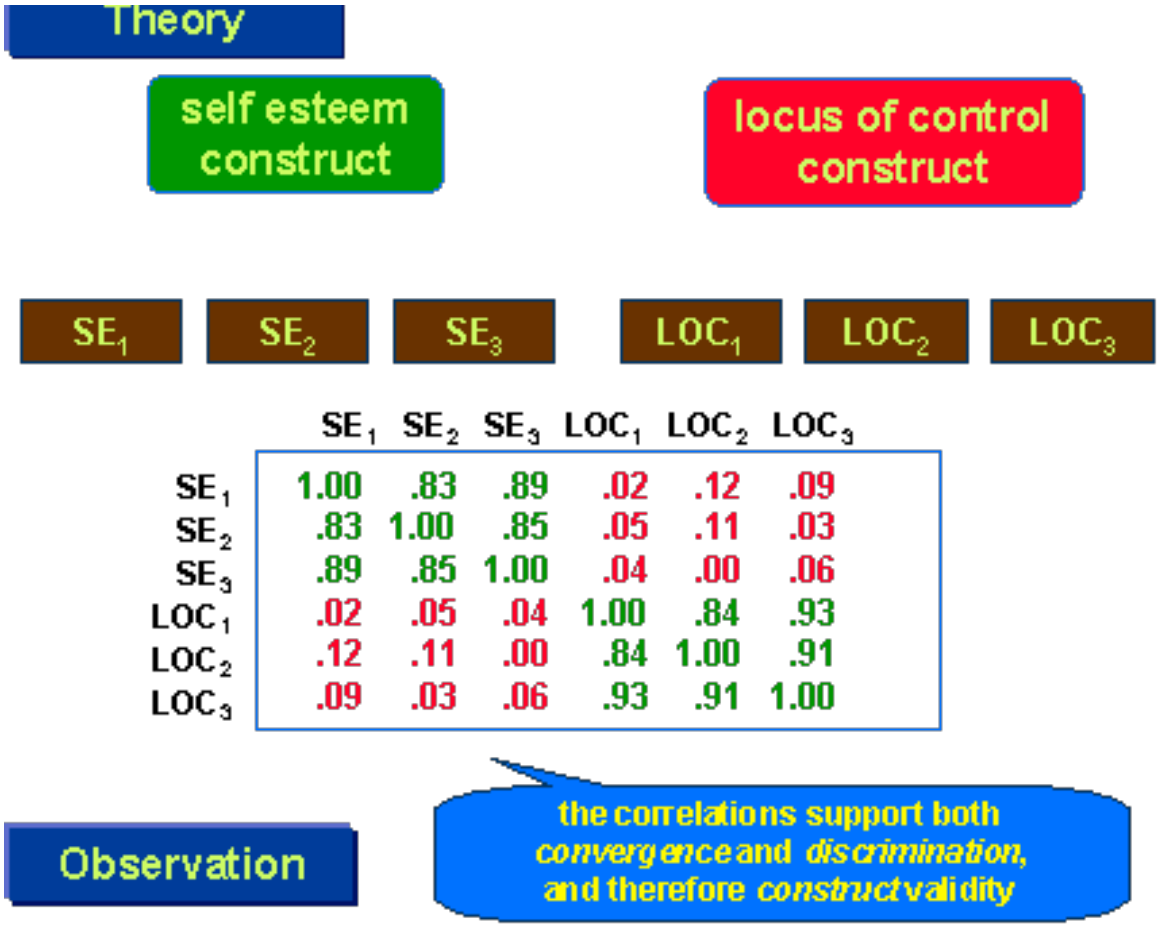
http://www.socialresearchmethods.net/kb/convdisc.php

Discriminant Validity

- "Measures of constructs that theoretically should not be related to each other are, in fact, observed to not be related to each other (that is, you should be able to discriminate between dissimilar constructs)"
- This that are different are NOT correlated

Convergent-Discriminant Validity

- The overall pattern of all of the measures shows two distinct constructs
- How do we build a measure of the underlying construct?
- Hint: It's an unsupervised technique that you saw in the machine learning class.



	SE_1	SE_2	SE_3	LOC_1	LOC_2	LOC_3
SE_1	1.00	.83	.89	.02	.12	.09
SE_2	.83	1.00	.85	.05	.11	.03
SE_3	.89	.85	1.00	.04	.00	.06
LOC_1	.02	.05	.04	1.00	.84	.93
LOC_2	.12	.11	.00	.84	1.00	.91
LOC_3	.09	.03	.06	.93	.91	1.00

http://www.socialresearchmethods.net/kb/convdisc.php

In-class exercise – Tweet Text Comprehensibility

- There's a dataset of different measures of "comprehensibility" for several thousand tweets about vaccines on Blackboard derived from Textstat: https://pypi.org/project/textstat/

1. Download the data and look up the corresponding measures on Wikipedia
2. Which constructs are we measuring?
3. Construct a correlation matrix (or even better, do a PCA)
4. What can we say about the convergent-discriminant validity of these constructs?
5. How would we automate this to do data-driven discovery?

Measures are:

- Syllable count
- Lexicon count
- Sentence count
- Flesch Reading Ease formula
- Flesch-Kincaid Grade Level
- Gunning FOG Formula
- SMOG Index
- Automated Readability Index
- Coleman-Liau Index
- Linsear Write Formula
- Dale-Chall Readability Score
- Difficult Words

Table 3 Results of PCA applied to measures of text comprehensibility.

	Readability (40%)	Verbatim (31%)	Sentences (17%)
Gunning-Fog Index	0.95		
Dale-Chall Readability Score	0.94		
Reading Ease	-0.91		
Flesch-Kincaid Index	0.88		
Automated Readability Index	0.8		
Difficult Words	0.72	0.52	
Length		0.96	
Syllable Count		0.95	
Lexicon Count		0.93	
Linsear Write Formula		0.73	-0.57
Sentence Count			0.89
SMOG Index			0.81

Note. Following Kaiser's criterion (retaining all eigenvalues ≥ 1.0) three factors, explaining 88% of the variance in the data, were retained. Factor loadings ≥ 0.40 are shown. SMOG = "Simple Measure of Gobbledygook."

Threats to construct validity

- Inadequate preoperational explication of constructs
  - When you aren't actually clear about what you're trying to measure
- Mono-operation bias
  - When you only use one measure for your construct and assume that it's the "right" one
  - Ex: measuring intelligence with the SAT only
- Interaction of different treatments
  - This is a form of selection bias, where your sample is exposed to other things that may affect what you're trying to measure
  - Ex: Athletes who can afford better training might be able to afford better technology. How will you tell the difference?

