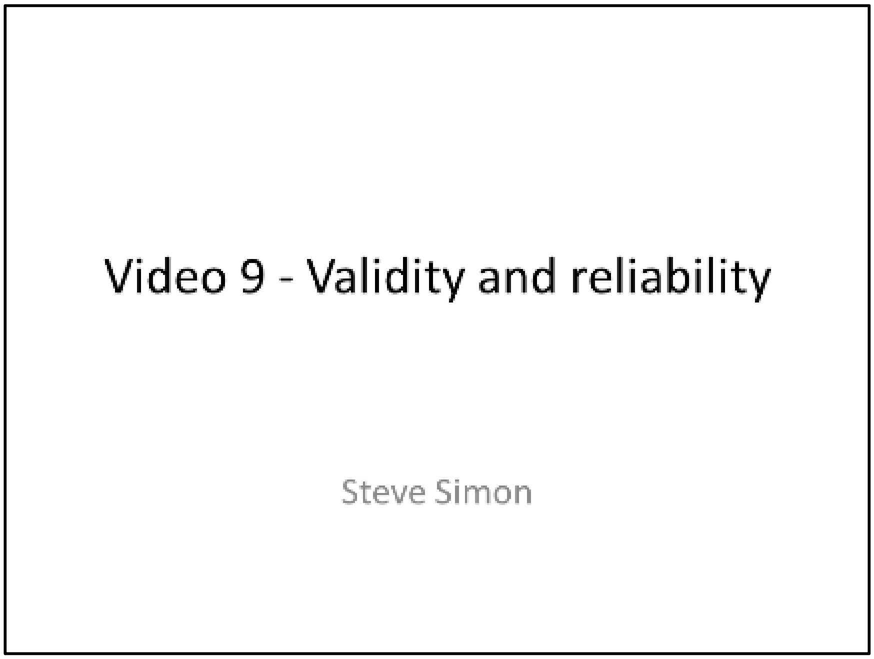
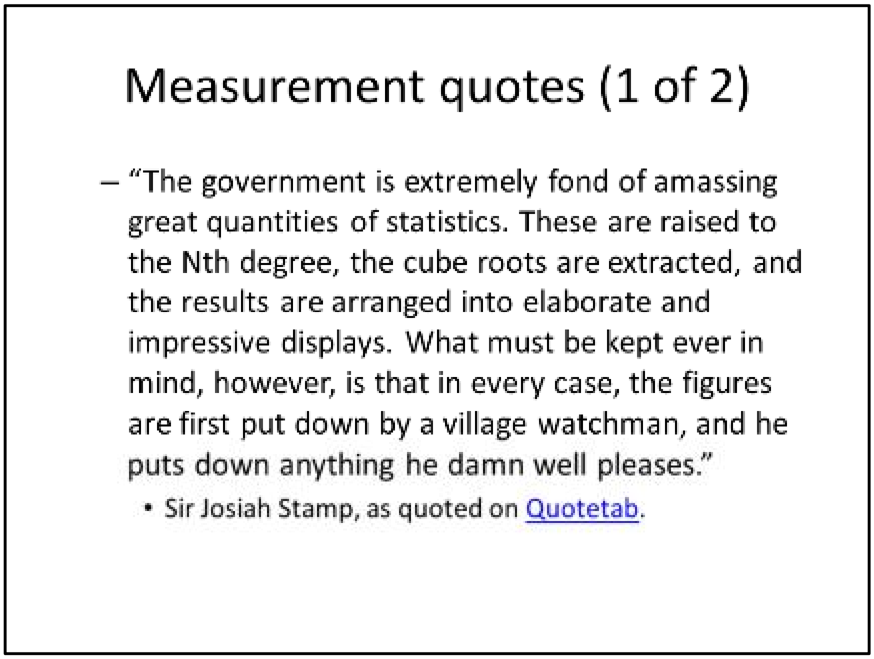


Video 9 - Validity and reliability

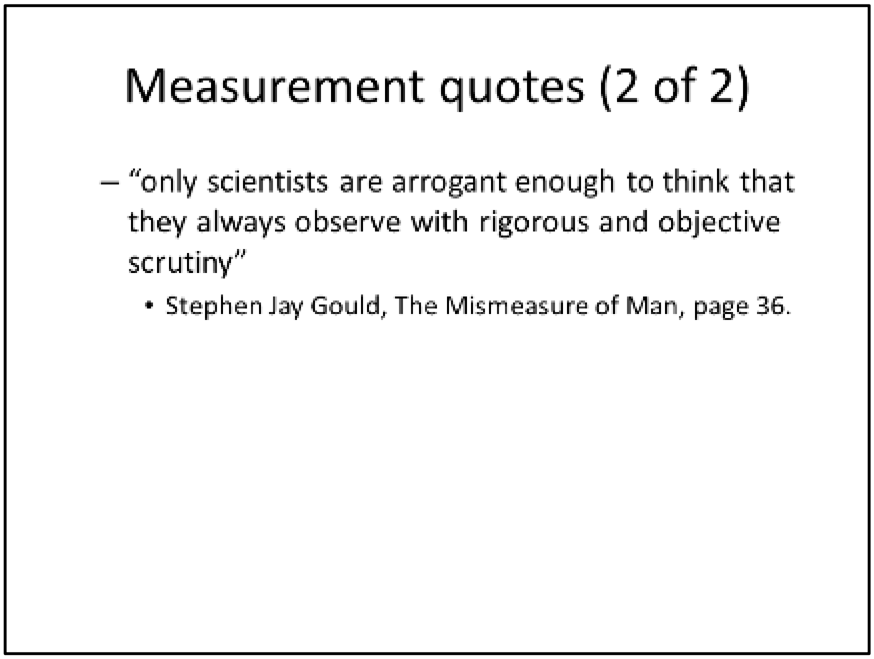
Steve Simon



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.

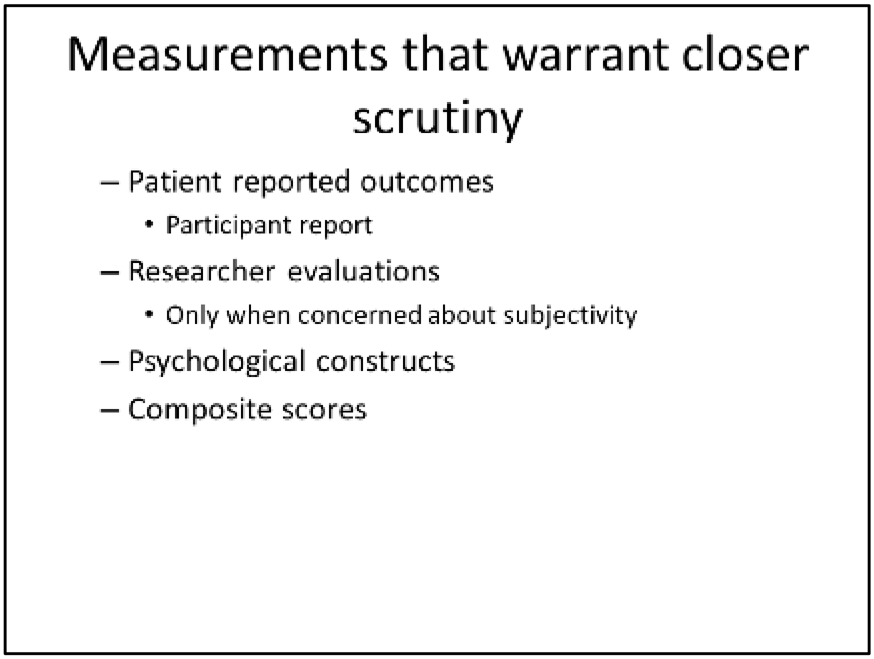
As much as I love numbers, I have to admit that they are often abused. Just because you can attach a number to something does not mean that the number is useful in any way. I want to talk about some of the problems associated with measurement and some of the great pains that you need to take to be sure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

