Measurement Validity

Qian-Li Xue

Biostatistics Program

Harvard Catalyst | The Harvard Clinical & Translational Science Center

Short course, October 27, 2016

Slides contributed by Jeannie-Marie Leoutsakos, Assistant Professor of Psychiatry & Mental Health, Johns Hopkins University

Objectives

- What is validity?
- Types of validity
  - Face & Content Validity
  - Criterion Validity
  - Construct Validity
  - Validity and Latent Variables
- Summing it up

What is validity?

- Two questions:
  - Are you measuring what you intend to measure?
  - What is the relationship of a scale to its purported cause?
- Critically important for concepts that are not readily observable
  - Social support
    - Number of people a person has contact with
    - Who is available in times of need
    - Reciprocity of the helping relationship

Definitions of Validity

- Accuracy of measurement
- Lack of bias

$$Bias\left[\hat{\theta}\right] = E\left[\hat{\theta}\right] - \theta$$

- Theta is “the truth”
- Theta hat is an estimate of theta, a measurement
- Expected value of theta hat is the long-run average

Validity in Classical Test Theory

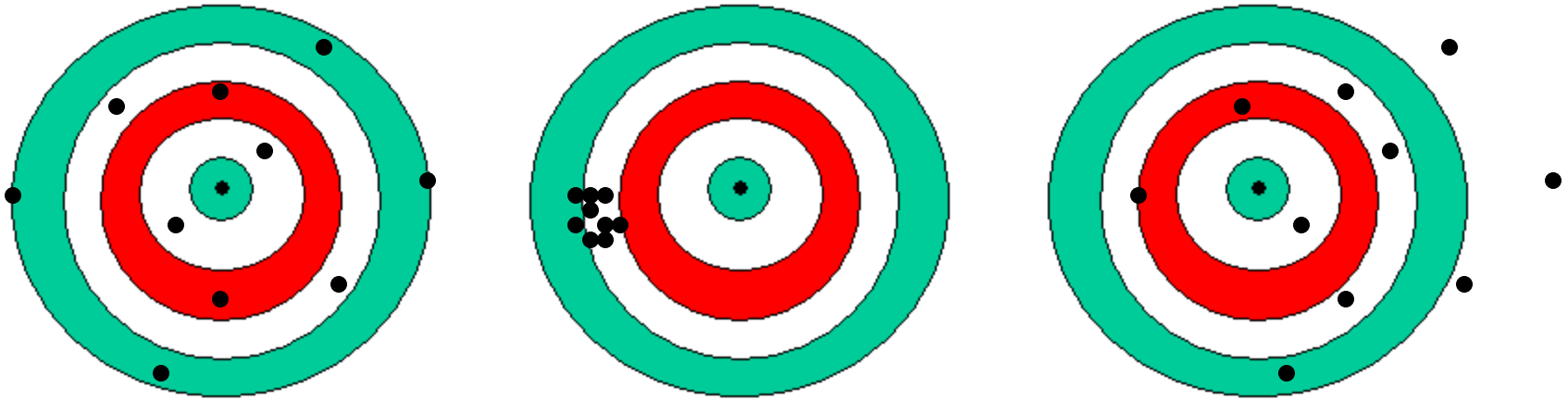
- Classical Test Theory

$$X = T_x + e$$

- Handled poorly, assumes no systematic bias.
- True Score T_x defined as $E\left[\hat{\theta}\right]$,not θ

$$X = T_x + e \Rightarrow X = \theta + Bias\left[\hat{\theta}\right] + e$$

Reliability vs. Validity



Valid, not Reliable Reliable, not Valid Neither

Face Validity

- How valid does the item *appear?*
- CESD: Depression: "I felt depressed"
  MMPI: Schizophrenia: "I believe in God"
- Increases motivation, decreases dissatisfaction in respondents
- Enhance credibility of results
- Improves public relations

Content Validity

- The extent to which one can generalize from a particular collection of items to all possible items that would be representative of a specified domain of items. (Nunnaly, 1978)

- Spelling among fourth graders: random sample of all words in widely-used fourth grade readers.

