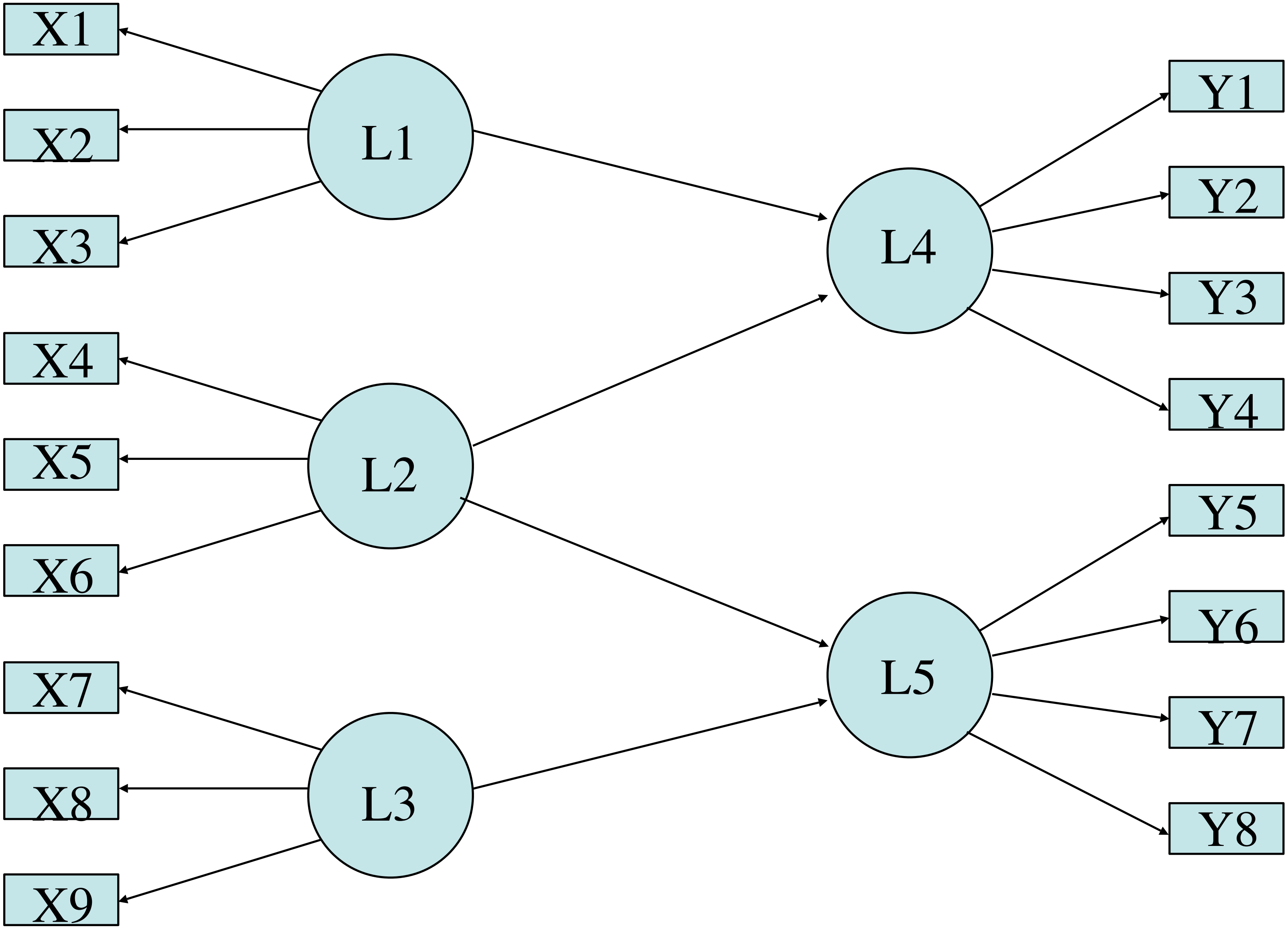
Psychometric Theory: A conceptual Syllabus



A Theory of Data: What can be measured

X1

What is measured?

Individuals

Objects

What kind of measures are taken?

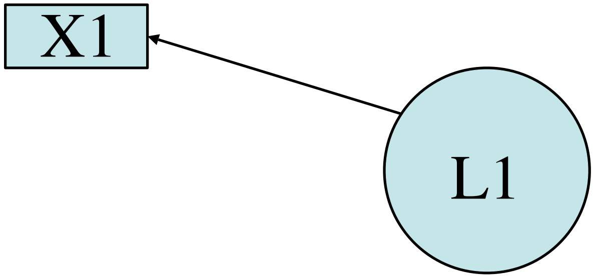
proximity

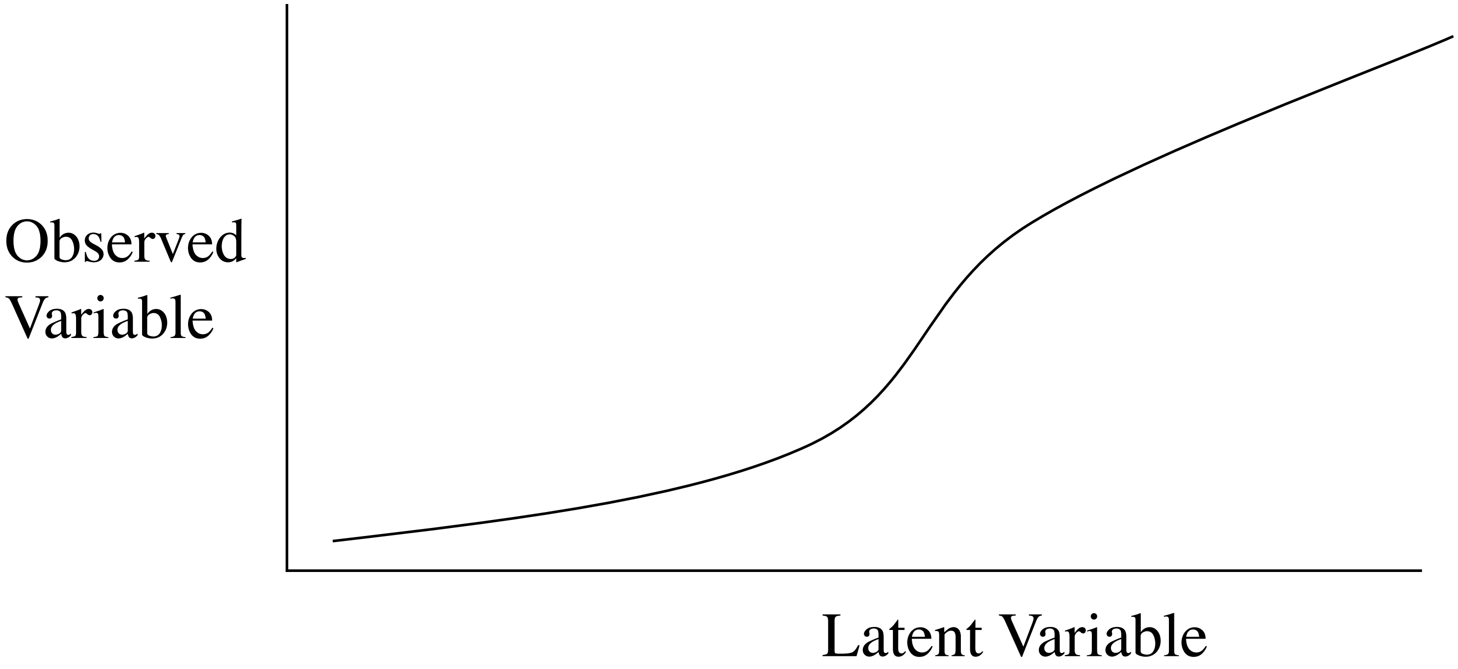
order

Comparisons are made on:

Single Dyads or Pairs of Dyads

Scaling: the mapping between observed and latent variables





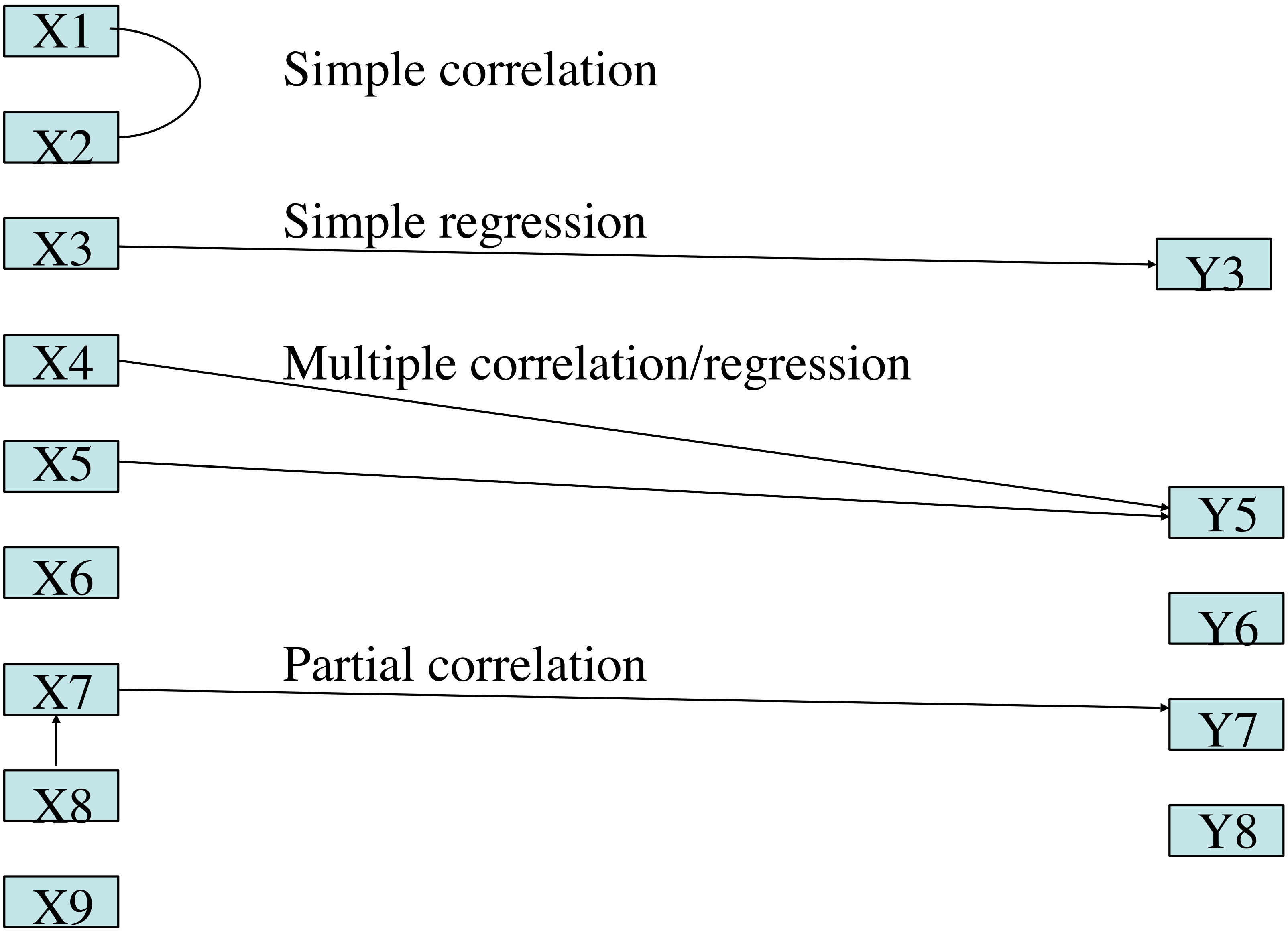
Where are we?

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

Measures of relationship

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

Variance, Covariance, and Correlation



Simple correlation

Simple regression

Multiple correlation/regression

Partial correlation

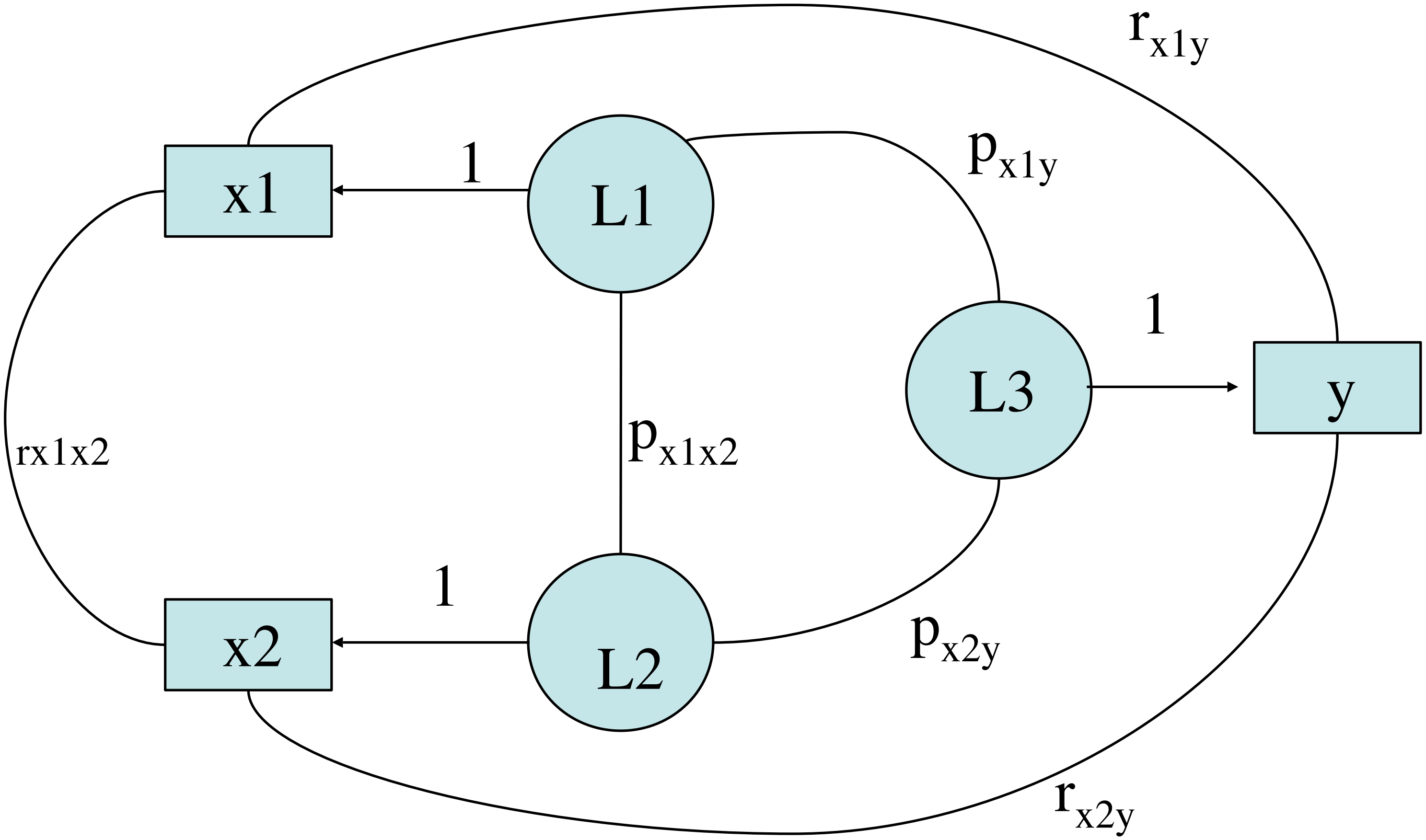
Measures of relationships with more than 2 variables

