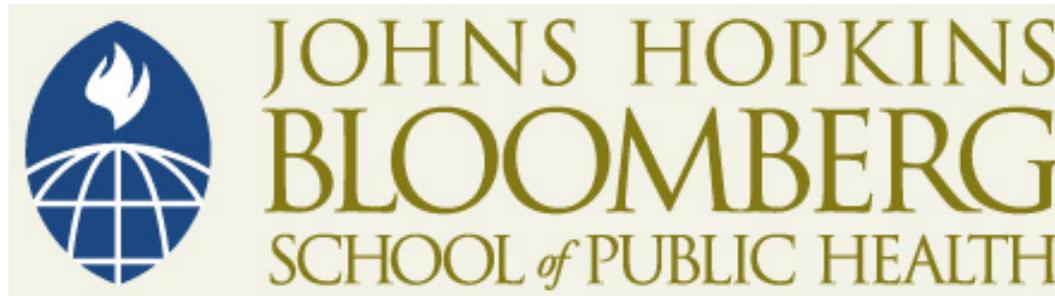


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and William Eaton. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Statistics in Psychosocial Research

Lecture 5

Validity I

Lecturer: William Eaton

Validity

The degree to which a test, scale, or assessment measures what it is supposed to measure.

Outline

- Face Validity
- Content Validity
- Construct Validity
 - Internal
 - External
- Criterion Validity
 - Concurrent
 - Predictive
- Validity and reliability
- Validity and utility

Classical Test Theory and Reliability:

$$x = T_x + e$$

The observed score equals the true score plus random measurement error

Validity

$$x = T_x + e_s + e_r$$

The observed score equals the true score plus systematic error (bias) plus random measurement error

Face Validity

Does the item seem right?

CESD: Depression: “I felt depressed”

MMPI: Schizophrenia: “I believe in God”

- Increases motivation among respondents
- Reduces dissatisfaction among respondents
- Enhances credibility of results
- Improves public relations

Face Validity: IQ Test Example

1. What would be the next number in this series?

1 ... 2 ... 3 ... 5 ... 8 ... 13 ... 21 ... 34 ... ??

47

53

55

62

65

I don't know

Face Validity: IQ Test

Example

2. If you rearrange the letters NLIRBE, you would have the name of a

River

Country

City

Animal

Plant

I don't know

Face Validity: IQ Test Example

3. Napoleon lost his final battle at

Moscow _ Waterloo _ Leipzig _ Berlin _ Paris

Moscow

Waterloo

Leipzig

Berlin

Paris

I don't know

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 (Nunnally, 1978)

- A final exam for this course should measure information about measurement that was presented in this course -- it should not be validated in terms of its correspondence to some other course, e.g., biostatistics, epidemiology courses
- Spelling among fourth graders: random sample of all words in widely used fourth grade readers, in groups of four: test is to circle one incorrectly spelled word
- “Big Five” Personality Traits: systematic sampling, over generations of research, of all words in the English dictionary which describe personality

Major Depressive Episode (MDE): Diagnostic Criteria

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

Content Validity of CESD-R for MDE

- 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
- Sleep disturbance
 - I slept much more than usual
 - I had a lot of trouble getting to sleep
- ETC. (five more symptom groups)

Construct: A Theory About How Experiences Are Organized

- Something . . . “scientists put together from their own imaginations....” Nunnally
- “A mini-theory to explain the relationships among various behaviors or attitudes.” Streiner and Norman
- “A latent variable . . . that causes items to be correlated.... An unobservable . . . that varies among persons”
paraphrased from DeVellis
- “A hypothesis that a variety of items will correlate with one another and will be similarly affected by such factors as experimental treatments, psychosocial factors, and biology” Miech

Constructs

- Intelligence
- Neuroticism
- Feminism
- Stress
- Distress
- Job Strain
- Social Class
- Dropsy
- Fever
- Gulf War Syndrome
- Interstitial Cystitis
- Rheumatoid Arthritis
- Planet
- Selenium

Construct Validity

“the extent to which an operational measure truly reflects the concept being investigated or 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”

Steps in Construct Validation

(Streiner and Norman from Cronbach and Meehl, 1955)

