Psychology 405: Psychometric Theory
Reliability Theory

William Revelle

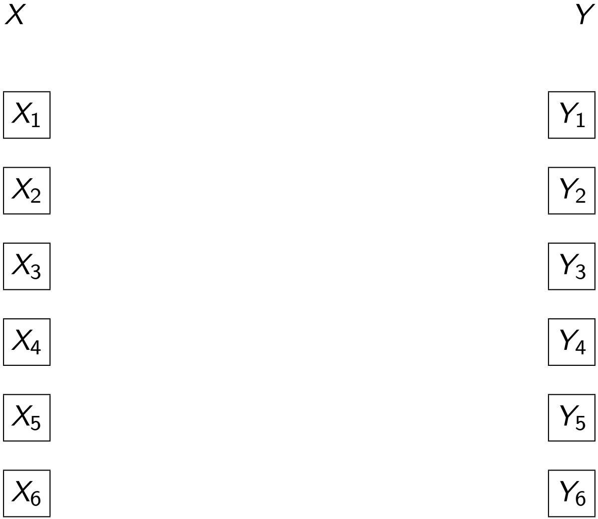
Department of Psychology
Northwestern University
Evanston, Illinois USA

May, 2014

Outline: Part I: Classical Test Theory

1. Preliminaries
   - Classical test theory
   - Congeneric test theory
2. Reliability and internal structure
   - Estimating reliability by split halves
   - Domain Sampling Theory
   - Coefficients based upon the internal structure of a test
   - Problems with α
3. Types of reliability
   - Alpha and its alternatives
4. Calculating reliabilities
   - Congeneric measures
   - Hierarchical structures
5. $2 \neq 1$
   - Multiple dimensions - falsely labeled as one
   - Using score.items to find reliabilities of multiple scales
   - Intraclass correlations

Observed Variables



X

Y

X_1

X_2

X_3

X_4

X_5

X_6

Y_1

Y_2

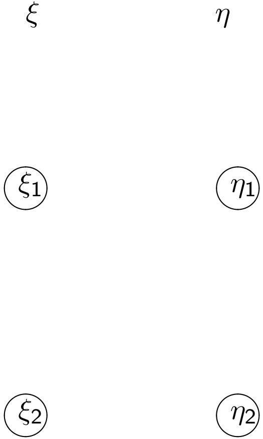
Y_3

Y_4

Y_5

Y_6

Latent Variables

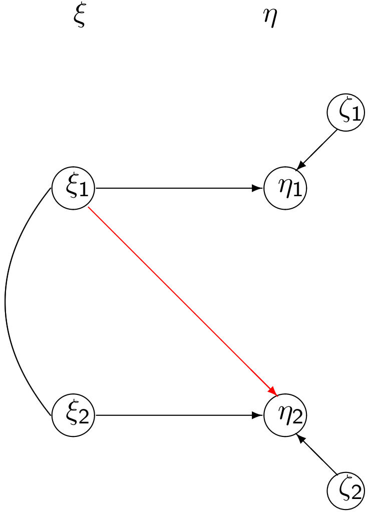


ξ η

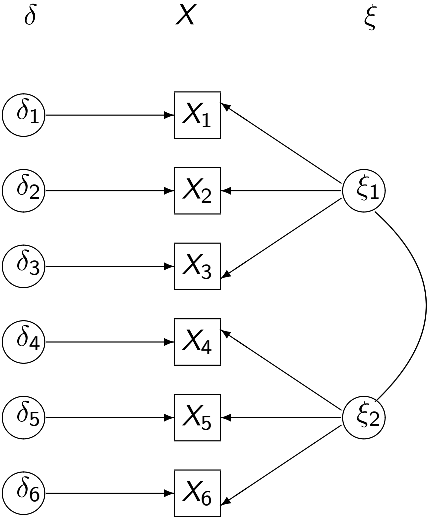
ξ_1 η_1

ξ_2 η_2

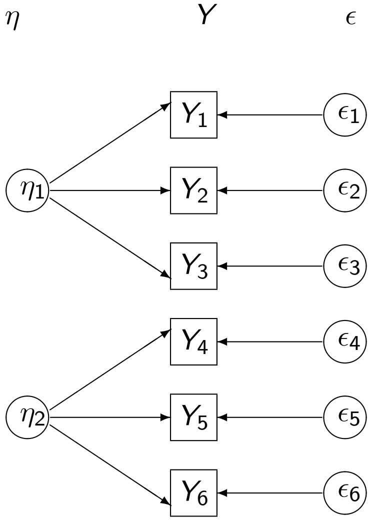
Theory: A regression model of latent variables