- Partial correlation
  - The relationship between x and y with z held constant (z removed)
- Multiple correlation
  - The relationship of x1 + x2 with y
  - Weight each variable by its independent contribution

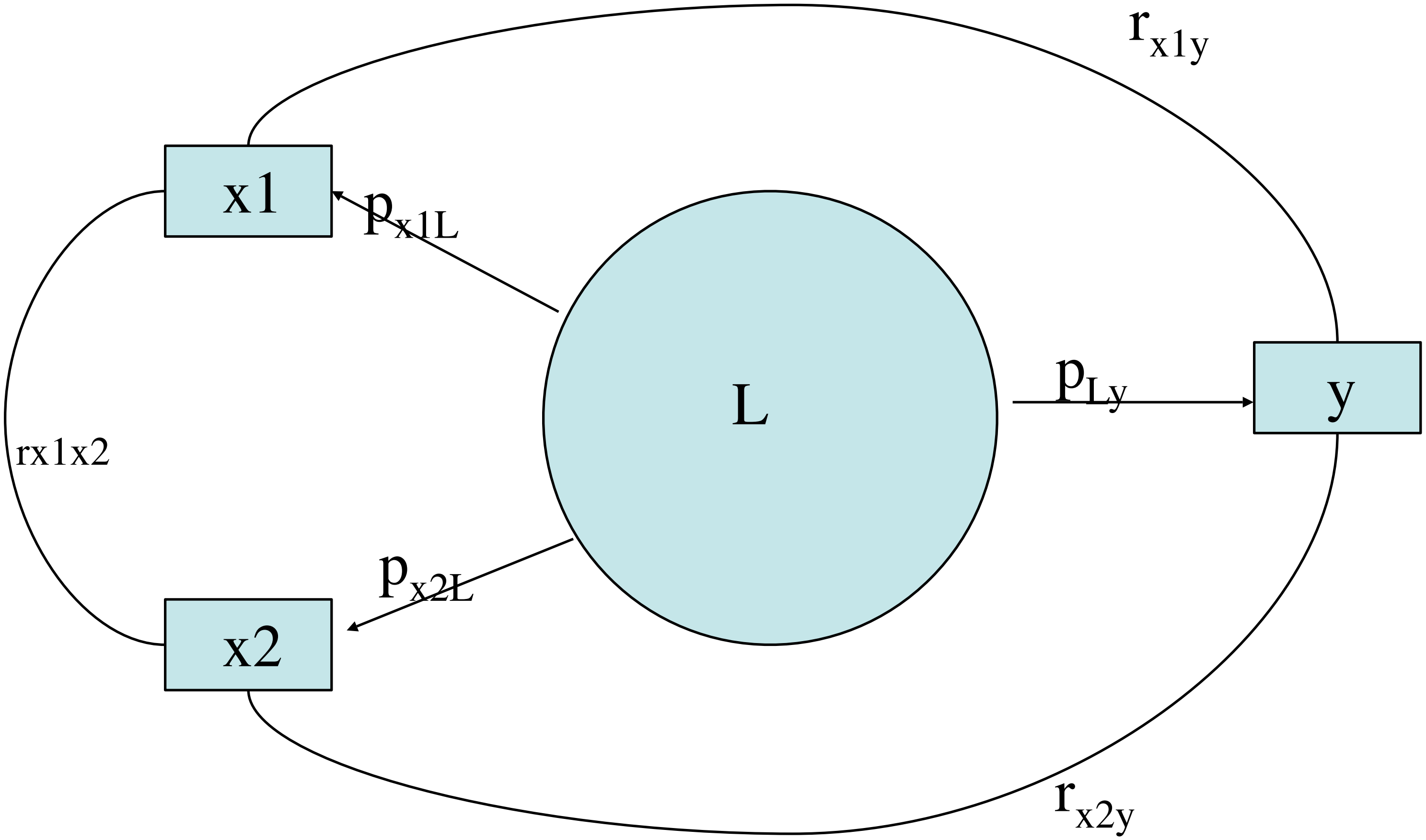
Problems with correlations

- 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_1	X_2	Y
X_1	1.00		
X_2	0.72	1.00	
Y	0.63	0.56	1.00

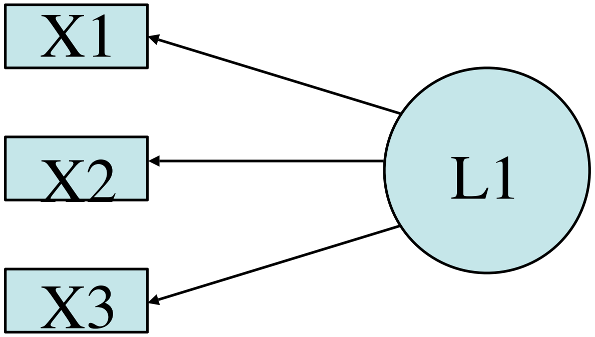
Partial r = $(r_{x1y}-r_{x1x2}*r_{x2y})/\mathrm{sqrt}((1-r_{x1x2}{}^2)*(1-r_{x2y}{}^2))$

Rx1y.x2 = .33 (traditional model) but = 0 with structural model

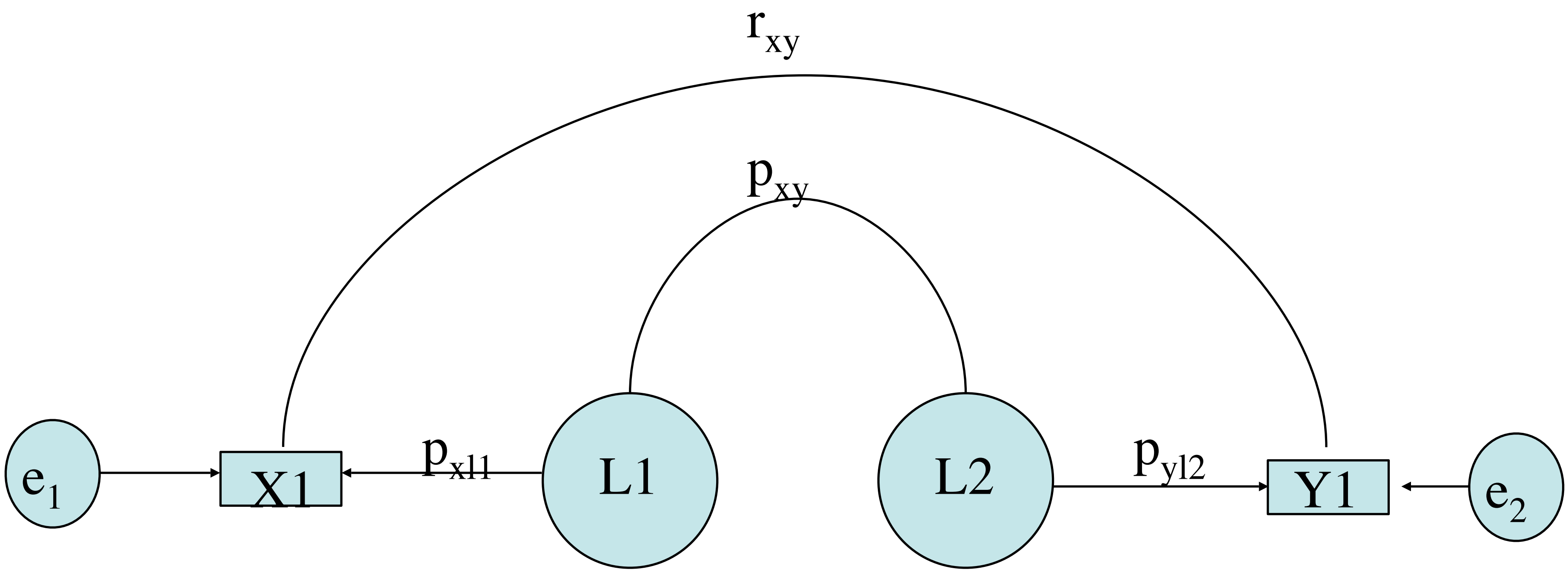
Reliability Theory

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



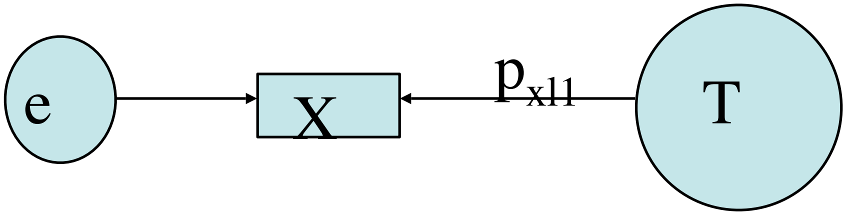
Classic Reliability Theory: How well do we measure what ever we are measuring

What is the relationship between X_1 and L_1?

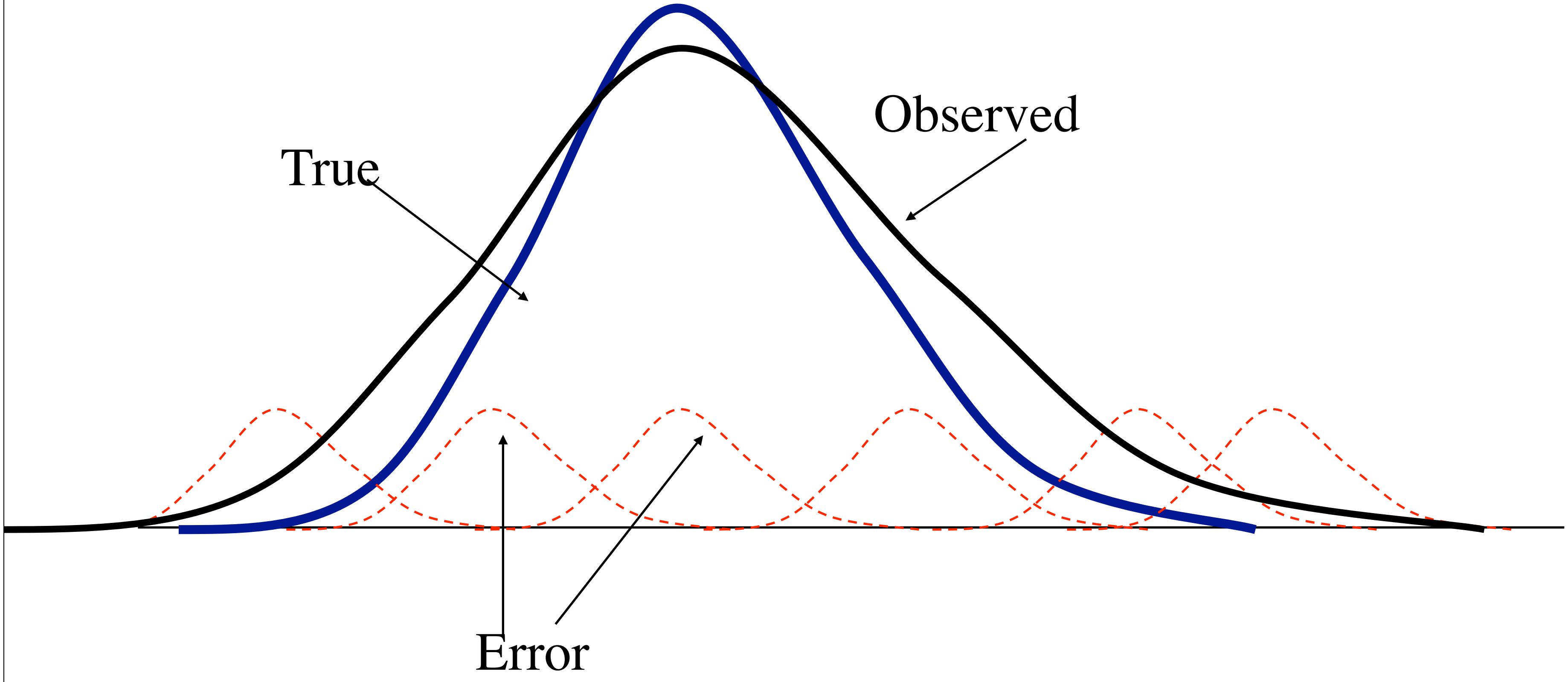
What is the variance of X_1, L_1, and E_1?