1. Spell out a set of theoretical constructs
2. State how they should be related
3. Develop measures for the constructs
4. Conduct studies to see whether the observed relationships between measures (step 3) agrees with the stated theories (step 2)

Two Major Aspects to Evaluate Construct Validity

- 1. Internal construct validity:** the degree to which items in the measure are associated with each other in the theoretically predicted direction
 - a. Discriminant validity:** the degree to which a scale is associated with measures of similar constructs and not associated with measures of distinct constructs
 - b. Convergent validity:** the degree to which a scale is associated with measures of similar constructs even when they are measured with a different modality
- 2. External/nomological construct validity:** the degree to which a scale is associated with other constructs in the theoretically predicted direction

Modalities of Measurement

- Clinical Rating
- Examination
- Self-report-- Structured Interview
- Telephone Interview
- Computer-assisted Interview
- Paper and pencil
- Informant Interview
- Biological assay

- Variable analysis– Herbert Blumer
- Monomethod bias
- Post-hoc validation

Failure of Discriminant Validity Estimates of Correlation between Scales if Keyed for the Same Time Period

Scale	Corr
Langner with CESD	.852
Langner with GWB	-.976
Langner with GWB	-.932
Langner with SCL-90	.802
HOS with CESD	.752
HOS with GWB	-.856
CESD with SCL-90	.800
CESD with SCL-90	.930
GWB with CES-D	-.796
GWB with SCL-90	-.760

source: *Mental Illness in the United States: Epidemiological Estimates*, Dohrenwend and Dohrenwend, eds. 1982.

