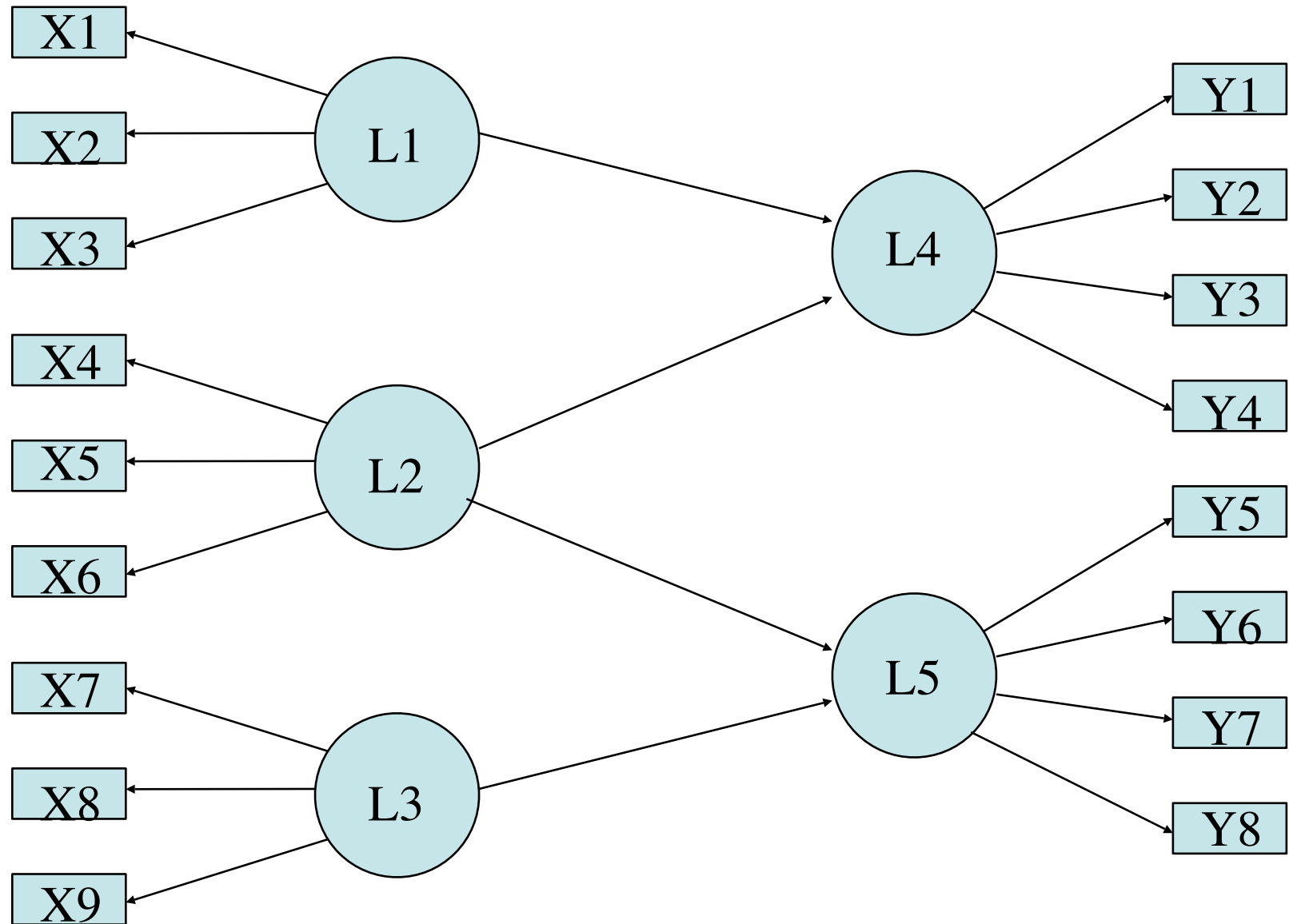


# Psychometric Theory: A conceptual Syllabus



# A Theory of Data: What can be measured

X1

What is measured?

Individuals

Objects

What kind of measures are taken?

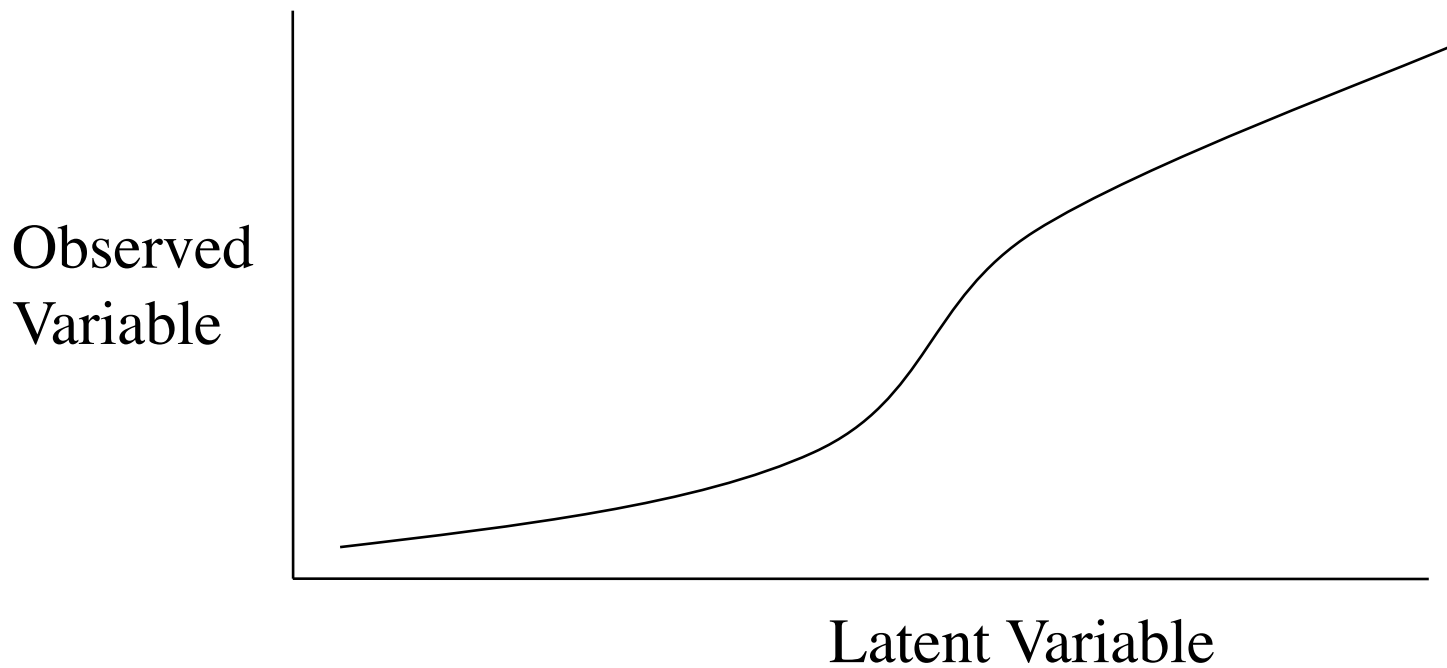
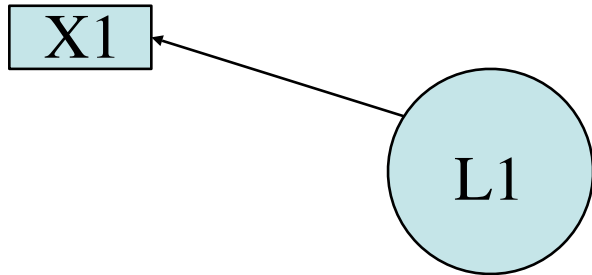
proximity

order

Comparisons are made on:

Single Dyads or Pairs of Dyads

# Scaling: the mapping between observed and latent variables



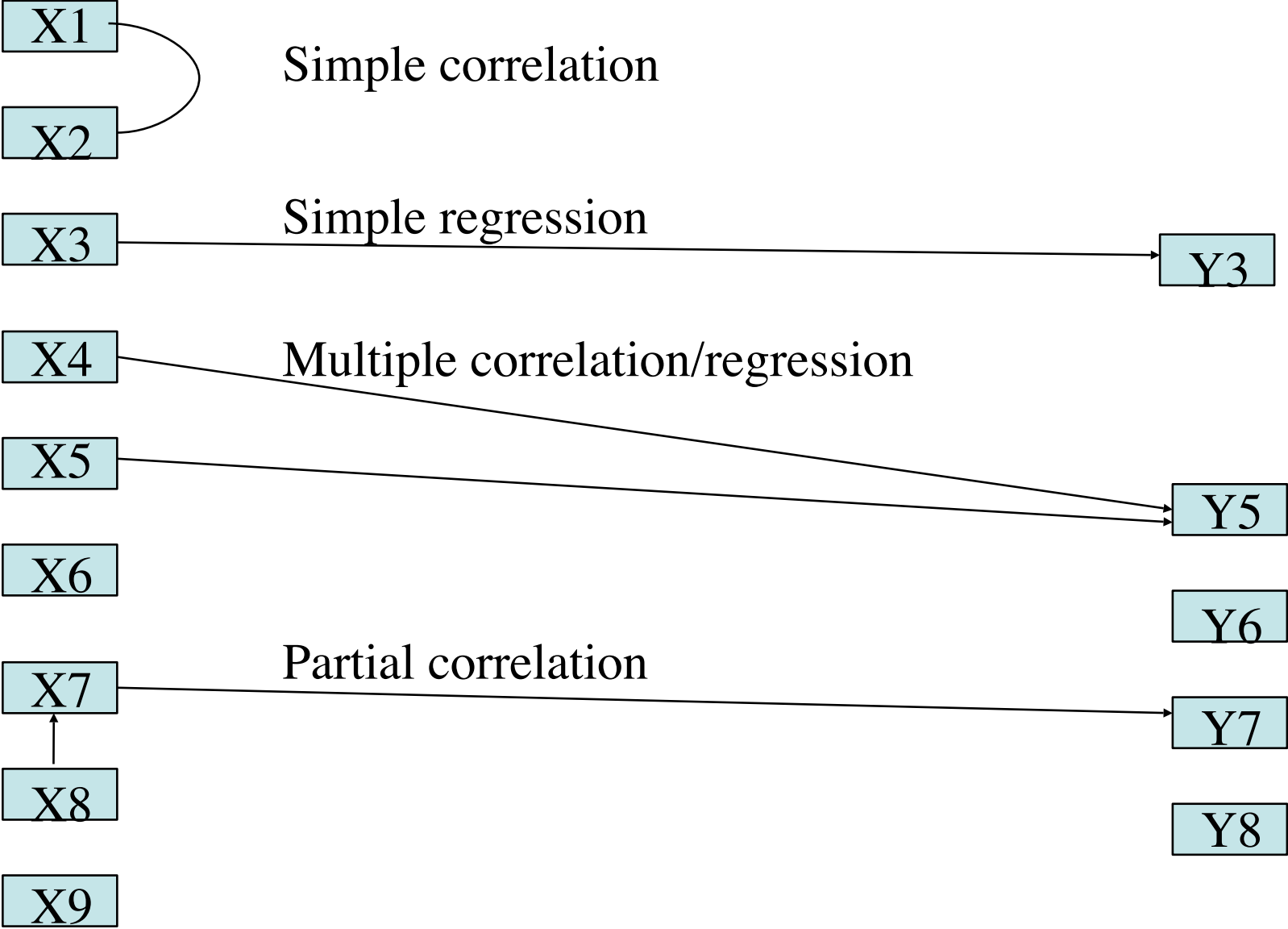
# Where are we?