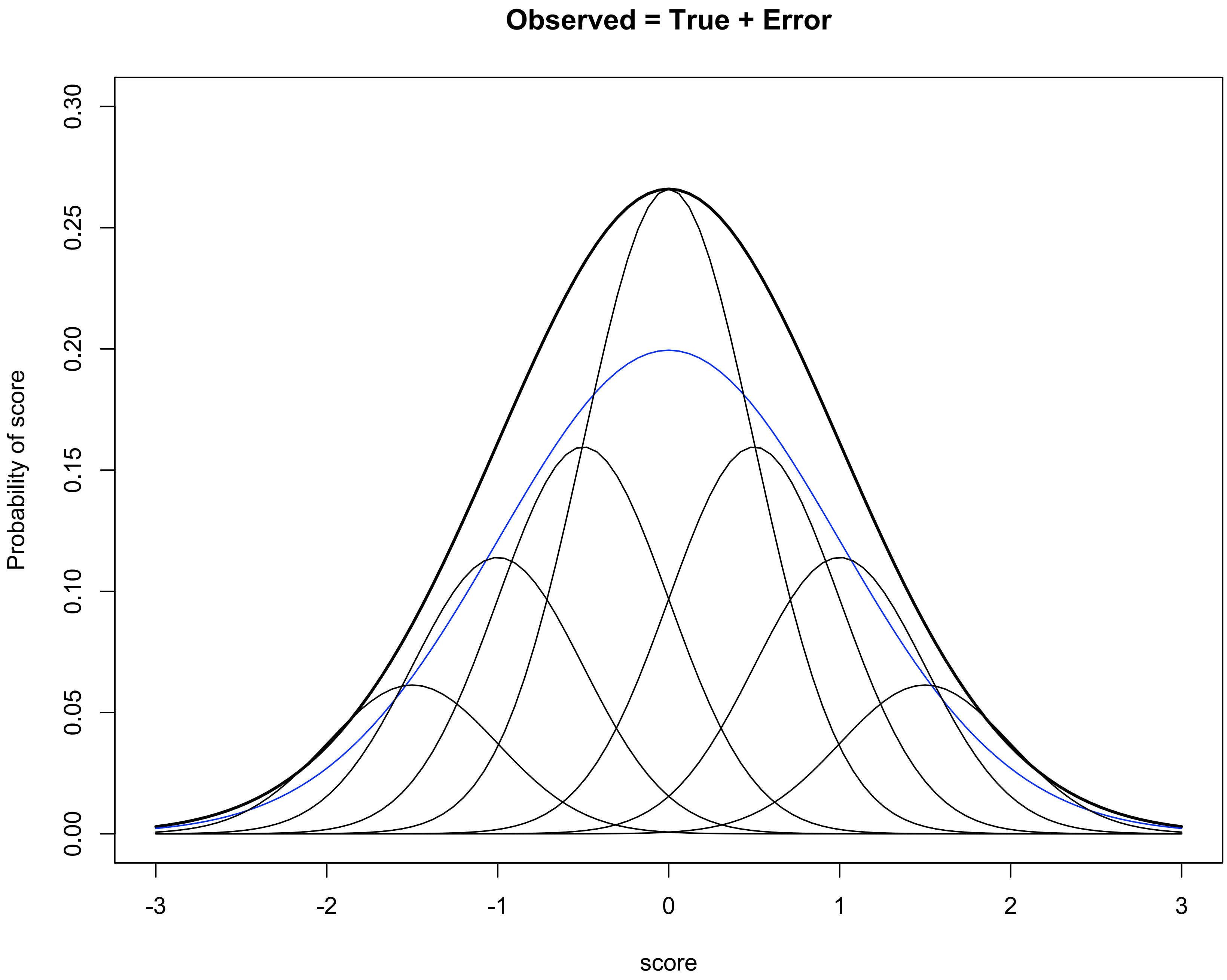
Let True Score for Subject I = expected value of X_i.

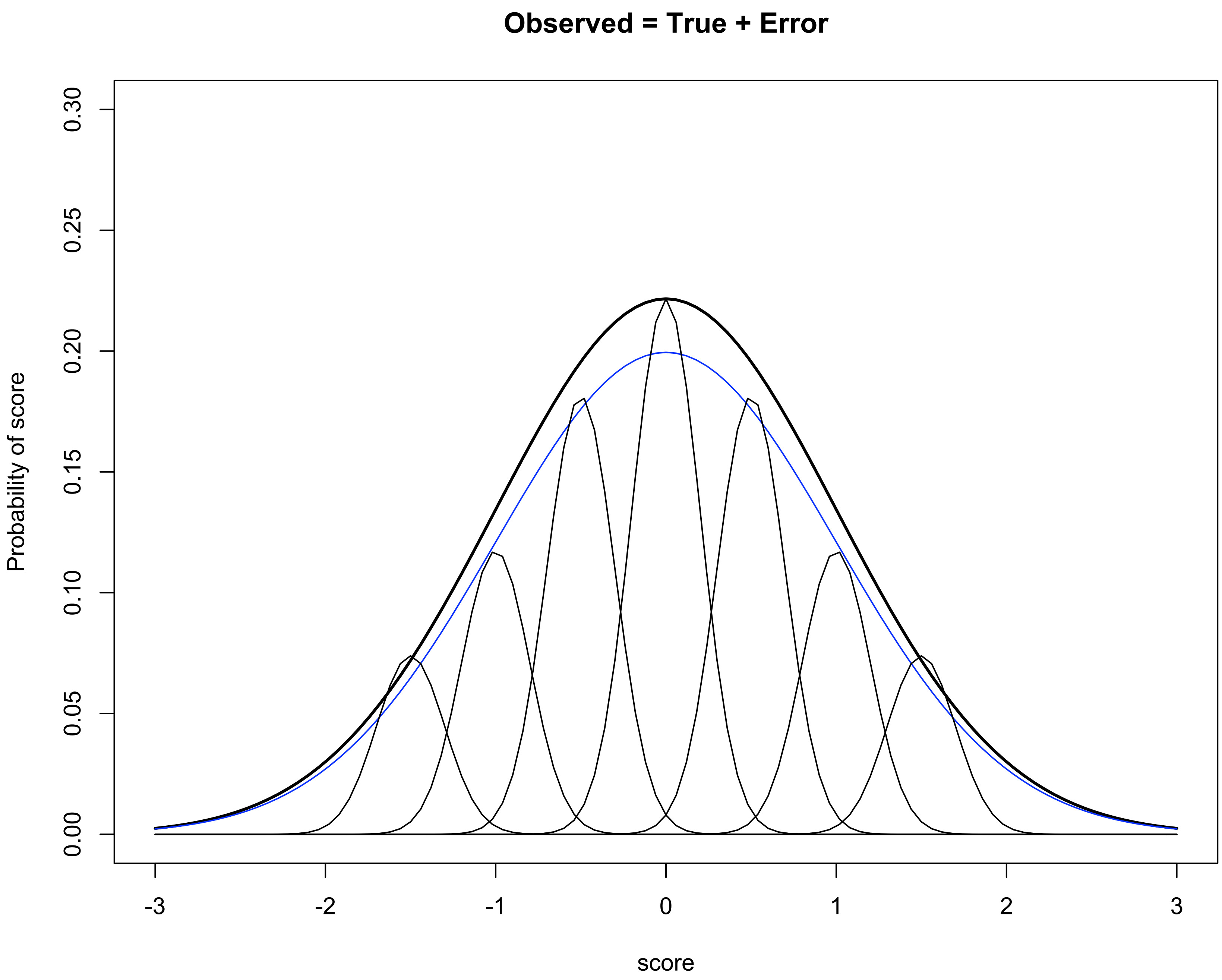
(note that this is not the Platonic Truth, but merely the average over an infinite number of trials.)



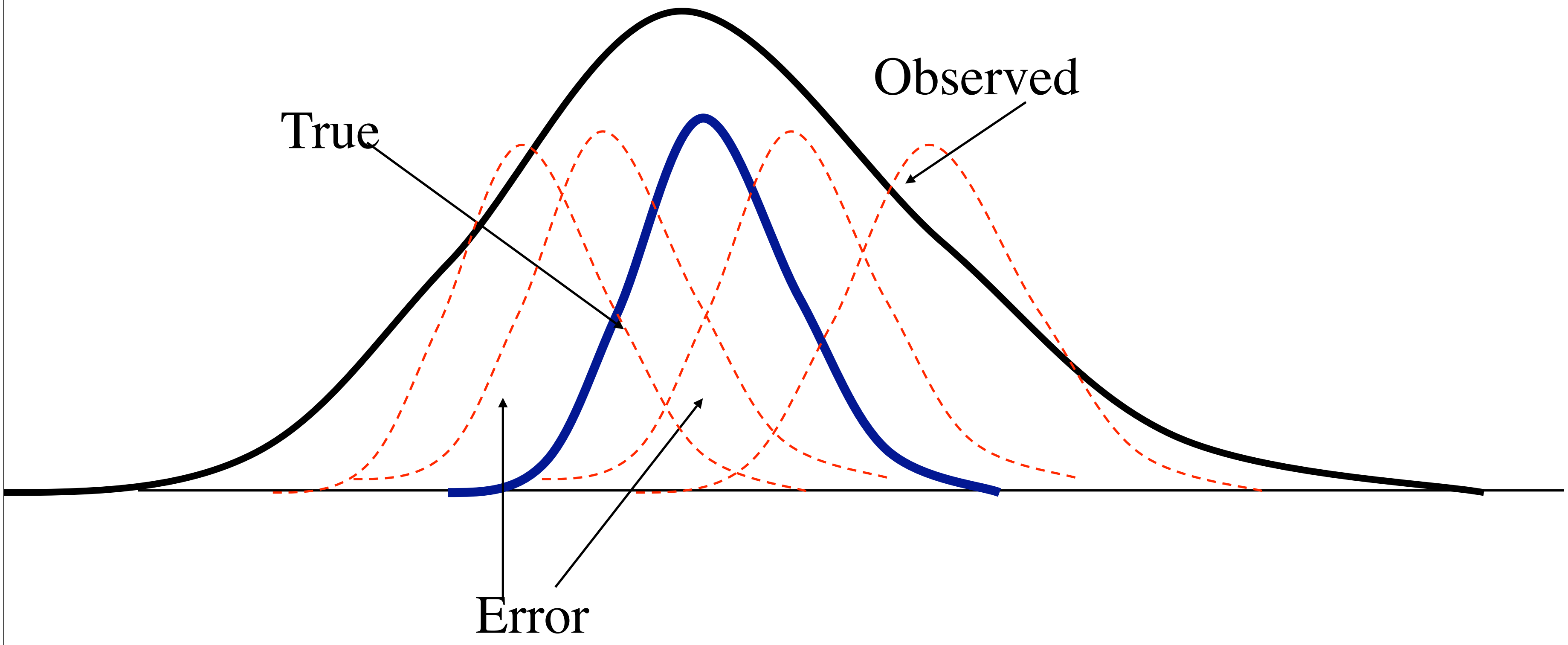
Observed= True + Error




Observed = True + Error
Probability of score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-3
-2
-1
0
1
2
3
score


Observed = True + Error
Probability of score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-3
-2
-1
0
1
2
3
score

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)

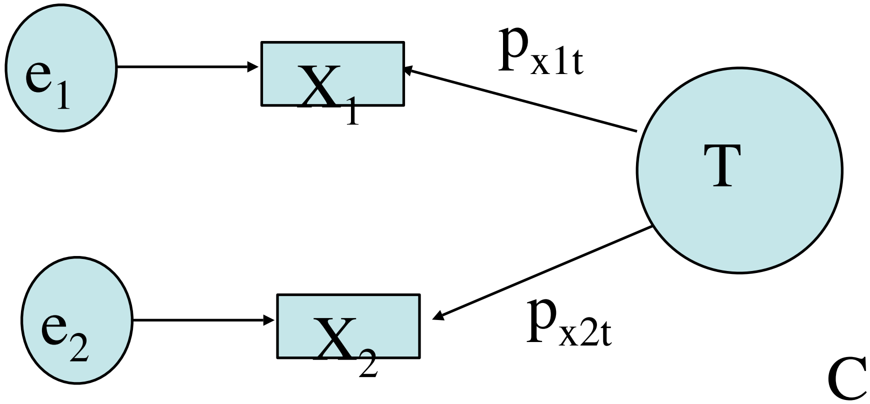
Estimating True score

- Given that $p^2_{ot} = V_t/V_o$ 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