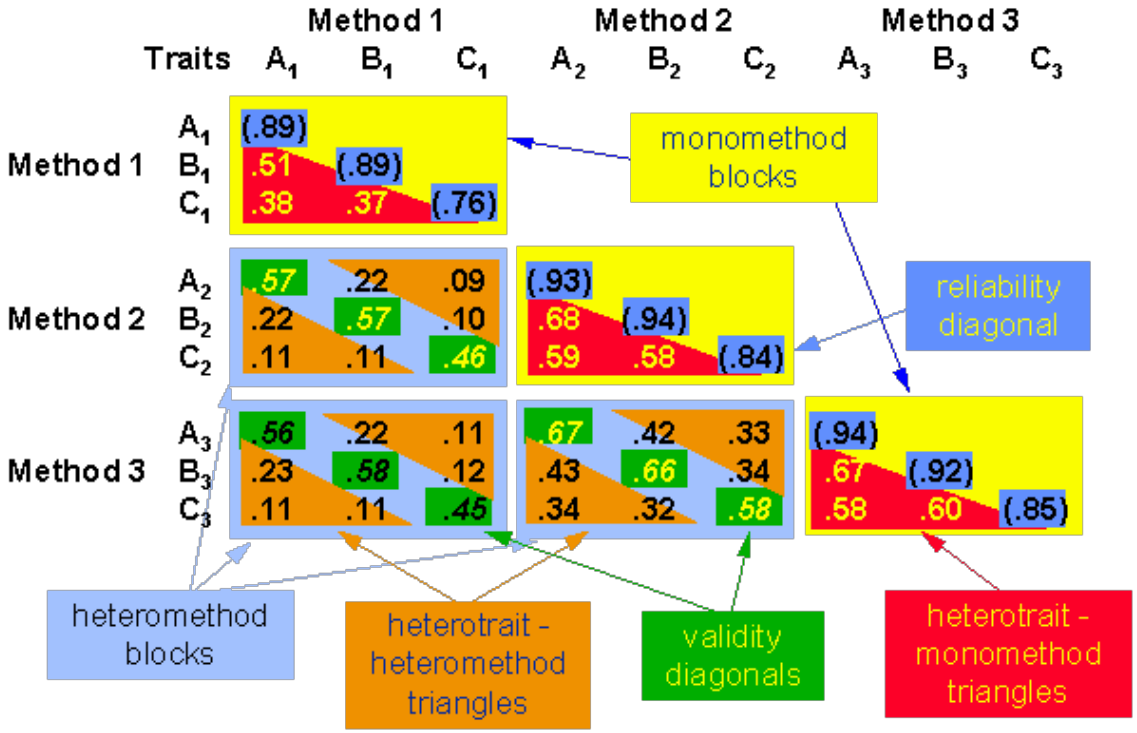
More threats to construct validity

- Interaction of treatment and testing:
  - The way you measure something may affect its result
  - Ex: Response bias – giving people a survey measuring their attitudes about vaccination may make them feel guilty for not vaccinating.
- Restricted generalizability across constructs:
  - You forgot to measure something else important
  - Ex: measuring the benefits of a medication without taking side effects into account
    - You get what you measure!
- Confounding constructs with levels of constructs
  - Measuring a range that is too narrow
  - Ex: Measuring how many older adults on Twitter post images of bungee jumping

“Social” threats to construct validity – applies to human studies

- Hypothesis guessing
  - When human participants in your study are trying to guess what the study is about
- Evaluation apprehension
  - People don’t like getting “graded” (unless they are graduate students)
- Experiment expectancies
  - This is basically confirmation bias or, if you’re interact with the subject, the subject may be trying to read the “right answer” from your behavior
  - Less of an issue for the sorts of analyses you are doing.

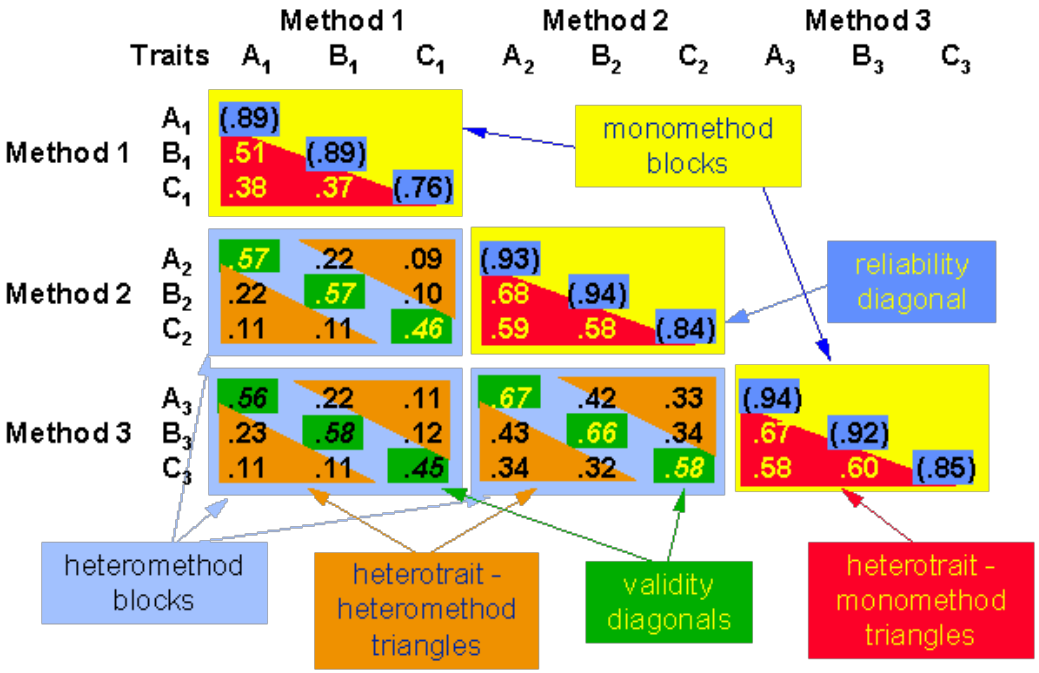
The Multi-Trait Multi-Method Matrix (MTMM)



https://socialresearchmethods.net/kb/mtmmmat.php

MTMM (cont.)

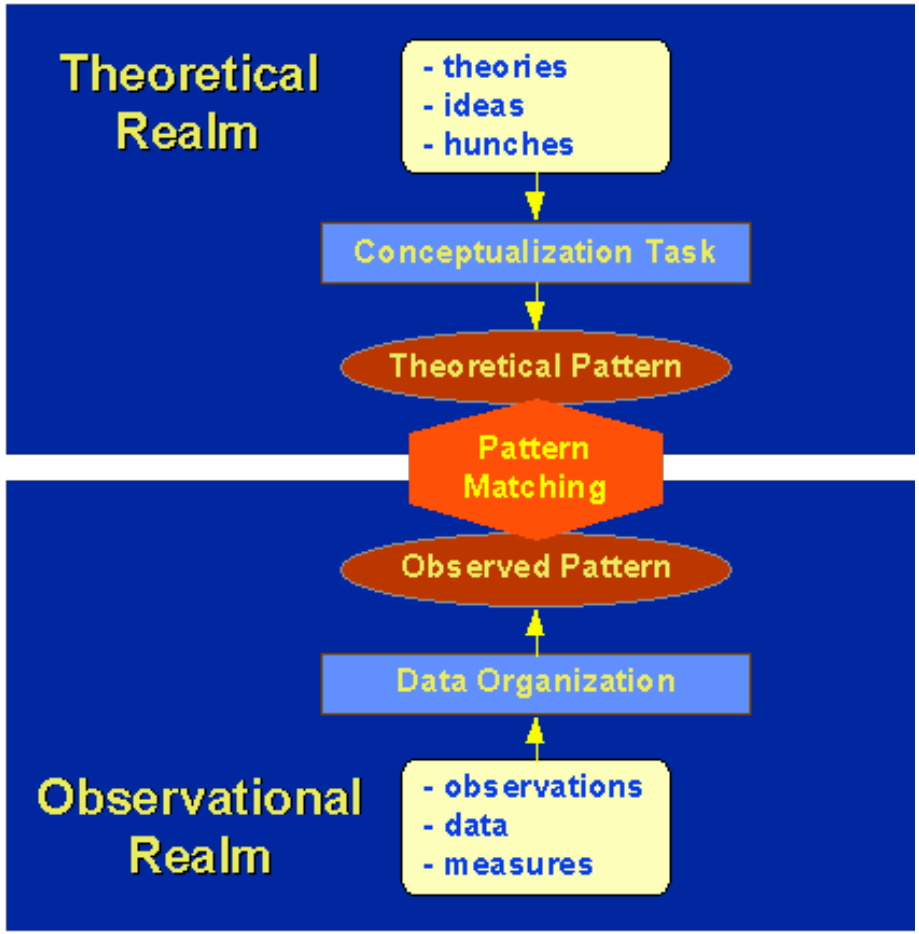
- Reliability: monotrait monomethod
- Validity: monotrait heteromethod
- What about the others? Trochim argues they aren't necessary for construct validity