- Issues in what types of measurements we can take (Theory of Data)
- Scaling and the shape of the relationship between latent variables and observed variables
- Measures of central tendency
- Measures of variability and dispersion
- Measures of relationships

# Measures of relationship

- Regression  $y = bx + c$ 
  - $b_{y.x} = \text{Cov}_{xy} / \text{Var}_x$
- Correlation
  - $r_{xy} = \text{Cov}_{xy} / \sqrt{V_x * V_y}$
  - Pearson Product moment correlation
    - Spearman (ppmc on ranks)
    - Point biserial (x is dichotomous, y continuous)
    - Phi (x, y both dichotomous)

# Variance, Covariance, and Correlation



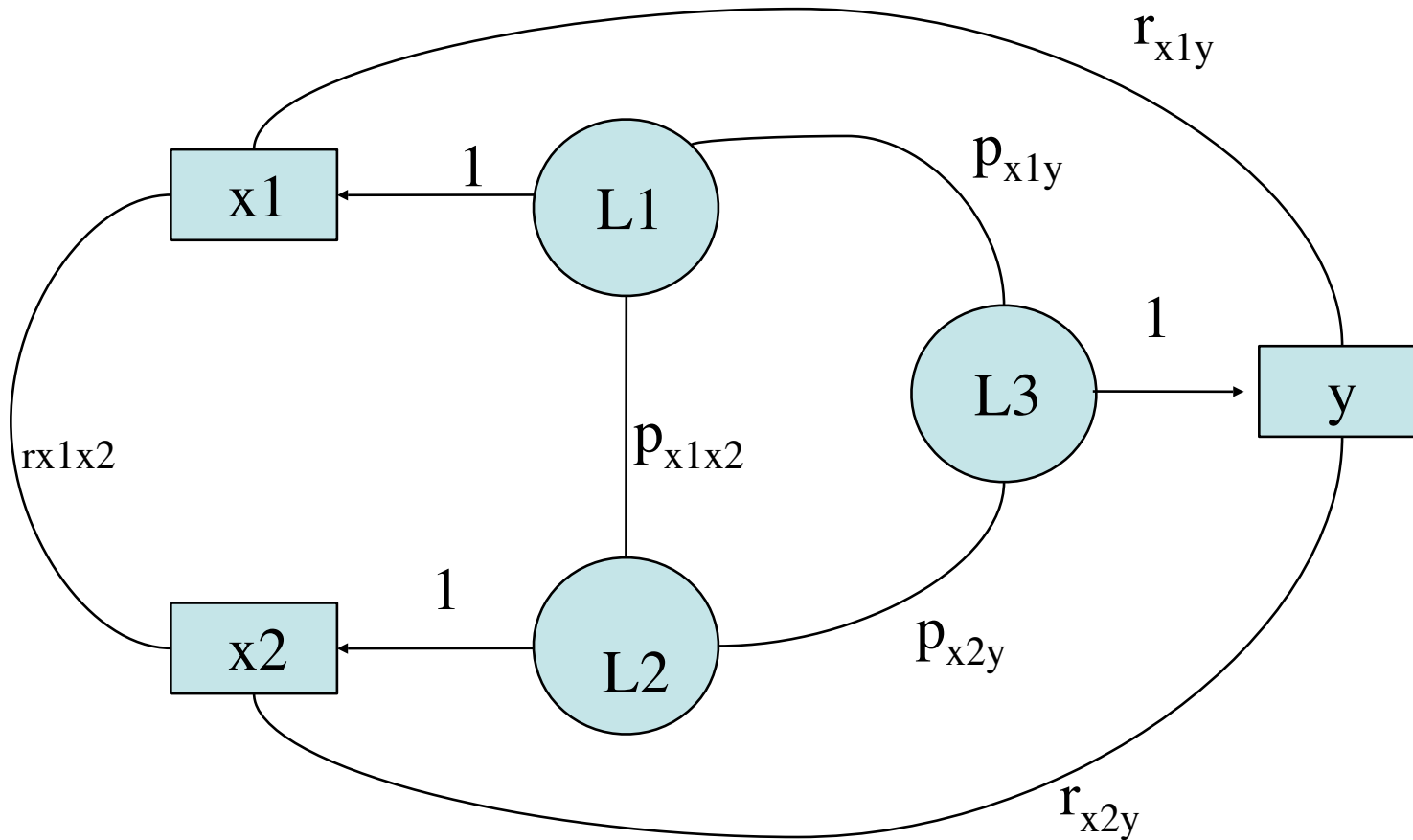
# Measures of relationships with more than 2 variables

- Partial correlation
  - The relationship between  $x$  and  $y$  with  $z$  held constant ( $z$  removed)
- Multiple correlation
  - The relationship of  $x_1 + x_2$  with  $y$
  - Weight each variable by its independent contribution

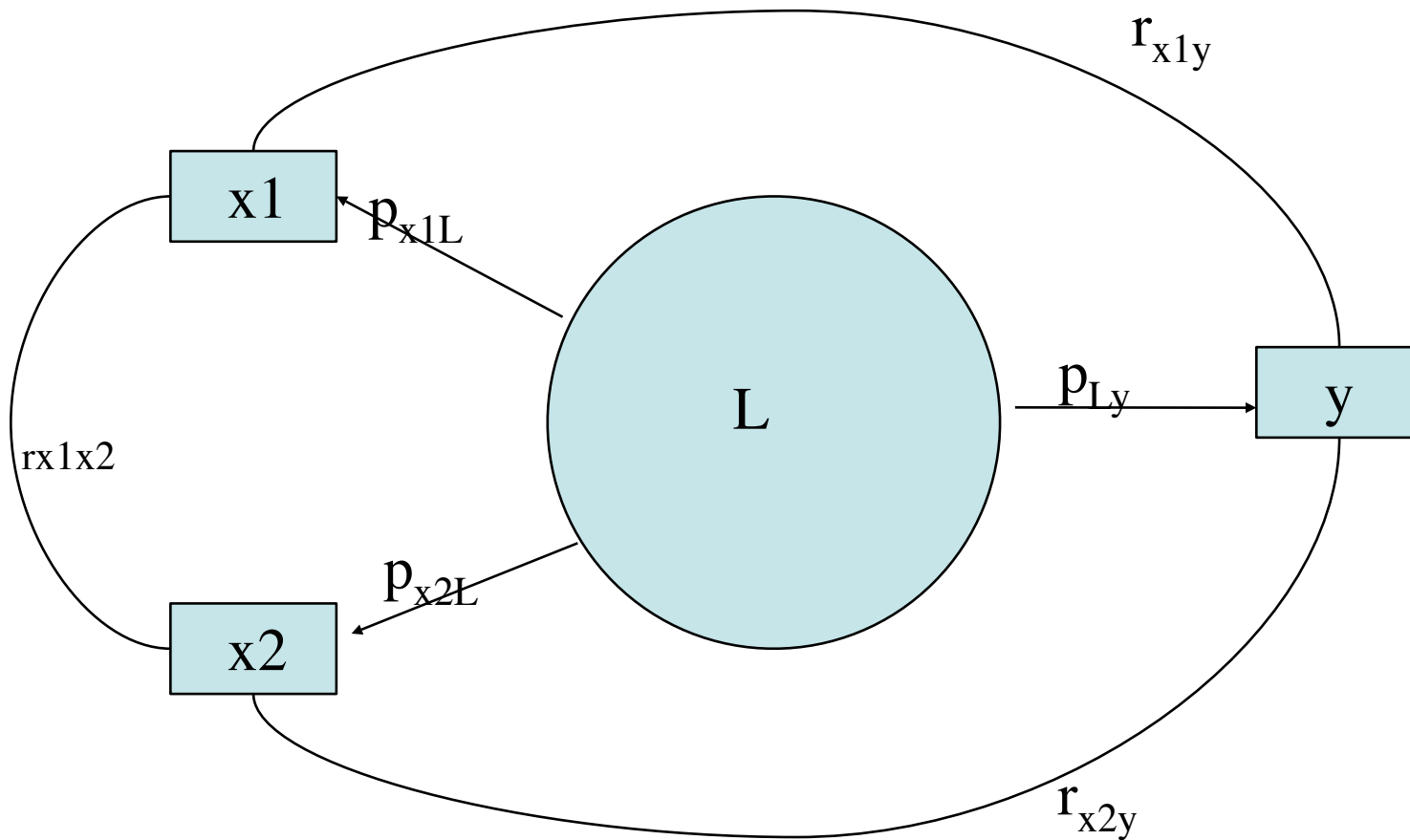
# Problems with correlations

- Simpson's paradox and the problem of aggregating groups
  - Within group relationships are not the same as between group or pooled relationships
- Phi coefficients and the problem of unequal marginals
- Alternative interpretations of partial correlations

