Video09 - Validity and reliability

Steve Simon

Three types of validity

- Internal validity
  - "The extent to which we can infer that the independent variable caused the dependent variable."
- External validity
  - "The extent to which the findings will generalize to other populations, settings, measures, and treatments."
- Measurement validity
  - "The quality of accuracy of individual measures or scores. The extent to which a score measures what it was intended to measure."
  - Also measurement reliability

Your book specifies three types of validity: internal validity, external validity, and measurement validity. I want to focus just on measurement validity in this lecture. The other types of validity are important, but the explanations are fairly simple to follow.

Measure validity is much harder to talk about, so I do want to spend a fair amount of time on it and on a closely related concept, measurement reliability.

Internal Validity

- "The extent to which we can infer that the independent variable *caused* the dependent variable."
- Three criteria for causality
  - IV *must precede* the outcome variable
  - IV *must be related* to the outcome
  - There must be no other variables that could explain why the IV is related to the outcome

Many people claim that it's really only randomized trials that can establish causality. But even in quasi-experimental and non-experimental studies, you can still talk about internal validity.

Establishing internal validity by the research approach

– Hierachy
  - Randomized Experiments
  - Quasi-Experimental studies
  - Comparative
  - Associational
  - Descriptive

Your book established a hierarchy and in fairness to the book, this is commonly accepted in most circles. But it is wrong.

For randomized designs, the cause precedes the effect, the cause is related to the outcome and there are no other variables that could explain the relationship. The major threats to internal validity are removed by randomization. There are still some threats, though, that you should be aware of.

For the other designs, you have to bring additional information into the analysis in order to establish internal validity.

Internal Validity is equivalence and control

– Evaluating the internal validity of a study –
- Equivalence of the groups on participant characteristics
- Control of extraneous experiences and environmental variables

There are two aspects of internal validity, equivalence and control.

Establishing equivalence

- Are groups equivalent prior to introduction of IV?
  - Assured without further work in randomized studies
  - Empirical comparisons for non-randomized studies
  - Can you use matching or statistical adjustments?

Random assignment assures equivalence of the two groups on average. A rule of thumb in your book is 30 people assigned to each group should give you confidence that random assignment will assure equivalence. A better reference demonstrates empirically that you need 20 to 80 per group.

For other studies, you have to establish equivalence empirically. This can only be done on measured variables, while randomization also assures equivalence on unmeasured variables.

For many non-randomized studies, you may need to rely on statistical adjustment or matching.

Establishing control

- Extraneous and environmental variables
  - Not of direct interest
  - Influence the outcome
  - Imbalanced
- Example: contamination
- Is one group affected more than the other?
  - Less of an issue for laboratory studies

Your study could be contaminated by other variables. These variables might be out of your control but which can influence the outcome.

Contamination: people in the intervention group are friends with the control group and share information.

The key issue is whether one group is more affected by extraneous variables.

In a controlled setting, there are fewer extraneous variables, but this changes in a field setting.

Other threats to internal validity

– Regression to the mean
– Dropouts/attrition
– Bias in assignment
– Carryover effects
– Changes in environment
– Instrument or observer inconsistency
– Patient expectations
– Observer bias

Extreme groups have issues with regression to the mean. Even without any intervention, the extremities will tend to lessen.

Dropouts or attrition means that you've designed a setting that is so difficult that no one can stay in. Differential attrition is especially troublesome.

Bias in assignment occurs when patients or their physicians directly or indirectly influence the assignment. Random assignment eliminates this bias, though sometimes physicians will try to subvert the randomization process.

Blinding helps control for expectation effects and observer bias.

Break #1

– What have you learned
  - Internal validity
– What's coming next
  - External validity

External Validity

- Population external validity
  - Representative sample
    - Few or no restrictions
    - Few or no dropouts
- Ecological external validity
  - Naturalness
    - Setting
    - Procedures

Distinction between internal and external validity

- Sampling process influences external validity
  - Random samples versus non-random samples
- Treatment allocation influences internal validity
  - Randomized design versus non-randomized design

External validity is influenced by the sampling process. Internal validity is influenced by the allocation of this sample to treatment groups.

Trade-offs between internal and external validity

- High degree of control
  - Avoid issues of equivalence
  - Unnatural setting
- Low degree of control
  - More chances of contamination
  - Closer to how medicine is practiced.

Break #2

- What have you learned
  - External validity
- What's coming next
  - Measurement validity

Measurement quotes (1 of 2)

- “The government is extremely fond of amassing great quantities of statistics. These are raised to the Nth degree, the cube roots are extracted, and the results are arranged into elaborate and impressive displays. What must be kept ever in mind, however, is that in every case, the figures are first put down by a village watchman, and he puts down anything he damn well pleases.”
  - Sir Josiah Stamp, as quoted on Quotetab.

As much as I love numbers, I have to admit that they are often abused. Just because you can attach a number to something does not mean that the number is useful in any way. I want to talk about some of the problems associated with measurement and some of the great pains that you need to take to be sure that your numbers have meaning.

Measurement quotes (2 of 2)

- "only scientists are arrogant enough to think that they always observe with rigorous and objective scrutiny"
  - Stephen Jay Gould, The Mismeasure of Man, page 36.

I also have to quote Stephen Jay Gould here as well. He wrote an excellent book, The Mismeasure of Man, that addresses many of the points I will talk about today from the perspective of intelligence tests. It is well worth reading because it helps you to resist the temptation to think that writing down a number and giving it a name is enough. You have to think long and hard about whether your measurements are of sufficient quality that you can rely on them to draw firm conclusions about the clinical care that you provide to your patients.

Measurements that warrant closer scrutiny

- Patient reported outcomes
  - Participant report
- Researcher evaluations
  - Only when concerned about subjectivity
- Psychological constructs
- Composite scores

For better or for worse, researchers tend to focus greater attention on certain types of measurements. There's no hard and fast rule here, but issues of measurement quality tend to appear most often in certain areas.

Don't think that if your measure is not on this list that it doesn't deserve careful scrutiny. There's really no consensus in the research community on what measurements require this extra level of attention.

I think it is a bit unfair, but there is a lot of distrust of patient reported outcomes among researchers. Why not believe what the patient says about himself or herself? Part of it might be that a patient's answers could potentially be influenced by their mood.

There is also a belief that patient reported outcome measures vary too much from one individual to another. Some people are stoic to a fault and others will complain endlessly at the drop of a hat.

It is worth noting that these factors also influence researcher observation, but researchers don't like it when you point this out to them. It's mostly a good thing that researchers require a high level of scrutiny of patient reported outcomes, but perhaps other measures deserve just as high a level of scrutiny.

There is a fair amount of scrutiny of researcher evaluations when these evaluations are perceived as having a high level of subjectivity. Now, our perceptions as to what is subjective are also subjective, so you need to be careful.

Psychological constructs are tools used to measure aspects of human behavior, such as intelligence, self-esteem, stress, and extraversion. In spite of recent advances in brain imaging, you cannot, for the most part, peek inside someone's mind and understand how they think.

Finally, many measures in clinical research are composites of one or more items. These individual items are scored and added up to get a total. If the individual items are chosen well, this can be a very effective approach, but you need to be careful.

Measurement Reliability

– Synoynms: consistency, precision, stability
– No measurement is perfectly reliable
– Dependent on the population
– Look for prior efforts in reliability

When you measure something, you want that measurement to be consistent, precise, and stable. You don't want something that changes as the phases of the moon change. You don't want a measurement that changes depending on who the attending physician is. You don't want a measurement that changes depending on any environmental factors that are extraneous to what you are measuring.

If your measure is not stable, then you have difficulty in assessing whether a change in that measurement is due to your intervention or due to the phases of the moon.

One thing you need to keep in mind is that the reliability coefficient is dependent on the population it is based on. Your book doesn't mention this, but it is important. Change the population and you change the reliability coefficient. Something with a great reliability coefficient in a population of college students might be terrible in a population with limited literacy skills, for example.

Since no measurement is ever conducted without some Reliability is usually established when a measure is developed. When you go about using a measure, look at what's already been published. Make sure it used in a context similar to yours. It's a whole lot easier to find a measurement that is already proven to be reliable than to develop your own measure and then establish its reliability.