https://socialresearchmethods.net/kb/mtmmmat.php

Pattern Matching

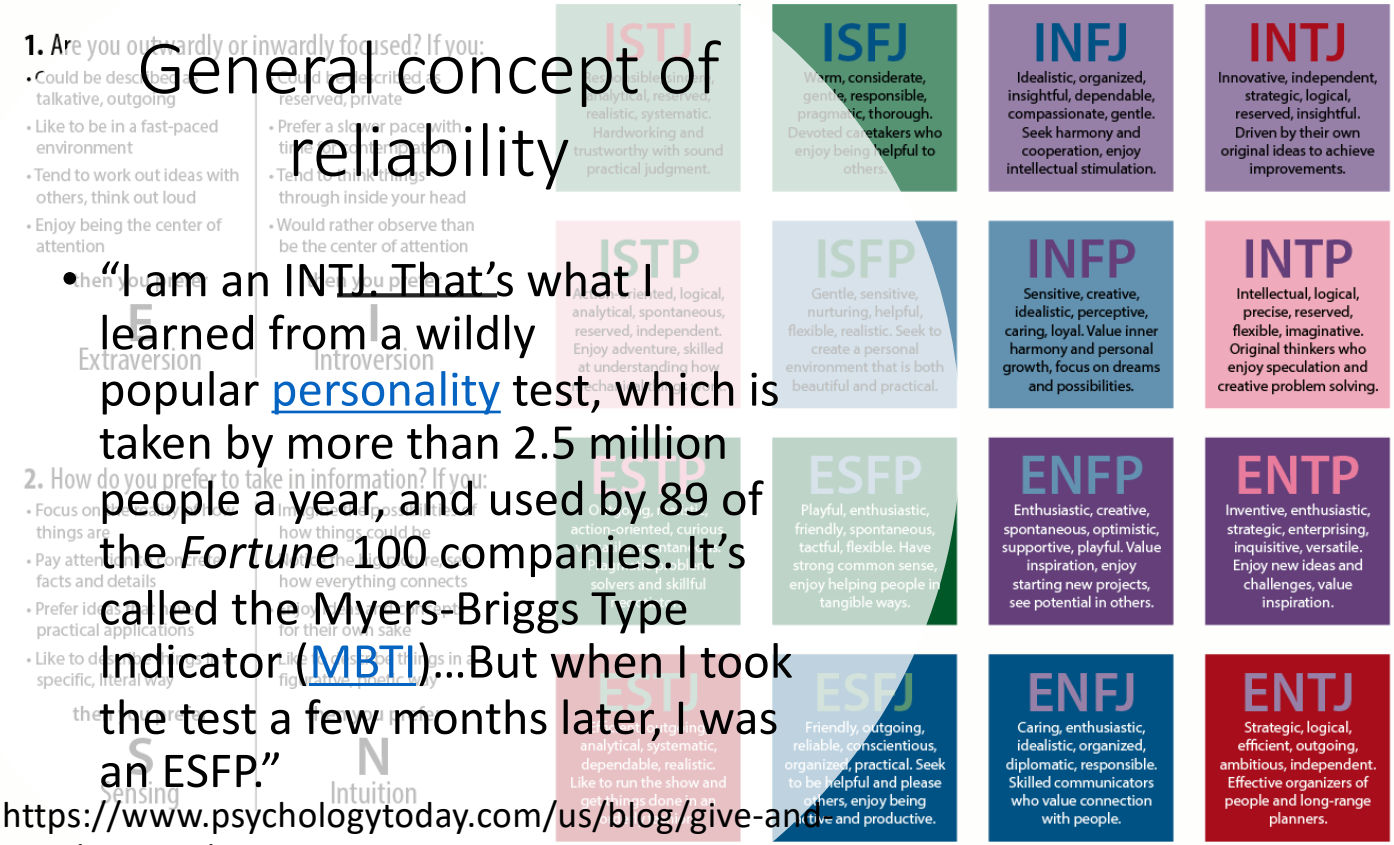
- Trochim, W., (1985). Pattern matching, validity, and conceptualization in program evaluation. Evaluation Review, 9, 5, 575-604 and Trochim, W. (1989). Outcome pattern matching and program theory. Evaluation and Program Planning, 12, 355-366.
- Compare to standard ML approach of finding patterns in data
- "Best" approach to construct validity is still under debate.
  - ML hasn't yet caught up!



Reliability

What's Your Personality Type?

Use the questions on the outside of the chart to determine the four letters of your Myers-Briggs type.
For each pair of letters, choose the side that seems most natural to you, even if you don't agree with every description.



3. How do you prefer to make decisions? If you:

- Make decisions in an impersonal way, using logical reasoning
- Value justice, fairness
- Enjoy finding the flaws in an argument
- Could be described as reasonable, level-headed

then you prefer **T** Thinking

- Base your decisions on personal values and how your actions affect others
- Value harmony, forgiveness
- Like to please others and point out the best in people
- Could be described as warm, empathetic

then you prefer **F** Feeling

4. How do you prefer to live your outer life? If you:

- Prefer to have matters settled
- Think rules and deadlines should be respected
- Prefer to have detailed, step-by-step instructions
- Make plans, want to know what you're getting into

then you prefer **J** Judging

- Prefer to leave your options open
- See rules and deadlines as flexible
- Like to improvise and make things up as you go
- Are spontaneous, enjoy surprises and new situations

then you prefer **P** Perceiving

General concept of reliability

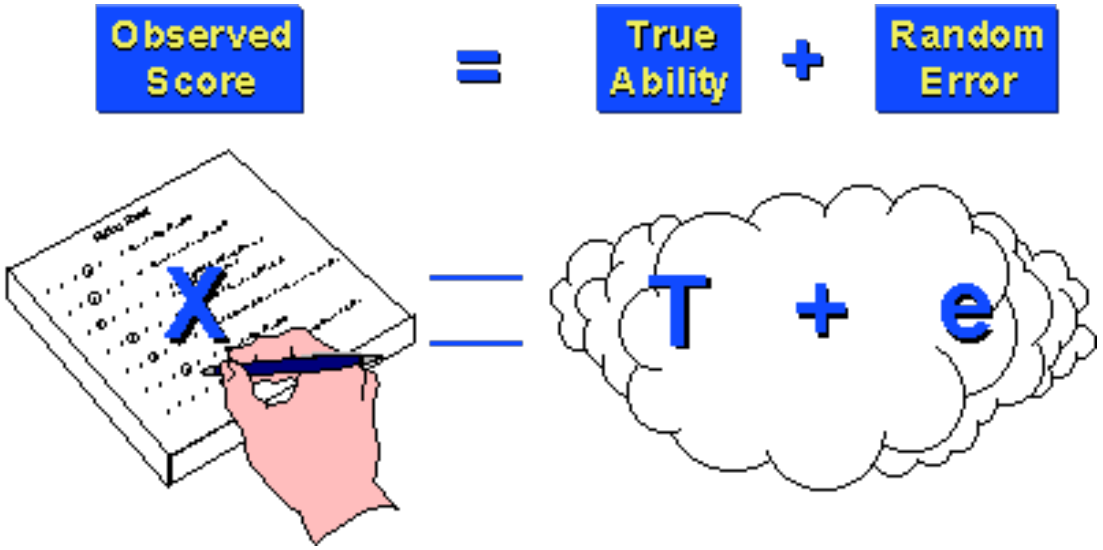
- "I am an INTJ. That's what I learned from a wildly popular personality test, which is taken by more than 2.5 million people a year, and used by 89 of the *Fortune* 100 companies. It's called the Myers-Briggs Type Indicator (MBTI)…But when I took the test a few months later, I was an ESFP."

https://www.psychologytoday.com/us/blog/give-and-take/201309/goodbye-mbti-the-fad-won-t-die

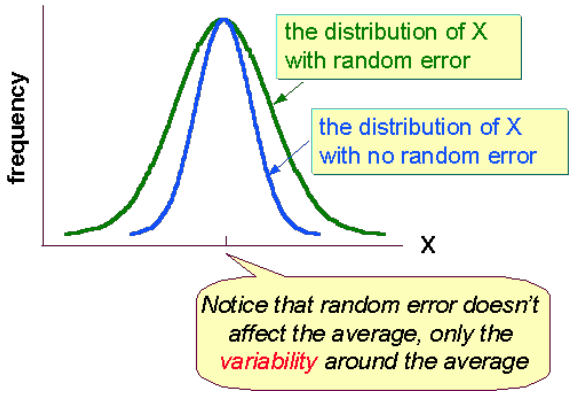
True score theory

Why do measures have low reliability?