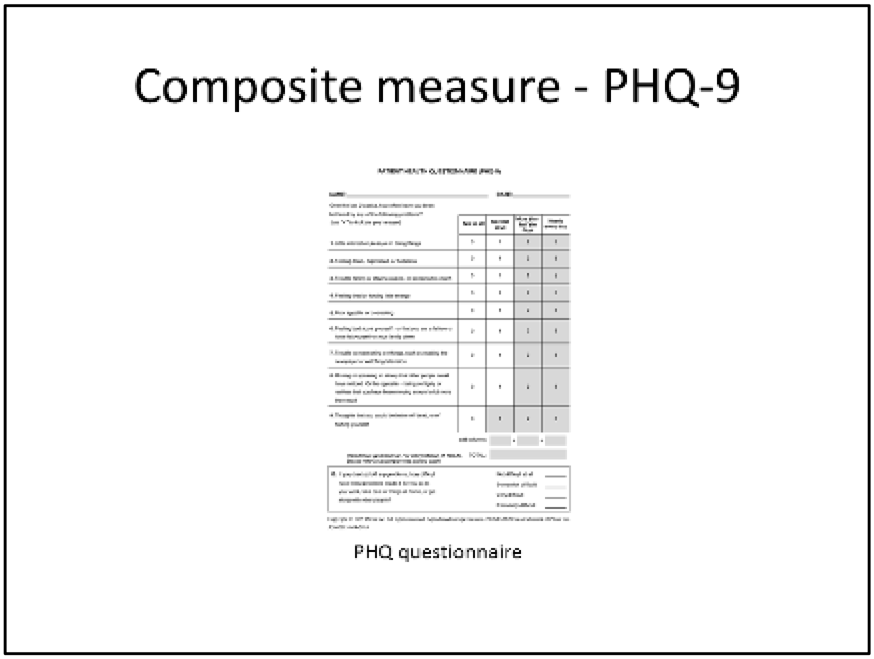
There is also a belief that patient reported outcome measures vary too much from one individual to another. Some people are stoic to a fault and others will complain endlessly at the drop of a hat.

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.



Composite measure - PHQ-9

PHQ questionnaire

The PHQ-9 questionnaire attempts to measure depression. You provide ansers to the following questions

“Over the last 2 weeks, how often have you been bothered by any of the following problems”

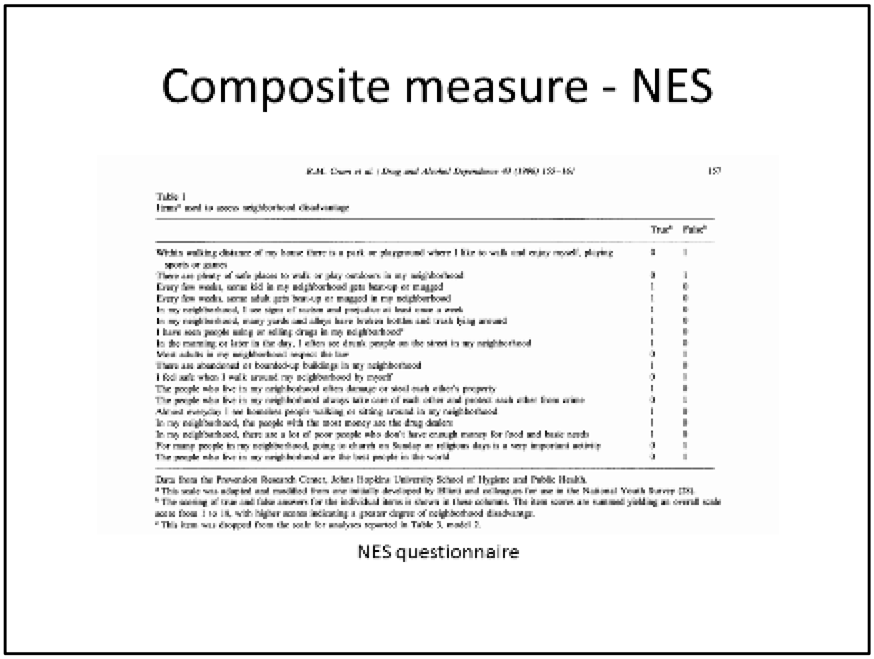
“1. Little interest or pleasure in doing things”

“2. Feeling down, depressed, or hopeless”

and seven more items. The possible responses are “Not at all” scored as a 0, “Several Days” scored as 1, “More than half the days” scored as a 2, and “Nearly every day” scored as a 3. Add up those scores to get a value between 0 and 27.

You might wonder why ask nine questions rather than one? Well, there is a single question, the Yale-Brown scale: “Do you often feel sad or depressed?” with a yes/no answer. It is very useful in a clinical setting where health care professionals might be too busy to ask a series of questions.

The reason to ask a series of questions, rather than one question is that it (frequently, but not always) reduces measurement error and produces better reliability. There are other reasons, as well, I suspect, such as the desire to look at depression on a continuum rather than as a binary outcome.



Composite measure - NES

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

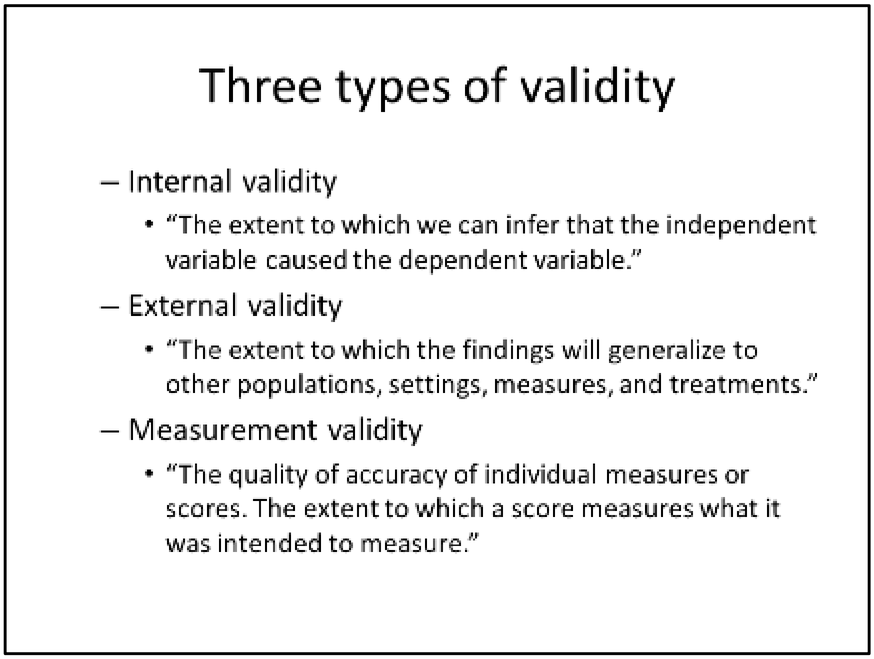
Data from the Prevention Research Center, Johns Hopkins University School of Hygiene and Public Health.
[a] This scale was adapted and modified from one initially developed by Elliott and colleagues for use in the National Youth Survey [28].
[b] The scoring of true and false answers for the individual 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 disadvantage.
[c] This item was dropped from the scale for analyses reported in Table 3, model 2.

NES questionnaire

I realize this image might be difficult to read on your computer. If you are curious, I have placed the image on the Canvas website.