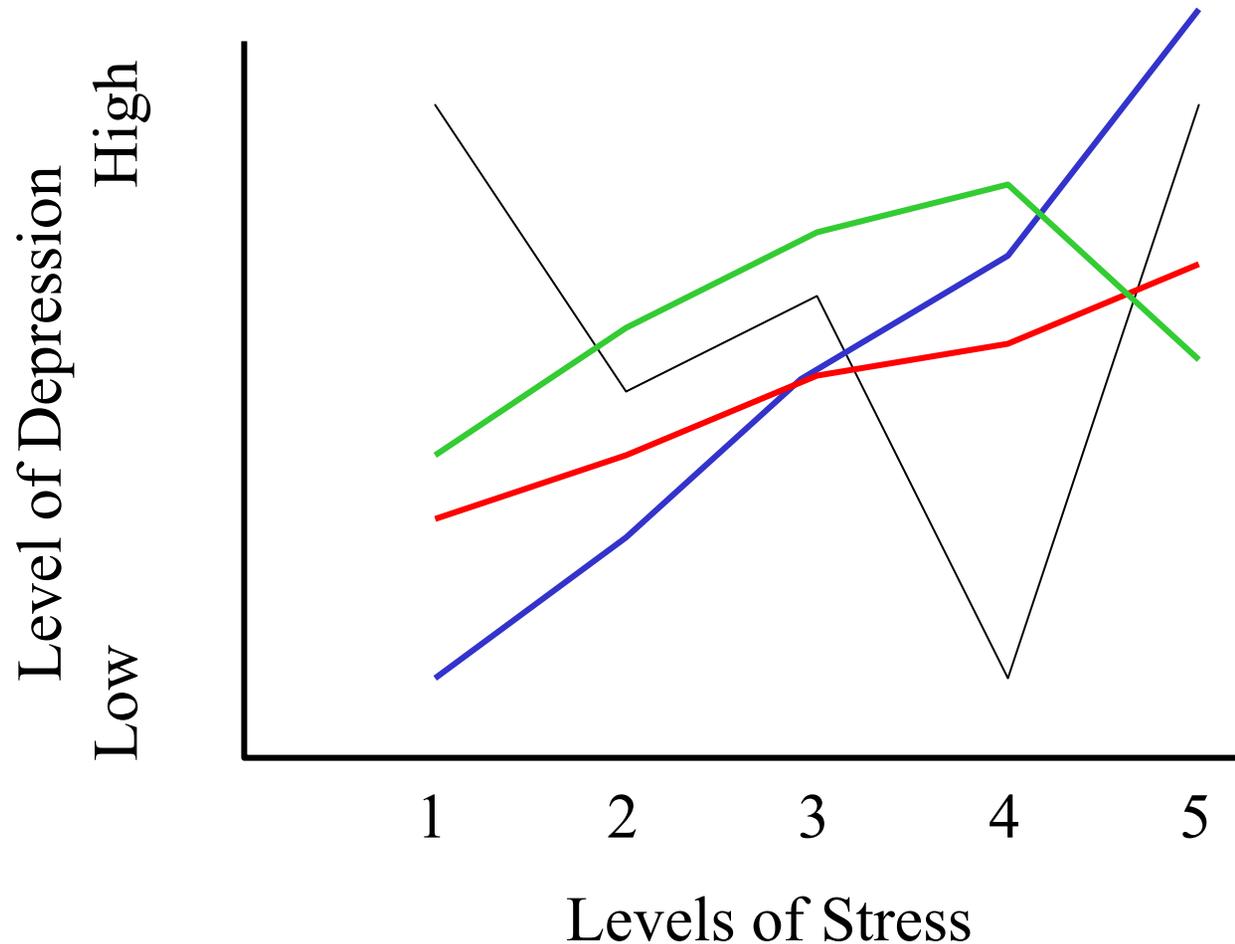
Known Groups Validation

Consumers' Need for Uniqueness Scale

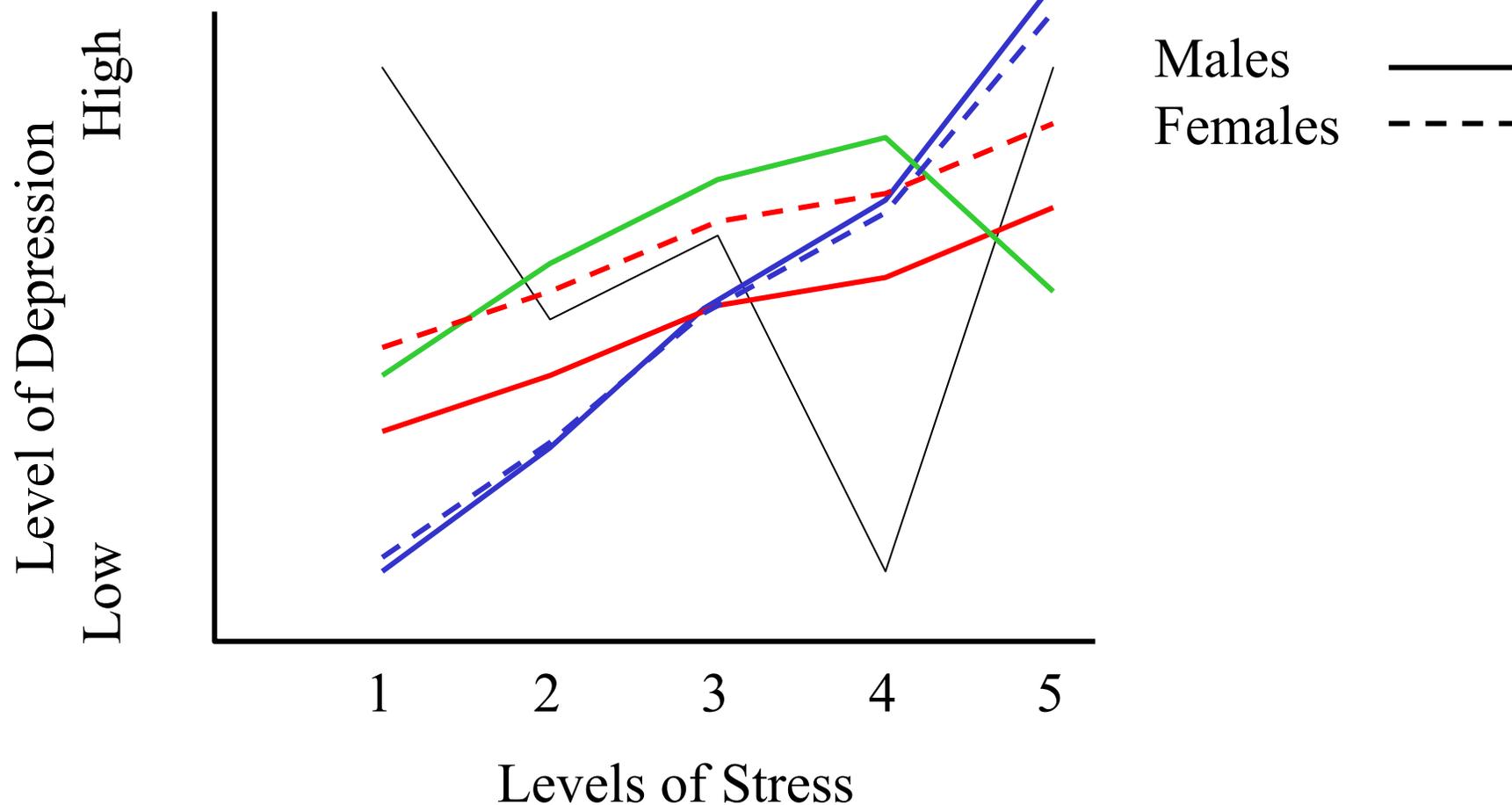
Group	Sample Size	Mean
Convenience Group Mail Survey	621	2.60
Tattoo artists	39	3.05
Owners of customized low-rider autos	22	2.99
Members of Society for Creative Anachronism	21	2.91
Art majors	22	3.06