Measurement Validity

- Reliability by itself is not enough.
  - Consistent measures of the “wrong thing” is bad
- Examples of the wrong thing
  - Measuring anxiety instead of stress
  - Measuring transient changes in a patient’s mood rather than chronic depression
- Validity
  - “Degree to which a measure … measures that which it was intended to measure”
- Reliability is a pre-requisite for validity
- Validity is a journey and not a destination

Reliability by itself is not enough. That seems a bit unfair. You had to do a lot of work to establish reliability. But you can’t stop there. If you have a reliable measurement, one that is consistent across time and between raters, then you could still have problems because you might be measuring the wrong thing.

This can happen very easily. You might think that you are measuring the stress that a patient is enduring, but it might be a measure of anxiety instead. Now these are often related, but people can experience one without the other very easily. Another example would be measuring transient changes in mood versus chronic depression.

So in addition to establishing that your measurement has good reliability, you also have to establish good validity.

Validity is, to quote from your book, the “degree to which a measure measures that which it was intended to measure.”

If you intend to measure A and you measure B instead, you have poor validity.

Now I talked about reliability first because it is a pre-requisite for validity. If a measurement is inconsistent across time or between raters, it can’t be measuring what you want it to measure. It needs stability and consistency first.

The other thing to keep in mind is that validity is not something that you establish and then you’re done. Validity is a journey and a never-ending journey at that. Each study in a series of studies that uses a particular measurement will contribute information about the validity about a measurement.

Break #3

- What have you learned
  - General concepts of measurement reliability/validity
- What's next
  - Five case studies

Case study #1 - Neighborhood Environment Survey

R.M. Crum et al. / Drug and Alcohol Dependence 43 (1996) 155–161 157

Table 1
Items[a] used to assess neighborhood disadvantage

	True[b]	False[b]
Within walking distance of my house there is a park or playground where I like to walk and enjoy myself, playing sports or games	0	1
There are plenty of safe places to walk or play outdoors in my neighborhood	0	1
Every few weeks, some kid in my neighborhood gets beat-up or mugged	1	0
Every few weeks, some adult gets beat-up or mugged in my neighborhood	1	0
In my neighborhood, I see signs of racism and prejudice at least once a week	1	0
In my neighborhood, many yards and alleys have broken bottles and trash lying around	1	0
I have seen people using or selling drugs in my neighborhood[c]	1	0
In the morning or later in the day, I often see drunk people on the street in my neighborhood	1	0
Most adults in my neighborhood respect the law	0	1
There are abandoned or boarded-up buildings in my neighborhood	1	0
I feel safe when I walk around my neighborhood by myself	0	1
The people who live in my neighborhood often damage or steal each other's property	1	0
The people who live in my neighborhood always take care of each other and protect each other from crime	0	1
Almost everyday I see homeless people walking or sitting around in my neighborhood	1	0
In my neighborhood, the people with the most money are the drug dealers	1	0
In my neighborhood, there are a lot of poor people who don't have enough money for food and basic needs	1	0
For many people in my neighborhood, going to church on Sunday or religious days is a very important activity	0	1
The people who live in my neighborhood are the best people in the world	0	1

Data from the Prevention Research Center, Johns Hopkins University School of Hygiene and Public Health.
[a] This scale was adapted and modified from one initially developed by Elliott and colleagues for use in the National Youth Survey (28).
[b] The scoring of true and false answers for the individual items is shown in these columns. The item scores are summed yielding an overall scale score from 1 to 18, with higher scores indicating a greater degree of neighborhood disadvantge.
[c] This item was dropped from the scale for analyses reported in Table 3, model 2.

The image on the left is a picture of the Boynton neighborhood in Detroit, Michigan. It was downloaded from

https://en.wikipedia.org/wiki/File:Boynton_neighborhood_Detroit_(Annabelle_Street).jpg

and is published under a public domain license.

The image on the right is a screenshot from the publication

Crum RM, Lillie-Blanton M, Anthony JC. Neighborhood environment and opportunity to use cocaine and other drugs in late childhood and early adolescence. Drug and Alcohol Dependence 43 (1996) 155-161.

It is copyrighted and made available here through the "Fair Use" provisions of copyright law. Your use of this image may or may not fall under the Fair Use provisions.

This is the Neighborhood Environment Survey.

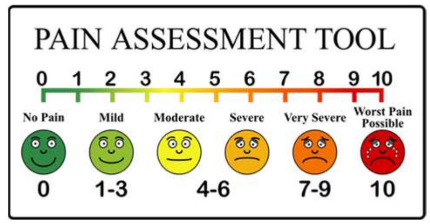
I realize this image might be difficult to read on your computer. I’m going to magnify this in a bit, but notice that this is a series of 18 true/false questions.

Case study #1 - Neighborhod Environment Survey

There are plenty of safe places to walk or play outdoors in my neighborhood
Every few weeks, some kid in my neighborhood gets beat-up or mugged
Every few weeks, some adult gets beat-up or mugged in my neighborhood
In my neighborhood, I see signs of racism and prejudice at least once a week
In my neighborhood, many yards and alleys have broken bottles and trash lying around

Here are some of the questions on the NES. “There are plenty of safe places to walk or play outdoors in my neighborhood” and “Every few weeks, some kid in my neighborhood gets beat-up or mugged.” These question are answered true/false and points of 0 or 1 are assigned. Notice that some of the questions get a 1 for true and some of the questions get a 1 for false. The total score is 0 to 18.

Case study #2 - Pain scale



The image on the left shows a dental procedure underway, hopefully with the patient getting appropriate analgesic medication. It was downloaded from

https://commons.wikimedia.org/wiki/File:Flickr_-_Official_U.S._Navy_Imagery_-_Dental_procedure_aboard_USS_Abraham_Lincoln.jpg

and is in the public domain.

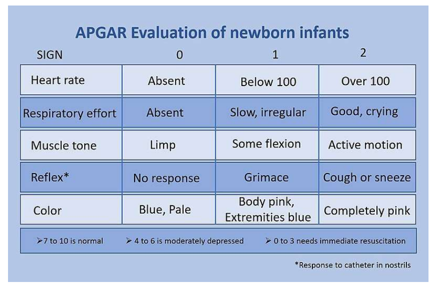
The image on the right is a pain scale. It was downloaded from

https://publicdomainvectors.org/en/free-clipart/Pain-scale/50153.html

and is in the public domain.

Pain is something that cannot be easily observed by an outsider. You are best off asking someone directly what their pain is like.

Case study #3 - Apgar score



APGAR Evaluation of newborn infants

SIGN	0	1	2
Heart rate	Absent	Below 100	Over 100
Respiratory effort	Absent	Slow, irregular	Good, crying
Muscle tone	Limp	Some flexion	Active motion
Reflex*	No response	Grimace	Cough or sneeze
Color	Blue, Pale	Body pink, Extremities blue	Completely pink

- 7 to 10 is normal
- 4 to 6 is moderately depressed
- 0 to 3 needs immediate resuscitation

*Response to catheter in nostrils

The image on the left of a newborn infant was found at

https://commons.wikimedia.org/wiki/File:HumanNewborn.JPG

and is published on an open source license.

The image on the right shows the five components of the Apgar score. It was found at

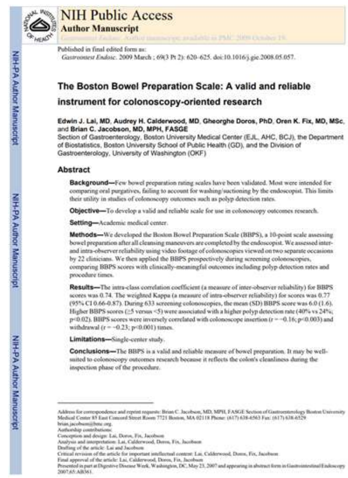
https://commons.wikimedia.org/wiki/File:APGAR_score.jpg

and is published on an open source license.

This is the Apgar score. It is a measurement taken one minute after a person is born. There are five components to the Apgar score, each rated as zero, one, or two. Just a hint here. If you happen to be born, being blue, limp, and not crying is a very bad thing.

This baby is the opposite. Screaming its lungs out, thrashing all around, nice pink color.

Case study #4 - Boston Bowel Prep Score



NIH Public Access
Author Manuscript

Published in final edited form as:
Gastrointest Endosc. 2009 March ; 69(3 Pt 2): 620–625. doi:10.1016/j.gie.2008.05.057.