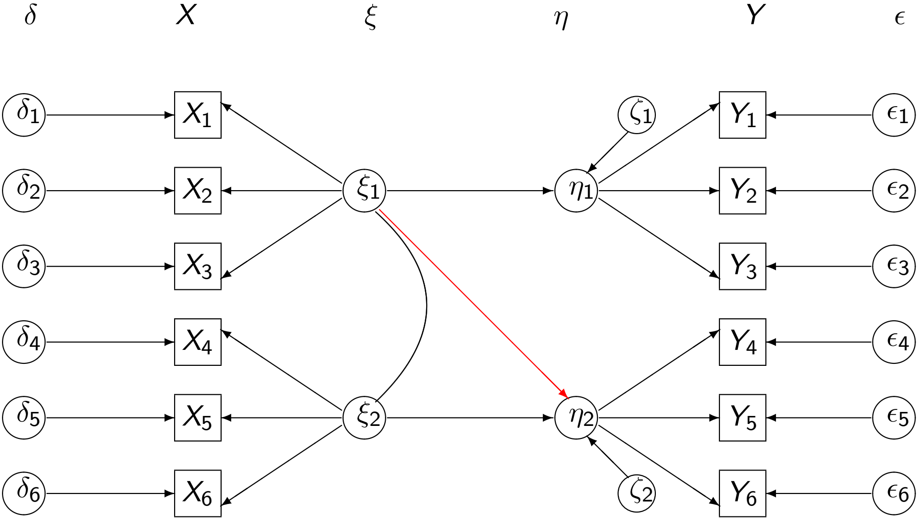
A measurement model for X – Correlated factors



A measurement model for Y - uncorrelated factors



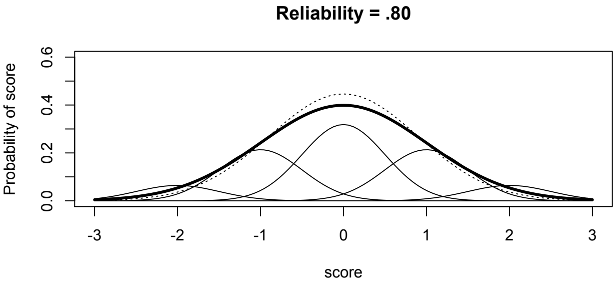
A complete structural model

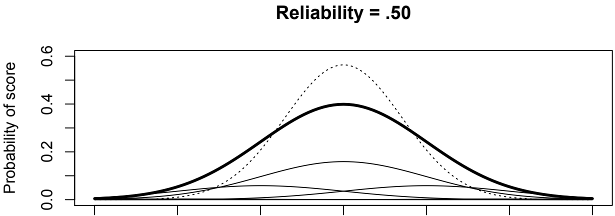


All data are befuddled with error

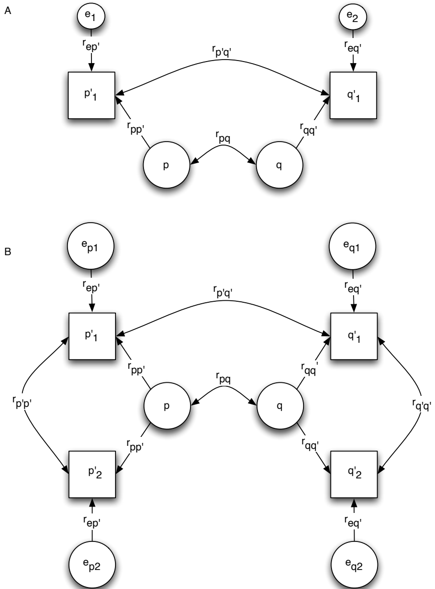
Now, suppose that we wish to ascertain the correspondence between a series of values, p, and another series, q. By practical observation we evidently do not obtain the true objective values, p and q, but only approximations which we will call p' and q'. Obviously, p' is less closely connected with q', than is p with q, for the first pair only correspond at all by the intermediation of the second pair; the real correspondence between p and q, shortly r_{pq} has been "attenuated" into $r_{p'q'}$ (Spearman, 1904, p 90).