Netemeyer et al, p. 81, from Tian et al, JCR, 2001

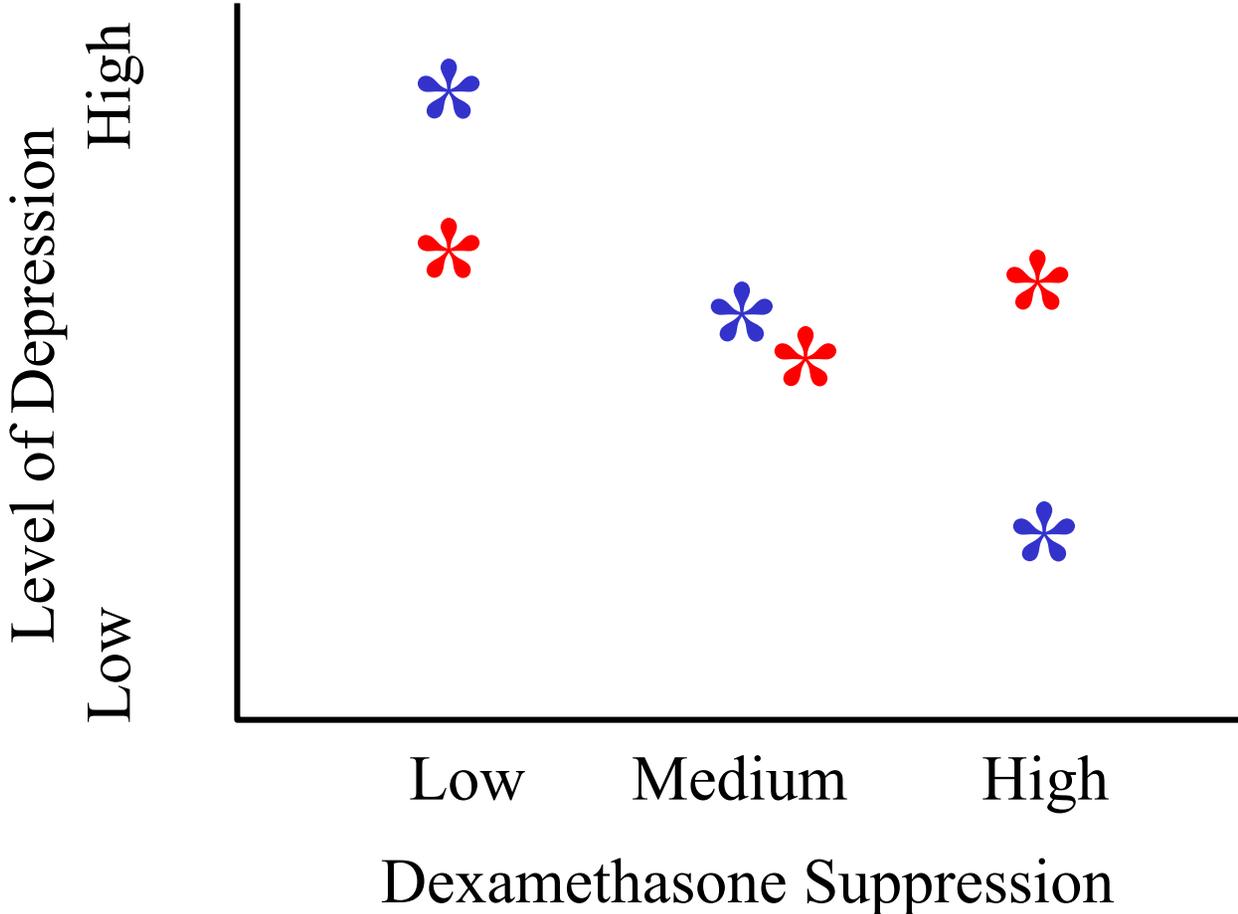
Construct Validation for Four Measures of Depression: Stress and Depression