The Boston Bowel Preparation Scale: A valid and reliable instrument for colonoscopy-oriented research

Edwin J. Lai, MD, **Audrey H. Calderwood, MD**, **Gheorghe Doros, PhD**, **Oren K. Fix, MD, MSc**, and **Brian C. Jacobson, MD, MPH, FASGE**
Section of Gastroenterology, Boston University Medical Center (EJL, AHC, BCJ), the Department of Biostatistics, Boston University School of Public Health (GD), and the Division of Gastroenterology, University of Washington (OKF)

Abstract

Background—Few bowel preparation rating scales have been validated. Most were intended for comparing oral purgatives, failing to account for washing/suctioning by the endoscopist. This limits their utility in studies of colonoscopy outcomes such as polyp detection rates.

Objective—To develop a valid and reliable scale for use in colonoscopy outcomes research.

Setting—Academic medical center.

Methods—We developed the Boston Bowel Preparation Scale (BBPS), a 10-point scale assessing bowel preparation after all cleansing maneuvers are completed by the endoscopist. We assessed inter- and intra-observer reliability using video footage of colonoscopies viewed on two separate occasions by 22 clinicians. We then applied the BBPS prospectively during screening colonoscopies, comparing BBPS scores with clinically-meaningful outcomes including polyp detection rates and procedure times.

Results—The intra-class correlation coefficient (a measure of inter-observer reliability) for BBPS scores was 0.74. The weighted Kappa (a measure of intra-observer reliability) for scores was 0.77 (95% CI 0.66-0.87). During 633 screening colonoscopies, the mean (SD) BBPS score was 6.0 (1.6). Higher BBPS scores (≥5 versus <5) were associated with a higher polyp detection rate (40% vs 24%; p<0.02). BBPS scores were inversely correlated with colonoscope insertion (r = −0.16; p<0.003) and withdrawal (r = −0.23; p<0.001) times.

Limitations—Single-center study.

Conclusions—The BBPS is a valid and reliable measure of bowel preparation. It may be well-suited to colonoscopy outcomes research because it reflects the colon's cleanliness during the inspection phase of the procedure.

Address for correspondence and reprint requests: Brian C. Jacobson, MD, MPH, FASGE Section of Gastroenterology Boston University Medical Center 85 East Concord Street Room 7721 Boston, MA 02118 Phone: (617) 638-6563 Fax: (617) 638-6529
Authorship contributions:
Conception and design: Lai, Doros, Fix, Jacobson
Analysis and interpretation: Lai, Calderwood, Doros, Fix, Jacobson
Drafting of the article: Lai and Jacobson
Critical revision of the article for important intellectual content: Lai, Calderwood, Doros, Fix, Jacobson
Final approval of the article: Lai, Calderwood, Doros, Fix, Jacobson
Presented in part at Digestive Disease Week, Washington, DC, May 23, 2007 and appearing in abstract form in Gastrointestinal Endoscopy 2007;65:AB361.

The image on the left is a colonoscope. It was downloeaded from

https://commons.wikimedia.org/wiki/File:PENTAX_Colonoscope001.jpg

and is published under an open source license.

On the right is the first page of a publication available through Pubmed Central. It is also available under an open source license.

This an except from an article establishing the reliability and validity of the Boston Bowel Prep score. When you get a colonoscopy done, you're supposed fast for a full day and also drink a foul tasting preparation. Almost a gallon of this stuff. When you're done, the next day, your bowel is cleaned out enough that the colonoscopist can go hunting for polyps in your colon.

An objective measure of how well you did with your fasting and with your drink preparation is the Boston Bowel Prep Score. This is an excerpt from the research article that established reliability and validity of this measure.

Lai EJ, Calderwood AH, Doros G, Fix OK, Jacobson BC. The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research. Gastrointest Endosc. 2009 Mar;69(3 Pt 2):620-5. doi: 10.1016/j.gie.2008.05.057. Epub 2009 Jan 10. PubMed PMID: 19136102; PubMed Central PMCID: PMC2763922.

Case study #4 - Boston Bowel Prep Score

0 = Unprepared colon segment with mucosa not seen due to solid stool that cannot be cleared.

1 = Portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid.

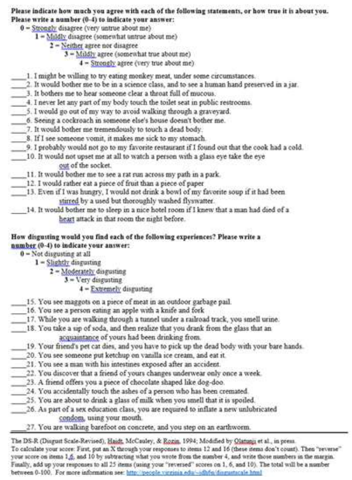
2 = Minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well.

3 = Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid. The wording of the scale was finalized after incorporating feedback from three colleagues experienced in colonoscopy.

Excerpt from Lai et al article

Here's a careful look at the four values. 0 means that you can't see anything because of the presence of solid stool (somebody didn't fast like they should have!) and 3 means only small fragments of stool or opaque liquid. You don't want a zero because they'll make you do the colonoscopy all over again.

Case study #5 - Disgust Scale Revised



Please indicate how much you agree with each of the following statements, or how true it is about you. Please write a number (0-4) to indicate your answer:

0 – Strongly disagree (very untrue about me)
1 – Mildly disagree (somewhat untrue about me)
2 – Neither agree nor disagree
3 – Mildly agree (somewhat true about me)
4 – Strongly agree (very true about me)

____1. I might be willing to try eating monkey meat, under some circumstances.
____2. It would bother me to be in a science class, and to see a human hand preserved in a jar.
____3. It bothers me to hear someone clear a throat full of mucous.
____4. I never let any part of my body touch the toilet seat in public restrooms.
____5. I would go out of my way to avoid walking through a graveyard.
____6. Seeing a cockroach in someone else's house doesn't bother me.
____7. It would bother me tremendously to touch a dead body.
____8. If I see someone vomit, it makes me sick to my stomach.
____9. I probably would not go to my favorite restaurant if I found out that the cook had a cold.
____10. It would not upset me at all to watch a person with a glass eye take the eye out of the socket.
____11. It would bother me to see a rat run across my path in a park.
____12. I would rather eat a piece of fruit than a piece of paper
____13. Even if I was hungry, I would not drink a bowl of my favorite soup if it had been stirred by a used but thoroughly washed flyswatter.
____14. It would bother me to sleep in a nice hotel room if I knew that a man had died of a heart attack in that room the night before.

How disgusting would you find each of the following experiences? Please write a number (0-4) to indicate your answer:

0 – Not disgusting at all
1 – Slightly disgusting
2 – Moderately disgusting
3 – Very disgusting
4 – Extremely disgusting

____15. You see maggots on a piece of meat in an outdoor garbage pail.
____16. You see a person eating an apple with a knife and fork
____17. While you are walking through a tunnel under a railroad track, you smell urine.
____18. You take a sip of soda, and then realize that you drank from the glass that an acquaintance of yours had been drinking from.
____19. Your friend's pet cat dies, and you have to pick up the dead body with your bare hands.
____20. You see someone put ketchup on vanilla ice cream, and eat it.
____21. You see a man with his intestines exposed after an accident.
____22. You discover that a friend of yours changes underwear only once a week.
____23. A friend offers you a piece of chocolate shaped like dog-doo.
____24. You accidentally touch the ashes of a person who has been cremated.
____25. You are about to drink a glass of milk when you smell that it is spoiled.
____26. As part of a sex education class, you are required to inflate a new unlubricated condom, using your mouth.
____27. You are walking barefoot on concrete, and you step on an earthworm.

The DS-R (Disgust Scale-Revised), Haidt, McCauley, & Rozin, 1994; Modified by Olatunji et al., in press.
To calculate your score: First, put an X through your responses to items 12 and 16 (these items don't count). Then "reverse" your score on items 1,6, and 10 by subtracting what you wrote from the number 4, and write those numbers in the margin. Finally, add up your responses to all 25 items (using your "reversed" scores on 1, 6, and 10). The total will be a number between 0-100. For more information see: http://people.virginia.edu/~jdh6n/disgustscale.html

The facial expression of disgust was downloaded from

https://commons.wikimedia.org/wiki/File:Disgust_expression_cropped.jpg

and was originally posted at

https://www.flickr.com/photos/iamagenious/2490996809

This image is published under an open source license.