All data are befuddled by error: Observed Score = True score + Error score





Spearman's parallell test theory



Classical True score theory

Let each individual score, x, reflect a true value, t, and an error value, e, and the expected score over multiple observations of x is t, and the expected score of e for any value of p is 0. Then, because the expected error score is the same for all true scores, the covariance of true score with error score (σ_{te}) is zero, and the variance of x, σ_x^2, is just

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 + 2\sigma_{te} = \sigma_t^2 + \sigma_e^2.$$

Similarly, the covariance of observed score with true score is just the variance of true score

$$\sigma_{xt} = \sigma_t^2 + \sigma_{te} = \sigma_t^2$$

and the correlation of observed score with true score is

$$\rho_{xt} = \frac{\sigma_{xt}}{\sqrt{(\sigma_t^2 + \sigma_e^2)(\sigma_t^2)}} = \frac{\sigma_t^2}{\sqrt{\sigma_x^2 \sigma_t^2}} = \frac{\sigma_t}{\sigma_x}. \tag{1}$$

Classical Test Theory

By knowing the correlation between observed score and true score, ρ_{xt}, and from the definition of linear regression predicted true score, $\hat{t}$, for an observed x may be found from

$$\hat{t} = b_{t.x}x = \frac{\sigma_t^2}{\sigma_x^2}x = \rho_{xt}^2 x. \tag{2}$$

All of this is well and good, but to find the correlation we need to know either σ_t^2 or σ_e^2. The question becomes how do we find σ_t^2 or σ_e^2?.

Regression effects due to unreliability of measurement

Consider the case of air force instructors evaluating the effects of reward and punishment upon subsequent pilot performance. Instructors observe 100 pilot candidates for their flying skill. At the end of the day they reward the best 50 pilots and punish the worst 50 pilots.

- Day 1
  - Mean of best 50 pilots 1 is 75
  - Mean of worst 50 pilots is 25
- Day 2
  - Mean of best 50 has gone down to 65 (a loss of 10 points)
  - Mean of worst 50 has gone up to 35 (a gain of 10 points)
- It seems as if reward hurts performance and punishment helps performance.
- If there is no effect of reward and punishment, what is the expected correlation from day 1 to day 2?

Correcting for attenuation

> *To ascertain the amount of this attenuation, and thereby discover the true correlation, it appears necessary to make two or more independent series of observations of both p and q. (Spearman, 1904, p 90)*

Spearman's solution to the problem of estimating the true relationship between two variables, p and q, given observed scores p' and q' was to introduce two or more additional variables that came to be called *parallel tests*. These were tests that had the same true score for each individual and also had equal error variances. To Spearman (1904b p 90) this required finding "the average correlation between one and another of these independently obtained series of values" to estimate the reliability of each set of measures $(r_{p'p'}, r_{q'q'})$, and then to find

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} r_{q'q'}}}. \tag{3}$$

Two parallel tests

The correlation between two parallel tests is the squared correlation of each test with true score and is the percentage of test variance that is true score variance

$$\rho_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \rho_{xt}^2. \tag{4}$$

Reliability is the fraction of test variance that is true score variance. Knowing the reliability of measures of p and q allows us to correct the observed correlation between p' and q' for the reliability of measurement and to find the unattenuated correlation between p and q.

$$r_{pq} = \frac{\sigma_{pq}}{\sqrt{\sigma_p^2 \sigma_q^2}} \tag{5}$$

and

$$r_{p'q'} = \frac{\sigma_{p'q'}}{\sqrt{\sigma_{p'}^2 \sigma_{q'}^2}} = \frac{\sigma_{p+e_1'} \sigma_{q+e_2'}}{\sqrt{\sigma_{p'}^2 \sigma_{q'}^2}} = \frac{\sigma_{pq}}{\sqrt{\sigma_{p'}^2 \sigma_{q'}^2}} \tag{6}$$

Modern “Classical Test Theory”