- "Big Five" Personality Traits: systematic sampling, over generations of research, of all words in the English dictionary which describe personality.

DSM Major Depressive Episode

At least five of the following symptoms have been present during the same two week depressed period

- Depressed mood
- Loss of all interest and pleasure
- Appetite or weight disturbance
- Sleep disturbance
- Agitation or slowing
- Fatigue or loss of energy
- Abnormal self-reproach or inappropriate guilt
- Poor concentration or indecisiveness
- Thoughts of death or suicide

CESD-R for Major Depression

- Depressed mood
  - I felt sad
  - I felt depressed
- Loss of all interest or pleasure
  - Nothing made me happy
  - I lost interest in my usual activities
- Appetite or weight disturbance
  - I lost a lot of weight without trying to
  - My appetite was poor

Face vs. Content Validity

- Both grouped under *translational validity* in some text books.
- Content validity stronger than face validity.
- Content validity relies on theory
  - e.g., in CESD-R example, one must accept the DSM definition of Major Depression, and that there are no other domains to be sampled from.

Criterion Validity

- Association of a test measure with a criterion variable.
- Ideally, you have a gold standard for a criterion variable.
- Criterion variable and test measure need to be ascertained independently
- Can be:
  - Concurrent: prostate cancer based on PSA
  - Predictive: college graduation based on SATs
  - Postdictive: MI based on creatine kinase and troponin

Sensitivity & Specificity

		Disease Status		
		Yes (D+)	No (D-)	
Test Result	T+	A	B	Pr(T+)=A+B/N
	T-	C	D	Pr(T-) =C+D/N
		Pr(D+)=A+C/N	Pr(D-)=B+D/N	N

- Sensitivity: Does the test identify cases?

$$Sens = P(T^+ \mid D^+) = A/(A+C)$$

- Specificity: Does the test identify noncases?

$$Spec = P(T^- \mid D^-) = D/(B+D)$$

Positive & Negative Predictive Value

		Disease Status		
		Yes (D+)	No (D-)	
Test Result	T+	A	B	Pr(T+)=A+B/N
	T-	C	D	Pr(T-) =C+D/N
		Pr(D+)=A+C/N	Pr(D-)=B+D/N	N

- PPV: What does a positive test result mean?

$$PPV = P(D^+ \mid T^+) = A/(A+B)$$

- NPV: What does a negative test result mean?

$$NPV = P(D^- \mid T^-) = D/(C+D)$$

Criterion Validity Example

Why should Ultra-Screen be offered to all pregnant women?

Patients are becoming increasingly sophisticated in their knowledge of risk assessment options, with study after study indicating that the vast majority of patients prefer to know their risk of a fetus with Down syndrome during their first trimester.

With the Ultra-Screen® protocol from NTD, physicians can provide patients with the standard of care in first trimester Down syndrome screening. This simple, non-invasive procedure enables a quick, efficient, reliable assessment of a patient's risk for Down syndrome more accurately, with detection rates as high as 95% and false positive rates as low as 2% with an optional fetal nasal bone assessment.

Patient FAQ (http://www.ntdlabs.com/patients/patient_faq.php)

Sensitivity & Specificity of UltraScreen®

	Trisomy 21+	Trisomy 21-	Total
Screen Positive	30	451	481
Screen Negative	3	5267	5270
Total	33	5718	5751

$$Sens = P(T^+ \mid D^+) = 30/(3+30) = 0.91$$

$$Spec = P(T^- \mid D^-) = 5267/(451+5267) = 0.92$$

$$FalsePositiveRate = 1 - Spec = 0.08$$

Data from Krantz et al., (2000) Obstet Gynecol 96(2):207-13.