Here's a second example of a composite score, the neighborhood environment survey. It is eighten questions long with questions like "There are plenty of safe places to walk or play outdorrs in my neighborhood" and "Every few weeks, some ki 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 (the paper incorrectly lists the range as 1 to 18).

Some of the approaches for establishing reliability and validity only work for composite measures.

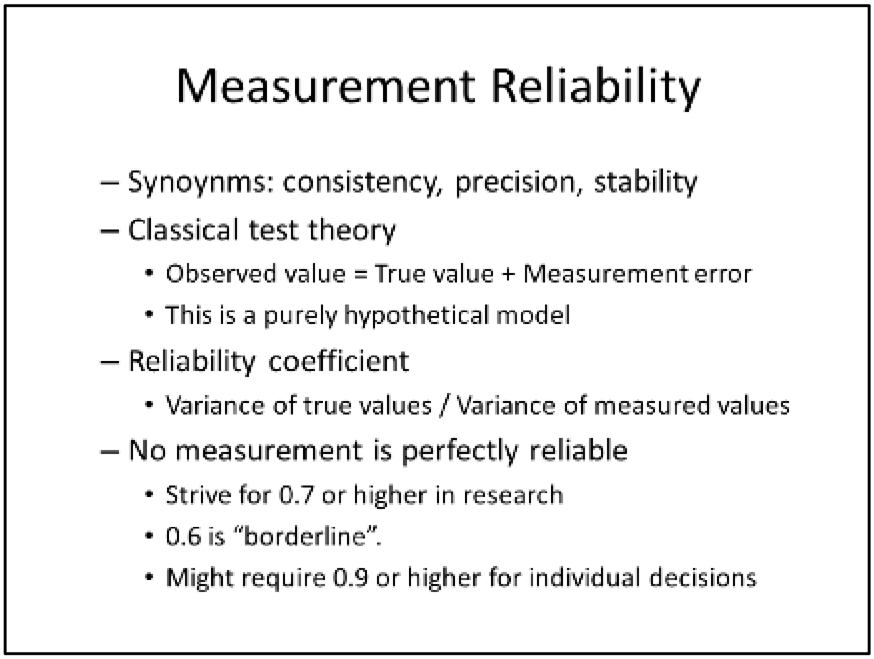


Three types of validity

- 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."

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.



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
- 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

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. 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 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.

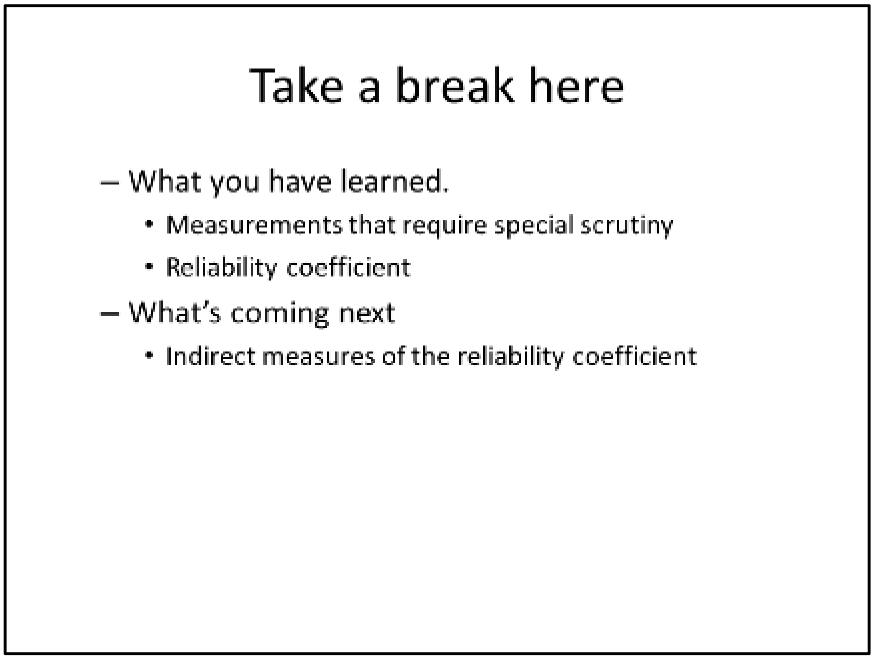
There are some informal standards for reliability. These choices can seem a bit arbitrary, but they are fairly well accepted in the research community.

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.

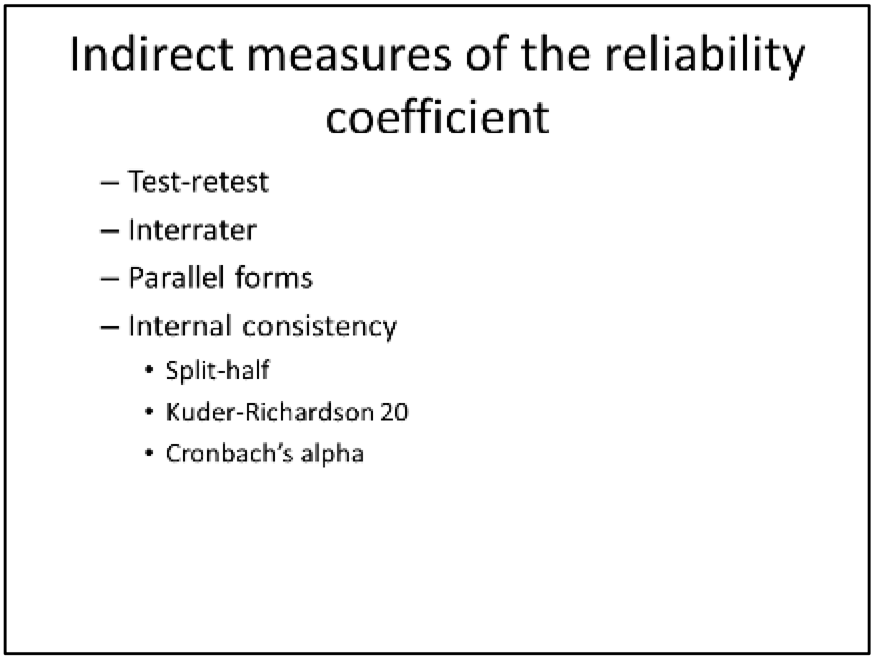


Take a break here

- What you have learned.
  - Measurements that require special scrutiny
  - Reliability coefficient
- What's coming next
  - Indirect measures of the reliability coefficient

Let's take a break here. We've talked about the types of measurements that typically draw extra concern. We also developed a hypothetical model relating the observed measurement to the true measurement plus measurement error. This allowed us to define the reliability coefficient.

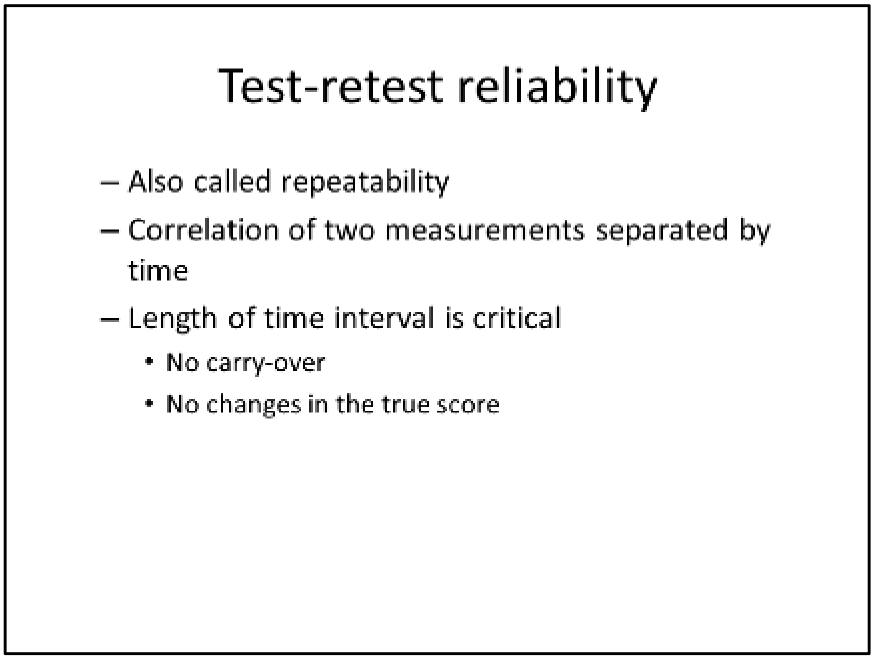
You can't measure the reliability coefficient directly, but in the next video, you'll see several approaches that can provide an indirect measure of this quantity.