Reliability is the correlation between two *parallel tests* where tests are said to be parallel if for every subject, the true scores on each test are the expected scores across an infinite number of tests and thus the same, and the true score variances for each test are the same ($\sigma^2_{p'_1} = \sigma^2_{p'_2} = \sigma^2_{p'}$), and the error variances across subjects for each test are the same ($\sigma^2_{e'_1} = \sigma^2_{e'_2} = \sigma^2_{e'}$) (see Figure 19), (Lord & Novick, 1968; McDonald, 1999). The correlation between two parallel tests will be

$$\rho_{p'_1 p'_2} = \rho_{p' p'} = \frac{\sigma_{p'_1 p'_2}}{\sqrt{\sigma^2_{p'_1}\sigma^2_{p'_2}}} = \frac{\sigma^2_p + \sigma_{pe_1} + \sigma_{pe_2} + \sigma_{e_1 e_2}}{\sigma^2_{p'}} = \frac{\sigma^2_p}{\sigma^2_{p'}}. \quad (7)$$

Classical Test Theory

but from Eq 4,

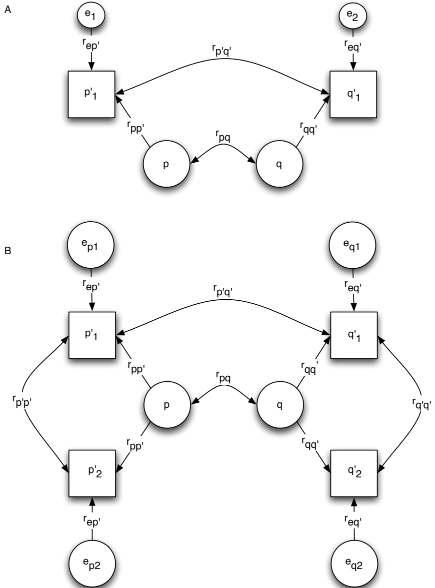
$$\sigma_p^2 = \rho_{p'p'}\sigma_{p'}^2 \tag{8}$$

and thus, by combining equation 5 with 6 and 8 the *unattenuated correlation* between p and q corrected for reliability is Spearman's equation 3

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'}r_{q'q'}}}. \tag{9}$$

As Spearman recognized, *correcting for attenuation* could show structures that otherwise, because of unreliability, would be hard to detect.

Spearman's parallell test theory



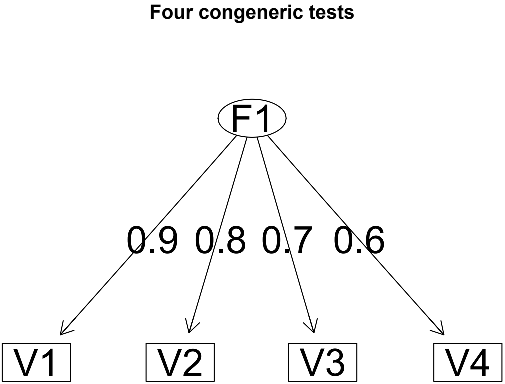
When is a test a parallel test?

But how do we know that two tests are parallel? For just knowing the correlation between two tests, without knowing the true scores or their variance (and if we did, we would not bother with reliability), we are faced with three knowns (two variances and one covariance) but ten unknowns (four variances and six covariances). That is, the observed correlation, $r_{p'_1 p'_2}$ represents the two known variances $s^2_{p'_1}$ and $s^2_{p'_2}$ and their covariance $s_{p'_1 p'_2}$. The model to account for these three knowns reflects the variances of true and error scores for p'_1 and p'_2 as well as the six covariances between these four terms. In this case of two tests, by defining them to be parallel with uncorrelated errors, the number of unknowns drop to three (for the true scores variances of p'_1 and p'_2 are set equal, as are the error variances, and all covariances with error are set to zero) and the (equal) reliability of each test may be found.

The problem of parallel tests

Unfortunately, according to this concept of parallel tests, the possibility of one test being far better than the other is ignored. Parallel tests need to be parallel by construction or assumption and the assumption of parallelism may not be tested. With the use of more tests, however, the number of assumptions can be relaxed (for three tests) and actually tested (for four or more tests).