Construct Validation for Four Measures of Depression: Gender and Depression



Construct Validation for Four Measures of Depression: Dexamethasone Suppression and Depression



Criterion Validation

- Requires a well-accepted criterion
 - “Gold Standard”
 - “Lead Standard”
- Useful only when new test has some outstanding characteristic in comparison to the accepted one, e.g.,
 - Expensive (Structured Diagnostic Interview versus Psychiatrist’s Examination)
 - Invasive (e.g., PSA versus digital examination)
 - Dangerous (Mantoux test for TB versus Chest X-ray)
 - Time-consuming
- Criterion validity highlights correspondence alone
- Generates quantitative measure of validity

Criterion validity

Association of a test measure with a criterion variable

- Concurrent validity
 - correlation of CESD with HAMD
 - correlation of “Do you feel hot?” with temperature from thermometer
 - Correlation of donation pledge with actual contribution
 - Correlation of self-report personality traits with peer reports of same traits
- Predictive Validity
 - success in college from SAT’s; grad school from GRE’s
 - Psychotic episode during young adulthood from psychosis proneness measures like magical ideation, recurrent illusions, social avoidance at beginning of college
 - General Health Questionnaire with visit to psychiatrist

CESD items

I felt sad

I felt depressed

Nothing made me happy

I lost interest in my usual
activities

I lost a lot of weight without
trying to

My appetite was poor

I slept much more than usual

I had a lot of trouble getting to
sleep

ETC. (five more symptom groups)

CAGE Screening for Alcoholism

- Cut down on drinking-have tried repeated without success (*Yes/No*)
- Annoyed by criticism about drinking habits (*Yes/No*)
- Guilty feelings about drinking (*Yes/No*)
- Eye opener drink needed in the morning (*Yes/No*)

Sensitivity and Specificity of CAGE

Pooled value	CAGE score	Sensitivity	Specificity
All studies	1	0.87	0.68
	2	0.71	0.90
	3	0.42	0.97
	4	0.20	0.99
Primary care	1	0.85	0.78
	2	0.71	0.91
	3	0.45	0.98
	4	0.23	0.99
Ambulatory medical patients	1	0.83	0.50
	2	0.60	0.92
	3	0.33	0.98
	4	0.13	0.99
Inpatients	1	0.98	0.56
	2	0.87	0.77
	3	0.50	0.92
	4	0.23	0.99

A diagnostic meta-analysis of 10 studies of general clinical populations. Criterion is DSM abuse or dependence. Adapted from Aertgeerts et al. The value of the CAGE in screening for alcohol abuse and alcohol dependence in general clinical populations: a diagnostic meta-analysis. *J Clin Epidemiol* 2004;57:30-39.