$$V_{x1}=V_t+V_{e1}$$
$$V_{x2}=V_t+V_{e2}$$

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

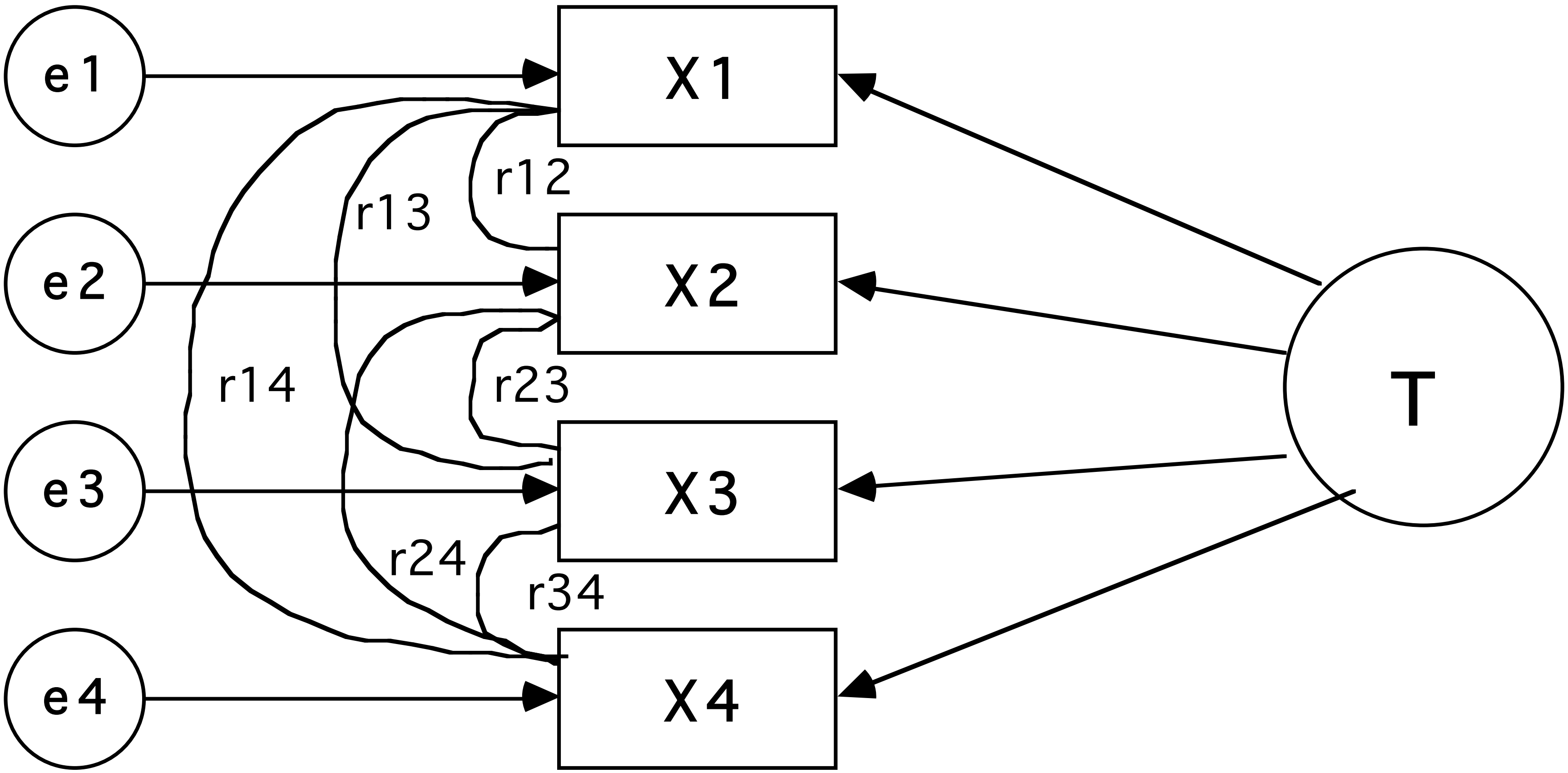
$$r_{xx}=C_{x1x2}/Sqrt(V_{x1}*V_{x2}) = V_t/V_x$$

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_{x1x2} =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} = sqrt(r_{xx}) ==> 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

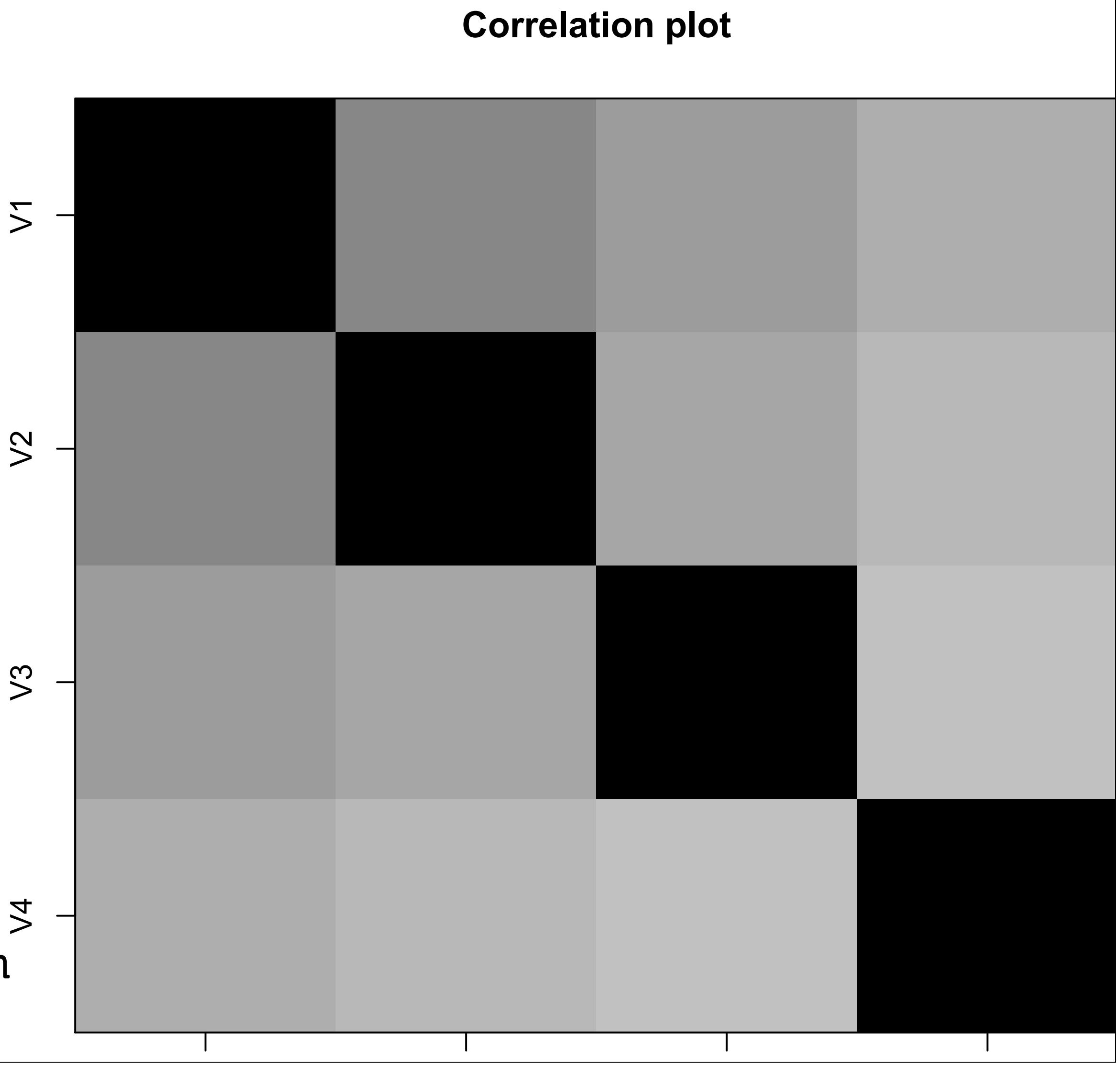


4 Congeneric measures

```
 > cong <-
sim.congeneric()
```

```
    V1   V2   V3   V4
V1 1.00 0.56 0.48 0.40
V2 0.56 1.00 0.42 0.35
V3 0.48 0.42 1.00 0.30
V4 0.40 0.35 0.30 1.00
```

```
cor.plot(cong,n=24,zlim
=c(0,1))
```



Observed Variances/Covariances

	x_1	x_2	x_3	x_4
x_1	Vx_1			
x_2	c_{x1x2}	Vx_2		
x_3	c_{x1x3}	c_{x2x3}	Vx_3	
x_4	c_{x1x4}	c_{x3x4}	c_{x3x4}	Vx_4

Model Variances/Covariances

	x_1	x_2	x_3	x_4
x_1	V_t+Ve_1			
x_2	$c_{x1t}c_{x2t}$	V_t+ Ve_2		
x_3	$c_{x1t}c_{x3t}$	$c_{x2t}c_{x3t}$	V_t+ Ve_3	
x_4	$c_{x1t}c_{x4t}$	$c_{x3t}c_{x4t}$	$c_{x3t}c_{x4t}$	V_t+ Ve_4

Observed and modeled Variances/Covariances

	x_1	x_2	x_3	x_4
x_1	Vx_1			
x_2	c_{x1x2}	Vx_2		
x_3	c_{x1x3}	c_{x2x3}	Vx_3	
x_4	c_{x1x4}	c_{x3x4}	c_{x3x4}	Vx_4

	x_1	x_2	x_3	x_4
x_1	V_t+Ve_1			
x_2	$c_{x1t}c_{x2t}$	$V_t+ Ve_2$		
x_3	$c_{x1t}c_{x3t}$	$c_{x2t}c_{x3t}$	$V_t+ Ve_3$	
x_4	$c_{x1t}c_{x4t}$	$c_{x3t}c_{x4t}$	$c_{x3t}c_{x4t}$	$V_t+ Ve_4$

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 ith 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 = Cov_{jd}/sqrt($(V_j*V_d$)
2. Cov_{jd}=Vj + $\sum c_{lj}$ = Vj + (k-1)*average cov_j
3. Domain variance (Vd) = sum of item variances + item covariances in domain =
4. Vd = k*(average variance) + k*(k-1) average covar
5. Let Va = average variance, Ca =average covariance
6. Then Vd = k(Va + (k-1)*Ca)

Correlating an Item with Domain

1. Assume that $V_j = V_a$ and $C_{jl}=C_a$
2. $r_{jd} = C_{jd}/\mathrm{sqrt}(V_j*V_d)$
3. $r_{jd}=(V_a+(k-1)C_a)/\mathrm{sqrt}(V_a*k(V_a + (k-1)*C_a))$
4. $r_{jd}^2=(V_a+(k-1)C_a)*(V_a+(k-1)C_a))/(V_a*k(V_a + (k-1)*C_a))$
5. Now, find the limit of r_{jd}^2 as k becomes large:
6. Lim k->∞ of $r_{jd}^{2\ a}= C_a/Vy$= av covar/av variance
7. 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 = $\sum$(variances) + $\sum$(covars)
3. Variance of n item test = $\sum v_j + \sum c_{jl} = V_n = n*V_a + n*(n-1)\ C_a$
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

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

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

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

Limit of squared r with domain

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

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

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.

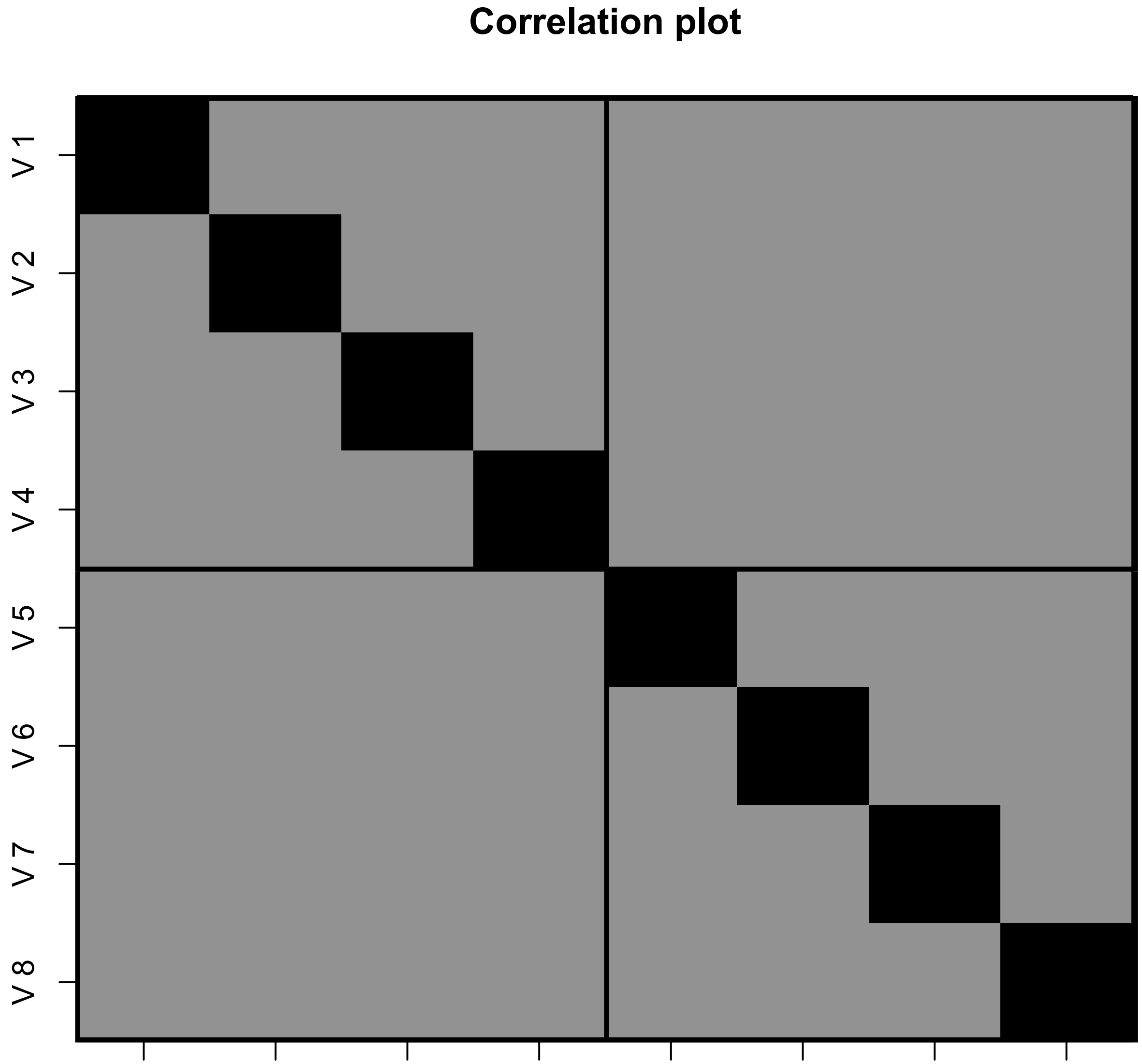
Coefficient Alpha

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

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

What is the correlation of this test with the domain?