The screenshot from "The Disgust Scale Home Page" at

http://people.stern.nyu.edu/jhaidt/disgustscale.html

is copyrighted and is covered under the "Fair use" provisions of copyright law. Your use of this image may or may not be covered.

There are 25 items, each rated between 0 (strongly disagree) and 4 (strongly agree).

Case study #5 - Disgust scale revised

0 = Strongly disagree (very untrue about me)
1 = Mildly disagree (somewhat untrue about me)
2 = Neither agree nor disagree
3 = Mildly agree (somewhat true about me)
4 = Strongly agree (very true about me)

____1. I might be willing to try eating monkey meat, under some circumstances.
____2. It would bother me to be in a science class, and to see a human hand preserved in a jar.
____3. It bothers me to hear someone clear a throat full of mucous.
____4. I never let any part of my body touch the toilet seat in public restrooms.
____5. I would go out of my way to avoid walking through a graveyard.
____6. Seeing a cockroach in someone else's house doesn't bother me.

Here is a closer look at the first six items on this scale. Monkey meat? Mucus? Cockroaches? They are all pretty disgusting if you ask me. But some people keep cockroaches as pets.

Break #4

- What you've seen so far
  - Five case studies
- What is coming next
  - Three dichotomies of measurement

So far, you've seen five case studies where you may want to assess reliability and validity. These are very different types of measurements and you can classify them into three important dichotomies. These three dichotomies will help you assess the best approach to take for assessing reliability and validity.

First dichotomy: Who is the measurer?

– Self reported outcomes
  - Also know as patient reported outcomes
– Researcher evaluations
  - Only when concerned about subjectivity

I think it is a bit unfair, but there is a lot of distrust of self reported outcomes among researchers. Why not believe what the patient says about himself or herself? Part of it might be that a patient's answers could potentially be influenced by their mood. They might also be influenced by a desire to look good. They may want to give an answer that they think the interviewer wants to hear.

There is also a belief that reported outcome measures vary too much from one individual to another. Some people are stoic to a fault and others will complain endlessly at the drop of a hat.

It is worth noting that these factors also influence researcher observation, but researchers don't like it when you point this out to them. It's mostly a good thing that researchers require a high level of scrutiny of reported outcomes, but perhaps other measures deserve just as high a level of scrutiny.

Second dichotomy: How many pieces?

- Composite scores
  - Sum or average
- Single measurement

Many measures in research are composites of one or more items. These individual items are scored and added up to get a total. Sometimes you average them instead of summing them. That's a minor distinction. In general, a composite measure involves evaluating something from many different angles. If you do this well,, this can be a very effective approach. But you do need to be careful.

In contrast, some measures are just a single number and there is no summing or averaging involved. There is just a single question or a single evaluation. It is a difficult decision to choose whether to develop a measurement using a composite scale versus a single measurement, and it depends a lot on what you are measuring.

The general consensus among researchers is that you get a better picture of something with a composite scale, but some measures do not warrant this extra effort.

Third dichotomy: is the measure soft or hard?

– Psychological or social constructs
  - Created and accepted by you and me
  - Impossible to observe directly
  - Examples: stress, anxiety

Constructs are measurements of something that is considered soft. They do not have an indepent existence but are created by society. Almost all measures in psychology, such stress and anxiety, are constructs because you do not have the ability to peer inside someone's brain.

There is a gray area, of course. We can indeed peek inside the brain to some extent with new imaging equipment, but I would argue that even these measures are not direct measures.

Third dichotomy: is the measure hard or soft

- Biological or physical measure
  - Has an objective reality
  - Potential for direct observation
  - Example: obesity, dementia

Some people argue that everything is socially constructed, but this is rather silly, in my opinion. Some things have an objective reality and were not created by us. Obesity is a physical concept that can be measured using body mass index, percentage of body fat, waist to hip ratio and many others. They all represent a physical manifestation of excess weight. Dementia also has a physical manifestion. You can’t measure it without autopsy perhaps, but there is a physical change in your brain that causes dementia.

Break #5

- What have you learned
  - Dichotomies of measurement
- What's coming next
  - Dichotomy examples

Examples of self reported outcomes

Neighborhood Environment Survey

Pain scale

The Neighborhood Environment Scale is a self report. You could get a researcher to drive through a neighborhood and assess its environment, but asking the residents themselves is faster, and possibly more accurate.

The pain scale is also a self report. If you are assessing pain in adults and children, you normally rely on self-reports. A researcher can assess pain in infants who are nonverbal by looking for facial expression, leg and arm movements, and other signs. In animals, it depends on the animal, but there are often clues in their appearance and behavior. Still, it is preferred to use self report for pain and you can get reliable measures from very young children.

Examples of researcher evaluation

Apgar score

Boston Bowel Prep Score

You cannot ask for a self report in a newborn infant. Actually, this newborn is offering a fairly strong assessment of his/her condition, but normally an outsider will assess how an infant is doing immediately after birth.

You also do not see self reports in assessing the bowel preparation, unless you are self-administering a colonoscopy. Definitely not recommended!

Examples of composite scores

Neighborhood Environment Survey

Apgar score

The neighborhood environment scale is a composite measurement because you ask 18 separate questions. The Apgar score is also a composite measurement because you rate the infant on five separate items.

Examples of single measurements

Pain scale

Boston Bowel Prep Score

The pain score is a single measurement. You do not ask about five different types of pain and add them up. Likewise, the Boston Bowel Prep score is a single measurement.

Examples of constructs

Apgar score

Neighborhood Environment Survey

There could be some debate about this, but I would call the Apgar score a social construct. You are measuring distress associated with the birth process. Distress is an abstraction that you cannot measure directly. The Neighborhood Environment Survey is also measuring a rather abstract concept, the quality of your living environment. What you and I would consider to be a good environment is something that we construct. This is an exmaple of a social construct.

Examples of biological or physical measurements

Boston Bowel Prep Score

Pain scale

The Boston Bowel Prep Score is a measure of a physical phenomenon, the visibility of the walls of your colon.

Pain is on the border, but I would call it a biological process rather than a psychological construct because it involves the transfer of information across your nervous system.

Pop quiz

- Is the disgust score
  - a self report or
  - a researcher evaluation?
- Is the disgust score
  - a composite measure or
  - a single measurement?
- Is the disgust score
  - a psychological or social construct or
  - a biological or physical measurement?

I put in the disgust scale because it is the perfect example of a psychological construct.

But is it a self report or a researcher evaluation?

Is is a composite measure or a single measurement?

Break #6

- What have you learned
  - Dichotomy examples
- What is coming next
  - Parallel forms
  - Split half reliability

You’ve seen five case studies and discussed where they fit among the three dichotomies. Next, you’ll see various approaches to reliability and when you can or cannot use them.

Measurement Reliability

- Synoynms: consistency, precision, stability
- Classical test theory
  - Observed value = True value + Measurement error
  - This is a purely hypothetical model
- Reliability coefficient
  - Variance of true values / Variance of measured values
- Depends on your population

When you measure something, you want that measurement to be consistent, precise, and stable. You don't want something that changes as the phases of the moon change. You don't want a measurement that changes depending on who the attending physician is. You don't want a measurement that changes depending on any environmental factors that are extraneous to what you are measuring.

If your measure is not stable, then you have difficulty in assessing whether a change in that measurement is due to your intervention or due to the phases of the moon.

Most measures of reliability rely on the true value model. This model says that the observed value of a measurement is equal to the true value plus measurement error. A measurement is reliable if the measurement error is small. Since the true value is almost always unknown, it is only a hypothetical model.

Your book talks about a reliability coefficient which is the variance of the true scores in a population divided by the variance of the observed scores in a population. Measurement error guarantees that the numerator is always less than or equal to the denominator. The reliability coefficient is equal to one only if there is no measurement error.

You should not be too surprised to find out that the reliability coefficient is a hypothetical value and can never be measured directly. But there are several indirect approaches.

One thing you need to keep in mind is that the reliability coefficient is dependent on the population it is based on. This is very important. Change the population and you change the reliability coefficient. Something with a great reliability coefficient in a population of college students might be terrible in a population with limited literacy skills, for example.