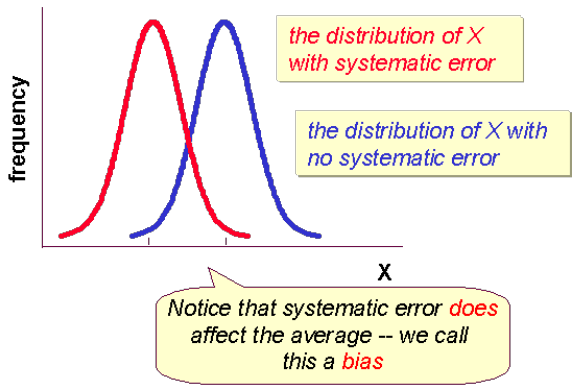
According to true score theory, it's when the "noise" overwhelms the "signal"



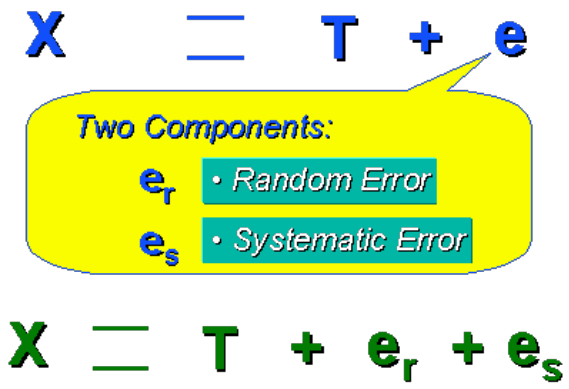
https://socialresearchmethods.net/kb/Assets/images/truescor.gif



VARIANCE



BIAS



Measurement error

But not all error is random.

Theory of reliability

- What do we mean by reliability anyway?
- We want to be able to say how much of a given measure is due to “truth” vs. “error”
  - For reliability = 1 – our measure is equal to the truth
  - For reliability = 0 – our measure is entirely error
- But we take multiple measurements
  - For reliability = 1 – the VARIANCE in our measure is equal to the truth
  - For reliability = 0 – the VARIANCE in our measure is entirely error
  - For reliability = 0.5 – HALF the VARIANCE in our measure is truth, half is error
- Variance due to truth: the actual construct is changing
- Variance due to error: the noise around the construct is changing

Theory of reliability (cont.)

- We can't measure truth!!
- BUT if two measures index the same construct, then they should COVARY.

$$\frac{\textbf{covariance}(X_1, X_2)}{\textbf{sd}(X_1) * \textbf{sd}(X_2)}$$

- This is the definition of a correlation!
- So: the reliability of a construct can be measured by the correlation between two measures of the same construct

Types of reliability

- Four types:
  1. Inter-Rater Reliability: Consistency across raters
  2. Test-Retest Reliability: Consistency across time
  3. Parallel-Forms Reliability: Consistency between two measures derived in the same way from different samples of the same data
  4. Internal Consistency Reliability: Consistency between items in the same tes
- Standard approach to measuring reliability:
  - Categorical: % agreement
  - Continuous: Correlation

Inter-coder reliability for categorical data

- Percentage agreement can be a misleading metric since agreement can occur by chance
- Assuming two raters, we can calculate Scott's π
- $\pi = \frac{p_o - pe}{1 - pe}$
- where p_o is the probability of observed agreement -- total number of items on which raters agree divided by total number of items
  - (1+5+9)/45 = 0.333
- p_e is the probability of expected agreement – squared sum of marginal responses.
  - $[(12+6)/90]^2 + [(15+15)/90]^2 + [(18+24)/90]^2 = 0.369$
- $\pi = \frac{0.333 - 0.369}{1 - 0.369} = -0.059$
- 1 means perfect agreement, zero or less means no more agreement than expected by chance

	Yes	No	Maybe	Marginal Sum
Yes	1	2	3	6
No	4	5	6	15
Maybe	7	8	9	24
Marginal Sum	12	15	18	45

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, 321-325.

Example Source: Scott's Pi. (2016, May 24). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Scott%27s_Pi&oldid=721888135

Cohen's κ: same idea but different p_e

- $\kappa = \frac{p_o - pe}{1 - pe}$
- where p_o is the probability of observed agreement -- total number of items on which raters agree divided by total number of items
  - (1+5+9)/45 = 0.333
- p_e is the probability of expected agreement – harmonic mean of marginal responses.
  - $[(12*6)/45] + [(15*15)/45]^{+}[(18*24)/45]/45 = 0.36$
- $\pi = \frac{0.333 - 0.36}{1 - 0.36} = -0.042$
- 1 means perfect agreement, zero or less means no more agreement than expected by chance

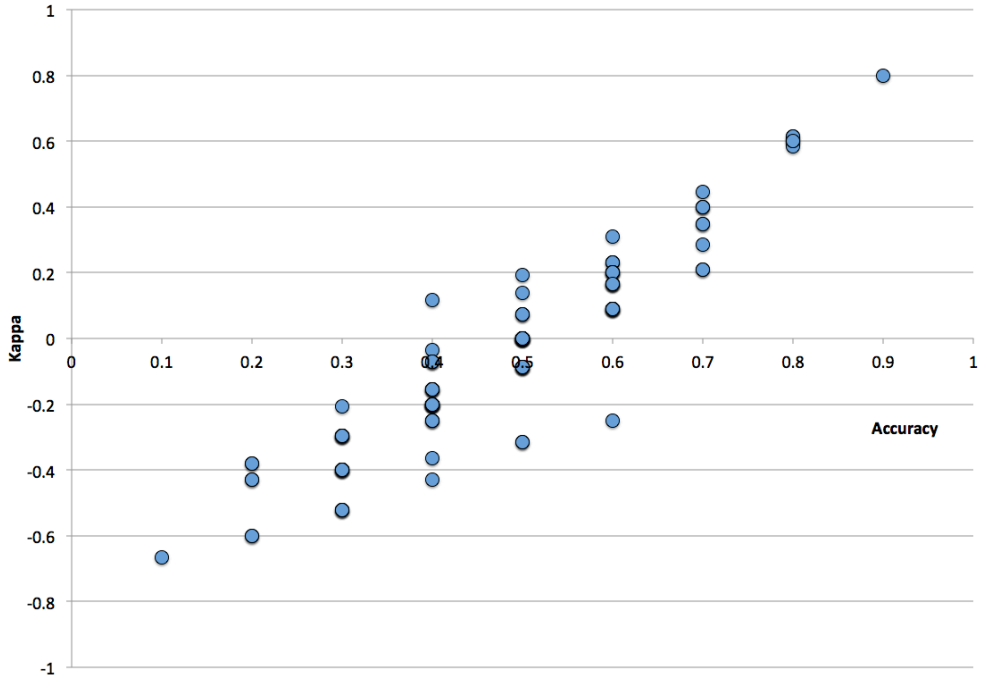
	Yes	No	Maybe	Marginal Sum
Yes	1	2	3	6
No	4	5	6	15
Maybe	7	8	9	24
Marginal Sum	12	15	18	45

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement. **20** (1): 37–46.

Example Source: Scott's Pi. (2016, May 24). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Scott%27s_Pi&oldid=721888135

Cohen's κ vs accuracy

- Such metrics are difficult to evaluate in the abstract – how good is "good enough" inter-rater reliability?
- $\kappa=1$ only if both raters have the exact same distribution over responses – this is very rare
  - Several refinements exist, including comparison to the theoretically maximum value that can be obtained given rater distributions.



By Genesis12 - Created in Excel using my own simulation, CC0, https://en.wikipedia.org/w/index.php?curid=43086928

Problem: what if there are more than two raters?