Two equivalent tests
k = 4

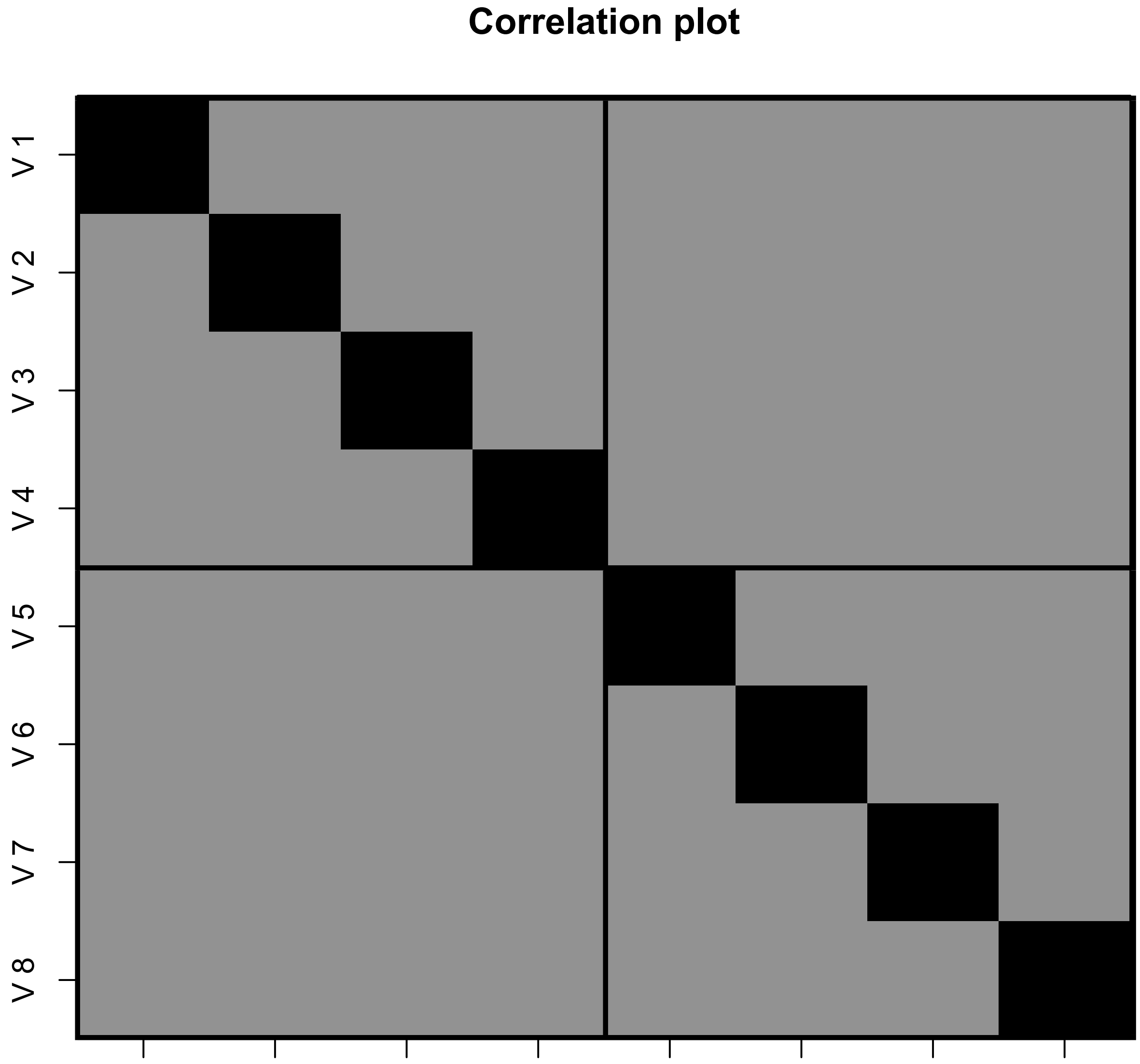


Coefficient alpha

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

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

Two equivalent tests
k = 4



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

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

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

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

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 v1. What is the correlation of this test with another test sample from the domain? What is the correlation of this test with the domain?

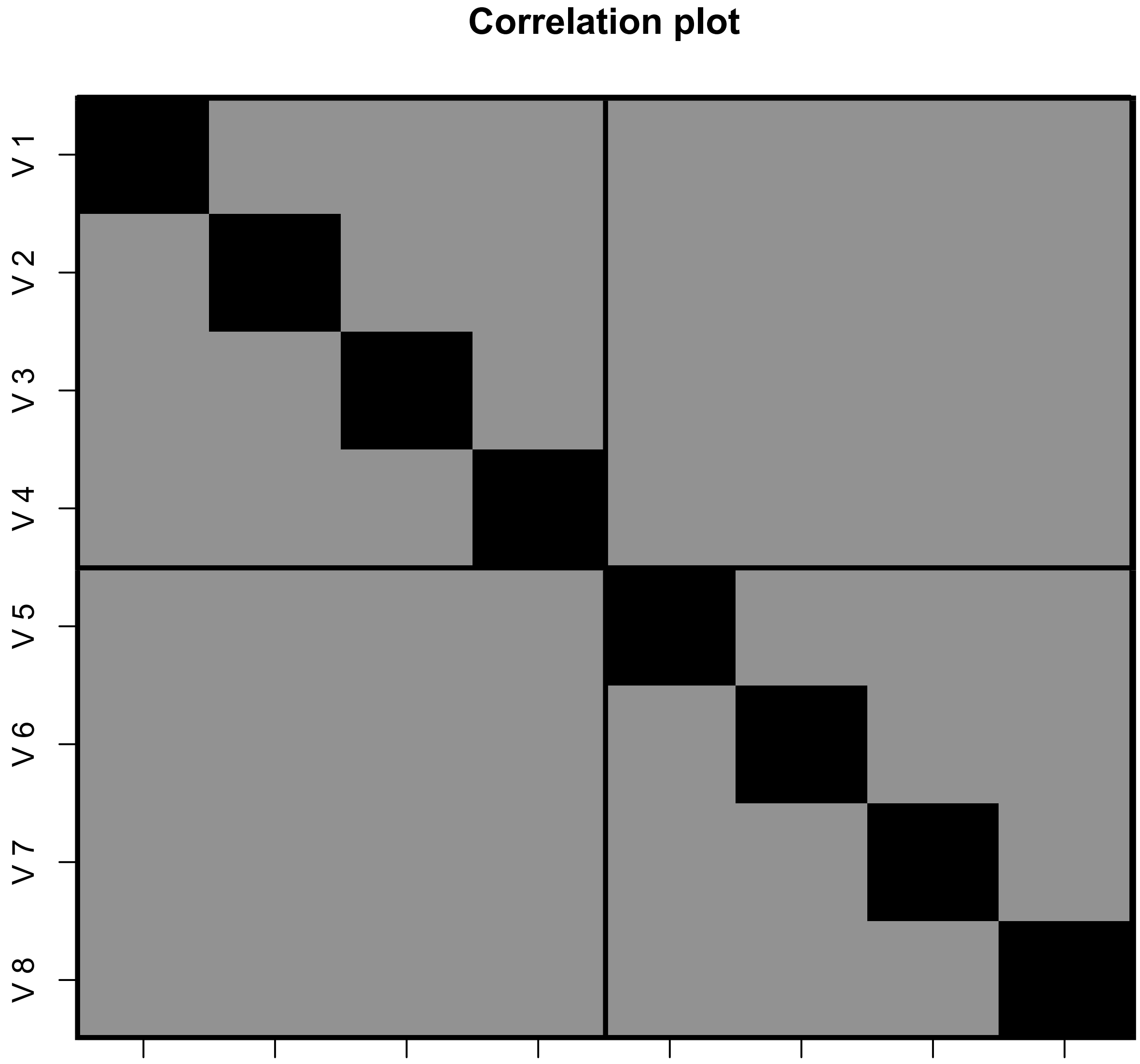
	Test 1	Test 2
Test 1	V_1	C_{12}
Test 2	C_{12}	V_2

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

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.

Two equivalent tests k = 4



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	$V_1 = k*[v_i+(k-1)*c1]$	$C_{12} = k*k*c_{12}$
Test 2	$C_{12} = k*k*c_{12}$	$V_2 = k*[v_i+(k-1)*c_2]$

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

Alpha from variances

- $Vt = V_1 = k*[v_i+(k-1)*c1]$ <=>
- $c_1 = (Vt - \sum v_i)/((k*(k-1))$
- $C_{12} = k^2 c_{12} = k^2*(Vt - \sum v_i)/((k*(k-1))$
- $rx_1x_2 =(k^2*(Vt - \sum v_i)/((k*(k-1)))/Vt =$
- $rx_1x_2 = [(Vt - \sum v_i)/Vt]*(k/(k-1)$
- 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	0.2	0.1
1	0.20	0.10
2	0.33	0.18
4	0.50	0.31
8	0.67	0.47
16	0.80	0.64
32	0.89	0.78
64	0.94	0.88
128	0.97	0.93

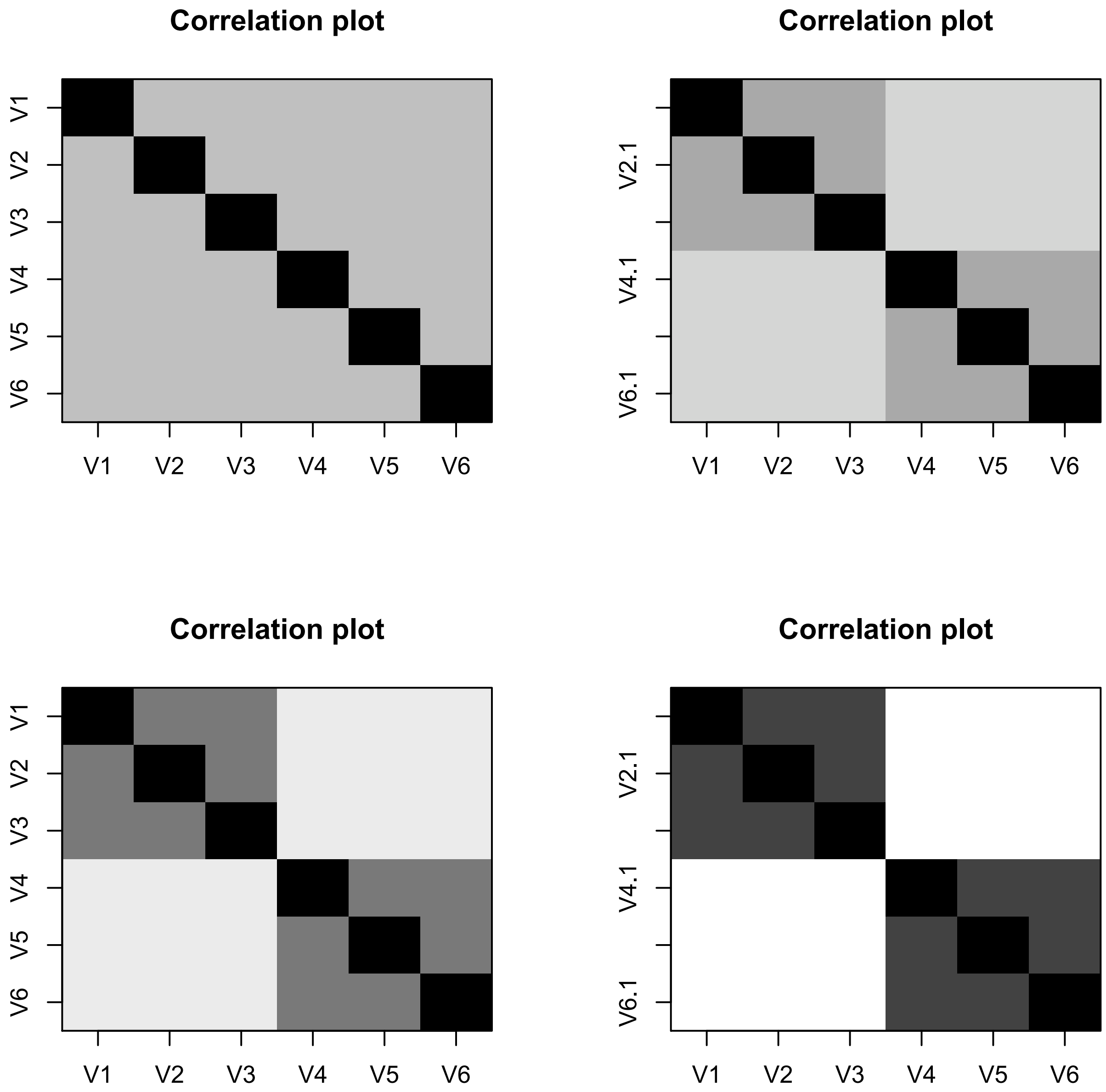
Alpha and test length

- 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. This is just the average intercorrelation within the test
- Average inter item r = alpha_1 =
  - alpha/(alpha + k*(1-alpha))
  - This allows us to find the average internal consistency

Problems with alpha

- Is the average intercorrelation representative of the shared item variance?
- Yes, if all items are equally correlated
- No, if items differ in their intercorrelations
- Particularly not if the test is “lumpy”
- Consider 4 correlation matrices with equal “average r” but drastically different structure.