Measurement Reliability

- No measurement is perfectly reliable
  - Strive for 0.7 or higher in research
  - 0.6 is “borderline”.
  - Might require 0.9 or higher for individual decisions

Since no measurement is ever conducted without some measurement error, no measurement has perfect reliability. You need to make a value judgement about whether the deviation from the truth is small enough that you can safely ignore it.

There are some informal standards for reliability. These choices can seem a bit arbitrary, but they are fairly well accepted in the research community.

In order for a measurement to be reliable enough to use in a research setting, where you are trying to characterize how a group of people are affected by an intervention, you would like a reliability coefficient of 0.7 or higher. It’s not perfect, but the individual measurement errors would be averaged out when you compute group means.

But if you are making decisions that might affect an individual, then you’d want a much higher level of reliability. Individual decisions might involve acceptance into a training program, for example. You would hate to see a large measurement error dominate the decision about an individual. In these settings, a reliability coefficient of 0.9 or higher might be asked for.

For the record, some sources say that your reliability could go as low as 0.6 and still be okay. Other sources disagree. If you have such a value, go ahead and report it using a term like “borderline” or “marginal” and hope that your peer-reviewer isn’t a stickler for this sort of thing.

Reliability is usually established when a measure is developed. When you go about using a measure, look at what’s already been published. Make sure it used in a context similar to yours. It’s a whole lot easier to find a measurement that is already proven to be reliable than to develop your own measure and then establish its reliability.

Indirect measures of the reliability coefficient

- Test-retest
- Interrater
- Internal consistency

Even though the reliability coefficient cannot be measured directly, you can usually get at it indirectly. What you do it take two measurements where the true value is expected to stay reasonably constant. If the two observed values correlate well, then you have indirect evidence that the measurement error is small.

Parallel forms

- "No man ever steps in the same river twice, for it's not the same river and he's not the same man."
  - Heraclitus
- Used when you can't run the same measurement twice.
- How to develop parallel forms
  - Change the question order
  - Minor changes to the wording
- Difficult to develop two parallel forms of the same measurement.

Sometimes the very act of measuring someone changes that person. I do this all the time. I put a quiz up each week, not to test you so much as to reinforce some of the key messages in my videos. The questions are not intended to challenge you and assess how much you've learned. Having come up with an answer, that helps you remember the key concepts better.

The opposite tendency can occur as well. The novelty of answering questions wears off over time and people may grow tired or bored and not answer the exact same questions a second time.

How likely is this to happen? It depends a lot on what is being measured. Measures of knowledge and understanding are more likely to have carry over effects.

In some settings, you can create a second version of your measurement by making minor changes. This could be in the wording or the ordering of the questions.

How much of a change do you want? Too little and you still have problems with carry over. Too much and you are no longer measuring the same thing.

The parallel forms measure of reliability is not used that frequently, because it just about kills you to get one version of a measurement up and running. Who wants to develop two parallel forms. It’s worth introducing here, though, because it helps you understand the next three forms of reliability.

Split half reliability

- Only used for composite measurements
- Split into halves, correlated
  - Odd-even split
  - Random split
- Brown-Spearman adjustment

If your measurement is a composite measure, then you can look at the correlation of the individual components to assess reliability.

You could split the measure in half, calling the even numbered items the first form and the odd numbered items the second form. The correlation between the odds and the evens is a measure of reliability.

It doesn’t have to be evens versus odds. You might want to assign items randomly to the first half versus the second half.

You do need to be careful, though. The reliability of a composite measurement is frequently thought to be related to the number of items in the composite. The greater the number of items, the greater the reliability. So if you artificially shorten the measurement, you are underestimating reliability. There is a simple adjustment, called the Spearman-Brown formula that most researchers use when looking at split half correlations.

Break #7

- What have you learned
  - Parallel forms
  - Split half reliability
- What's next
  - KR-20
  - Cronbach's alpha

Kuder-Richardson 20 formula

- Only for composite measures with binary items
- Book's formula is confusing
  - S^2 and σ^2 used interchangably
  - Subscripts are missing
- Correct formula
  - $X_i = \Sigma_i B_{ij}$
    - Where B_{ij} is binary (0 or 1)
  - $KR - 20 = \frac{1}{I-1}(1 - \frac{\Sigma_i \hat{p}_i \hat{q}_i}{S^2})$
    - where $S^2 = \Sigma_i (X_i - \bar{X})^2$,
    - $\hat{p}_j = \Sigma_i B_{ij}$,
    - $\hat{q}_j = \Sigma_i \left(1 - B_{ij}\right)$

Another measure for reliability, Kuder-Richardson 20, is used for composite measures, but only those composite measures that have binary items. Your book does a poor job explaining this, and the notation is inconsistent.

Kuder-Richardson 20 interpretation

– KR-20

- $\frac{1}{I-1}(1-\frac{\Sigma_i \hat{p}_i \hat{q}_i}{S^2})$
  - $\Sigma_i \hat{p}_i \hat{q}_i$ is a theoretical minimum variation
  - S^2 is observed variation
  - $S^2 = \Sigma_i \hat{p}_i \hat{q}_i$ implies randomness
  - $S^2 > \Sigma_i \hat{p}_i \hat{q}_i$ implies internal consistency

If you are curious, the formula is comparing a theoretical minimum variation, a variation computed using independent Bernoulli random variables, but with different p's and q's for each item. Strictly speaking, this is not accurate, but a smaller value than the sum of the pq's could only occur if there is negative correlation among the individual items, and this implies almost a conspiracy among the individual items to make things as bad as possible for you.

You compare this to the variation observed among the total scores in the sample. If the observed variation is equal to the theoretical minimum, the individual items are behaving randomly, with no internal consistency. This means that any split halves that you could compute would have next to no correlation.

If there is much more observed variation, that means that people show positive correlations. Low on one item means low on most of the other items and high on one item means high on the other items. This positive correlation is a measure of internal consistency.

Keep in mind that if you have a different population, the minimum variation would stay the same, but the observed variation might change. So a measure that is reliable in one population might prove to be not reliable in a different population.

Cronbach's alpha, formula

- Used for composite measurements with continuous items
- Book's formula is confusing
  - ΣS^2 should be ΣS_i^2
- Correct formula
  - $X_i = \Sigma_j X_{ij}$
  - $\alpha = \frac{1}{I-1}(1 - \frac{\Sigma_i S_i^2}{S^2})$
    - where $S_i^2 = \Sigma_j \left(X_{ij} - \bar{X}_j\right)^2$, and
    - $S^2 = \Sigma_i (X_i - \bar{X})^2$, and

A similar measure, Cronbach's alpha is used for composite measures, but does not require the individual items to be binary.

Again, your book does a poor job explaining this, and the notation is confusing.

Cronbach's alpha, interpretation

– Cronbach's α
  - $\frac{1}{I-1}(1-\frac{\Sigma_i S_i^2}{S^2})$
    – $\Sigma_i S_i^2$ is a theoretical minimum variation
    – S^2 is observed variation
    – $S^2 = \Sigma_i S_i^2$ implies randomness
    – $S^2 > \Sigma_i S_i^2$ implies internal consistency

– Cronbach's alpha is NOT a measure of unidimensionality

Just like Kuder-Richardson 20, Cronbach's alpha computes a a theoretical minimum variation. This time it is a sum of the variances for the individual items. Strictly speaking, this is not accurate, but a smaller value than the sum of the variances implies a deep and dark conspiracy against you by the individual items in your composite.

You compare this to the variation observed in the total score. If the two values are close, that tells you that the individual items are more or less independent of each other and that any split halves that you might compute would have little or no correlation.

If there is much more observed variation, that means that people show positive correlations. Low on one item means low on most of the other items and high on one item means high on the other items. This positive correlation is a measure of internal consistency.

Some people confuse the concept of internal consistency with uni-dimensionality. Uni-dimensionality means that all of the items are measuring the same construct. If they are, then Cronbach's alpha will be large. But you can also get a large value for

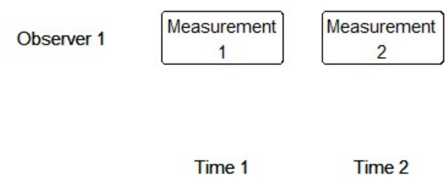
Cronbach's alpha, even when the items are measuring multiple constructs, especially when you have a large number of items. Dimensionality can only be measured using some form of factor analysis.