- Fleiss' κ generalizes Scott's π NOT Cohen's κ…
- $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$
- $p_j=\frac{1}{Nn}\sum_{i=1}^{N} n_{ij,}$ where N = total number of items rated, n is number of ratings per item, and n_{ij} is the number of times that item i was assigned to category j
- $P_i=\frac{1}{n(n-1)}\left[\sum_{j=1}^{k} {n_{ij}}^2 - n\right]$, where k is the total number of categories
- $\bar{P}=\frac{1}{N}\sum_{i=1}^{N} P_i$ –– the average of all the P_i
- $\bar{P}_e=\frac{1}{N}\sum_{j=1}^{k} {p_j}^2$

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters.*Psychological bulletin*, *76*(5), 378.

Fleiss' kappa. (2016, July 15). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Fleiss%27_kappa&oldid=729953115

A worked example

- $\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$ = (1/10) (1+0.253+0.308+0.440+0.330+0.462 +0.242+0.176+0.286+0.286) = 0.378
- $\bar{P}_e = 0.143^2 + 0.200^2 + 0.279^2 + 0.150^2 + 0.229^2 = 0.213$
- $\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} = \frac{0.378 - 0.213}{1 - 0.213} = 0.210$

n_{ij}	1	2	3	4	5	P_i
1	0	0	0	0	14	1.000
2	0	2	6	4	2	0.253
3	0	0	3	5	6	0.308
4	0	3	9	2	0	0.440
5	2	2	8	1	1	0.330
6	7	7	0	0	0	0.462
7	3	2	6	3	0	0.242
8	2	5	3	2	2	0.176
9	6	5	2	1	0	0.286
10	0	2	2	3	7	0.286
Total	20	28	39	21	32	
$\mathbf{p_j}$	0.143	0.200	0.279	0.150	0.229	

Fleiss' kappa. (2016, July 15). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Fleiss%27_kappa&oldid=729953115

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters.*Psychological bulletin*, *76*(5), 378.

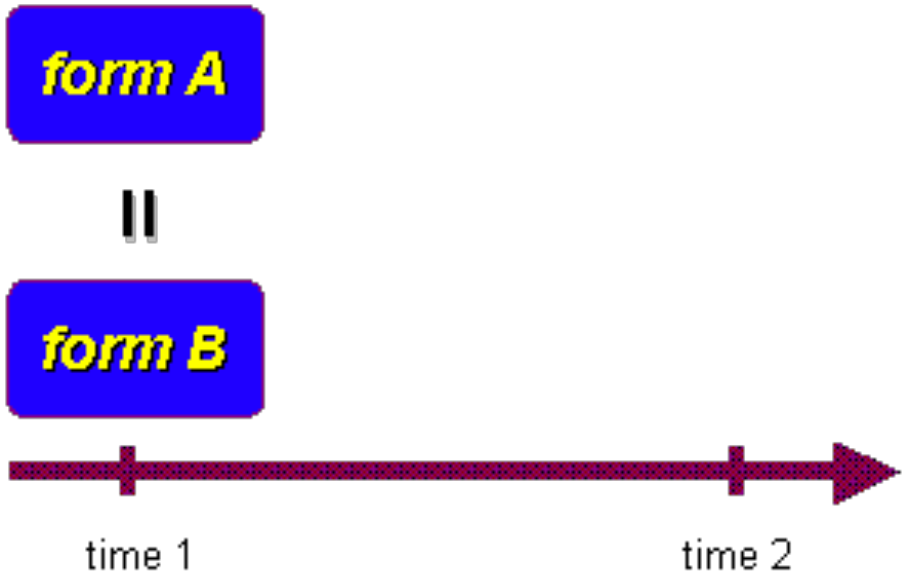
How to interpret Fleiss' κ?

- Very carefully!
- The table presents the interpretations of Landis & Koch but there is no consensus on these values
- >0 means some agreement
- <= 0 means no significant agreement
- Larger is better

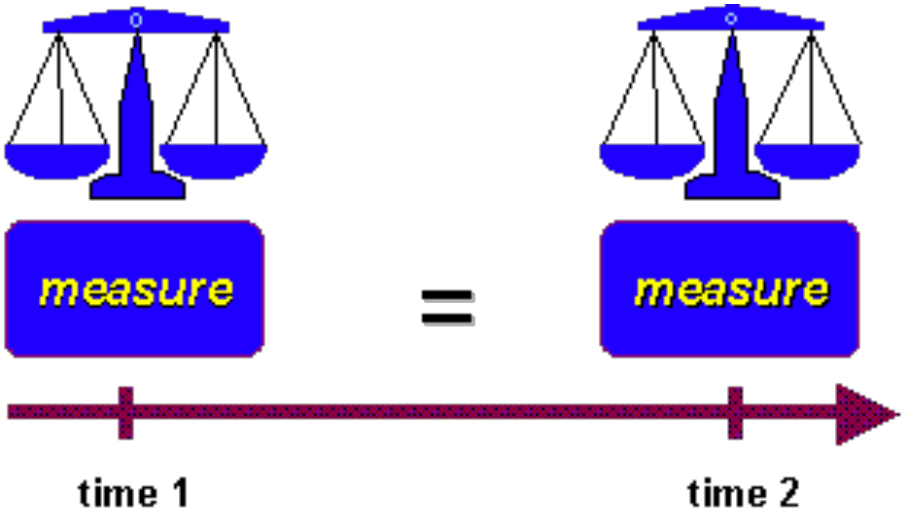
κ	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.
cited in Fleiss' kappa. (2016, July 15). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Fleiss%27_kappa&oldid=729953115

Measures of reliability



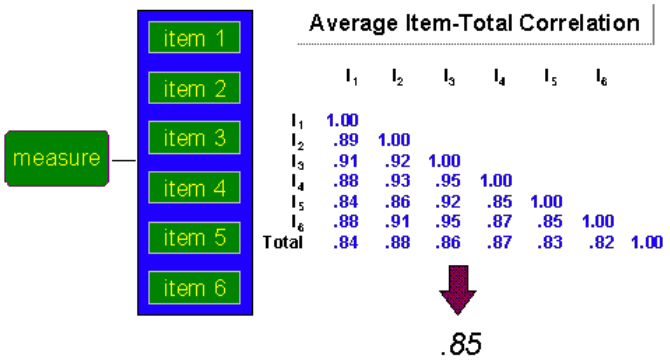
Parallel forms reliability;
requires creating many different items

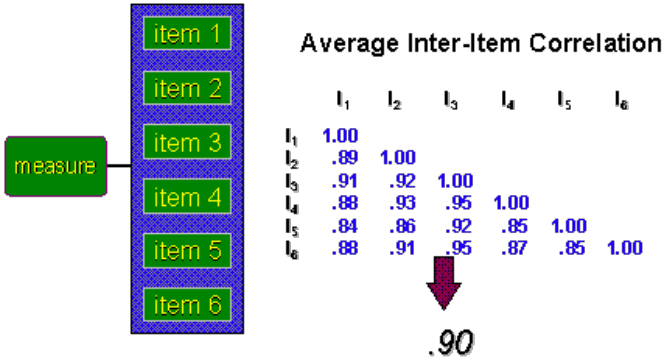


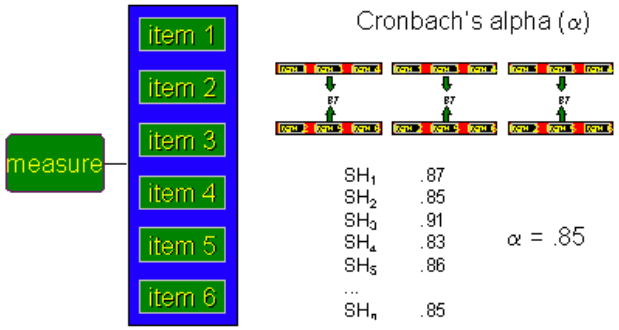
Test-retest reliability; declines with time,
but how much?