4 correlation matrices



alpha$_1$ = .3 and alpha = .72 for all 4 sets

```
> S1
    V1  V2  V3  V4  V5  V6
V1 1.0 0.3 0.3 0.3 0.3 0.3
V2 0.3 1.0 0.3 0.3 0.3 0.3
V3 0.3 0.3 1.0 0.3 0.3 0.3
V4 0.3 0.3 0.3 1.0 0.3 0.3
V5 0.3 0.3 0.3 0.3 1.0 0.3
V6 0.3 0.3 0.3 0.3 0.3 1.0
```

```
> S2
     V1   V2   V3   V4   V5   V6
V1 1.00 0.45 0.45 0.20 0.20 0.20
V2 0.45 1.00 0.45 0.20 0.20 0.20
V3 0.45 0.45 1.00 0.20 0.20 0.20
V4 0.20 0.20 0.20 1.00 0.45 0.45
V5 0.20 0.20 0.20 0.45 1.00 0.45
V6 0.20 0.20 0.20 0.45 0.45 1.00
```

```
> S3
    V1  V2  V3  V4  V5  V6
V1 1.0 0.6 0.6 0.1 0.1 0.1
V2 0.6 1.0 0.6 0.1 0.1 0.1
V3 0.6 0.6 1.0 0.1 0.1 0.1
V4 0.1 0.1 0.1 1.0 0.6 0.6
V5 0.1 0.1 0.1 0.6 1.0 0.6
V6 0.1 0.1 0.1 0.6 0.6 1.0
```

```
> S4
     V1   V2   V3   V4   V5   V6
V1 1.00 0.75 0.75 0.00 0.00 0.00
V2 0.75 1.00 0.75 0.00 0.00 0.00
V3 0.75 0.75 1.00 0.00 0.00 0.00
V4 0.00 0.00 0.00 1.00 0.75 0.75
V5 0.00 0.00 0.00 0.75 1.00 0.75
V6 0.00 0.00 0.00 0.75 0.75 1.00
```

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’

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

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

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

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

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

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-h and Omega

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

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-h are estimates of general variance.

Calculating alpha

- round(cor(items),2)   #what are their correlations?

Find Alpha from correlations

	q_262	q_1480	q_819	q_1180	1742
q_262	1	0.26	0.41	0.51	0.48
q_1480	0.26	1	0.66	0.52	0.47
q_819	0.41	0.66	1	0.41	0.65
q_1180	0.51	0.52	0.41	1	0.49
q_1742	0.48	0.47	0.65	0.49	1

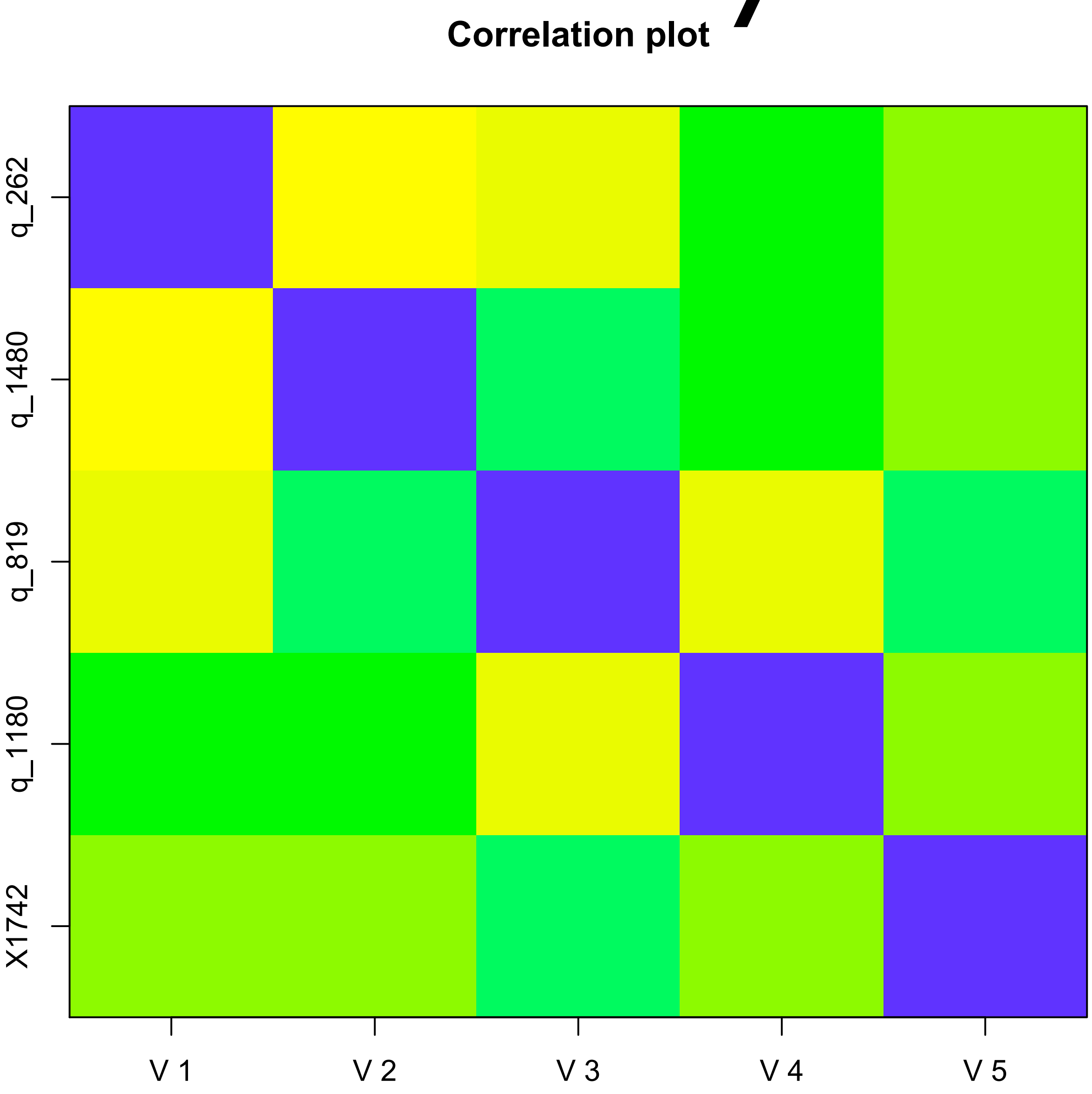
Alpha from correlations

- Total variance = sum of all item correlations
  - = sum(item) = 14.72
- total covariances = Vt - ∑item variance
  - = sum(item) - tr(item) = 9.72
- average covariance =
  - (Vt - ∑item variance)/(nvar *(nvar-1)) = .486
- alpha = ((Vt - ∑item variance)/Vt)*(nvar *(nvar-1))
  - = alpha = .83

The items

```
> item <- read.clipboard()
> item
  q_262 q_1480 q_819 q_1180 X1742
1  1.00   0.26  0.41   0.51  0.48
2  0.26   1.00  0.66   0.52  0.47
3  0.41   0.66  1.00   0.41  0.65
4  0.51   0.52  0.41   1.00  0.49
5  0.48   0.47  0.65   0.49  1.00
```

Visually



```
cor.plot(item,TRUE,12,zlim=c(0,1))
```

alpha

```
> alpha(item)

Reliability analysis
Call: alpha(x = item)

  raw_alpha std.alpha G6(smc) average_r
      0.83      0.83    0.83      0.49

 Reliability if an item is dropped:
       raw_alpha std.alpha G6(smc) average_r
q_262       0.82      0.82    0.80      0.53
q_1480      0.79      0.79    0.76      0.49
q_819       0.77      0.77    0.73      0.46
q_1180      0.79      0.79    0.77      0.49
X1742       0.77      0.77    0.77      0.46
```

Items with total scale

```
Item statistics
          r r.cor
q_262  0.69  0.58
q_1480 0.76  0.70
q_819  0.82  0.78
q_1180 0.76  0.68
X1742  0.81  0.74
```

Reliability: multiple estimates

- $r_{xx} = V_t/V_x = 1 - V_e/V_x$
- but what is V_e ?
- Trace of X
- Trace of X - (sum(average C_{xx}) (alpha)
- Trace of X - sum(sqrt(average(C_{xx}^2)))
- Trace of X - sum(smc X) (G6)

Squared Multiple Correlations

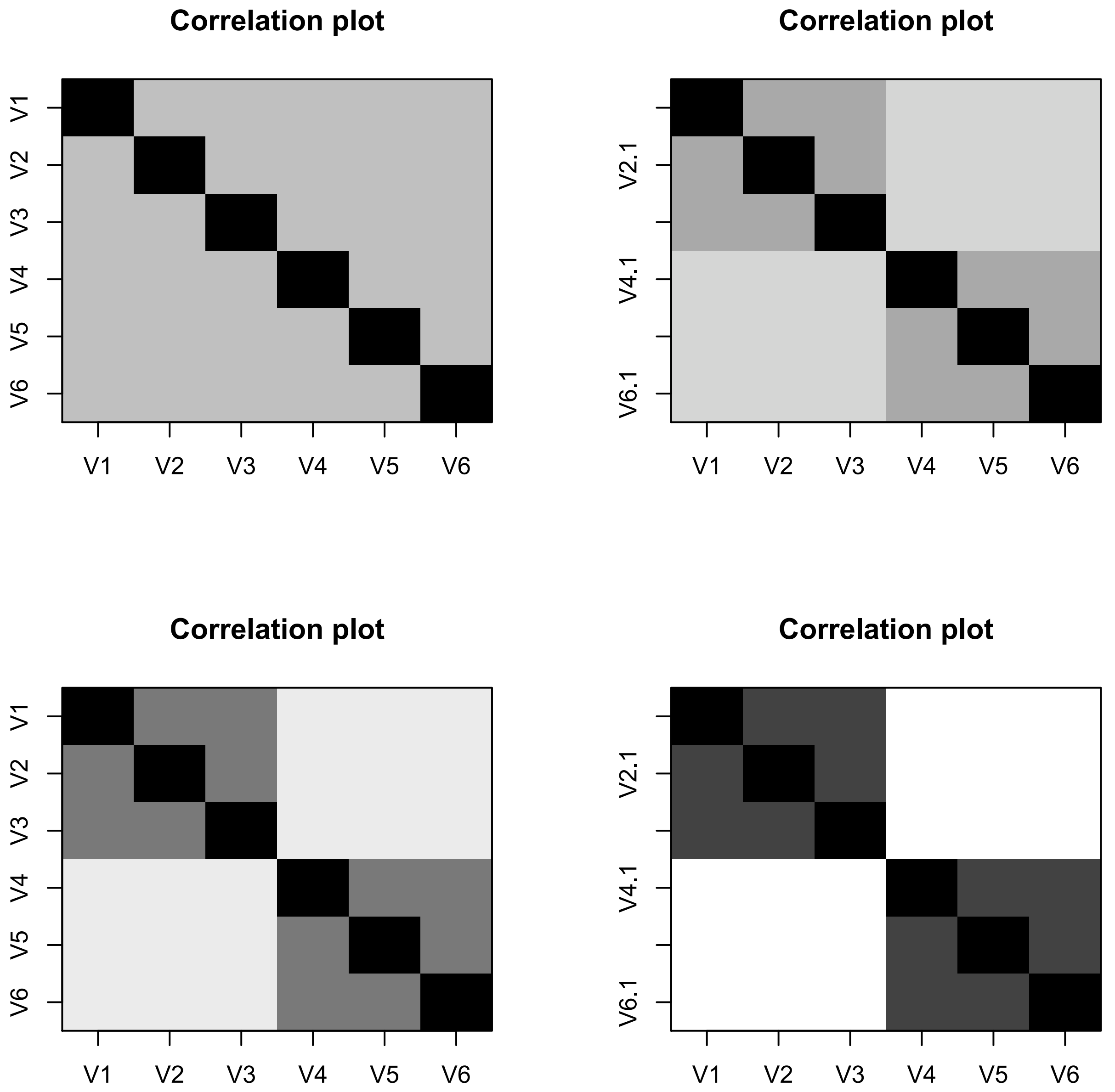
$$smc = 1 - diag(R^{-1})$$

```
> round(solve(item),2)
       [,1]  [,2]  [,3]  [,4]  [,5]
q_262   1.56  0.34 -0.37 -0.65 -0.35
q_1480  0.34  2.12 -1.24 -0.78  0.03
q_819  -0.37 -1.24  2.51  0.30 -1.02
q_1180 -0.65 -0.78  0.30  1.81 -0.41
X1742  -0.35  0.03 -1.02 -0.41  2.02
> round(1/diag(solve(item)),2)
[1] 0.64 0.47 0.40 0.55 0.50
> round(1-1/diag(solve(item)),2)
[1] 0.36 0.53 0.60 0.45 0.50
> round(smc(item),2)
[1] 0.36 0.53 0.60 0.45 0.50
```

Alternative estimates

```
>  guttman(item)
Alternative estimates of reliability
Beta =  0.73  This is an estimate of the worst
split half reliability
Guttman bounds
 L1 =  0.66
 L2 =  0.83
 L3 (alpha) =  0.83
 L4 (max) =  0.91
 L5 =  0.81
 L6 (smc) =  0.83
 alpha of first PC =  0.83
 estimated glb =  0.91
 beta estimated by first and second PC =  0.64
This is an exploratory statistic
```

4 correlation matrices



Alternative reliabilities

	S1	S2	S3	S4
alpha	0.72	0.72	0.72	0.72
G6:smc	0.68	0.72	0.78	0.86
G4: max	0.72	0.76	0.83	0.89
glb	0.72	0.76	0.83	0.89
beta	0.62*	0.48	0.24	0.00

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

$$r12 = C12/(sqrt(V1*V2) = 2rab/(1+rab)$$

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

Alpha and Beta with general and group factors

		Test Size = 10 items		Test Size = 20 items	
General Factor	Group Factor	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