Break #8

- What have you learned so far.
  - KR-20
  - Cronbach's alpha
- What is coming next
  - Test-retest reliability
  - Inter-rater reliability

Wow. That's a lot to digest. Don't be afraid to ask me questions about reliability. Let's take a break here. We talked about measures of internal consistency and why I don't like them. We also talked about some practical advice: report reliability measures from your literature review and measure reliability, if you can, in your current study.

Test-retest reliability



Test-retest reliability (also called repeatability)

- Correlation of two measurements separated by time
- Length of time interval is critical
  - No carry-over
  - No changes in the true score
- Useful for composite scores and single values
- Useful for self-report and researcher evaluation
- Not possible for some measures

Test-retest reliability is the correlation coefficient of two measurements taken at different times. This is also known as repeatability.

The correlation coefficient between the two measurments is an estimate of the reliability coefficient.

The time interval is critical here. You don’t want two measurements that are so close together that the measurement error for the first measurement is correlated with the measurement error of the second measurement.

Suppose you are measuring your patient’s knowledge about a disease. If you give the same test only a few minutes apart, your patient will remember his/her answers to the first test when answering the second test.

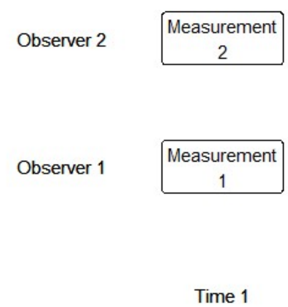
So you want a long enough time interval that there is no carry over effect.

But too large an interval is also problematic. You want to make sure that the true score is the same (or very close)

Over what size interval would you expect the measure to be stable? It depends on what you are measuring. Intelligence is likely to be stable along long time frames but mood changes rapidly.

Test-retest reliability works well for any type of measure. But it can’t always be applied. The Apgar score has to be measured at the time of birth. If the baby is blue, limp, and unresponsive after three hours, that’s not quite the same as the same occuring right after birth.

Inter-rater reliability



Inter-rater reliability

- Used for researcher evaluations only
- Simplest case
  - Two independent raters
  - Ratings for every patient
- Analysis
  - Intraclass correlation or Cohen's Kappa
- Extensions

When the researcher does the evaluation and there is concern that a subjective element may creep in and cause measurement error. Your observed score might be higher or lower depending only on who is rating you.

Reliability is pretty simple to measure in this setting, if you have the resources. Just get two raters and have them both compute the measurement. If the correlation between the two raters is high, you have good reliability.

Rather than computing a direct correlation, inter-rater reliability is usually computed as an intra-class correlation. The intraclass correlation generalizes naturally to more complex settings.

If your measurement is binary (note the entire measurement is binary, which is not the same as saying that the individual components of a composite score are binary), then a different statistic, Cohen's Kappa is used. Like the intraclass correlation, there are extensions of Kappa to multiple raters.

You can't always have all the raters rate all the patients, especially if you have more than two raters. There are extensions to cases where you have random assignments of patients to different raters, but the formulas are tricky.

Practical guidance on reliability

- Is there previous literature?
  - Report their reliability coefficients
- Is your setting similar?
  - Different demographics?
  - Different cultural norms?
  - Different literacy?
  - Different language?
- Compare to reliability in your sample
  - Test-retest and inter-rater reliability preferred.
  - 0.7 or higher

You should include a discussion of reliability in your literature review. Cite the reliability coefficients in previous work, as it adds to the credibility of your proposed research.

But take a step back and ask if you can extrapolate safely to the research setting that you propose. Recall the hypothetical reliability coefficient. It compared the variation of the true score to the variation of the observed score across patients in the population you are studying. If your population is quite different than the populations in your literature review, you have no guarantee that a measurement proven to be reliable in previous work will continue to stay reliable in your setting.

Some differences to look for are differences in the demographics of your population, differences in cultural norms and expectations, differences in literacy levels (especially for measurements that require your patients to read and respond to a questionnaire).

If you are measuring something that requires translation to a different language, keep in mind that not all concepts translate well from one language to another. Sometimes it helps to pay for a second and independent translation back to the original language. If there are discrepancies, then maybe it was in the back-translation, but more likely, you are asking for a different type of information in your new language without realizing it.

If you can, incorporate a measure of reliability into your study. There are two reasons for this. First, your setting may be different enough to raise concerns. Getting a current measure of reliability helps to allay those concerns. Second, reliability is never quite perfect, because all of the measures of reliability are indirect measures. Your effort to assess reliability will supplement the previous work on reliability and make things a bit easier for future researchers.

I have a strong preference for test-restest reliability or inter-rater reliability, if you can get it. The other measures of reliability, parallel forms, the split half correlation, Kuder-Richardson 20, and Cronbach's alpha are measures of the internal homogeneity of your composite measure. In my mind they are a poor substitute for test-retest reliability or inter-rater reliability.

I do not like these measures. Let me restate that. I despise these measures. They are simplistic and fail to measure what I think are the important features of reliability (stability over time and consistency between raters). I think people use them mindlessly and fail to recognize that they are measuring something very limited.

If you can't measure reliability using a test-retest approach or using inter-rater reliability, then go ahead and use these other approaches. But they are a pale substitute in my opinion.

The general target value for a reliability coefficient is 0.7 or higher. You might get by with a reliability coefficient of 0.6, but don't count on it.

Break #9

- What have you learned so far.
  - Test-retest reliability
  - Interrater reliability
- What is coming next
  - Face validity
  - Content validity
  - Response process evidence

Types of measurement validity

– Face validity
– Content validity
– Response process evidence
– Criterion validity
– Construct validity

There are several different ways to establish validity. I’ll talk about each of these in turn.

Face validity and content validity

- Face validity
  - Opinions from your patients
  - Subjective and unquantifiable
- Only used for composite measures

- Content validity
  - Opinions from outside experts
  - Subjective and unquantifiable
- Only used for composite measures

There are varying definitions of face validity and content validity. Let me share the defintions that I like. This is my class and I get to dictate the rules. But I'll let you know what others define these two terms as.

Face validity is information from your patients, typically for a composite measurement. They look at the individual items in your composite measurement and tell you the ones that don't really belong. They should also tell you about items that are missing in your composite measurement that you should include. Face validity is a totally subjective approach and to some people it seems like letting the inmates run the asylum.

To be fair, face validity is an important step in establishing validity, but it should probably not be the only step.

Content validity is information from content experts rather than your patients. But otherwise, it is the exact same thing. The experts look at your composite measurement and tell you that certain things need to go and other things need to be added.

Now, who is an expert? It can be anyone, really. Normally, you would use credentials like a degree and a publication record in the area to establish that someone is qualified to tinker with your measurement.

Both face validity and content validity are purely qualitative. There is no numeric measure or score that you get from these types of validity. You do have to establish consensus, if you seek face validity or content validity from more than one source, but this is usually established qualitatively.

There are structured ways to get information about face and content validity from your patients and from an expert panel, such as the Delphi method. You can use these methods, or you could just use a structured interview.

Even though these approaches are soft, they are well worth the effort.

Now some people use the terms face validity and content validity interchangably. Your book says that face validity is just looking at the measure and giving a general impression while content validity requires delving into the individual items of a composite measure.

I won't say that your book is wrong, but your book is wrong. Actually, I'm probably wrong, but I'm your teacher and you're stuck with my opinion, at least until the semester ends.

Seriously, if there is a disagreement in the research community about how to establish validity, what you do is you do it your way, but with the expectation that when you submit your paper to peer review, plan for the possibility that the peer reviewer will ask you to define things their way. It's normally not a good idea to fight a peer-reviewer, especially when there is no consensus in the research community, unless they are asking for something that is seriously wrong and misleading.

Now your teacher, on the other hand, you can argue with him until the cows come home. He actually will enjoy the argument and you won't get him to shut up about the varying research perspectives.

Question.

– Should Statisticians work on problems that are subjective and unquantifiable?

Answer.

– Yer darn tootin!