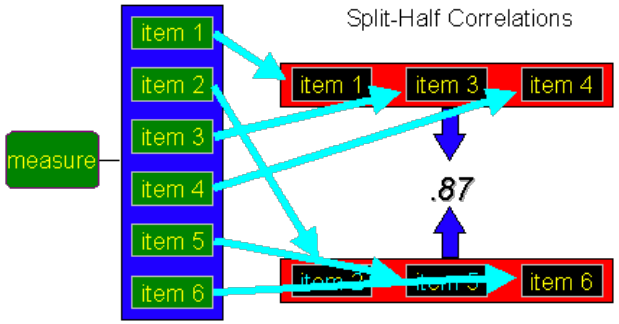
Internal Consistency Reliability

- Fundamental idea: find the correlation of all of the items that are supposed to measure the same construct
- Average Inter-Item: Calculate the average correlation
- Average Item-Total: Sum all of the measures and then calculate the average correlation with this sum





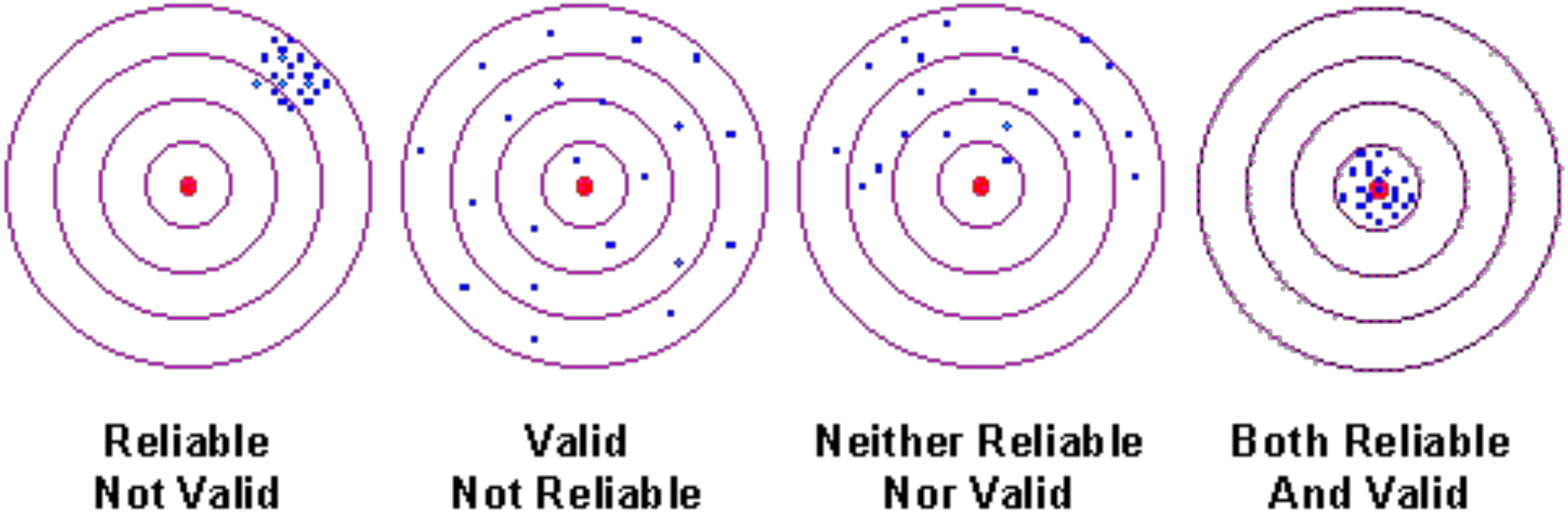




Cronbach's Alpha

- Split-half reliability: Take all items, divide them in two, and measure correlation between the sums of each half
- Cronbach's Alpha: Calculate the average split-half reliability for all possible splits
- Especially useful for figuring out what items to drop from a scale to improve reliability.