Indirect measures of the reliability coefficient

- Test-retest
- Interrater
- Parallel forms
- Internal consistency
  - Split-half
  - Kuder-Richardson 20
  - Cronbach's alpha

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.



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

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.

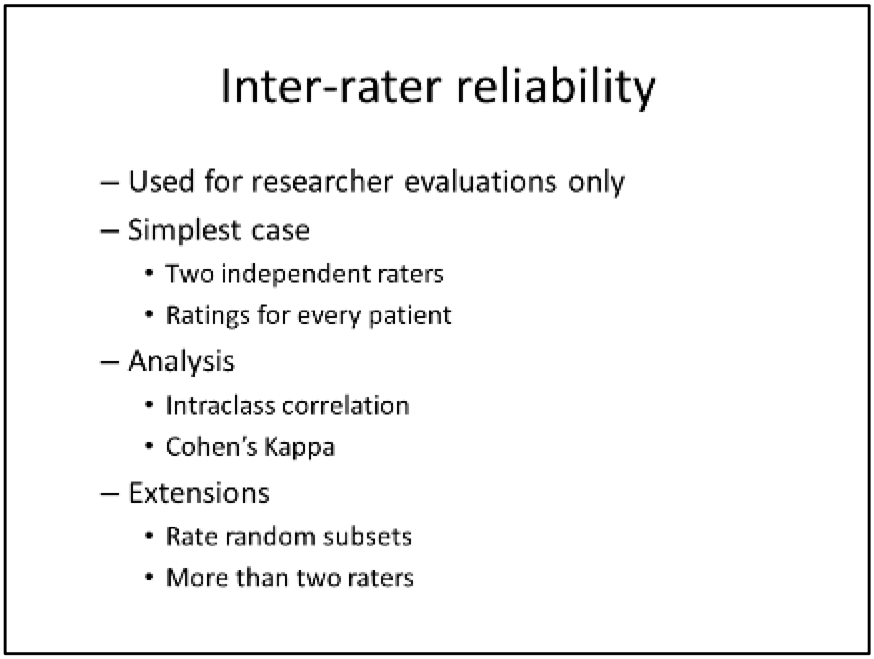
The time interval is critical here. You don’t want two measurements that are so close together that the measurement error for the first measurement is correlated with the measurement error of the second measurement.

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.



Inter-rater reliability

- Used for researcher evaluations only
- Simplest case
  - Two independent raters
  - Ratings for every patient
- Analysis
  - Intraclass correlation
  - Cohen's Kappa
- Extensions
  - Rate random subsets
  - More than two raters

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.

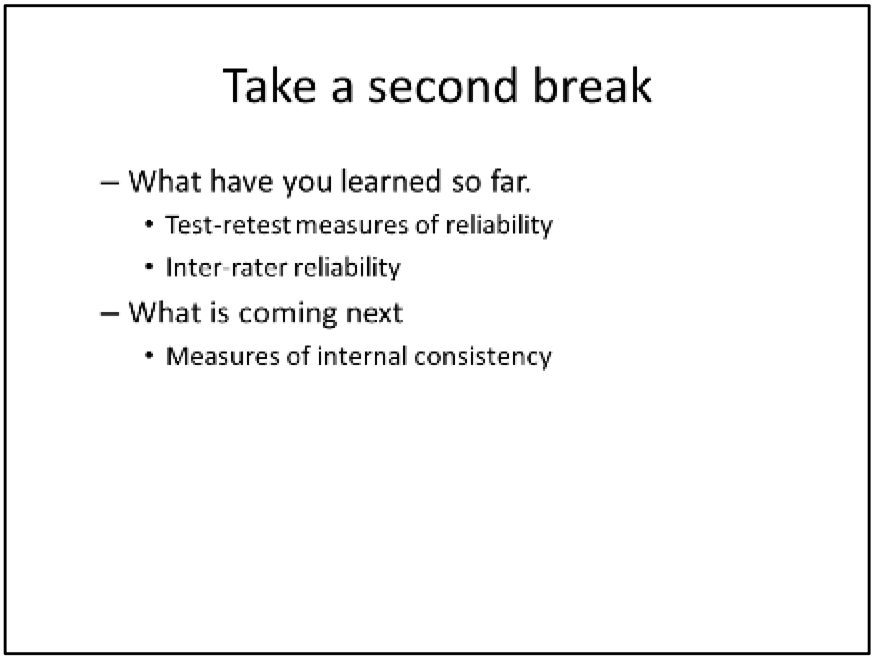
Reliability is pretty simple to measure in this setting, if you have the resources. Just get two raters and have them both compute the measurement. If the correlation between the two raters is high, you have good reliability.

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.

I say that, even though I like formulas. It's not tricky so much as tedious. So I do not want to share all the details here. If you are interested in looking at inter-rater reliability in a more complex setting than just two raters, I'd love to talk to you about it.

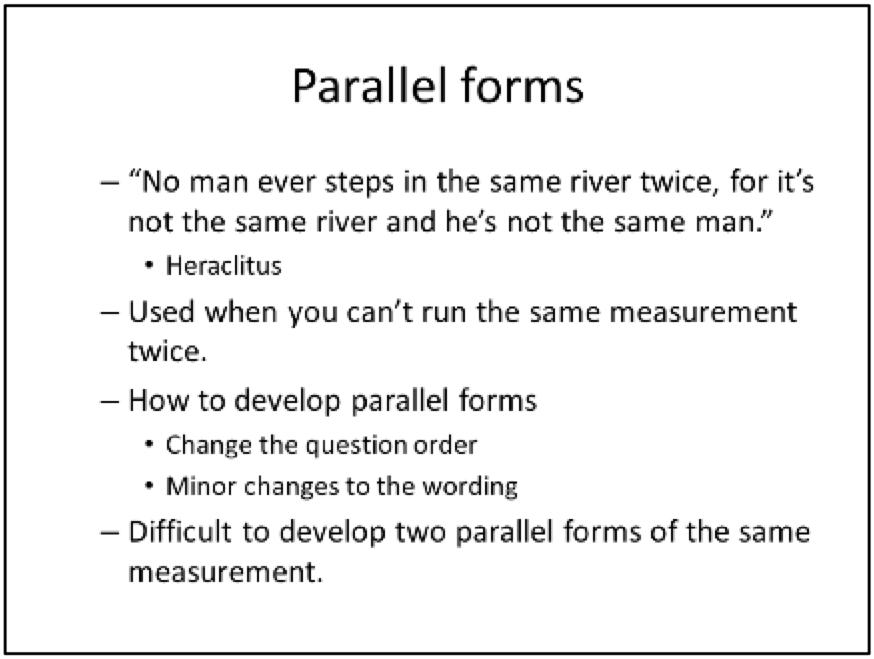


Take a second break

- What have you learned so far.
  - Test-retest measures of reliability
  - Inter-rater reliability
- What is coming next
  - Measures of internal consistency

Time for another break. We’ve talked about test-retest reliability. The tricky part here is deciding how far apart in the time the test and retest have to be. You also learned about inter-rater reliability, which is used for researcher evaluations where you are concerned about subjectivity in the measurement process.