# Partial correlation: conventional model



# Partial correlation: Alternative model



# Partial Correlation: classical model

	X <sub>1</sub>	X <sub>2</sub>	Y
X <sub>1</sub>	1.00		
X <sub>2</sub>	.72	1.00	
Y	.63	.56	1.00

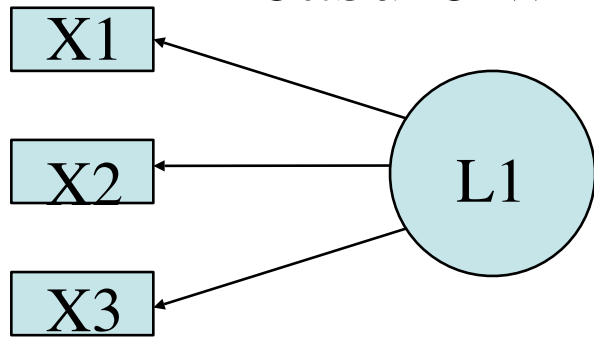
$$\text{Partial } r = (r_{x_1y} - r_{x_1x_2} * r_{x_2y}) / \sqrt{((1 - r_{x_1x_2}^2) * (1 - r_{x_2y}^2))}$$

$R_{x_1y.x_2} = .33$  (traditional model) but = 0 with structural model

# Reliability Theory

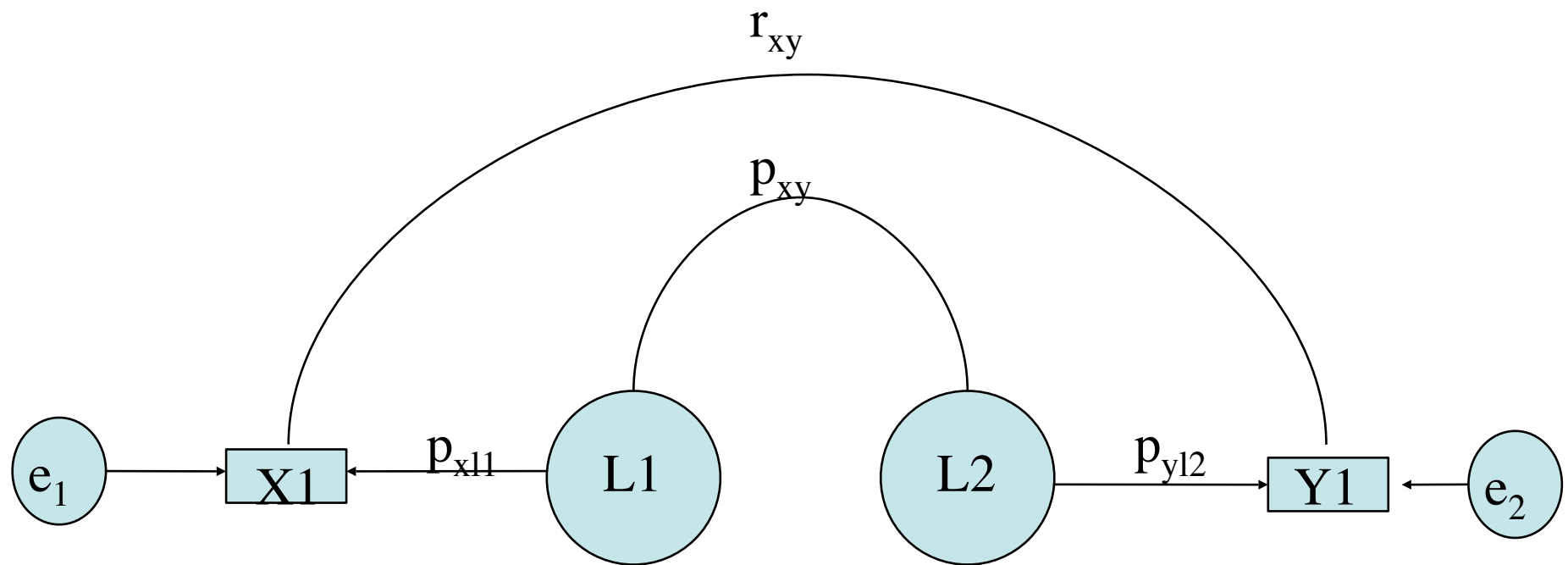
Classical and modern approaches

# Classic Reliability Theory: How well do we measure what ever we are measuring



## Classic Reliability Theory:

How well do we measure what ever we are measuring and what is the relationships between latent variables



## Classic Reliability Theory:

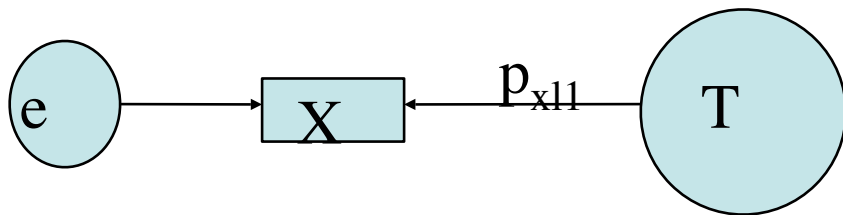
How well do we measure what ever we are measuring

What is the relationship between  $X_1$  and  $L_1$ ?

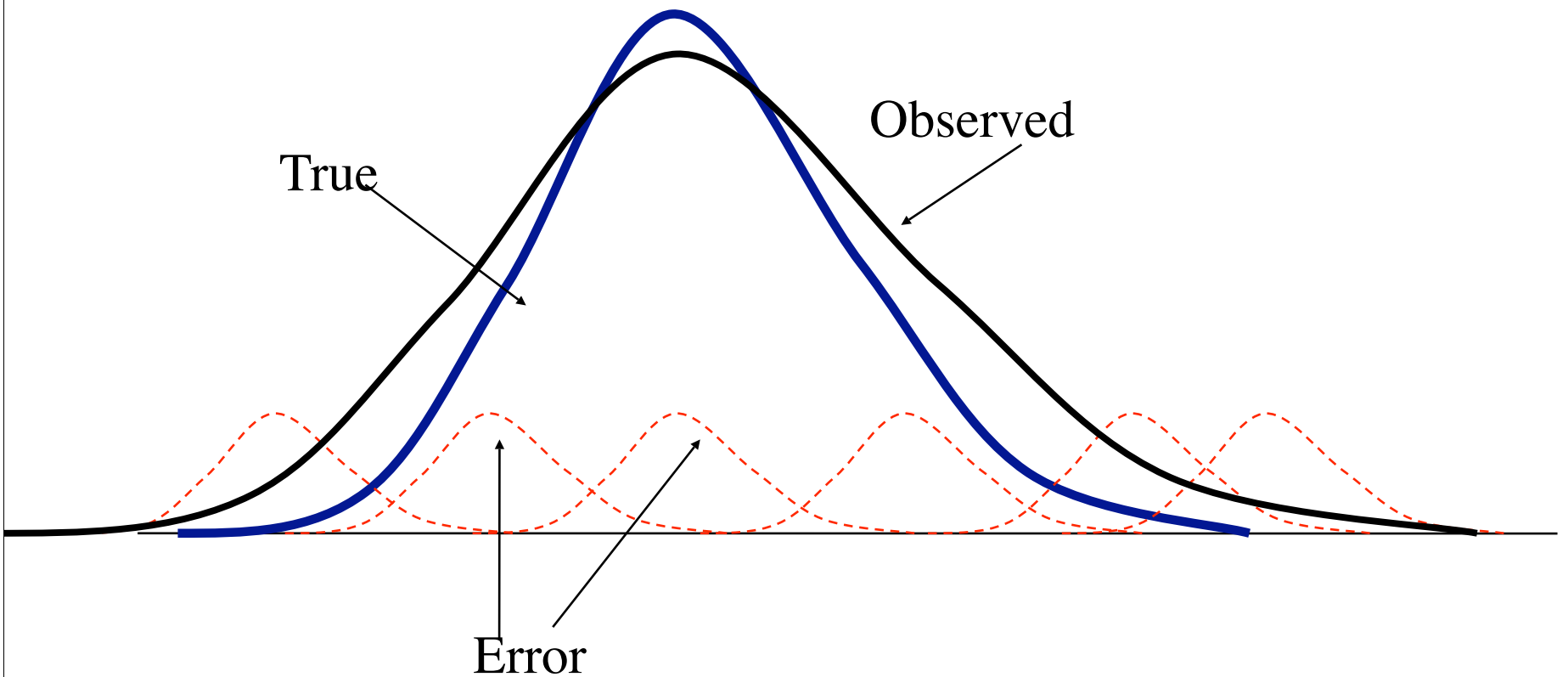
What is the variance of  $X_1$ ,  $L_1$ , and  $E_1$ ?

Let True Score for Subject I = expected value of  $X_i$ .