Relationship between reliability and validity

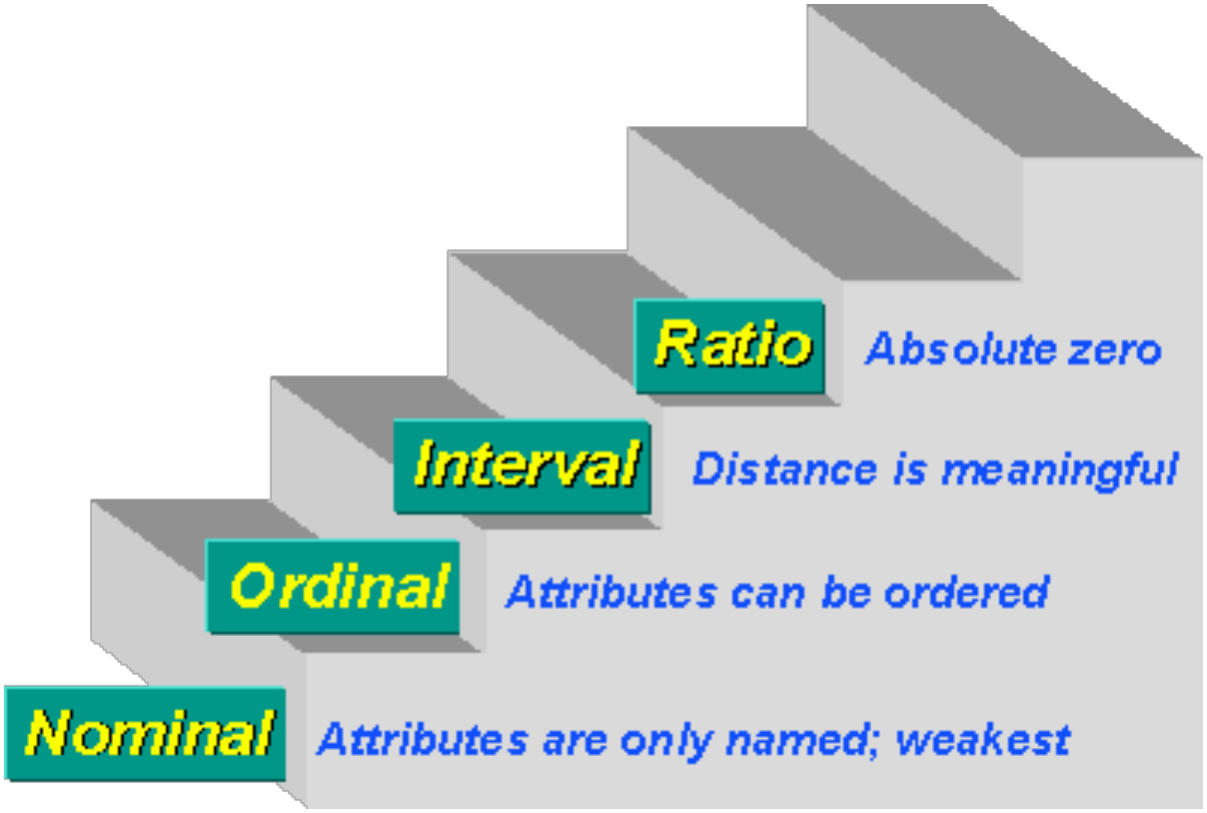


https://socialresearchmethods.net/kb/Assets/images/rel&val1.gif

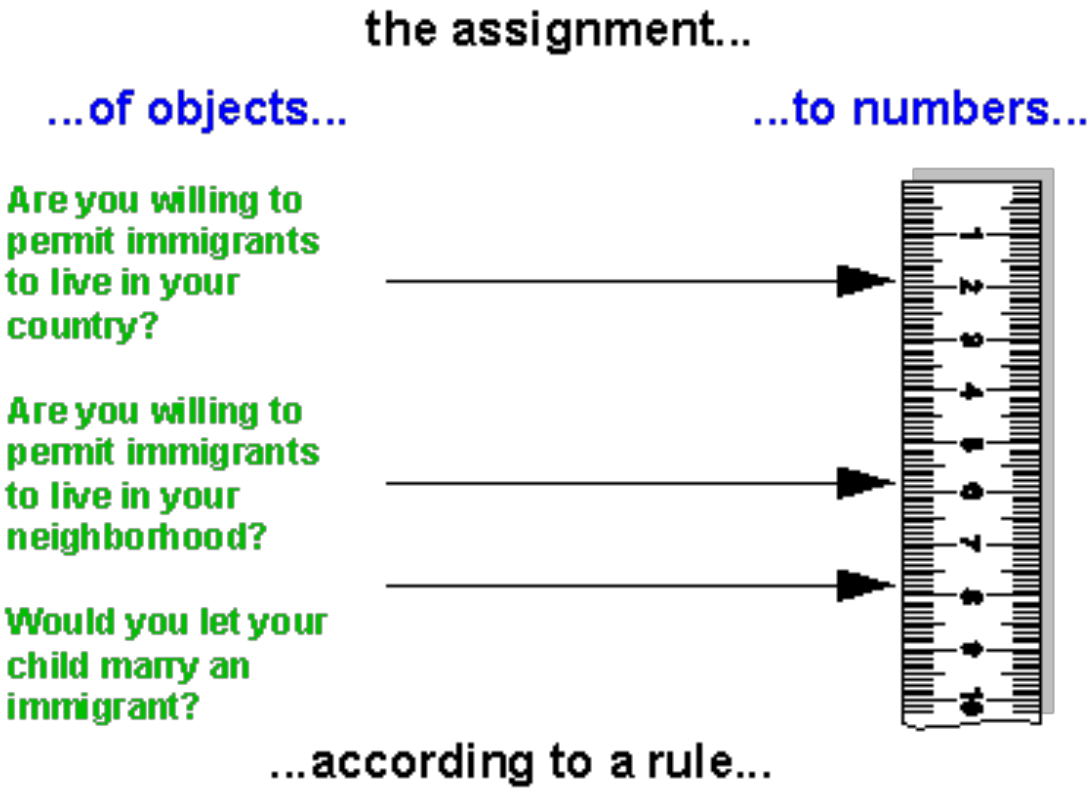
Relationship to bias/variance tradeoff

- In-class discussion
- Is BVT more similar to External vs. internal or reliability vs. construct validity?

Levels of Measurement



Scaling: General issues



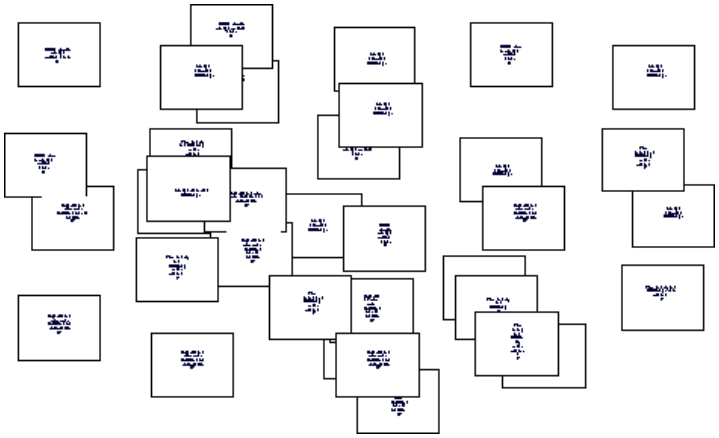
Thurstone Scaling

1. Figure out what construct you're trying to measure
2. Generate potential scale items
3. Rate the scale items (and ensure reliability!)
4. Analyze rating data to compute scale score values
5. Administer the scale

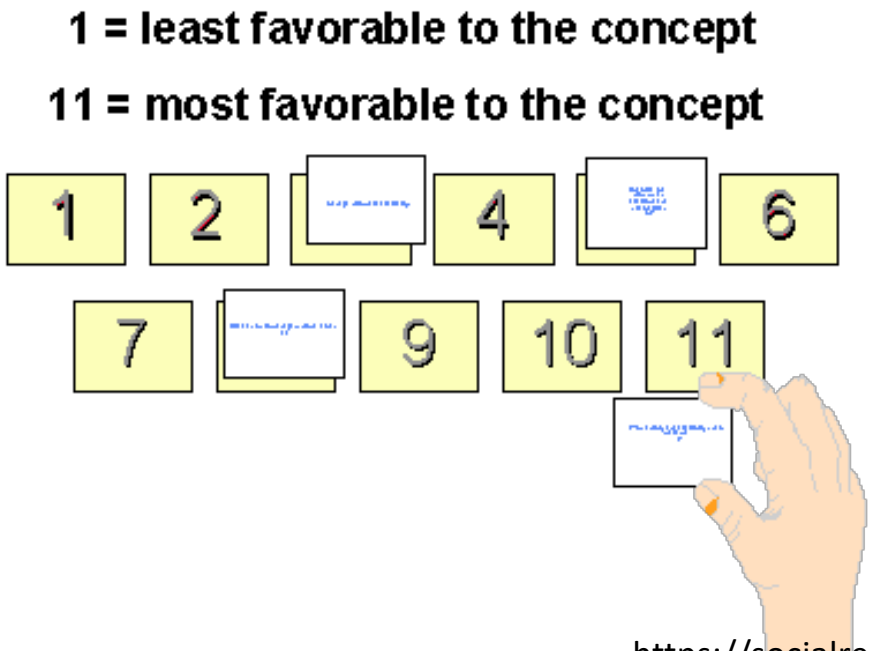
- Can possibly apply this to analysis of Twitter data.
- Once the scale exists, what kinds of features and regression models might be used to measure the construct on hold-out data?

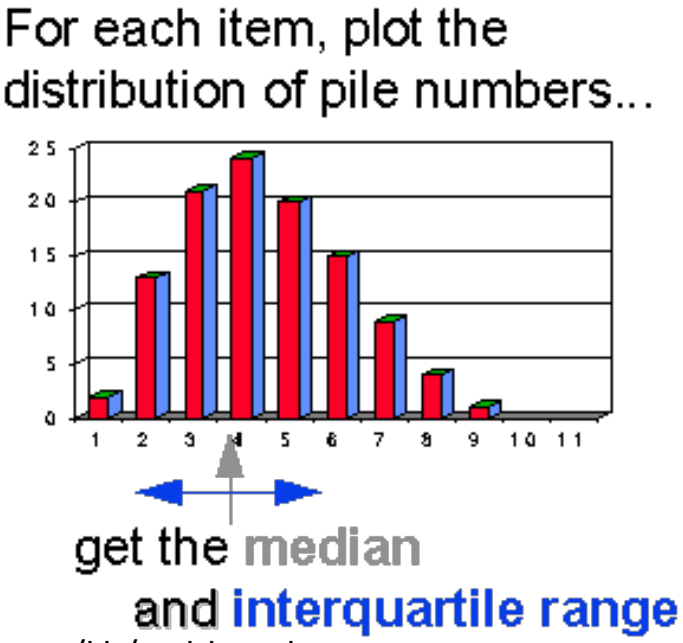
Generate potential scale items

- people get AIDS by engaging in immoral behavior
- you can get AIDS from toilet seats
- AIDS is the wrath of God
- anybody with AIDS is either gay or a junkie
- AIDS is an epidemic that affects us all
- people with AIDS are bad
- people with AIDS are real people
- AIDS is a cure, not a disease
- you can get AIDS from heterosexual sex



https://socialresearchmethods.net/kb/scalthur.php





https://socialresearchmethods.net/kb/scalthur.php

Rate the scale items

Use several raters and measure inter-rater reliability

Select and Administer Final Scale Items

- Select statements at equal intervals across range.
- Average responses across all items
- Done!

Agree	Disagree	People with AIDS are like my parents.
Agree	Disagree	Because AIDS is preventable, we should focus our resources on prevention instead of curing.
Agree	Disagree	People with AIDS deserve what they got.
Agree	Disagree	Aids affects us all.
Agree	Disagree	People with AIDS should be treated just like everybody else.
Agree	Disagree	AIDS will never happen to me.
Agree	Disagree	It's easy to get AIDS.
Agree	Disagree	AIDS doesn't have a preference, anyone can get it.
Agree	Disagree	AIDS is a disease that anyone can get if they are not careful.
Agree	Disagree	If you have AIDS, you can still lead a normal life.
Agree	Disagree	AIDS is good because it helps control the population.
Agree	Disagree	I can't get AIDS if I'm in a monogamous relationship.