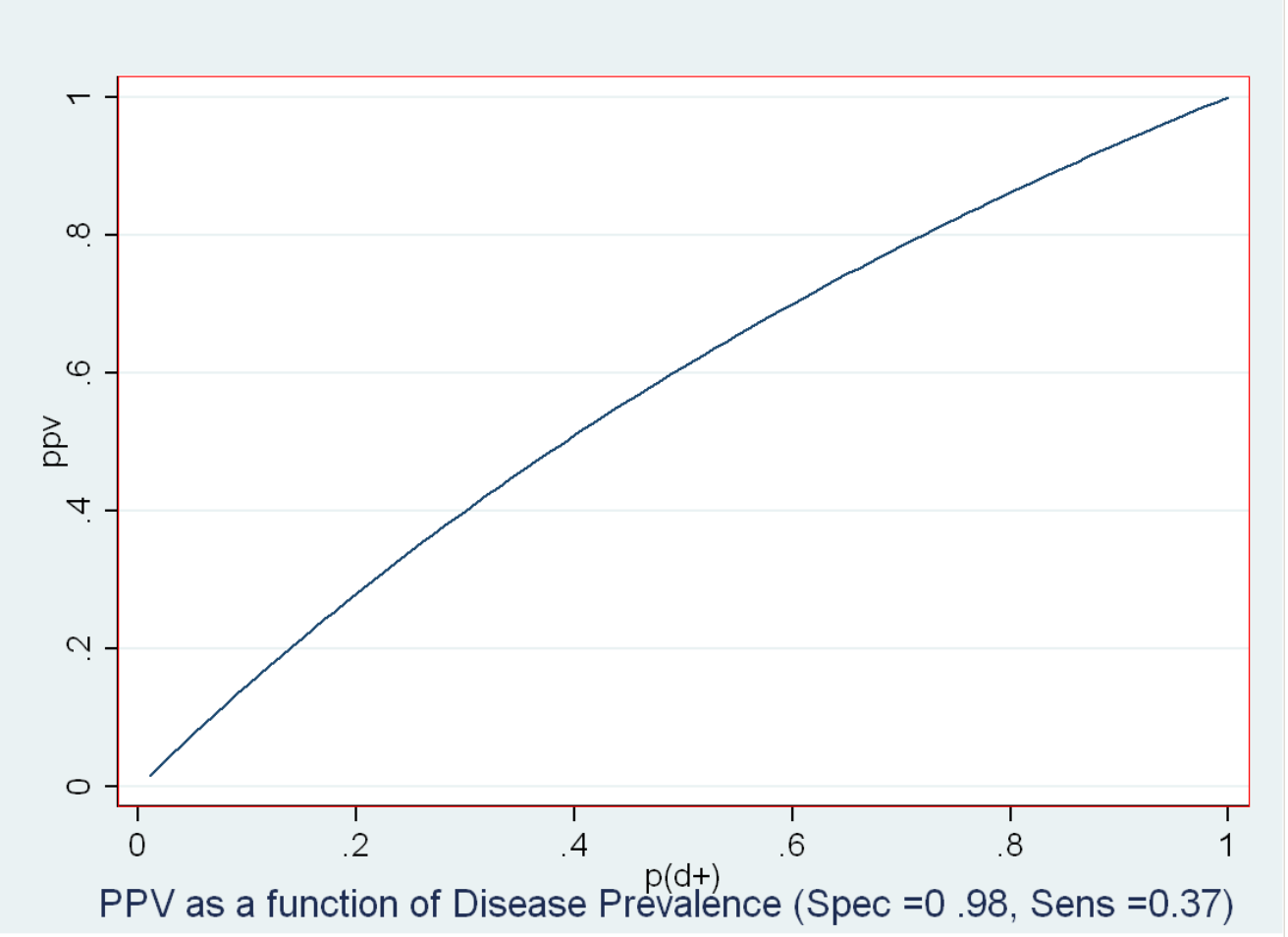
PPV & NPV of UltraScreen®

	Trisomy 21+	Trisomy 21-	Total
Screen Positive	30	451	481
Screen Negative	3	5267	5270
Total	33	5718	5751

$$NPV = P(D^- \mid T^-) = 5267/(3+5267) \approx 1$$

$$PPV = P(D^+ \mid T^+) = 30/(30+451) = 0.06$$

PPV and Disease Prevalence



Caveats

- PPV and NPV are affected by disease prevalence.
- The key question: what is your denominator?