The Patient Health Questionnaire - 9 (PHQ-9)

- Developed by Spitzer et al., a *self-administered* version of the depression module of the PRIME-MD
- Designed to be used in clinical settings so primary care practitioners can efficiently screen for depression
- 9 symptom items and 2 questions about functional impairment

Phrasing of the PHQ-9

- Over the past **2 weeks**, how often have you been bothered by any of the following problems?
 - *For each item, the answer choice are “Not at all” - 0 points, “Several days” - 1 point, “More than half the days” - 2 points and “Nearly every day” - 3 points.*

Symptoms from the PHQ-9

- 1. Feeling down, depressed or hopeless?**
- 2. Little interest or pleasure in doing things?**
- 3. Trouble falling/staying asleep, sleeping too much?**
- 4. Feeling tired or having little energy?**
- 5. Poor appetite or overeating?**
- 6. Feeling bad about yourself - or that you are a failure or have let yourself or your family down?**
- 7. Trouble concentrating on things, such as reading the newspaper or watching television?**
- 8. Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual?**
- 9. Thoughts that you would be better off dead or of hurting yourself in some way?**

PHQ-9: Validity and Reliability

Reliability		Validity	
Internal Consistency – Cronbach’s α	0.86	Sensitivity*	98%
Test-Retest Correlation	0.84	Specificity*	80%

*Compared to the “lead standard” in psychiatry - the Structured Clinical Interview for the *DSM* (SCID)

Agreement Between DIS and SCAN for Lifetime Panic Disorder: Baltimore ECA Follow-Up

Interview Using DIS	Psychiatrist Using SCAN		
	Never a case	Positive Diagnosis	Total
Never a case	305	18	323
Positive diagnosis	1	7	8
Total	306	25	331

$$P_o = (305+7)/331 = .94$$

$$P_e = (323/331) * (306/331) = .90$$

Continued

Agreement between DIS and SCAN for Lifetime Panic Disorder: Baltimore ECA Follow-Up

- Sensitivity = $7/25 = 0.28$
- Specificity = $305/306 = 0.99$
- Kappa = $(P_o - P_e)/(1 - P_e)$
= $(.94 - .90)/(1 - .90) = 0.40$

Relationship of Kappa to Sensitivity and Specificity

Prevalence	Sensitivity	Specificity	Kappa
0.01	0.95	0.95	0.14
0.01	0.50	0.99	0.16
0.01	0.99	0.50	0.01
0.10	1.00	0.90	0.47
0.10	0.95	0.95	0.61
0.10	0.90	0.90	0.39
0.25	0.95	0.95	0.76
0.25	0.50	0.99	0.39
0.25	0.99	0.50	0.19

The Relationship of Reliability to Validity

- Internal consistency ~ ? Internal construct validity
- Reliability sets a maximum for validity
 - $\rho_{xy} \leq \sqrt{\rho_{xx}}$
- Validity establishes reliability
- Value of positive versus negative findings in a new science
- Where there is no gold standard:
 - Reliability = consistency on measurement of identical construct with maximally *similar* methods
 - Validity = consistency of measurement of identical construct with maximally *different* methods

Validity and Utility

- Measures versus Constructs
- Internal versus External construct validity
- Bandwidth versus Fidelity
- Range of difficulty
 - CAT for SAT's
 - ADL's
- Syndrome heterogeneity
 - Rheumatoid arthritis
 - Depressive Disorder

Effect of Invalidity on Prevalence

$$P_o = \text{Sens} \times P_t + (1-\text{Spec}) \times (1-P_t)$$

The observed prevalence equals the sum of the product of the sensitivity multiplied by the true prevalence (cases) and one minus the specificity multiplied by one minus the true prevalence (non-cases).

After Rogan and Gladen, AJE, 1978)

Criterion Validity and Prevalence

True Prevalence	Sensitivity	Specificity	Observed Prevalence
0.01	0.95	0.95	.059
0.01	0.50	0.99	.015
0.01	0.99	0.50	.505
0.10	1.00	0.90	.190
0.10	0.95	0.95	.140
0.10	0.90	0.90	.180
0.25	0.95	0.95	.275
0.25	0.50	0.99	.132
0.25	0.99	0.50	.622

Measurement and Statistical Parameters

- “Random” error and
 - Correlation coefficient-- attenuated
 - Covariance -- unaffected
 - Regression coefficient -- affected by errors in x
 - Prevalence – usually overestimated
 - Incidence – usually overestimated
 - Odds ratios– usually attenuated

Empirical Tests to Help Assess Validity (next lecture)

1. Criterion validity
 - a. sensitivity and specificity
 - b. ROC curves
2. Construct validity
 - a. multitrait-multimethod matrix
 - b. factor analysis