Next, we’ll talk about some very different measures of reliability, measures of internal consistency. I don’t like these measures nearly as much, but you do need to know about them because they are used quite often.



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.

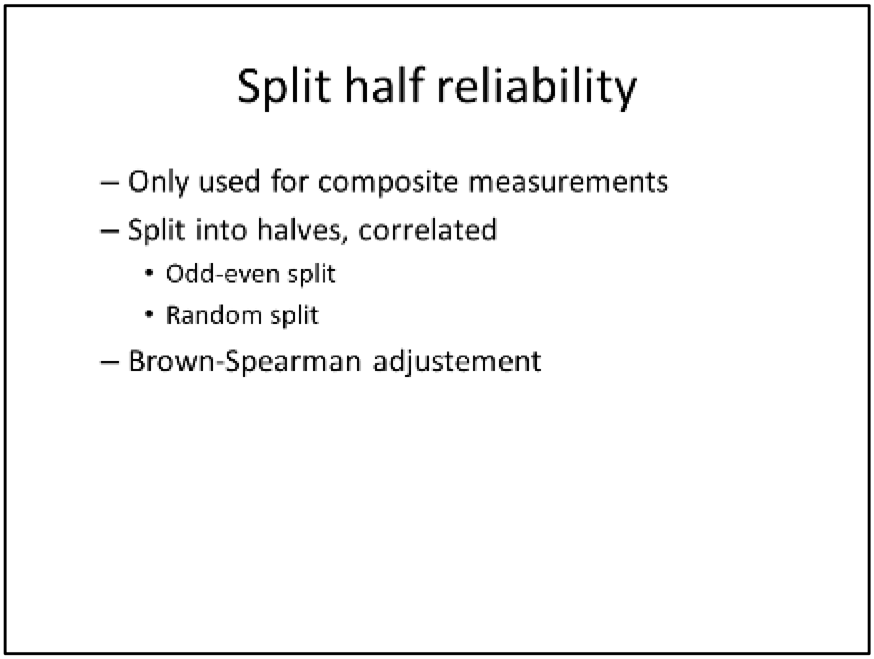
The opposite tendency can occur as well. The novelty of answering questions wears off over time and people may grow tired or bored and not answer the exact same questions a second time.

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

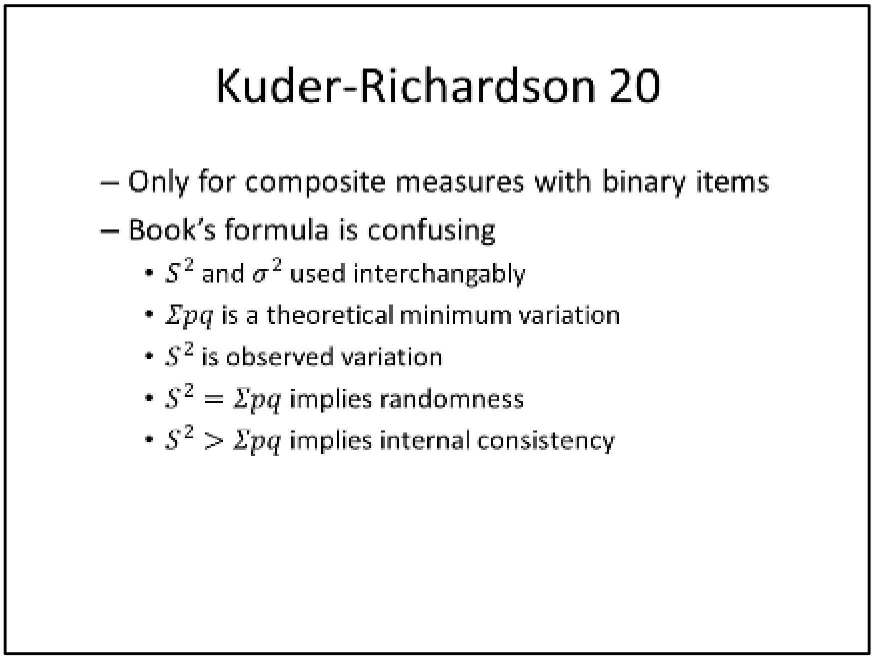
- Only used for composite measurements
- Split into halves, correlated
  - Odd-even split
  - Random split
- Brown-Spearman adjustement

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.



Kuder-Richardson 20

- Only for composite measures with binary items
- Book's formula is confusing
  - S^2 and σ^2 used interchangably
  - Σpq is a theoretical minimum variation
  - S^2 is observed variation
  - $S^2 = \Sigma pq$ implies randomness
  - $S^2 > \Sigma pq$ implies internal consistency

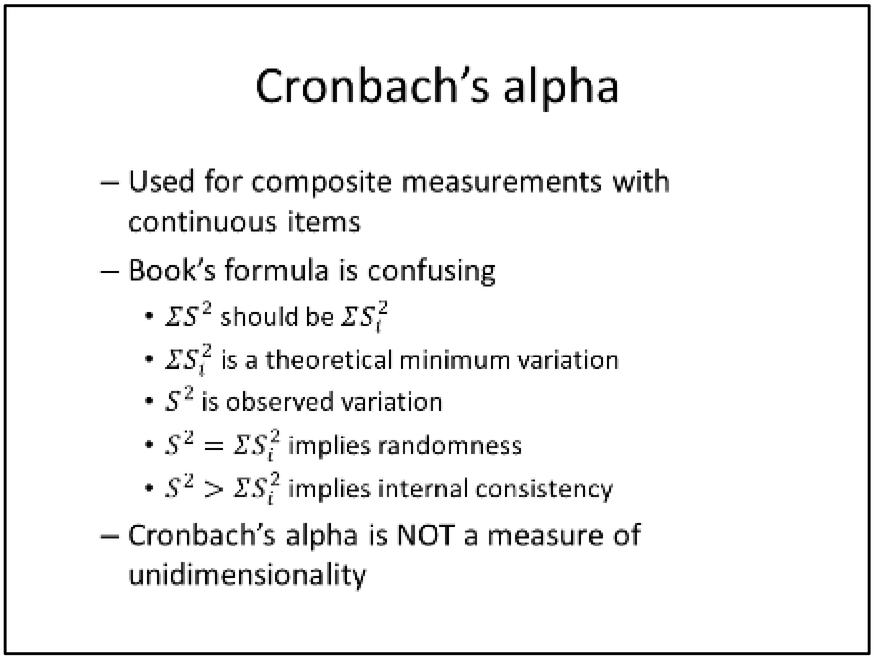
Another measure for reliability, Kuder-Richardson 20, is used for composite measures, but only those composite measures that have binary items. Your book does a poor job explaining this, and the notation is inconsistent.