Criterion Validity without Gold Standards?

- LEAD Standard:
  - Longitudinal – clinical evaluation at more than one point in time.
  - Expert – clinicians with demonstrated reliability
  - All Data – collateral informants, past history and records, etc.
- See Spitzer (1983)

Thorny Question

- How do we talk about validity of things we can't measure directly and for which we don't have a gold standard?
- How do we calculate bias without knowing theta?

$$Bias\left[\hat{\theta}\right] = E\left[\hat{\theta}\right] - \theta$$

$$X = T_x + e \Rightarrow X = \theta + Bias\left[\hat{\theta}\right] + e$$

Some Options

- Fiat. Declare a Gold Standard.
- Define theta relative to something measurable, like an outcome (e.g., *since he committed suicide, he must have been depressed)*.
- Construct Validity/Nomological Network
- Latent Variable Models

The Problem with Fiat

HOUSE JOINT MEMORIAL 54

48TH LEGISLATURE - STATE OF NEW MEXICO - FIRST SESSION, 2007

INTRODUCED BY

Joni Marie Gutierrez

A JOINT MEMORIAL

DECLARING PLUTO A PLANET AND DECLARING MARCH 13, 2007, "PLUTO PLANET DAY" AT THE LEGISLATURE.

NOW, THEREFORE, BE IT RESOLVED BY THE LEGISLATURE OF THE STATE OF NEW MEXICO that, as Pluto passes overhead through New Mexico's excellent night skies, it be declared a planet and that March 13, 2007 be declared "Pluto Planet Day" at the legislature.

What is a Construct?

- Something . . . “scientists put together from their own imaginations….” -*Nunnaly*
- “A mini-theory to explain the relationships among various behaviors or attitudes.” *Streiner and Norman*
- “A name for patterns of covariation of traits or characteristics in individuals” *W. Eaton*
- Similar to the concept of “syndrome” in medicine

Example: Irritable bowel syndrome

Signs and symptoms

- Abdominal pain or cramping
- A bloated feeling
- Gas
- Diarrhea or constipation
- Mucus in the stool
- Anxiety
- Lower back pain
- Reduced sexual desire
- Sleep disturbance

Why is it a construct?

- Not directly observable
- No definitive diagnostic test
- Diagnosis by exclusion
- No known pathogen

Construct Validity

"…the extent to which operational variables used to observe covariation in and between constructs can be interpreted in terms of theoretical constructs ." *(Calder et al, 1982, in Netemeier et al, page 71)*