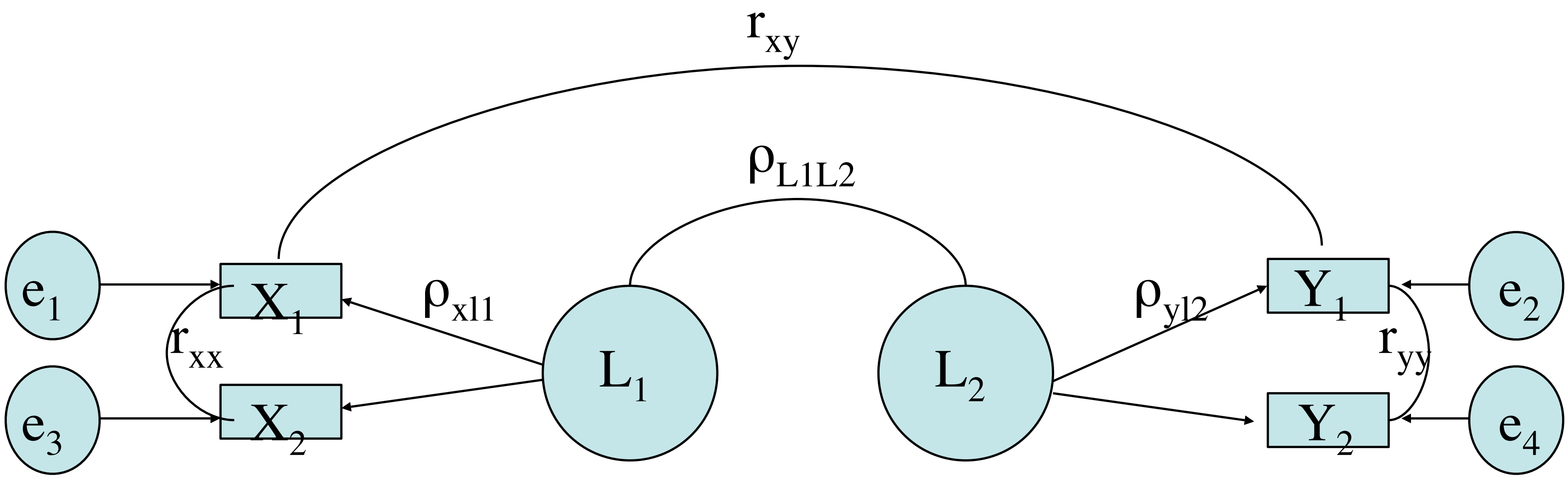
Generalizability Theory Reliability across facets

- 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



$\rho_{xL1} = \text{sqrt}(r_{xx})$ $\rho_{yL2} = \text{sqrt}(r_{yy})$

$$\rho_{L1L2} = r_{xy}/\rho_{x1l}\rho_{yl2}$$

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

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

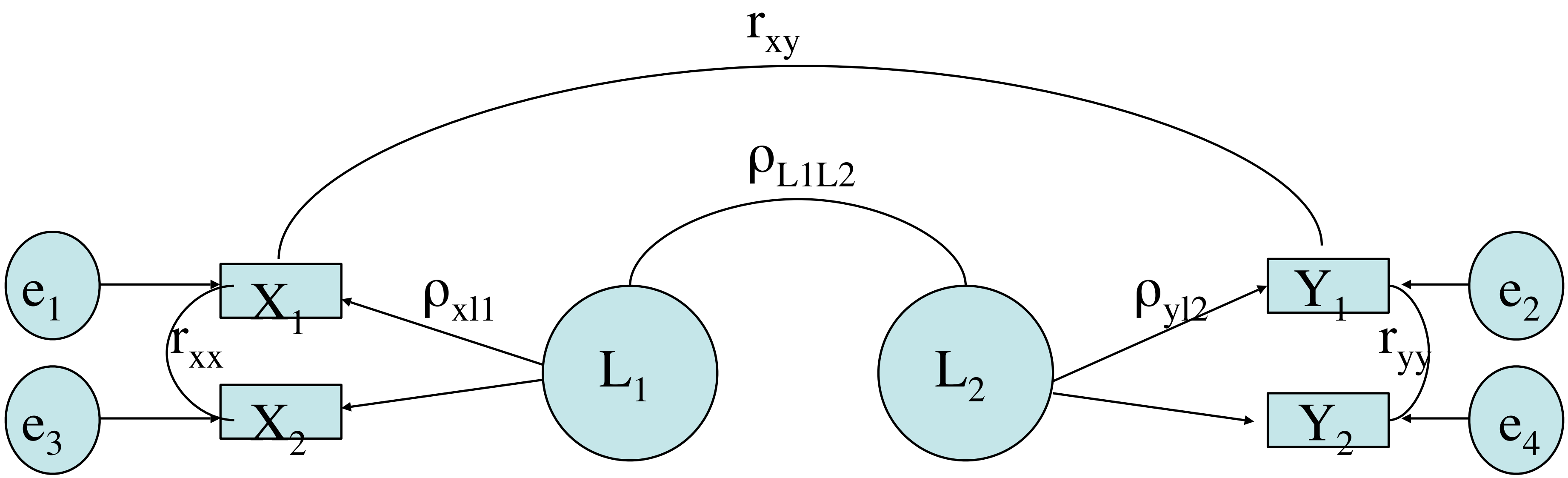
Correcting for attenuation

	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	ρ_{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