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

- Used for composite measurements with continuous items
- Book's formula is confusing
  - ΣS^2 should be ΣS_i^2
  - ΣS_i^2 is a theoretical minimum variation
  - S^2 is observed variation
  - $S^2 = \Sigma S_i^2$ implies randomness
  - $S^2 > \Sigma S_i^2$ implies internal consistency
- Cronbach's alpha is NOT a measure of unidimensionality

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

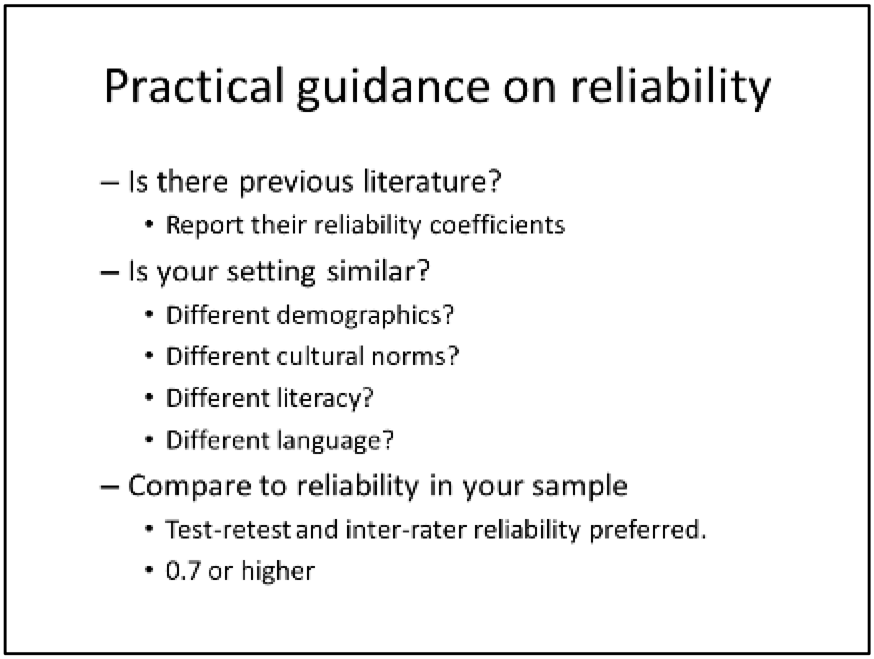
Again, your book does a poor job explaining this, and the notation is confusing.

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.



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.

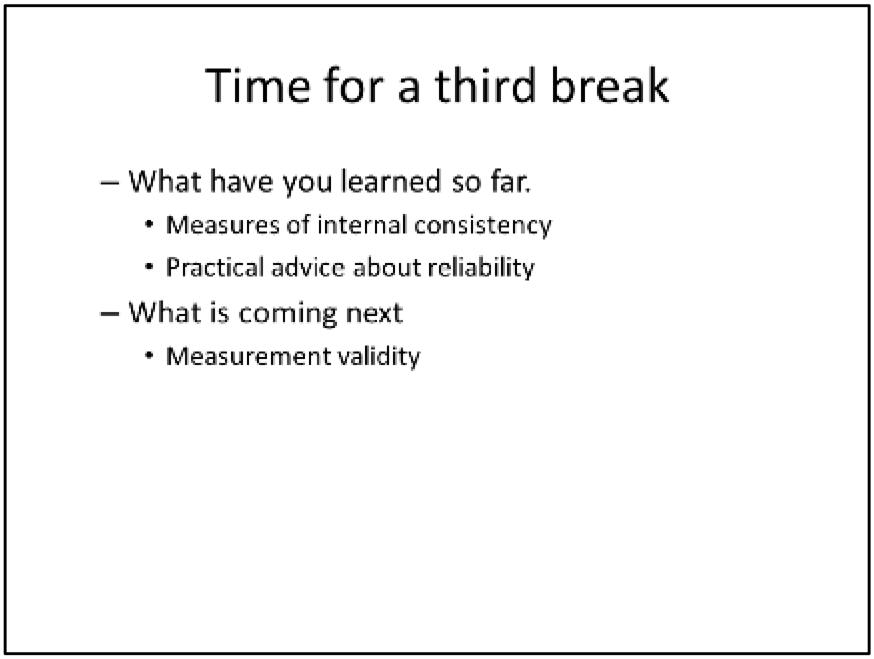
If you can, incorporate a measure of reliability into your study. There are two reasons for this. First, your setting may be different enough to raise concerns. Getting a current measure of reliability helps to allay those concerns. Second, reliability is never quite perfect, because all of the measures 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.

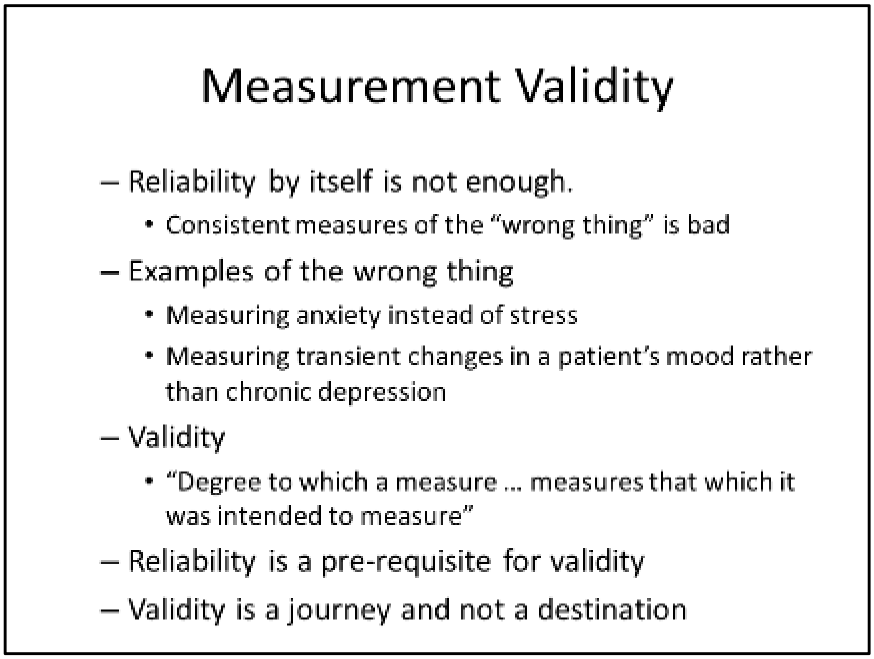


Time for a third break

- What have you learned so far.
  - Measures of internal consistency
  - Practical advice about reliability
- What is coming next
  - Measurement validity

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.

Next, we’ll tackle measurement validity.



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.

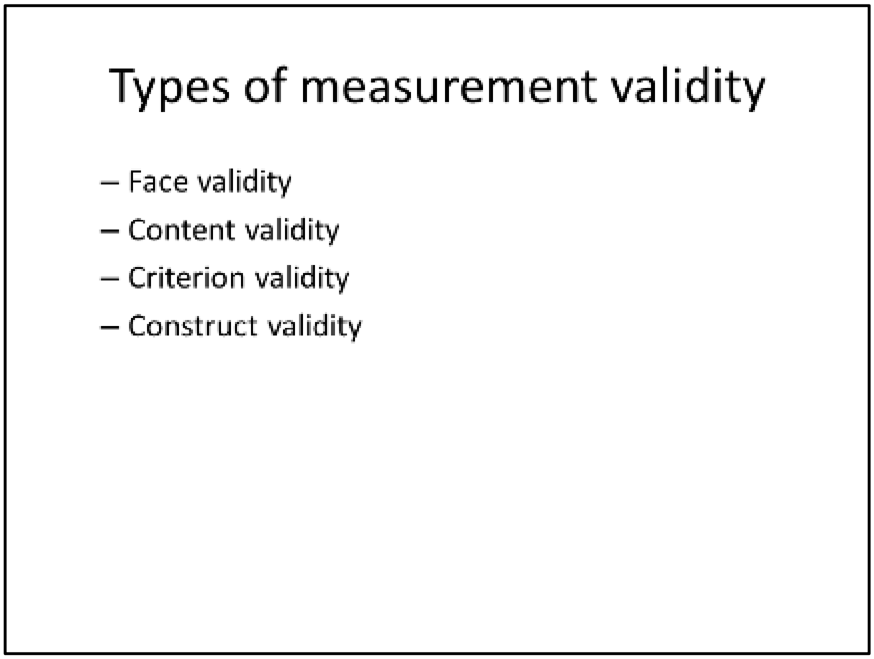
So in addition to establishing that your measurement has good reliability, you also have to establish good validity.