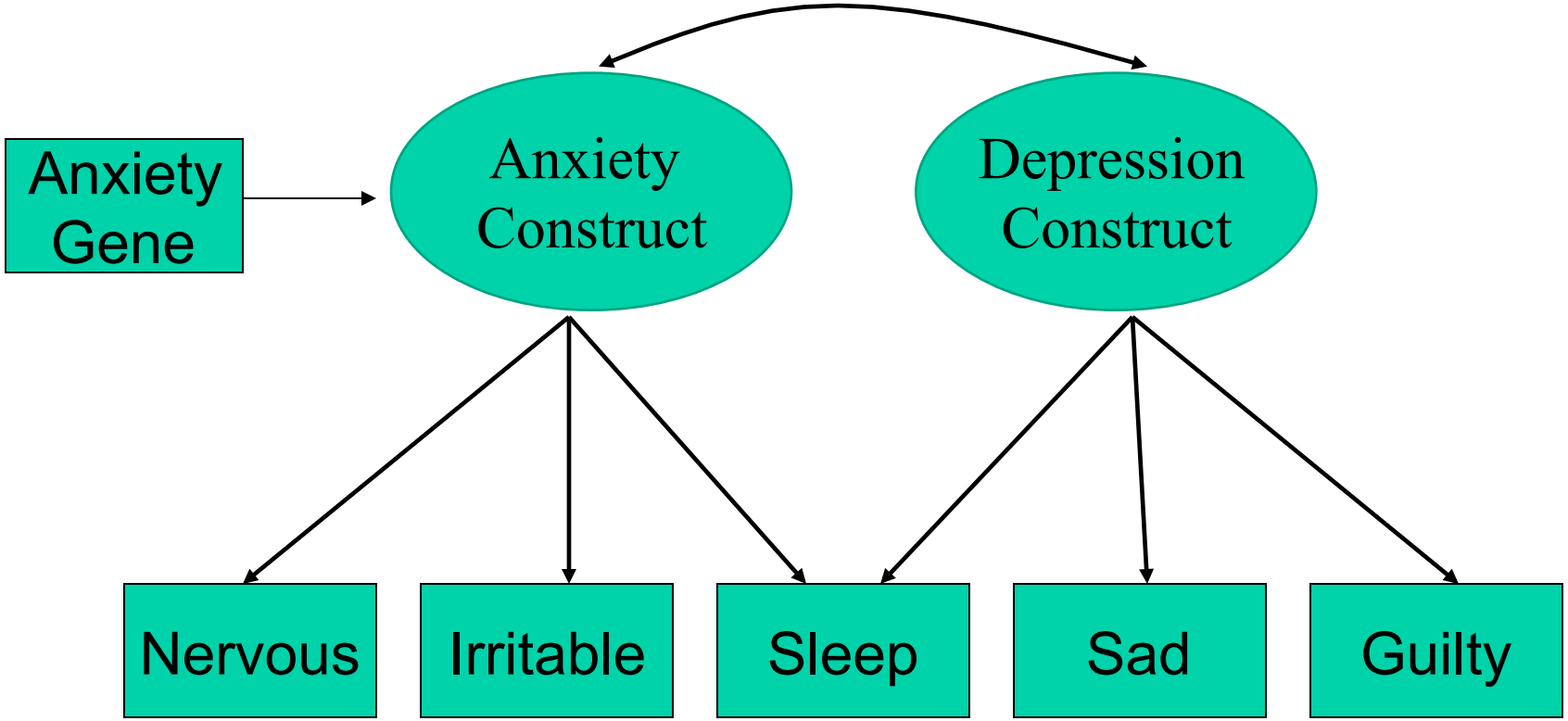
"the degree to which a measure satisfies theoretical predictions about the construct, across a range of theories, and with a range of modalities of measurement" *W. Eaton*

How well does your measure fit into your nomological network?

Nomological Network

In the absence of a gold standard or absolute point of reference, attempts to "nail down" construct through "triangulation" – fitting it into a network of related constructs and measures.

Nomological Network Example



Nomological Networks

Per Cronbach and Meehl (1955):

- 1. Defines the construct by defining the laws in which it occurs.
- 2. The laws govern relationships among constructs and observables (indicators)
- 3. for a construct to be “scientifically admissible” it must occur in a network with connections to observables.

Nomological Networks

- 4. Learning more about a theoretical construct is a matter of elaborating its network, or increasing components' definiteness.

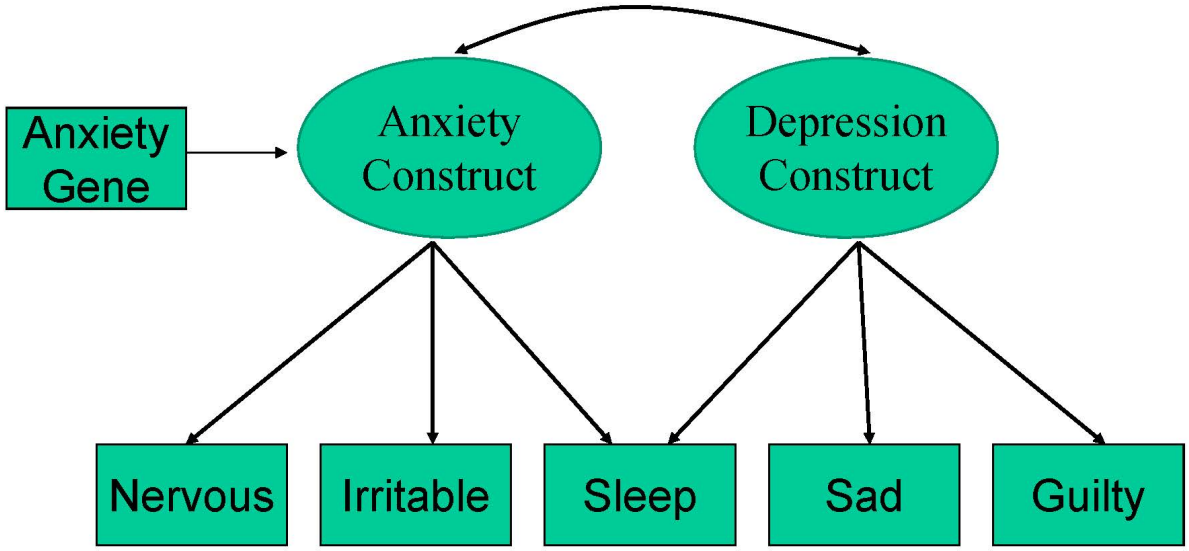
- 5. adding elements if confirmed by observation or if it reduces the number of nomologicals required to predict the same observations.