$\rho_{xL1} = sqrt(r_{xx})$ $\rho_{yL2} = sqrt(r_{yy})$

$\rho_{L1L2} = r_{xy}/\rho_{x1l}\rho_{yl2}$

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

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

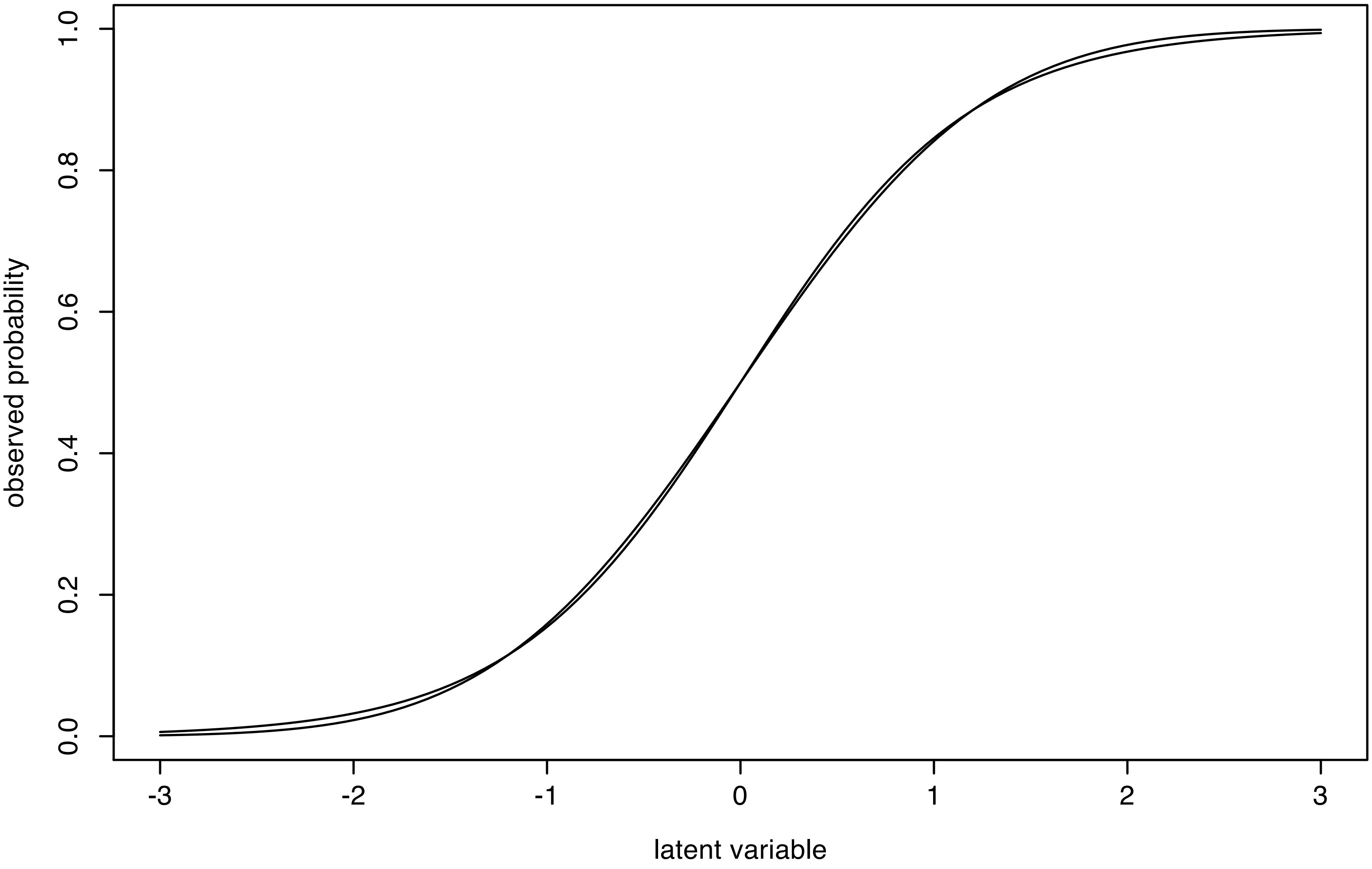
Item Response Theory

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

Item Response Theory

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

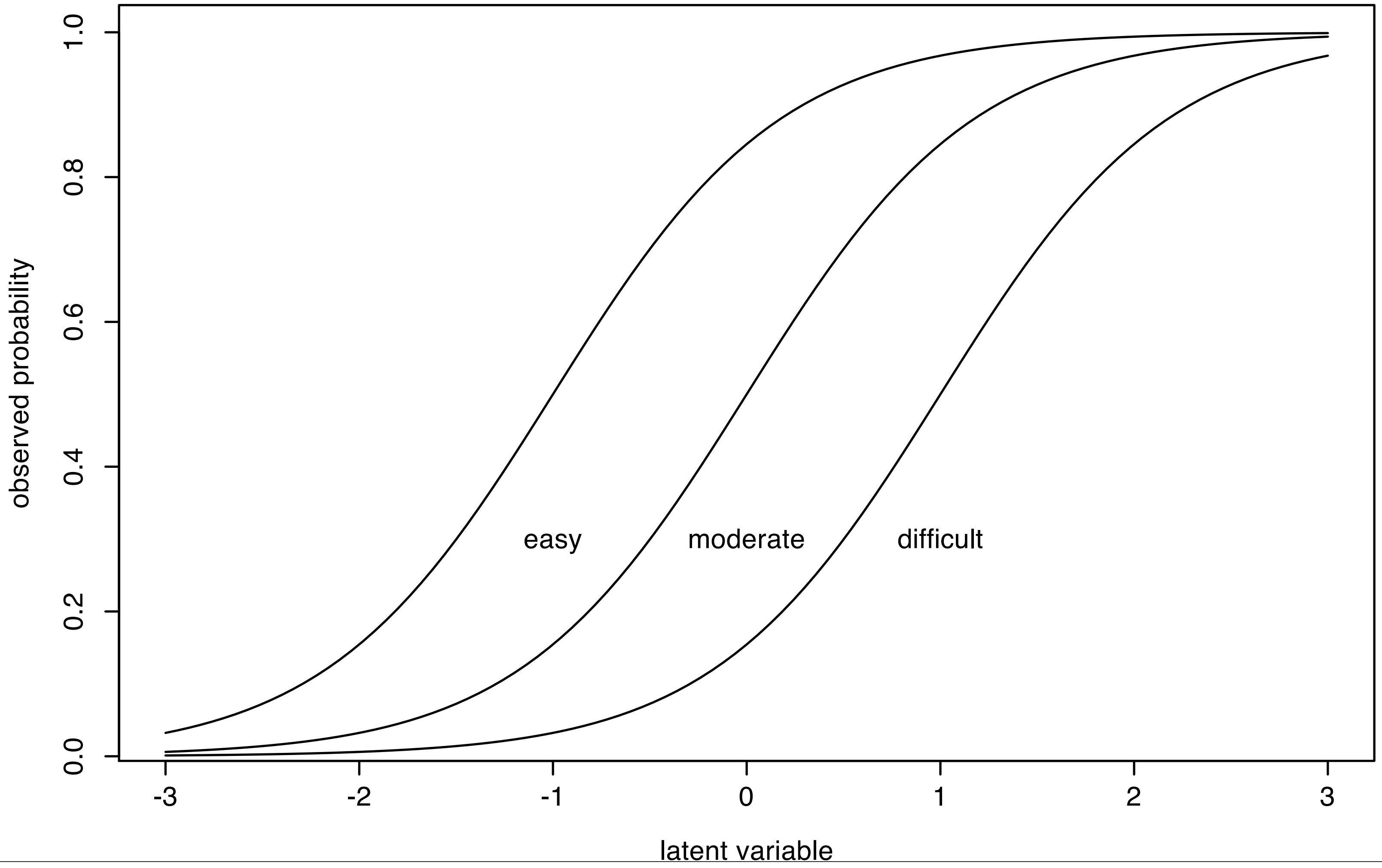
Logistic and cumulative normal



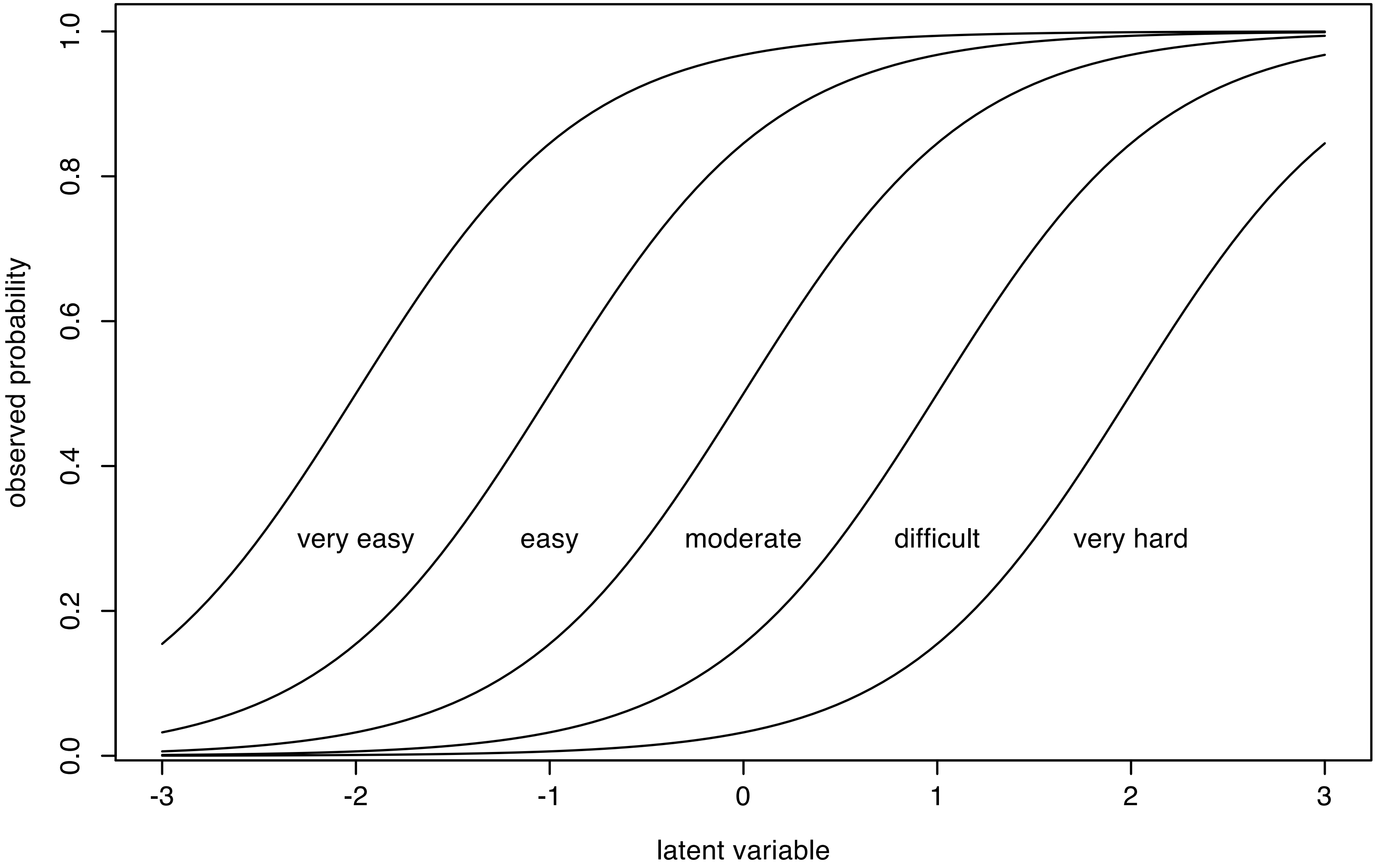
Item difficulty and ability

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

IRT of three item difficulties



item difficulties = -2, -1, 0 , 1, 2



Estimation of ability for a particular person for known item difficulty

- The probability of any pattern of responses (x1, x2, x3, …. Xn) is the product of the probabilities of each response ∏(p(xi)).
- Consider the odds ratio of a response
  - p/(1-p) = 1/(1+exp(1.7*(δ_j- θ_i))) /(1- 1/(1+exp(1.7*(δ_j- θ_i)))) =
  - p/(1-p) = exp(1.7*(δ_j- θ_i))) and therefore:
  - Ln(odds) = 1.7* (θ_i - δ_j) and
  - Ln (odds of a pattern) = 1.7∑ (θ_i - δ_j) 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.

IRT using R

- Use the ltm package (requires MASS)
- example data sets include LSAT and Abortion attitudes
- Lsat[1:10,] shows some data
- describe(LSAT) (means and sd)
- m1 <- rasch(LSAT)

Consider data from the LSAT

	Item 1	Item 2	Item 3	Item 4	Item 5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	1
5	0	0	0	0	1
6	0	0	0	0	1
7	0	0	0	0	1
8	0	0	0	0	1
9	0	0	0	0	1
10	0	0	0	1	0

...

Descriptive stats

```
                    describe(LSAT)
          n mean   sd median min max range  skew   se
Item 1 1000 0.92 0.27      1   0   1     1 -3.20 0.01
Item 2 1000 0.71 0.45      1   0   1     1 -0.92 0.01
Item 3 1000 0.55 0.50      1   0   1     1 -0.21 0.02
Item 4 1000 0.76 0.43      1   0   1     1 -1.24 0.01
Item 5 1000 0.87 0.34      1   0   1     1 -2.20 0.01
```

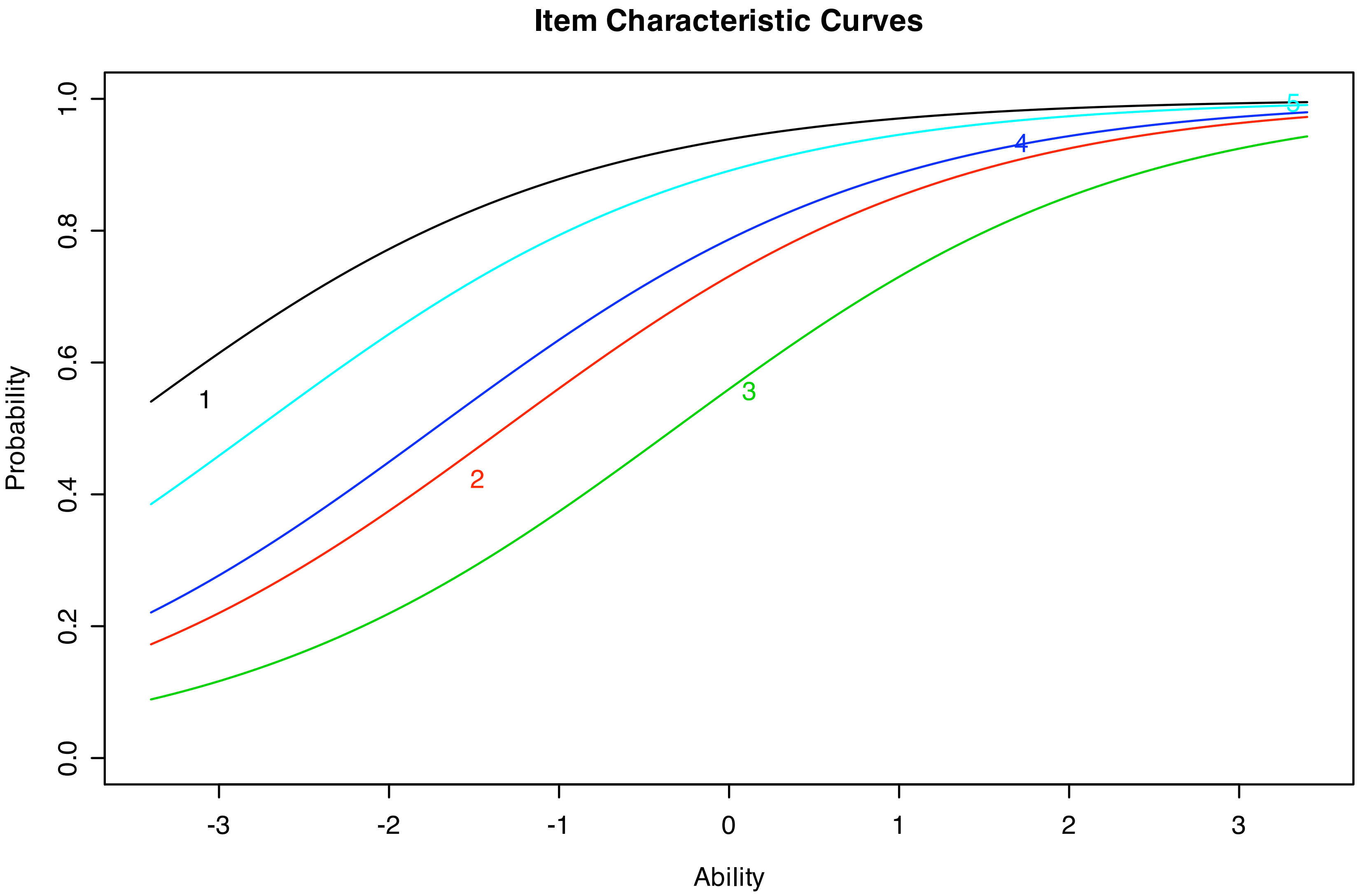
Correlations and alpha

```
       Item 1 Item 2 Item 3 Item 4 Item 5
Item 1   1.00   0.07   0.10   0.04   0.02
Item 2   0.07   1.00   0.11   0.06   0.09
Item 3   0.10   0.11   1.00   0.11   0.05
Item 4   0.04   0.06   0.11   1.00   0.10
Item 5   0.02   0.09   0.05   0.10   1.00
```

```
cl <- cor(LSAT)
Vt <- sum(cl)                             6.53
iv <- sum(diag(cl))   (or tr(cl))  = 5
alpha <- ((Vt-iv)/Vt)*(5/4) (6.53-5)*5/4
 alpha
[1] 0.29
```

Rasch model

```
m1 <- rasch(Lsat)
 coef(m1,TRUE)
       Dffclt Dscrmn P(x=1|z=0)
Item 1 -3.615  0.755      0.939
Item 2 -1.322  0.755      0.731
Item 3 -0.318  0.755      0.560
Item 4 -1.730  0.755      0.787
Item 5 -2.780  0.755      0.891
```



Item Characteristic Curves

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(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

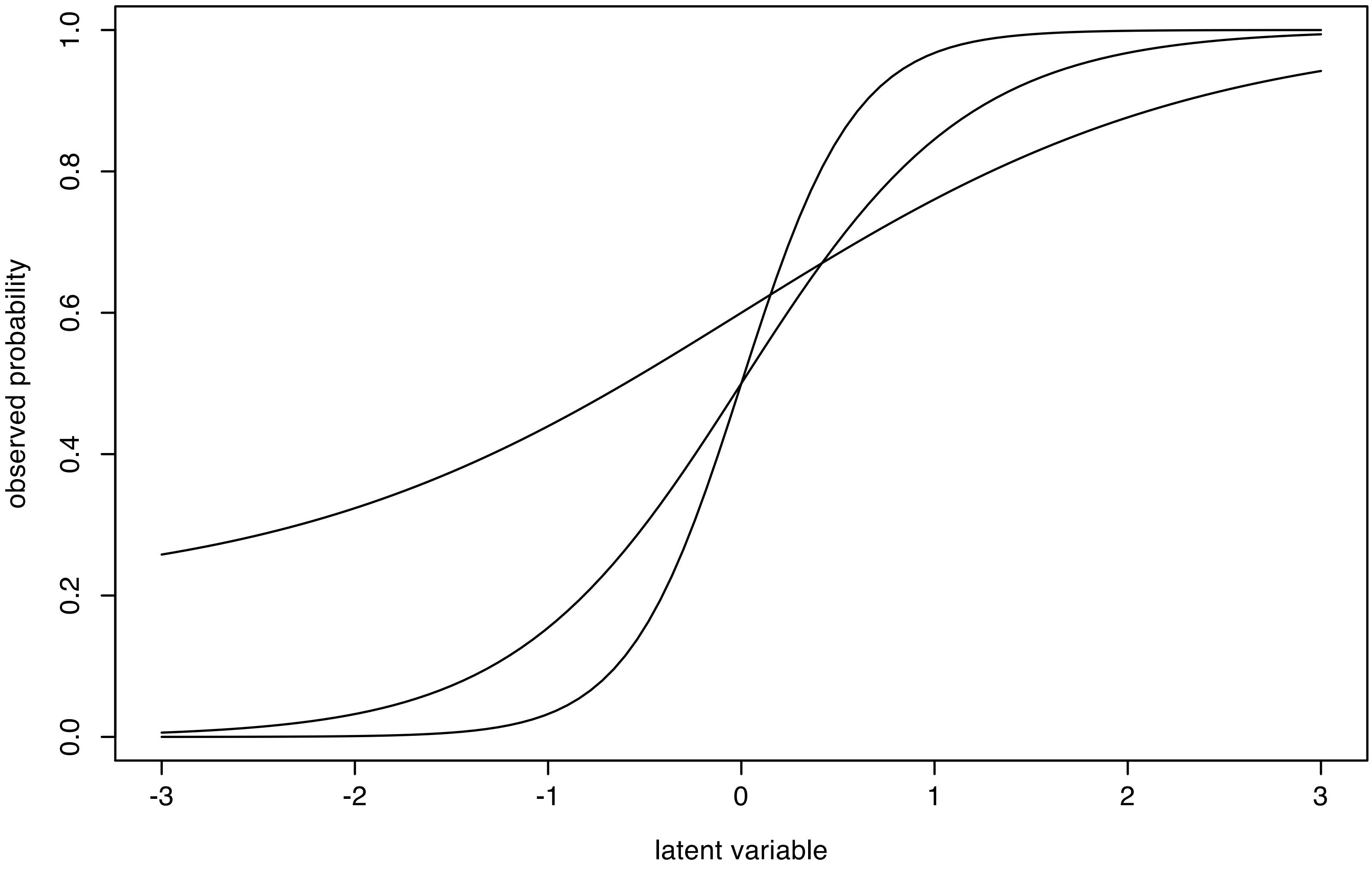
- CAT items have difficulty level tailored to individual so that each person passes about 50% of the items.
- This increases the subjective feeling of failure and interacts with test anxiety
- Anxious people quit after failing and try harder after success -- their pattern on CAT is to do progressively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

3 parameter IRT

slope, location, guessing



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.