Response process evidence

– Observe the process
  - Watch as patients fill out the form
  - Ask questions along the way
  - Monitor response times
  - Encourage them to think aloud
– Supplement with interview
– Goal is to identify problematic elements
  - Confusion, misunderstandings, language issues

Response process validity is the direct observation of patients as they fill out the survey that you are using for measurement. You can think of it as part of the face validation, or you can call it an additional type of validation. I like the latter because it sounds more impressive.

There's nothing too difficult about this. As you observe the process, ask questions, see if there are any items that seem to take too long to answer. Encourage your patients to talk aloud as they are working. If you want to get really fancy, you can use eye tracking to see if someone is losing focus or getting distracted.

You can supplement this process with an interview afterwards. Your goal in this exercise is to identify items that are confusing or ambiguous, or which seem to draw the wrong type of response. Look especially for issues which may come from the use of excessively technical language.

You can do this sort of exercise with experts as well as patients. Ask your experts to pretend that they are patients and get them to fill things out, talk aloud, and ask them questions along the way.

Break #10

– What have you learned so far.
  - Face validity
  - Content validity
  - Response process evidence
– What is coming next
  - Criterion validity
  - Construct validity
  - Validity of diagnostic tests

Let's stop again. You've seen a general definition of validity and specific examples of face and content validity. Next we'll discuss criterion validity and construct validity.

Criterion validity

- Comparison to external criterion
  - Represents “truth”
  - Not always available
- Predictive evidence
  - Measurement in the future
  - Be careful about dropouts
- Concurrent evidence
  - Measured at the same time

Criterion validity is the most straightforward approach to establishing validity. You want to see how well your measurement corresponds with what it’s supposed to measure. So include what your supposed to measure and see how strongly it correlates with what you want to measure.

This isn’t always possible, of course, but if you can measure truth then go for it!

Now, why, might you ask yourself, would you use a measurement that correlates well with truth when you can measure truth directly? It probably has something to do with time or money. You can measure the truth but it costs too much to do it in a big study. Or it takes way too long. So you run a smaller study where you measure truth, show that your cheaper and faster measurement correlates well with the truth, and then you can save a whole bunch of time and money in the big study.

Your evidence for validity is predictive evidence if the truth represents something that occurs in the future, meaning after your measurement is taken. In the big study, you can’t wait around and wait for the truth to reveal itself. But in a smaller study, you might have that luxury.

It’s important in using predictive evidence that you don’t have dropouts, especially if those dropouts tend to differ from those who do provide you with data.

Your book offers an interesting example of this with standardized testing for college admission. A school might want to correlate an SAT score, for example, with the grades that a student gets after one year of college. Easy to do, but think about the dropouts. A college, for the most part, is going to admit only those people who score above a cutoff for the SAT. You lose information about those who scored low on the SAT and are left only with those students in a narrow range of SAT scores. It’s even worse if the students who score super high on the SAT decide to attend a more prestigious university than your little podunk college.

Another example is using criterion validity for a test intended to diagnose disease. Suppose you have a test that can predict appendicitis. Patients who score high on the measurement, you send them straight to the OR, so you can cut out the appendix before it ruptures. But what about the patients who score low. They probably don’t have appendicitis, but you don’t know. They won’t volunteer to get cut open in the name of science.

Predictive evidence can sometimes take too long, so you may want to use concurrent evidence, evidence that you can collect at the same time as your measurement. Your book suggests that you ask colllege students at the end of their first year to re-take the SAT and see how that re-take correlates with the grades they are just receiving. It’s not perfect, but it certainly takes less time.

The other application of concurrent evidence is when you don’t have a direct measure of truth, but you have an already validated measure of truth that you can collect concurrently with your new measure. The assumption here, as earlier is that your new measure is cheaper or faster than the currently used and validated measure. If you correlate well with an already validated measure, and that validated measure has already been shown to correlate well with the truth, then you have indirectly established criterion validity.

Now this approach has limits. You can never get quite as much evidence of validity as the already validated measurement has.

Construct validity

- Used for a psychological construct
- No direct measure of the truth exists
- Define associations consistent with your constuct
  - Does your measurement show the expected association?
  - Known as convergent evidence
- Define non-associations with your construct
  - Does your measurement also show non-association?
  - Known as discriminant or divergent evidence

Construct validity is when you are developing a psychological construct and you don't have a direct measure of the construct you are trying to measure. What you do have is various associations and non-associations that your construct is expected to have. You develop these using your deep thinking power or maybe just a bit of common sense. If your measurement shows the same associations and non-associations that you would expect your construct to have, you have established construct validity.

Alternative framework for validity

- – Content
- – Response processes
- – Internal structure
- – Relations to other variables
- – Consequences

Your book cites a different standard for establishing validity. It's a good standard, but not used that commonly in my experience. Read this on your own.

Validity of diagnostic tests

- Sensitivity
  - A test's ability to obtain a positive result when the target condition is really present
- Specificity
  - A test's ability to obtain a negative result when the target condition is really absent

Diagnostic tests are a special example of validation. It is essentially criterion validity using predictive evidence. Since the diagnostic measurement is binary and the criterion is binary, you can summarize the results using a two by two table. I won't go into any detail on sensitivity and specificity except to mention that I can never remember which is sensitivity and which is specificity.

Break #11

- What have you learned so far?
  - Criterion validity
  - Construct validity
  - Validity of diagnostic tests
- What's next
  - Revisit the case studies

Case study #1

Here is the Neighborhood Environment Survey again.

This is a self report and it is also a composite measure.

Case study #1 - Neighborhood environment survey

R.M. Crum et al. / Drug and Alcohol Dependence 43 (1996) 155–161 157

Table 1
Items[a] used to assess neighborhood disadvantage

	True[b]	False[b]
Within walking distance of my house there is a park or playground where I like to walk and enjoy myself, playing sports or games	0	1
There are plenty of safe places to walk or play outdoors in my neighborhood	0	1
Every few weeks, some kid in my neighborhood gets beat-up or mugged	1	0
Every few weeks, some adult gets beat-up or mugged in my neighborhood	1	0
In my neighborhood, I see signs of racism and prejudice at least once a week	1	0
In my neighborhood, many yards and alleys have broken bottles and trash lying around	1	0
I have seen people using or selling drugs in my neighborhood[c]	1	0
In the morning or later in the day, I often see drunk people on the street in my neighborhood	1	0
Most adults in my neighborhood respect the law	0	1
There are abandoned or boarded-up buildings in my neighborhood	1	0
I feel safe when I walk around my neighborhood by myself	0	1
The people who live in my neighborhood often damage or steal each other's property	1	0
The people who live in my neighborhood always take care of each other and protect each other from crime	0	1
Almost everyday I see homeless people walking or sitting around in my neighborhood	1	0
In my neighborhood, the people with the most money are the drug dealers	1	0
In my neighborhood, there are a lot of poor people who don't have enough money for food and basic needs	1	0
For many people in my neighborhood, going to church on Sunday or religious days is a very important activity	0	1
The people who live in my neighborhood are the best people in the world	0	1

Data from the Prevention Research Center, Johns Hopkins University School of Hygiene and Public Health.
[a] This scale was adapted and modified from one initially developed by Elliott and colleagues for use in the National Youth Survey (28).
[b] The scoring of true and false answers for the individual items is shown in these columns. The item scores are summed yielding an overall scale score from 1 to 18, with higher scores indicating a greater degree of neighborhood disadvantge.
[c] This item was dropped from the scale for analyses reported in Table 3, model 2.

Case study #1 - Neighborhood environment survey

– Reliability - What you can't do
  - Inter-rater reliability
– Reliability - What you can do
  - Test-retest reliability
  - Cronbach's alpha

This is a composite measure and a self report. You can't compare two or more raters because there is only one "self."

Cronbach's alpha is used for components that are continuous and this scale has binary (true/false) statements.

It is easy to run a test-retest reliability study. Neighborhoods don't change overnight, so it would be fine to wait a week or so.

You can also run Cronbach's alpha here. If you are a stickler for detail, the reliability measure in this case is better described as Kuder-Richardson 20 because the individual components are binary.