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.

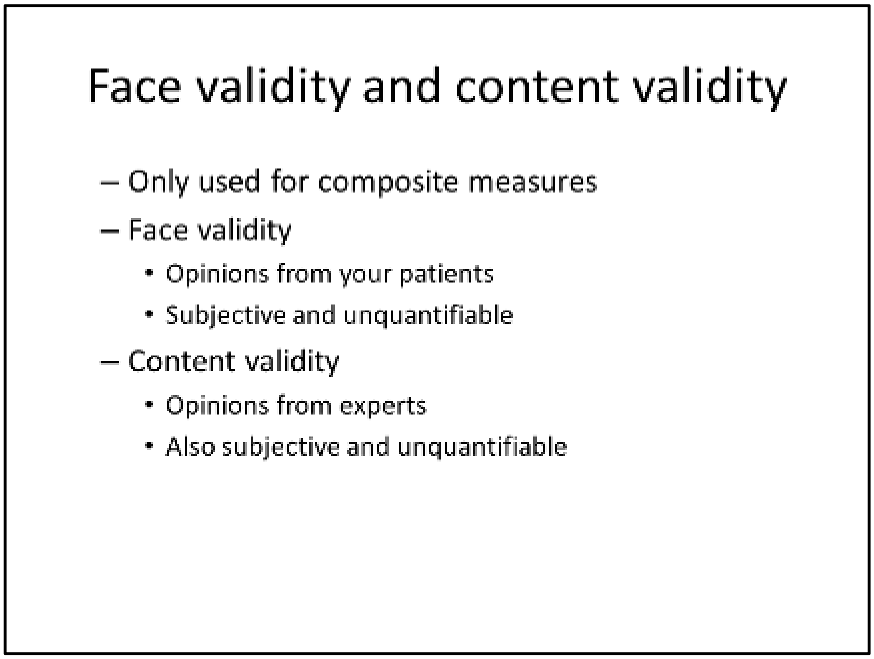
The other thing to keep in mind is that validity is not something that you establish and then you’re done. Validity is a journey and a never-ending journey at that. Each study in a series of studies that uses a particular measurement will contribute information about the validity about a measurement.



Types of measurement validity

- – Face validity
- – Content validity
- – 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

- Only used for composite measures
- Face validity
  - Opinions from your patients
  - Subjective and unquantifiable
- Content validity
  - Opinions from experts
  - Also subjective and unquantifiable

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.

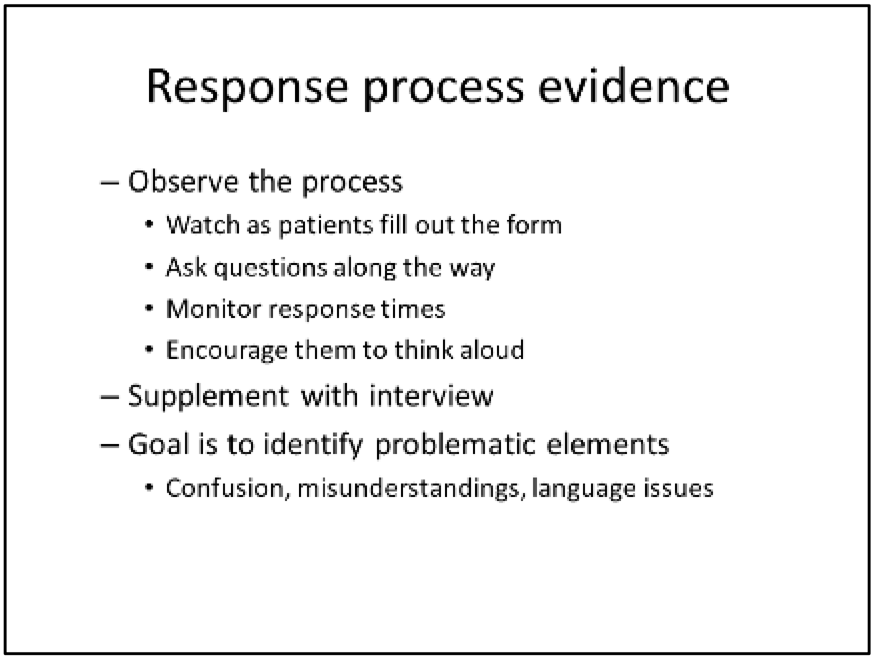
Even though these approaches are soft, they are well worth the effort.

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.



Response process evidence

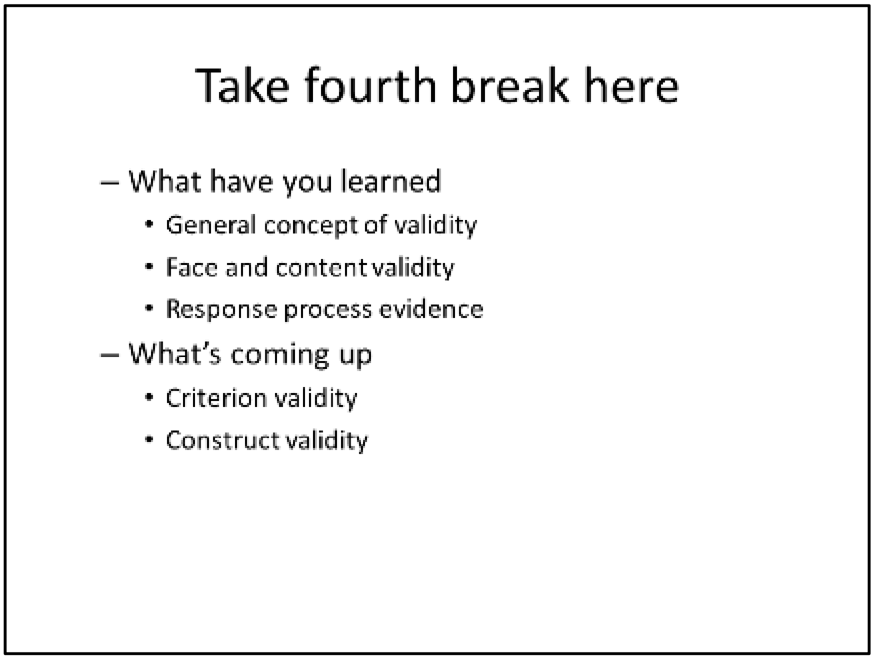
- Observe the process
  - Watch as patients fill out the form
  - Ask questions along the way
  - Monitor response times
  - Encourage them to think aloud
- Supplement with interview
- Goal is to identify problematic elements
  - Confusion, misunderstandings, language issues

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.