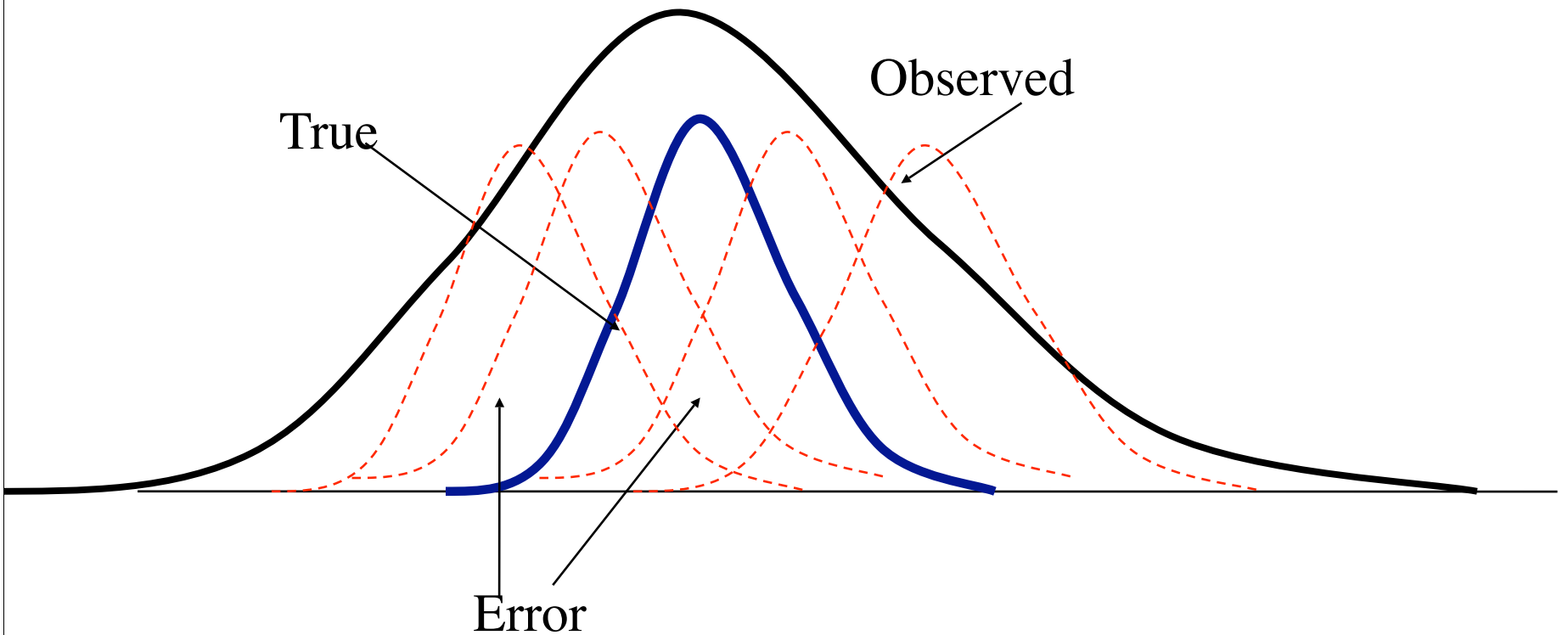
(note that this is not the Platonic Truth, but merely the average over an infinite number of trials.)



Observed = True + Error



Observed = True + Error



# Observed = Truth + Error

- Define True score as expected observed score. Then Truth is uncorrelated with error, since the mean error for any True score is 0.
- Variance of Observed = Variance (T+E)=  
 $V(T) + V(E) + 2Cov(T,E) = V_t + V_e$
- Covariance O,T =  $Cov_{(T+E),T} = V_t$
- $p_{ot} = C_{ot} / \sqrt{V_o * V_t} = V_t / \sqrt{V_o * V_t} = \sqrt{V_t / V_o}$
- $p^2_{ot} = V_t / V_o$  (the squared correlation between observed and truth is the ratio of true score variance to observed score variance)

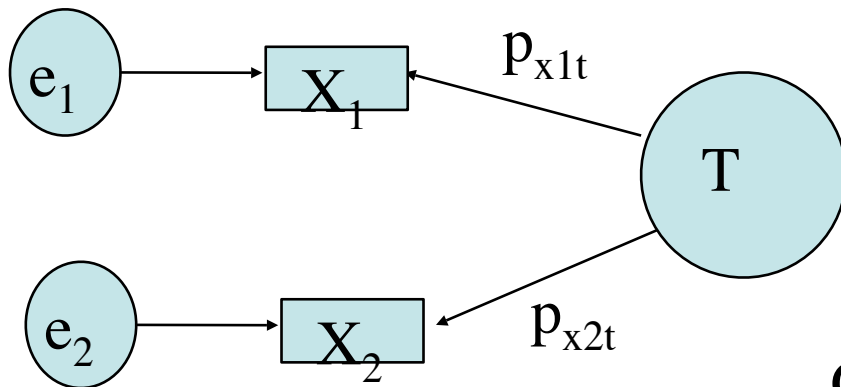
# Estimating True score

- Given that  $p_{ot}^2 = V_t/V_o$  and  $p_{ot} = \sqrt{V_t/V_o}$ , then for an observed score  $x$ , the best estimate of the true score can be found from the prediction equation:
- $Z_t = p_{ox}Z_x$
- The problem is, how do we find the variance of true scores and the variance of error scores?

# Estimating true score: regression artifacts

- Consider the effect of reward and punishment upon pilot training:
  - From 100 pilots, reward the top 50 flyers, punish the worst 50.
  - Observation: praise does not work, blame does!
  - Explanation?

# Parallel Tests



$$V_{x1} = V_t + V_{e1}$$

$$V_{x2} = V_t + V_{e2}$$

$$C_{x1x2} = V_t + C_{te1} + C_{te2} + C_{e1e2} = V_t$$

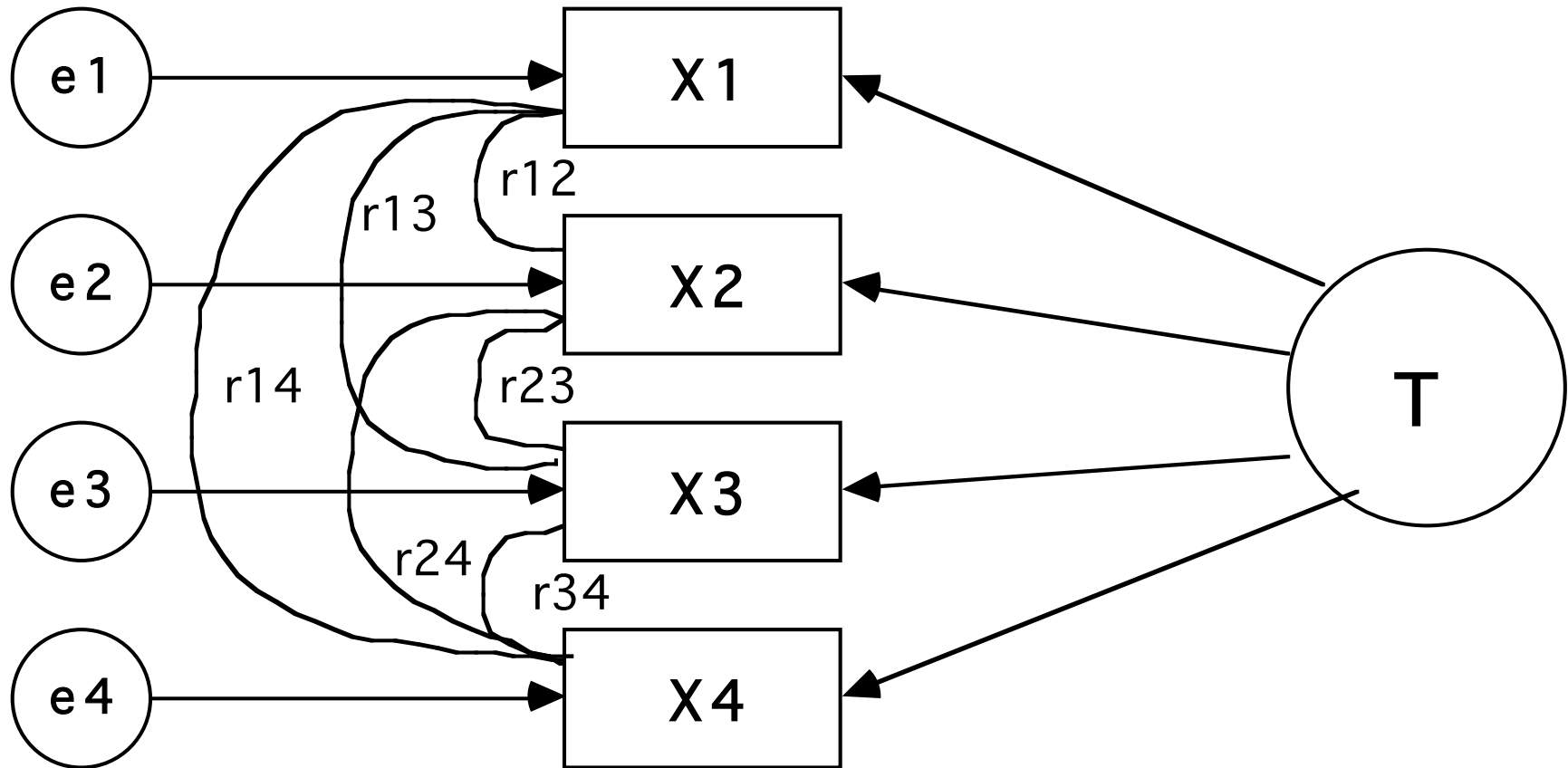
$$r_{xx} = C_{x1x2} / \text{Sqrt}(V_{x1} * V_{x2}) = V_t / V_x$$

The reliability of a test is the ratio of the true score variance to the observed variance = the correlation of a test with a test “just like it”

# Reliability and parallel tests

- $r_{x_1x_2} = V_t/V_x = r_{xt}^2$
- The reliability is the correlation between two parallel tests and is equal to the squared correlation of the test with the construct.  $r_{xx} = V_t/V_x =$  percent of test variance which is construct variance.
- $r_{xt} = \text{sqrt}(r_{xx}) \implies$  the validity of a test is bounded by the square root of the reliability.
- How do we tell if one of the two “parallel” tests is not as good as the other? That is, what if the two tests are not parallel?

# Congeneric Measurement



# Observed Variances/Covariances

	$X_1$	$X_2$	$X_3$	$X_4$
$X_1$	$V_{X_1}$			
$X_2$	$c_{x_1x_2}$	$V_{X_2}$		
$X_3$	$c_{x_1x_3}$	$c_{x_2x_3}$	$V_{X_3}$	
$X_4$	$c_{x_1x_4}$	$c_{x_3x_4}$	$c_{x_3x_4}$	$V_{X_4}$

# Model Variances/Covariances

	$x_1$	$x_2$	$x_3$	$x_4$
$x_1$	$V_t + Ve_1$			
$x_2$	$c_{x_1tx_2t}$	$V_t + Ve_2$		
$x_3$	$c_{x_1tx_3t}$	$c_{x_2tx_3t}$	$V_t + Ve_3$	
$x_4$	$c_{x_1tx_4t}$	$c_{x_3tx_4t}$	$c_{x_3tx_4t}$	$V_t + Ve_4$

# Observed and modeled Variances/Covariances

	$x_1$	$x_2$	$x_3$	$x_4$
$x_1$	$V_{x_1}$			
$x_2$	$c_{x_1x_2}$	$V_{x_2}$		
$x_3$	$c_{x_1x_3}$	$c_{x_2x_3}$	$V_{x_3}$	
$x_4$	$c_{x_1x_4}$	$c_{x_3x_4}$	$c_{x_3x_4}$	$V_{x_4}$
	$x_1$	$x_2$	$x_3$	$x_4$
$x_1$	$V_t + Ve_1$			
$x_2$	$c_{x_1tx_2t}$	$V_t + Ve_2$		
$x_3$	$c_{x_1tx_3t}$	$c_{x_2tx_3t}$	$V_t + Ve_3$	
$x_4$	$c_{x_1tx_4t}$	$c_{x_3tx_4t}$	$c_{x_3tx_4t}$	$V_t + Ve_4$

# Estimating parameters of the model

1. Variances:  $V_t, Ve_1, Ve_2, Ve_3, Ve_4$
2. Covariances:  $Ctx_1, Ctx_2, Ctx_3, Ctx_4$
3. Parallel tests: 2 tests, 3 equations, 5 unknowns,  
assume  $Ve_1 = Ve_2,$   $Ctx_1 = Ctx_2$
4. Tau Equivalent tests: 3 tests, 6 equations, 7  
unknowns, assume
  1.  $Ctx_1 = Ctx_2 = Ctx_3$  but allow unequal error variance
5. Congeneric tests: 4 tests, 10 equations, 9 unknowns!