- 6. Indicators "overlap" or "measure the same thing" if their positions in the nomological net tie them to the same construct variable.

Construct Validity

- Internal Construct Validity – involves relationships between indicators/ measures and constructs

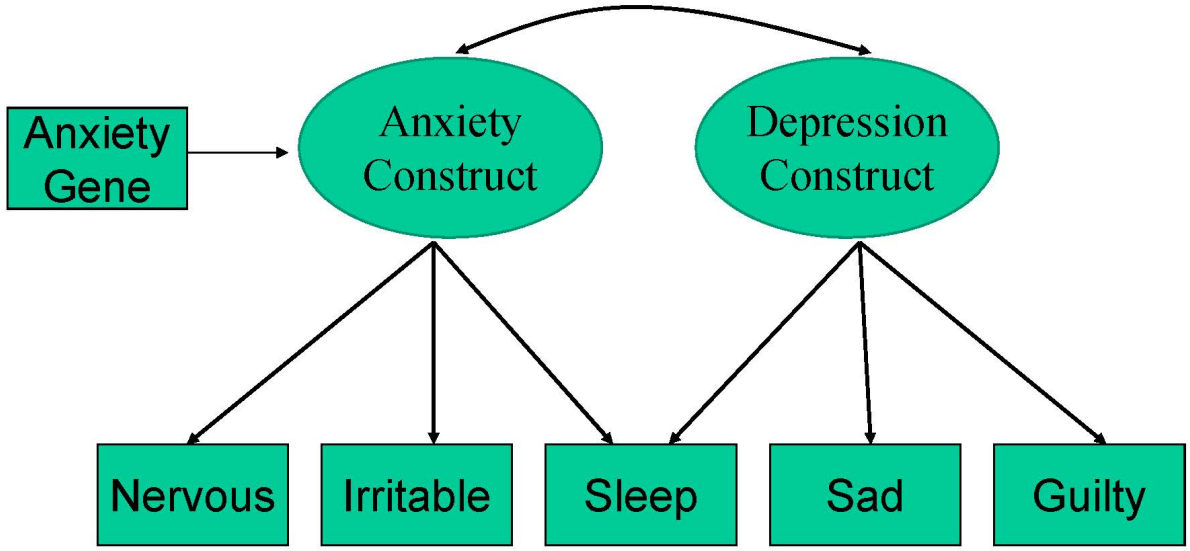
- External Construct Validity AKA Nomological Validity – involves relationships between constructs and other constructs or variables.



Internal Construct Validity

- Discriminant: degree to which indicators/measures are associated with indicators of the same construct and **not** with indicators of different constructs.

- Convergent: degree to which indicators/measures are associated with other indicators of the same construct, even when measured with a different modality.



Review

Lecture		Type
	Translational Validity	Face Validity
		Content Validity
	Criterion Validity	Predictive Validity
		Concurrent Validity
		Postdictive Validity
Construct Validity	Internal Construct Validity	Convergent Validity
		Discriminant Validity
		Known Group Validity
	External Construct Validity	Nomological Validity