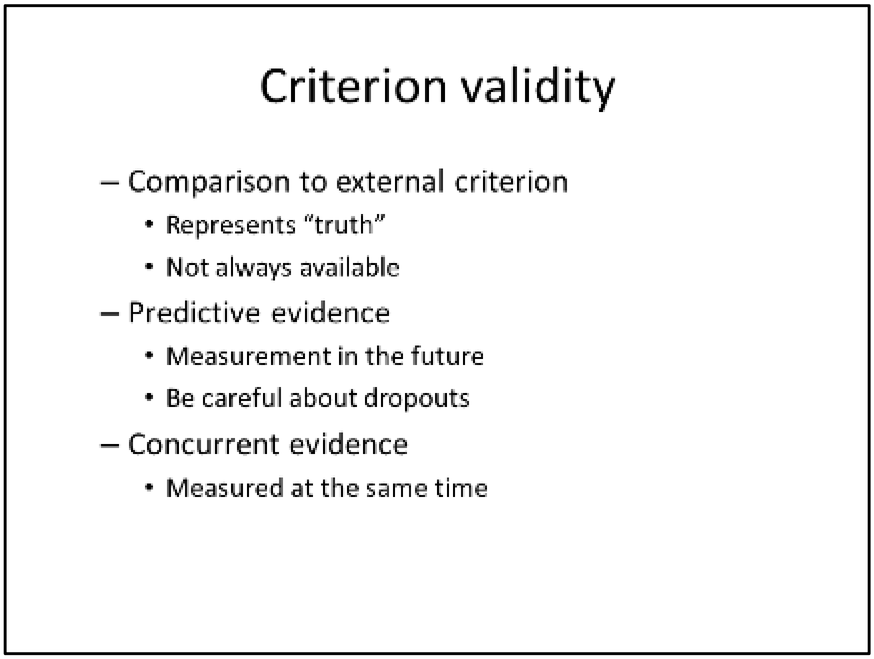
You can do this sort of exercise with experts as well as patients. Ask your experts to pretend that they are patients and get them to fill things out, talk aloud, and ask them questions along the way.



Take fourth break here

- What have you learned
  - General concept of validity
  - Face and content validity
  - Response process evidence
- What's coming up
  - Criterion validity
  - Construct validity

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.

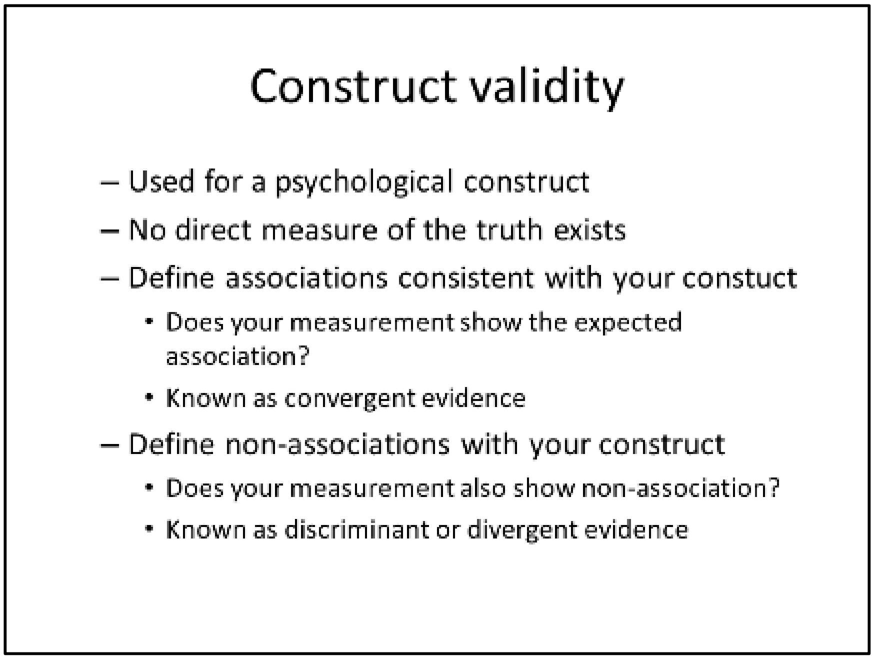
Your book offers an interesting example of this with standardized testing for college admission. A school might want to correlate an SAT score, for example, with the grades that a student gets after one year of college. Easy to do, but think about the dropouts. A college, for the most part, is going 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.

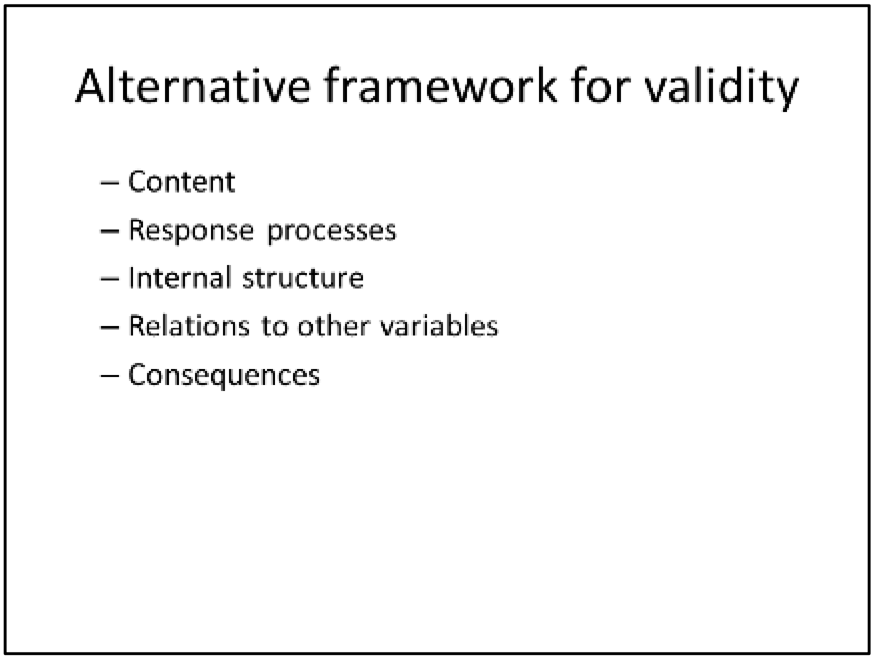
Now this approach has limits. You can never get quite as much evidence of validity as the already validated measurement has.



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

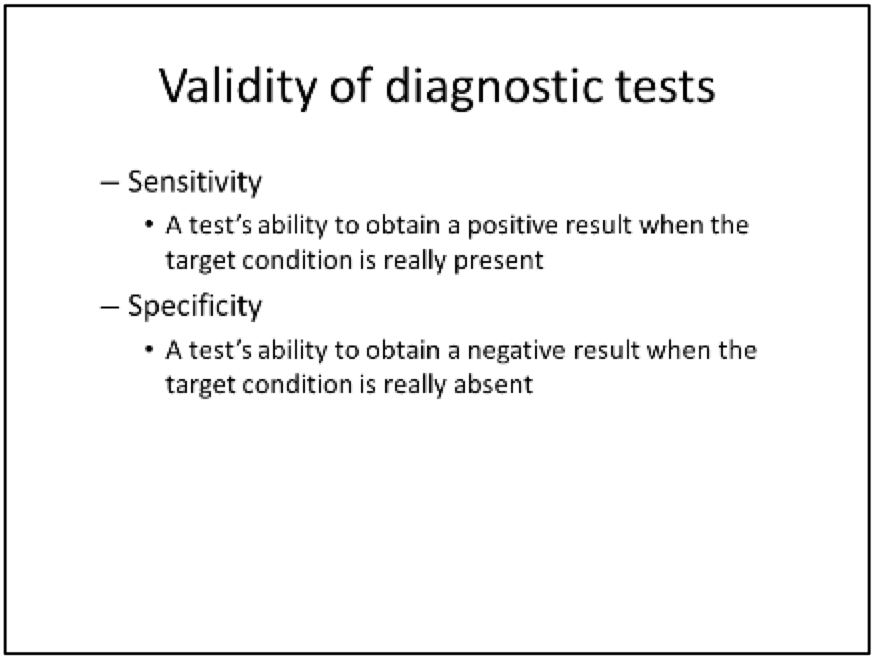
Construct validity is when you are developing a psychological construct and you don’t have a direct measure of the construct you are trying to measure. What you do have is various associations and non-associations that your construct is expected to have. You develop these using your deep thinking power 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.

Stop here

- What have you learned so far?
  - Criterion validity
  - Construct validity
- Next week
  - Spring break
  - After that, data collection issues

–