# Domain Sampling theory

- Consider a domain ( $D$ ) of  $k$  items relevant to a construct. (E.g., English vocabulary items, expressions of impulsivity). Let  $D_i$  represent the number of items in  $D$  which the  $i$ th subject can pass (or endorse in the keyed direction) given all  $D$  items. Call this the domain score for subject  $I$ . What is the correlation (across subjects) of scores on an item  $j$  with the domain scores?

# Correlating an Item with Domain

1. Correlation =  $\text{Cov}_{id} / \sqrt{(V_j * V_d)}$
2.  $\text{Cov}_{id} = V_j + \sum c_{lj} = V_j + (k-1) * \text{average cov}_j$
3. Domain variance ( $V_d$ ) = sum of item variances + item covariances in domain =
4.  $V_d = k * (\text{average variance}) + k * (k-1) \text{ average covar}$
5. Let  $V_a$  = average variance,  $C_a$  = average covariance
6. Then  $V_d = k(V_a + (k-1) * C_a)$

# Correlating an Item with Domain

1. Assume that  $V_j = V_a$  and  $C_{jl} = C_a$
2.  $r_{jd} = C_{jd} / \sqrt{V_j * V_d}$
3.  $r_{jd} = (V_a + (k-1)C_a) / \sqrt{V_a * k(V_a + (k-1)C_a)}$
4.  $r_{jd}^2 =$
5.  $(V_a + (k-1)C_a) * (V_a + (k-1)C_a) / (V_a * k(V_a + (k-1)C_a))$
6. Now, find the limit of  $r_{jd}^2$  as  $k$  becomes large:
7.  $\lim_{k \rightarrow \infty} r_{jd}^2 = C_a / V_a = \text{av covar} / \text{av variance}$
8. I.e., the amount of domain variance in an average item (the squared correlation of an item with the domain) is the average intercorrelation in the domain

# Domain Sampling 2: correlating an n item test with the domain

1. What is the correlation of a test with n items with the domain score?
2. Domain variance =  $\Sigma(\text{variances}) + \Sigma(\text{covars})$
3. Variance of n item test =  $\Sigma v_j + \Sigma c_{jl} =$ 
  1.  $V_n = n * V_a + n * (n-1) C_a$
4.  $r_{nd} = C_{nd} / \text{sqrt}(V_n * V_d)$      $r_{nd}^2 = C_{nd}^2 / (V_n * V_d)$

# Squared correlation with domain

$$r_{nd}^2 = \frac{\{n \cdot V_a + n \cdot (k-1)C_a\} \cdot \{n \cdot V_a + n \cdot (k-1)C_a\}}{\{n \cdot V_a + n \cdot (n-1) \cdot C_a\} \cdot \{k(V_a + (k-1)C_a)\}}$$

$$r_{nd}^2 = \frac{\{V_a + (k-1)C_a\} \cdot \{n \cdot V_a + n \cdot (k-1)C_a\}}{\{V_a + (n-1) \cdot C_a\} \cdot \{k(V_a + (k-1)C_a)\}} \implies$$

$$r_{nd}^2 = \frac{\{n \cdot V_a + n \cdot (k-1)C_a\}}{\{V_a + (n-1) \cdot C_a\} \cdot \{k\}}$$

## Limit of squared r with domain

$$r_{nd}^2 = \frac{\{n * V_a + n * (k-1) C_a\}}{\{V_a + (n-1) * C_a\} * \{k\}}$$

$$\lim \text{ as } k \rightarrow \infty \text{ of } r_{nd}^2 = \frac{n * C_a}{V_a + (n-1) C_a}$$

The amount of domain variance in a n-item test (the squared correlation of the test with the domain) is a function of the number of items in the test and the average covariance within the test.

# Coefficient Alpha

Consider a test made up of  $k$  items with an average intercorrelation  $r$

2. What is the correlation of this test with another test sampled from the same domain of items?
3. What is the correlation of this test with the domain?

# Coefficient alpha

	Test 1	Test 2
Test 1	$V_1$	$C_{12}$
Test 2	$C_{12}$	$V_2$

$$r_{x_1x_2} = \frac{C_{12}}{\sqrt{V_1 * V_2}}$$

# Coefficient alpha

Let  $r_1$  = average correlation within test 1

Let  $r_2$  = average correlation within test2

Let  $r_{12}$  = average correlation between items in test 1 and test 2

	Test 1	Test 2
Test 1	$V_1 = k * [1 + (k-1) * r_1]$	$C_{12} = k * k * r_{12}$
Test 2	$C_{12} = k * k * r_{12}$	$V_2 = k * [1 + (k-1) * r_2]$

$$r_{x_1x_2} = \frac{k * k * r_{12}}{\sqrt{k * [1 + (k-1) * r_1] * k * [1 + (k-1) * r_2]}}$$

# Coefficient Alpha

$$r_{X_1X_2} = \frac{k * k * r_{12}}{\sqrt{k * [1+(k-1) * r_1] * k * [1+(k-1) * r_2]}}$$

But, since the two tests are composed of randomly equivalent items,  $r_1=r_2=r_{12}$  and

$$r_{X_1X_2} = \frac{k * r}{1+(k-1)r} = \text{alpha} = \alpha$$

# Coefficient alpha

Let  $r_1$  = average correlation within test 1 =  $r$  (by sampling)

Let  $r_2$  = average correlation within test2 =  $r$  (by sampling)

Let  $r_{12}$  = average correlation between items in test 1 and test 2 =  $r$

	Test 1	Test 2
Test 1	$V_1 = k * [1 + (k-1) * r]$	$C_{12} = k * k * r$
Test 2	$C_{12} = k * k * r$	$V_2 = k * [1 + (k-1) * r]$

$$r_{x_1x_2} = \frac{k * r}{1 + (k-1)r} = \text{alpha} = \alpha$$

# Coefficient alpha and domain sampling

$$r_{X_1X_2} = \frac{k * r}{1 + (k-1)r} = \text{alpha} = \alpha$$

Note that this is the same as the squared correlation of a test with a test with the domain. Alpha is the correlation of a test with a test just like it and is the the percentage of the test variance which is domain variance (if the test items are all made up of just one domain).

# Coefficient alpha - another approach

Consider a test made up of  $k$  items with average variance  $v_1$ . What is the correlation of this test with another test sample from the domain?  
What is the correlation of this test with the domain?

	Test 1	Test 2
Test 1	$V_1$	$C_{12}$
Test 2	$C_{12}$	$V_2$

$$r_{x_1x_2} = \frac{C_{12}}{\sqrt{V_1 * V_2}}$$

# Coefficient alpha - from variances

- Let  $V_t$  be the total test variance test 1 = total test variance for test 2.
- Let  $v_i$  be the average variance of an item within the test.
- To find the correlation between the two tests, we need to find the covariance with the other test.

# Coefficient alpha

Let  $r_1$  = average correlation within test 1

Let  $r_2$  = average correlation within test 2

Let  $r_{12}$  = average correlation between items in test 1 and test 2

	Test 1	Test 2
Test 1	$V_1 = k * [v_i + (k-1) * c_1]$	$C_{12} = k * k * c_{12}$
Test 2	$C_{12} = k * k * c_{12}$	$V_2 = k * [v_i + (k-1) * c_2]$

$V_t = V_1 = V_2 \iff c_1 = c_2 = c_{12}$  (from our sampling assumptions)

# Alpha from variances

- $V_t = V_1 = k * [v_i + (k-1) * c_1] \Leftrightarrow$
- $c_1 = (V_t - \sum v_i) / ((k * (k-1)))$
- $C_{12} = k^2 c_{12} = k^2 * (V_t - \sum v_i) / ((k * (k-1)))$
- $r_{X_1 X_2} = (k^2 * (V_t - \sum v_i) / ((k * (k-1)))) / V_t =$
- $r_{X_1 X_2} = [(V_t - \sum v_i) / V_t] * (k / (k-1))$
- This allows us to find coefficient alpha without finding the average interitem correlation!

# The effect of test length on internal consistency

	Average r	Average r
Number of items	.2	.1
1	.20	.10
2	.33	.18
4	.50	.31
8	.67	.47
16	.80	.64
32	.89	.78
64	.94	.88
128	.97	.93

# Alpha and test length

- Estimates of internal consistency reliability reflect both the length of the test and the average inter-item correlation. To report the internal consistency of a domain rather than a specific test with a specific length, it is possible to report the “alpha 1” for the test:
- Average inter item  $r = \alpha_1 =$ 
  - $\alpha / (\alpha + k * (1 - \alpha))$
  - This allows us to find the average internal consistency of a scale independent of test length

# Split half estimates

	Xa	Xb	Xa'	Xb'
Xa	Va	Cab	Caa'	Cba'
Xb	Cab	Vb	Cab'	Cbb'
Xa'	Caa'	Cba'	Va'	Ca'b'
Xb'	Cab'	Cbb'	Ca'b'	Vb'

$$r_{12} = C_{12}/\text{sqrt}(V_1 * V_2)$$

$$C_{12} = C_{aa'} + C_{ba'} + C_{ab'} + C_{bb'} \approx 4 * C_{ab}$$

$$V_1 = V_2 = Va + Vb + 2C_{ab} \approx 2(V_a + C_{ab})$$

$$r_{12} = 2C_{ab}/(Va + Cab)$$

$$r_{12} = 2r_{ab}/(1 + r_{ab})$$

# Reliability and components of variance

- Components of variance associated with a test score include
- General test variance
- Group variance
- Specific item variance
- Error variance (note that this is typically confounded with specific)

# Components of variance - a simple analogy

- Height of Rockies versus Alps
- Height of base plateau
- Height of range
- Height of specific peak
- Snow or tree cover

# Coefficients Alpha, Beta, Omega

Test	General	Group	Specific	Error
Reliable	General	Group	Specific	
Common Shared	General	Group		
Alpha	General	< group		
Beta	≈general			
Omega	general			

# Alpha and reliability

- Coefficient alpha is the average of all possible splits and overestimates the general but underestimates the total common variance. It is a lower bound estimate of reliable variance.
- Beta and Omega are estimates of general variance.