Likert Scaling

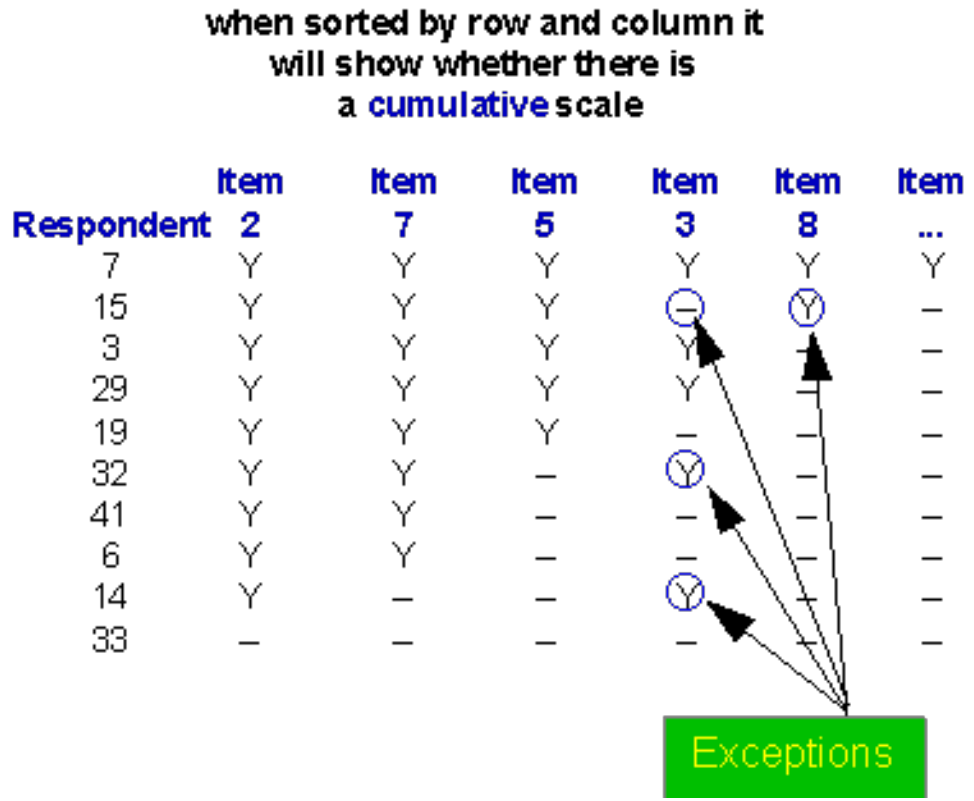
- Similar to Thurstone scaling, but judges rate every item on a pre-defined scale (e.g., 1-5)
- Calculate correlations between all items and sum of all items
- Throw out items that do not correlate
- Throw out items with limited variance (responses all cluster around same number)

Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	1. I feel good about my work on the job.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	2. On the whole, I get along well with others at work.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	3. I am proud of my ability to cope with difficulties at work.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	4. When I feel uncomfortable at work, I know how to handle it.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	5. I can tell that other people at work are glad to have me there.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	6. I know I'll be able to cope with work for as long as I want.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	7. I am proud of my relationship with my supervisor at work.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	8. I am confident that I can handle my job without constant assistance.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	9. I feel like I make a useful contribution at work.
Strongly Disagree	SomewhatDisagree	Somewhat Agree	Strongly Agree	10. I can tell that my coworkers respect me.

https://socialresearchmethods.net/kb/scallik.php

Guttmann Scaling

- Similar to Thurstone and Likert scaling, but judges rate every item as Y or N
- Weight of each item is determined by how many people agreed with it
- Select final items according to consistency
- Present checklist to subjects and calculate weighted sum



https://socialresearchmethods.net/kb/scallik.php

npj | Digital Medicine

www.nature.com/npjdigitalmed

BRIEF COMMUNICATION OPEN

Don't quote me: reverse identification of research participants in social media studies

John W. Ayers[1], Theodore L. Caputi[2], Camille Nebeker[3] and Mark Dredze[4]

We investigated if participants in social media surveillance studies could be reverse identified by reviewing all articles published on PubMed in 2015 or 2016 with the words "Twitter" and either "read," "coded," or "content" in the title or abstract. Seventy-two percent (95% CI: 63–80) of articles quoted at least one participant's tweet and searching for the quoted content led to the participant 84% (95% CI: 74–91) of the time. Twenty-one percent (95% CI: 13–29) of articles disclosed a participant's Twitter username thereby making the participant immediately identifiable. Only one article reported obtaining consent to disclose identifying information and institutional review board (IRB) involvement was mentioned in only 40% (95% CI: 31–50) of articles, of which 17% (95% CI: 10–25) received IRB-approval and 23% (95% CI:16–32) were deemed exempt. Biomedical publications are routinely including identifiable information by quoting tweets or revealing usernames which, in turn, violates ICMJE ethical standards governing scientific ethics, even though said content is scientifically unnecessary. We propose that authors convey aggregate findings without revealing participants' identities, editors refuse to publish reports that reveal a participant's identity, and IRBs attend to these privacy issues when reviewing studies involving social media data. These strategies together will ensure participants are protected going forward.

npj Digital Medicine (2018)1:30 ; doi:10.1038/s41746-018-0036-2

Unobtrustive Measures and Ethics

- Most of you will be doing work that doesn't require actual interaction with subjects
  - Your measures are "unobtrusive" because they don't interfere with natural proceses.
- Pros: Reduces bias!
- Cons: Are people aware that they are being monitored? They cannot provide consent. This may be ethically problematic.
  - Ex: sorting through someone's trash to collect financial data
- Must always consider the risk to the participants

Rubric

- 1) Revisit your research question. Restate the question as a hypothesis, and list the key constructs and any proposed relationship (causal or otherwise) between them.
- 2) Revisit your pilot data. List all of the features in your pilot data. Create a table relating your features to your constructs. Explicitly highlight those constructs for which you don't have features, and those features that don't correspond to constructs in your question.
- 3) For each construct missing a feature, come up with a plan to measure that construct (or otherwise develop a scale to measure that construct).

Construct (Hypothesis: More readable tweets, with the same level of emotionality, are more compelling)	**Feature**
Readability	Dale-Chall Index, Flesch-Kincaid Index, etc.
Compelling-ness of Tweet	Number of retweets Number of @mentions
Emotionality of tweet	
	Number of sentences

Rubric (cont.)

- 4) For each feature, indicate whether it is nominal, ordinal, interval, or ratio. Use your constructs to justify these selections.
- 5) Use your pilot data to demonstrate as much validity of each feature/construct pair as you can. For each feature/construct pair, justify why this is a valid mapping. Specifically indicate whether it has face validity, content validity, predictive validity, and/or convergent-discriminant validity.
- 6) For each feature/construct pair, test its convergent-discriminant validity using the data that you have available. If you can't do so using your available data, develop a plan to gather data to do so, and indicate the resources that would be required to implement this plan. If you can, implement this plan. If you can't, justify your continued use of this feature.

Rubric (cont.)

- 7) For each feature, justify why this is a reliable measure by quantifying measurement error. If you cannot do so, come up with a plan to collect data to do so.
- 8) For each feature, indicate any potential sources of methodological bias. Come up with (and if possible, implement) a plan to mitigate this bias.
- 9) For each feature/construct pair, indicate the appropriate measures of reliability: Cronbach's alpha, internal consistency, Fleiss' Kappa, etc. All of these measures depend on multiple features per construct. If you don't have this for some constructs, develop a plan to gather data to do get the appropriate features, and indicate the resources that would be required to implement this plan. If you can implement this plan, do so.
- 10) For each existing *and proposed* feature and construt, list the threats to construct validity for your study and indicate any potential source of ethical concerns that may arise by collecting your data. What is your plan to fix these?

Conclusion

- Construct validity is core to research validity
  - Without it, you have no guarantee that you're actually measuring anything useful
- Can define two major aspects:
  - Convergent-discriminant validity -- are your measures biased – in other words does your theory match the data?
  - Reliability – are your measures noisy? – in other words, does variance dominate
- There are several ways to generate metrics from all sorts of data types
  - Do it ethically