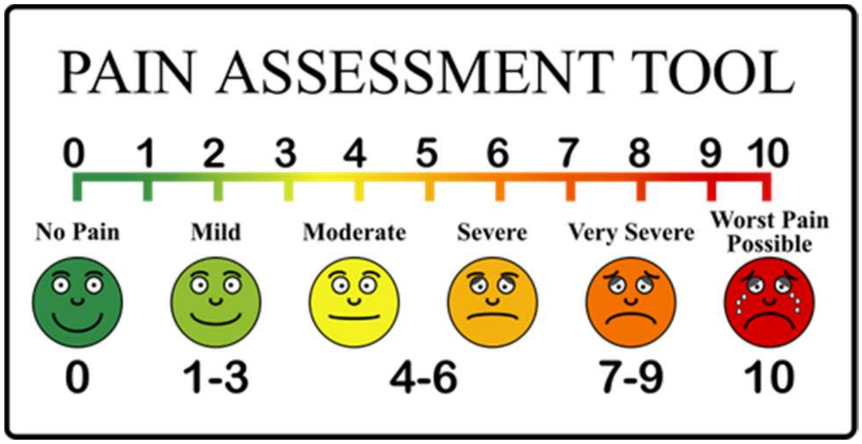
Case study #1 - Neighborhood environment survey

- Validity - What you can't do
  - Criterion validity
- Validity - What you can do
  - Face/content validity
  - Response process validity
  - Factor analysis
  - Construct validity

Case study #2

This is a pain scale again. It is a self report, but it is a single measure rather than a composite.

Case study #2 - Pain scale



Case study #2 - Pain scale

- Reliability - What you can't do
  - Inter-rater reliability
  - Cronbach's alpha
- Reliability - What you can do
  - Test-retest reliability

Like the first measurement, this self report scale cannot compare two or more independent raters. It is a single measurement, so you can't apply Cronbach's alpah or KR-20.

Test-retest reliability works well here, but you have to make sure that you are quick. A narrow time interval between the test and the re-test is important if you are looking at acute pain.

Case study #2 - Pain scale

- Validity - What you can't do
  - Face/content validity
  - Response process validity
  - Factor analysis
- Validity - What you can do
  - Criterion validity
  - Construct validity

Case study #3

This is the Apgar score again. It is a composite measure collected by the researcher and not by self report.

Case study #3 - Apgar score

APGAR Evaluation of newborn infants

SIGN	0	1	2
Heart rate	Absent	Below 100	Over 100
Respiratory effort	Absent	Slow, irregular	Good, crying
Muscle tone	Limp	Some flexion	Active motion
Reflex*	No response	Grimace	Cough or sneeze
Color	Blue, Pale	Body pink, Extremities blue	Completely pink

- 7 to 10 is normal
- 4 to 6 is moderately depressed
- 0 to 3 needs immediate resuscitation

*Response to catheter in nostrils

Case study #3 - Apgar score

– Reliability - What you can't do
  • Test-retest reliability
– Reliability - What you can do
  • Inter-rater reliability
  • Cronbach's alpha

Because timing is important, you cannot evaluate the Apgar score at one minue and at two hours. You also can't use KR-20 because it is not binary.

Inter-rater reliability is very easy and very useful here. Have two raters at the scene of the birth and ask them to estimate the Apgra score. No peeking! Then correlate the responses.

Cronbach's alpha is really intended for continuous components, and values of 0, 1, and 2 are not really on a continuum. But there is nothing terribly wrong with pretending it is continuous.

Case study #3 - Apgar score

- Validity - What you can't do
  - Construct validity
- Validity - What you can do
  - Face/content validity
  - Response process validity
  - Criterion validity

Construct validity is hard to apply here. There is no theory of what an Apgar is or is not associated with.

It is a composite measure so you can have experts review the individual components. You can also watch as someone answers the five components of the Apgar score. There are several predictive criterion. Does a low Apgar score predict infant mortality?

You can run a factor analysis on the Apgar score because it has multiple comonents.

Case study #4

This is an example of a physician report. No self report is available here. But you still want to examine reliability and validity because this does have the potential to be perceived as subjective.

Note also that, unlike the Apgar score, this is not a composite measure. There is a single number that you get.

Case study #4 - Boston Bowel Prep Score

0 = Unprepared colon segment with mucosa not seen due to solid stool that cannot be cleared.

1 = Portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen due to staining, residual stool and/or opaque liquid.

2 = Minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well.

3 = Entire mucosa of colon segment seen well with no residual staining, small fragments of stool or opaque liquid. The wording of the scale was finalized after incorporating feedback from three colleagues experienced in colonoscopy.

Case study #4 - Boston Bowel Prep Score

- Reliability - What you can't do
  - Test-retest reliability
  - Cronbach's alpha
- Reliability - What you can do
  - Inter-rater reliability

Case study #4 - Boston Bowel Prep Score

- Validity - What you can't do
  - Face/content validity
  - Response process validity
  - Criterion validity
- Validity - What you can do
  - Construct validity

Case study #5 - Disgust Scale Revised

Please indicate how much you agree with each of the following statements, or how true it is about you. Please write a number (0-4) to indicate your answer:

0 – Strongly disagree (very untrue about me)
1 – Mildly disagree (somewhat untrue about me)
2 – Neither agree nor disagree
3 – Mildly agree (somewhat true about me)
4 – Strongly agree (very true about me)

____1. I might be willing to try eating monkey meat, under some circumstances.
____2. It would bother me to be in a science class, and to see a human hand preserved in a jar.
____3. It bothers me to hear someone clear a throat full of mucous.
____4. I never let any part of my body touch the toilet seat in public restrooms.
____5. I would go out of my way to avoid walking through a graveyard.
____6. Seeing a cockroach in someone else's house doesn't bother me.
____7. It would bother me tremendously to touch a dead body.
____8. If I see someone vomit, it makes me sick to my stomach.
____9. I probably would not go to my favorite restaurant if I found out that the cook had a cold.
____10. It would not upset me at all to watch a person with a glass eye take the eye out of the socket.
____11. It would bother me to see a rat run across my path in a park.
____12. I would rather eat a piece of fruit than a piece of paper
____13. Even if I was hungry, I would not drink a bowl of my favorite soup if it had been stirred by a used but thoroughly washed flyswatter.
____14. It would bother me to sleep in a nice hotel room if I knew that a man had died of a heart attack in that room the night before.

How disgusting would you find each of the following experiences? Please write a number (0-4) to indicate your answer:

0 – Not disgusting at all
1 – Slightly disgusting
2 – Moderately disgusting
3 – Very disgusting
4 – Extremely disgusting

____15. You see maggots on a piece of meat in an outdoor garbage pail.
____16. You see a person eating an apple with a knife and fork
____17. While you are walking through a tunnel under a railroad track, you smell urine.
____18. You take a sip of soda, and then realize that you drank from the glass that an acquaintance of yours had been drinking from.
____19. Your friend's pet cat dies, and you have to pick up the dead body with your bare hands.
____20. You see someone put ketchup on vanilla ice cream, and eat it.
____21. You see a man with his intestines exposed after an accident.
____22. You discover that a friend of yours changes underwear only once a week.
____23. A friend offers you a piece of chocolate shaped like dog-doo.
____24. You accidentally touch the ashes of a person who has been cremated.
____25. You are about to drink a glass of milk when you smell that it is spoiled.
____26. As part of a sex education class, you are required to inflate a new unlubricated condom, using your mouth.
____27. You are walking barefoot on concrete, and you step on an earthworm.

The DS-R (Disgust Scale-Revised), Haidt, McCauley, & Rozin, 1994; Modified by Olatunji et al., in press.
To calculate your score: First, put an X through your responses to items 12 and 16 (these items don't count). Then "reverse" your score on items 1,6, and 10 by subtracting what you wrote from the number 4, and write those numbers in the margin. Finally, add up your responses to all 25 items (using your "reversed" scores on 1, 6, and 10). The total will be a number between 0-100. For more information see: http://people.virginia.edu/~jdh6n/disgustscale.html

Case study #5

– What do you think?

– What measures of reliability?
  - Test-retest reliability
  - Inter-rater reliability
  - Measures of internal consistency

– What measures of validity?
  - Face/content validity
  - Response process validity
  - Criterion validity
  - Construct validity

Conclusion

- What you've seen today
  - Internal validity
  - External validity
  - Measurement reliability
  - Measurement validity
  - Three dichotomies of measurement
  - Five case studies

Different measures of reliability and validity apply depending on whether your measurement is a self report or not and depending on whether it is a composite measure or not and whether is is a construct or not.