# Alpha and Beta

## Find the least related subtests

	Subtest A	Subtest B	Subtest A'	Subtest B'
A	$g+G1+S+E$	$g$	$g$	$g$
B	$g$	$g+G2+S+E$	$g$	$g$
A'	$g$	$g$	$g+G3+S+E$	$g$
B'	$g$	$g$	$g$	$g+G4+S+E$

$$r_{12} = C_{12} / (\sqrt{V_1 * V_2}) = 2r_{ab} / (1 + r_{ab})$$

Beta is the worst split half reliability while alpha is the average

# Alpha and Beta with general and group factors

General Factor	Group Factor	Test Size = 10 items		Test Size = 20 items	
		Alpha	Beta	Alpha	Beta
0.25	0.00	0.77	0.77	0.87	0.87
0.20	0.05	0.75	0.71	0.86	0.83
0.15	0.10	0.73	0.64	0.84	0.78
0.10	0.15	0.70	0.53	0.82	0.69
0.05	0.20	0.67	0.34	0.80	0.51
0.00	0.25	0.63	0.00	0.77	0.00

# Generalizability Theory

## Reliability across facets

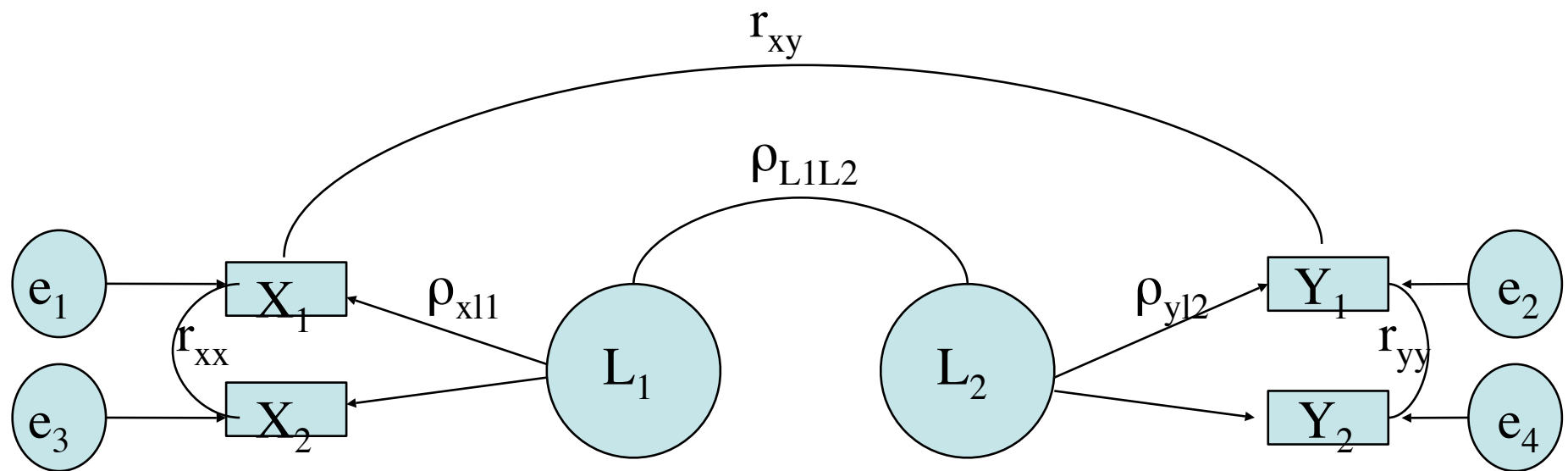
- The consistency of individual differences across facets may be assessed by analyzing variance components associated with each facet. I.e., what amount of variance is associated with a particular facet across which one wants to generalize.
- Generalizability theory is a decomposition of variance components to estimate sources of particular variance of interest.

# Facets of reliability

Across items	Domain sampling Internal consistency
Across time	Temporal stability
Across forms	Alternate form reliability
Across raters	Inter-rater agreement
Across situations	Situational stability
Across “tests” (facets unspecified)	Parallel test reliability

# Classic Reliability Theory: correcting for attenuation

How well do we measure what ever we are measuring  
and what is the relationships between latent variables



$$\rho_{xL1} = \text{sqrt}(r_{xx})$$

$$\rho_{yL2} = \text{sqrt}(r_{yy})$$

$$\rho_{L1L2} = r_{xy} / \rho_{xL1} \rho_{yL2}$$

$$\rho_{L1L2} = r_{xy} / \text{sqrt}(r_{xx} * r_{yy})$$

Disattenuated (unattenuated) correlation is observed correlation corrected for unreliability of observed scores

# Correcting for attenuation

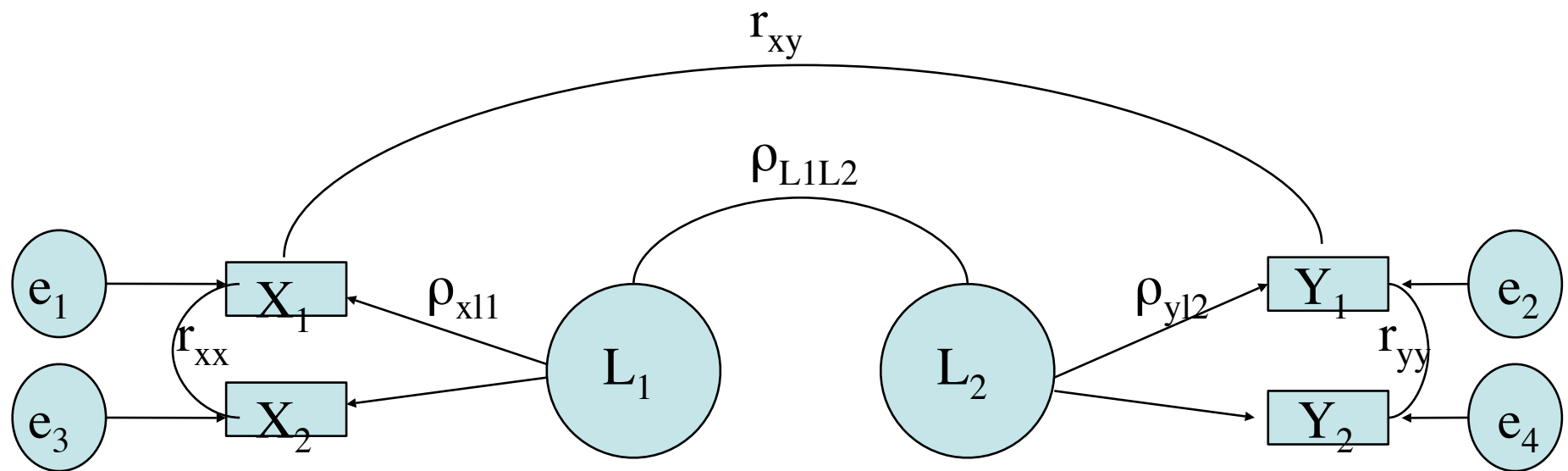
	$L_1$	$L_2$	$X_1$	$X_2$	$Y_1$	$Y_2$
$L_1$	$V_{L1}$					
$L_2$	$C_{L1L2}$	$V_{L2}$				
$X_1$	$C_{L1X}$	$C_{L1L2}^*$ $C_{L1X}$	$V_{L1}+V_{e1}$			
$X_2$	$C_{L1X}$	$C_{L1L2}^*$ $C_{L1X}$	$C_{L1X}^2$	$V_{L1}+V_{e3}$		
$Y_1$	$C_{L1L2}^*C_{L2Y}$	$C_{L2Y}$	$C_{L1X}^*C_{L1L2}$ $*C_{L2Y}$	$C_{L1X}^*C_{L1L2}$ $*C_{L2Y}$	$V_{L2}+V_{e2}$	
$Y_2$	$C_{L1L2}^*C_{L2Y}$	$C_{L2Y}$	$C_{L1X}^*C_{L1L2}$ $*C_{L2Y}$	$C_{L1X}^*C_{L1L2}$ $*C_{L2Y}$	$C_{L2Y}^2$	$V_{L2}+V_{e4}$

# Correcting for attenuation

	$L_1$	$L_2$	$X_1$	$X_2$	$Y_1$	$Y_2$
$L_1$	1					
$L_2$	$\rho_{L1L2}$	1				
$X_1$	$\rho_{L1X} = \sqrt{r_{xx}}$	$\rho_{L1L2}^*$ $\rho_{L1X}$	1			
$X_2$	$\rho_{L1X} = \sqrt{r_{xx}}$	$\rho_{L1L2}^*$ $\rho_{L1X}$	$\rho_{L1X}^2 = r_{xx}$	1		
$Y_1$	$\rho_{L1L2}^* \rho_{L2Y}$	$\rho_{L2Y}$ $= \sqrt{r_{yy}}$	$\rho_{L1X}^* \rho_{L1L2}^*$ $\rho_{L2Y}$	$\rho_{L1X}^* \rho_{L1L2}$ $* \rho_{L2Y}$	1	
$Y_2$	$\rho_{L1L2}^* \rho_{L2Y}$	$\rho_{L2Y}$ $= \sqrt{r_{yy}}$	$\rho_{L1X}^* \rho_{L1L2}^*$ $\rho_{L2Y}$	$\rho_{L1X}^* \rho_{L1L2}$ $* \rho_{L2Y}$	$\rho_{L2Y}^2$ $= r_{yy}$	1

# Classic Reliability Theory: correcting for attenuation

How well do we measure what ever we are measuring  
and what is the relationships between latent variables



$$\rho_{xL1} = \text{sqrt}(r_{xx})$$

$$\rho_{yL2} = \text{sqrt}(r_{yy})$$

$$\rho_{L1L2} = r_{xy} / \rho_{x11} \rho_{y12}$$

$$\rho_{L1L2} = r_{xy} / \text{sqrt}(r_{xx} * r_{yy})$$

Disattenuated (unattenuated) correlation is observed correlation  
corrected for unreliability of observed scores

# Classic reliability - limitation

All of the conventional approaches are concerned with generalizing about individual differences (in response to an item, time, form, rater, or situation) between people. Thus, the emphasis is upon consistency of rank orders. Classical reliability is a function of large between subject variability and small within subject variability. It is unable to estimate the within subject precision for a single person.

# The New Psychometrics- Item Response Theory

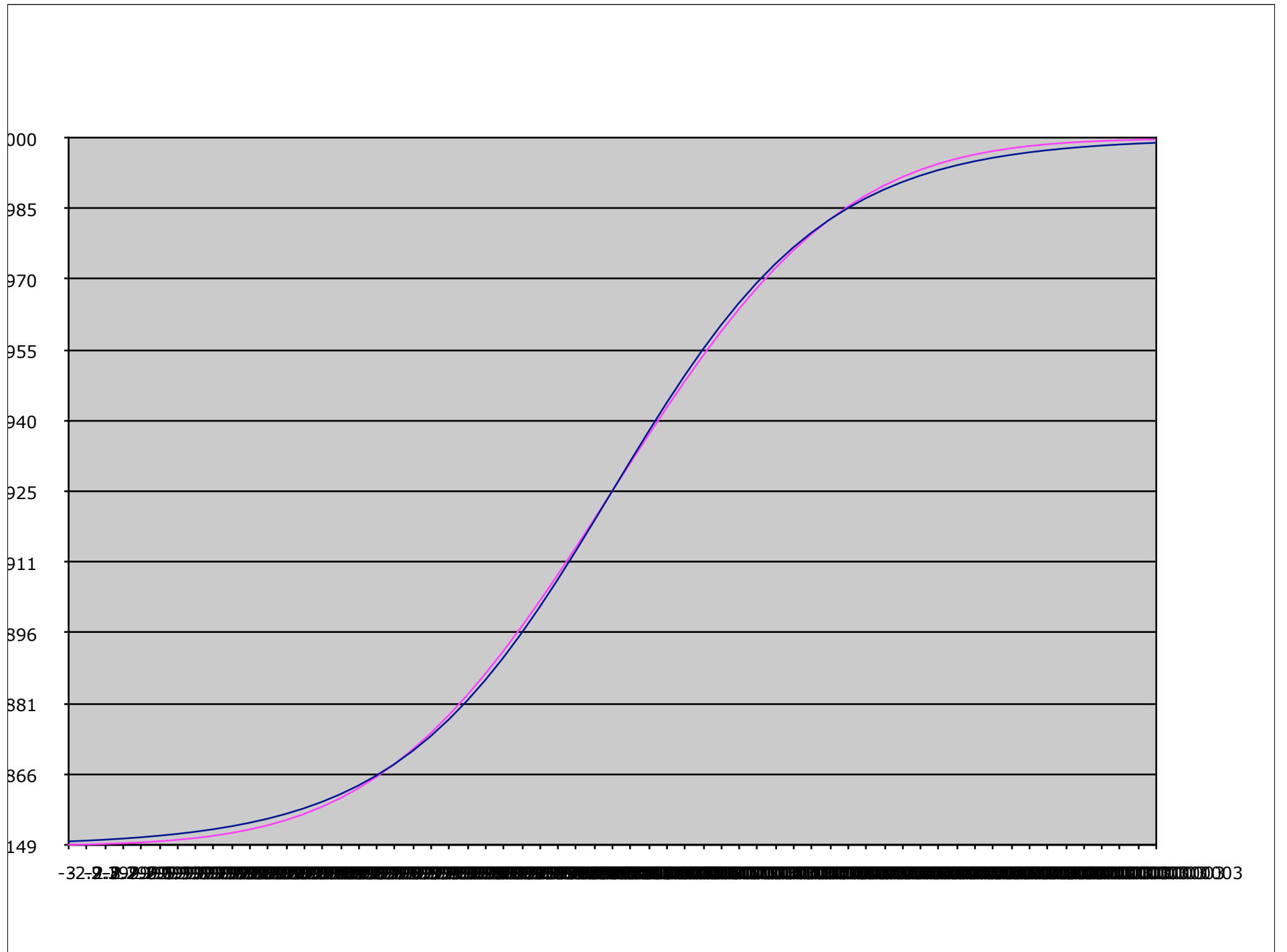
- Classical theory estimates the correlation of item responses (and sums of items responses, i.e., tests) with domains.
- Classical theory treats items as random replicates but ignores the specific difficulty of the item, nor attempts to estimate the probability of endorsing (passing) a particular item

# Item Response Theory

- Consider the person's value on an attribute dimension ( $\theta_i$ ).
- Consider an item as having a difficulty  $\delta_j$
- Then the probability of endorsing (passing) an item  $j$  for person  $i = f(\theta_i, \delta_j)$
- $p(\text{correct} \mid \theta_i, \delta_j) = f(\theta_i, \delta_j)$
- What is an appropriate function?
- Should reflect  $\delta_j - \theta_i$  and yet be bounded 0,1.

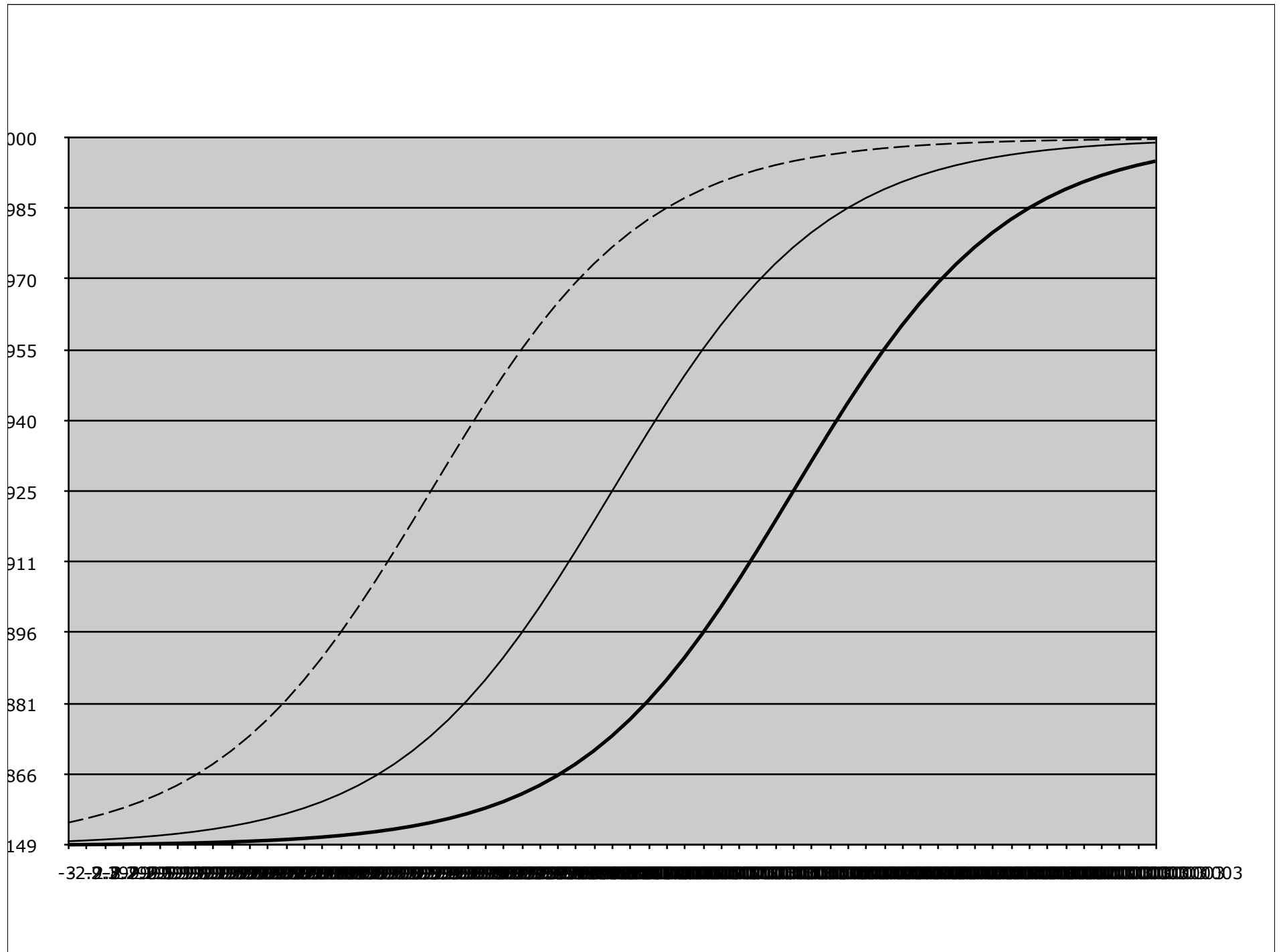
# Item Response Theory

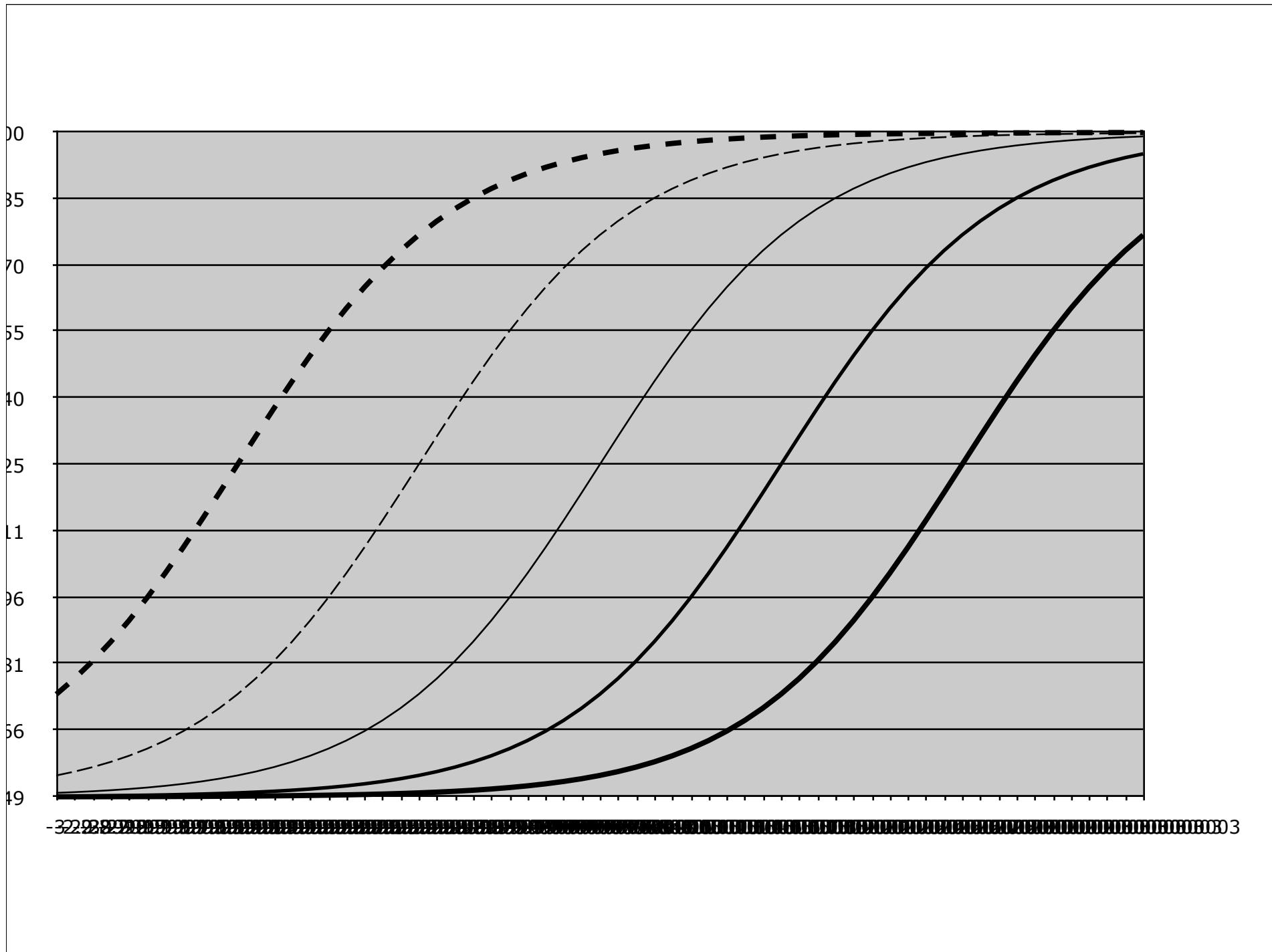
- $p(\text{correct} \mid \theta_i, \delta_j) = f(\theta_i, \delta_j) = f(\delta_j - \theta_i)$
- Two logical functions:
  - Cumulative normal (see, e.g., Thurstonian scaling)
  - Logistic =  $1/(1+\exp(\delta_j - \theta_i))$  (the Rasch model)
  - Logistic with weight of 1.7
    - $1/(1+\exp(1.7*(\delta_j - \theta_i)))$  approximates cumulative normal



# Item difficulty and ability

- Consider the probability of endorsing an item for different levels of ability and for items of different difficulty.
- Easy items ( $\delta_j = -1$ )
- Moderate items ( $\delta_j = 0$ )
- Difficulty items ( $\delta_j = 1$ )





# Estimation of ability for a particular person for known item difficulty

- The probability of any pattern of responses ( $x_1, x_2, x_3, \dots, X_n$ ) is the product of the probabilities of each response  $\prod(p(x_i))$ .
- Consider the odds ratio of a response
  - $p/(1-p) = 1/(1+\exp(1.7*(\delta_j - \theta_i))) / (1 - 1/(1+\exp(1.7*(\delta_j - \theta_i)))) =$
  - $p/(1-p) = \exp(1.7*(\delta_j - \theta_i))$  and therefore:
  - $\text{Ln}(\text{odds}) = 1.7 * (\delta_j - \theta_i)$  and
  - $\text{Ln}(\text{odds of a pattern}) = 1.7 \sum (\delta_j - \theta_i)$  for known difficulty

# Unknown difficulty

- Initial estimate of ability for each subject (based upon total score)
- Initial estimate of difficulty for each item (based upon percent passing)
- Iterative solution to estimate ability and difficulty (with at least one item difficulty fixed).

# Classical versus the “new”

- Ability estimates are logistic transform of total score and are thus highly correlated with total scores, so why bother?
- IRT allows for more efficient testing, because items can be tailored to the subject.
- Maximally informative items have  $p(\text{passing given ability and difficulty})$  of .5
- With tailored tests, each person can be given items of difficulty appropriate for them.

# Computerized adaptive testing

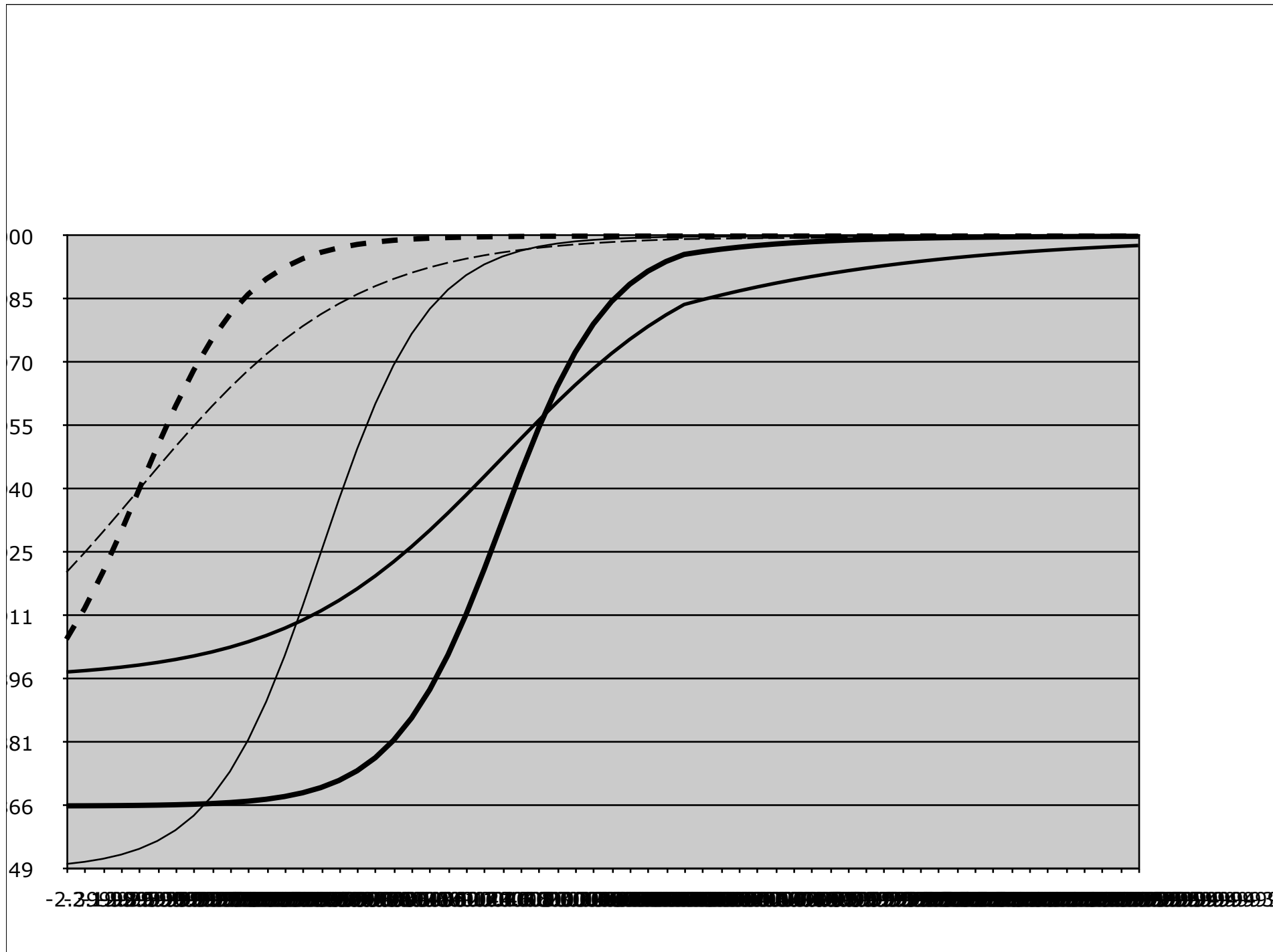
- CAT allows for equal precision at all levels of ability
- CAT/IRT allows for individual confidence intervals for individuals
- Can have more precision at specific cut points (people close to the passing grade for an exam can be measured more precisely than those far (above or below) the passing point).

# Psychological (non-psychometric) problems with CAT

- CAT items have difficulty level tailored to individual so that each person passes about 50% of the items.
- This increases the subjective feeling of failure and interacts with test anxiety
- Anxious people quit after failing and try harder after success -- their pattern on CAT is to do progressively worse as test progresses (Gershon, 199x, in preparation)

# Generalizations of IRT to 2 and 3 item parameters

- Item difficulty
- Item discrimination (roughly equivalent to correlation of item with total score)
- Guessing (a problem with multiple choice tests)
- 2 and 3 parameter models are harder to get consistent estimates and results do not necessarily have monotonic relationship with total score



# Item Response Theory

- Can be seen as a generalization of classical test theory, for it is possible to estimate the correlations between items given assumptions about the distribution of individuals taking the test
- Allows for expressing scores in terms of probability of passing rather than merely rank orders (or even standard scores). Thus, a 1 sigma difference between groups might be seen as more or less important when we know how this reflects chances of success on an item
- Emphasizes non-linear nature of response scores.