Four congeneric tests – 1 latent factor



Observed variables and estimated parameters of a congeneric test

Observed correlations and modeled parameters

Variable	*Test*$_1$	*Test*$_2$	*Test*$_3$	*Test*$_4$
Test$_1$	$\sigma^2_{x_1} = \lambda_1\sigma^2_\theta + \epsilon^2_1$			
Test$_2$	$\sigma_{x_1x_2} = \lambda_1\sigma_\theta\lambda_2\sigma_\theta$	$\sigma^2_{x_2} = \lambda_2\sigma^2_\theta + \epsilon^2_2$		
Test$_3$	$\sigma_{x_1x_3} = \lambda_1\sigma_\theta\lambda_3\sigma_\theta$	$\sigma_{x_2x_3} = \lambda_2\sigma_\theta\lambda_3\sigma_\theta$	$\sigma^2_{x_3} = \lambda_3\sigma^2_\theta + \epsilon^2_3$	
Test$_4$	$\sigma_{x_1x_4} = \lambda_1\sigma_\theta\lambda_4\sigma_t$	$\sigma_{x_2x_4} = \lambda_2\sigma_\theta\lambda_4\sigma_\theta$	$\sigma_{x_3x_4} = \lambda_3\sigma_\theta\lambda_4\sigma_\theta$	$\sigma^2_{x_4} = \lambda_4\sigma^2_\theta + \epsilon^2_4$

Observed variables and estimated parameters of a congeneric test

	V1	V2	V3	V4	V1	V2	V3	V
V1	s_1^2				$\lambda_1\sigma_t^2 + \sigma_{e_1}^2$			
V2	s_{12}	s_2^2			$\lambda_1\lambda_2\sigma_t^2$	$\lambda_2\sigma_t^2 + \sigma_{e_2}^2$		
V3	s_{13}	s_{23}	s_3^2		$\lambda_1\lambda_3\sigma_t^2$	$\lambda_2\lambda_3\sigma_t^2$	$\lambda_3\sigma_t^2 + \sigma_{e_3}^2$	
V4	s_{14}	s_{24}	s_{34}	s_4^2	$\lambda_1\lambda_4\sigma_t^2$	$\lambda_2\lambda_3\sigma_t^2$	$\lambda_3\lambda_4\sigma_t^2$	$\lambda_4\sigma_t^2$

But what if we don't have three or more tests?

Unfortunately, with rare exceptions, we normally are faced with just one test, not two, three or four. How then to estimate the reliability of that one test? Defined as the correlation between a test and a test just like it, reliability would seem to require a second test. The traditional solution when faced with just one test is to consider the internal structure of that test. Letting reliability be the ratio of true score variance to test score variance (Equation 1), or alternatively, 1 - the ratio of error variance to true score variance, the problem becomes one of estimating the amount of error variance in the test. There are a number of solutions to this problem that involve examining the internal structure of the test. These range from considering the correlation between two random parts of the test to examining the structure of the items themselves.

Split halves

$$\Sigma_{XX'} = \begin{pmatrix} \mathbf{V_x} & \vdots & \mathbf{C_{xx'}} \\ \dots & \dots & \dots \\ \mathbf{C_{xx'}} & \vdots & \mathbf{V_{x'}} \end{pmatrix} \tag{10}$$

and letting $V_{\mathbf{x}} = \mathbf{1V_x1}'$ and $C_{\mathbf{XX'}} = \mathbf{1C}_{XX'}\mathbf{1}'$ the correlation between the two tests will be

$$\rho = \frac{C_{xx'}}{\sqrt{V_x V_{x'}}}$$

But the variance of a test is simply the sum of the true covariances and the error variances:

$$V_{\mathbf{x}} = \mathbf{1V_x1}' = \mathbf{1C_t1}' + \mathbf{1V_e1}' = V_t + V_e$$

Split halves

and the structure of the two tests seen in Equation 10 becomes

$$\Sigma_{XX'} = \left(\begin{array}{ccc} \mathbf{V_X} = \mathbf{V_t} + \mathbf{V_e} & \vdots & \mathbf{C_{xx'}} = \mathbf{V_t} \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\ \mathbf{V_t} = \mathbf{C_{xx'}} & \vdots & \mathbf{V_{t'}} + \mathbf{V_{e'}} = \mathbf{V}_{X'} \end{array}\right)$$

and because $\mathbf{V}_t = \mathbf{V}_{t'}$ and $\mathbf{V}_e = \mathbf{V}_{e'}$ the correlation between each half, (their reliability) is

$$\rho = \frac{C_{XX'}}{V_X} = \frac{V_t}{V_X} = 1 - \frac{V_e}{V_t}.$$

Split halves

The split half solution estimates reliability based upon the correlation of two random split halves of a test and the implied correlation with another test also made up of two random splits:

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc}
\mathbf{V}_{\mathbf{x}_1} & \vdots & \mathbf{C}_{\mathbf{x}_1\mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_1\mathbf{x}'_1} & \vdots & \mathbf{C}_{\mathbf{x}_1\mathbf{x}'_2} \\
\multicolumn{3}{c|}{\cdots\cdots\cdots\cdots} & \multicolumn{3}{c}{\cdots\cdots\cdots\cdots} \\
\mathbf{C}_{\mathbf{x}_1\mathbf{x}_2} & \vdots & \mathbf{V}_{\mathbf{x}_2} & \mathbf{C}_{\mathbf{x}_2\mathbf{x}'_1} & \vdots & \mathbf{C}_{\mathbf{x}_2\mathbf{x}'_1} \\
\hline
\mathbf{C}_{\mathbf{x}_1\mathbf{x}'_1} & \vdots & \mathbf{C}_{\mathbf{x}_2\mathbf{x}'_1} & \mathbf{V}_{\mathbf{x}'_1} & \vdots & \mathbf{C}_{\mathbf{x}'_1\mathbf{x}'_2} \\
\mathbf{C}_{\mathbf{x}_1\mathbf{x}'_2} & \vdots & \mathbf{C}_{\mathbf{x}_2\mathbf{x}'_2} & \mathbf{C}_{\mathbf{x}'_1\mathbf{x}'_2} & \vdots & \mathbf{V}_{\mathbf{x}'_2}
\end{array}\right)$$

Split halves

Because the splits are done at random and the second test is parallel with the first test, the expected covariances between splits are all equal to the true score variance of one split ($\mathbf{V_{t_1}}$), and the variance of a split is the sum of true score and error variances:

$$\Sigma_{XX'} = \left(\begin{array}{c:c|c:c} \mathbf{V_{t_1}} + \mathbf{V_{e_1}} & \mathbf{V_{t_1}} & \mathbf{V_{t_1}} & \mathbf{V_{t_1}} \\ \hdashline \mathbf{V_{t_1}} & \mathbf{V_{t_1}} + \mathbf{V_{e_1}} & \mathbf{V_{t_1}} & \mathbf{V_{t_1}} \\ \hline \mathbf{V_{t_1}} & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}} + \mathbf{V_{e'_1}} & \mathbf{V_{t'_1}} \\ \mathbf{V_{t_1}} & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}} & \mathbf{V_{t'_1}} + \mathbf{V_{e'_1}} \end{array}\right)$$

The correlation between a test made of up two halves with intercorrelation ($r_1 = V_{t_1}/V_{x_1}$) with another such test is

$$r_{xx'} = \frac{4V_{t_1}}{\sqrt{(4V_{t_1} + 2V_{e_1})(4V_{t_1} + 2V_{e_1})}} = \frac{4V_{t_1}}{2V_{t_1} + 2V_{x_1}} = \frac{4r_1}{2r_1 + 2}$$

and thus

The Spearman Brown Prophecy Formula

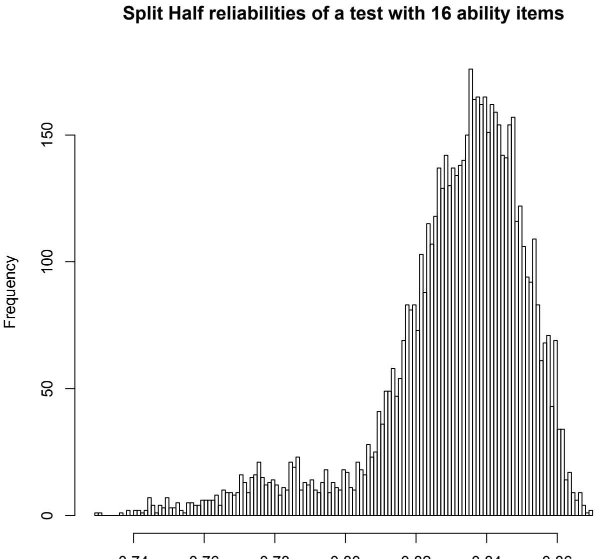
The correlation between a test made of up two halves with intercorrelation ($r_1 = V_{t_1}/V_{x_1}$) with another such test is

$$r_{xx'} = \frac{4V_{t_1}}{\sqrt{(4V_{t_1} + 2V_{e_1})(4V_{t_1} + 2V_{e_1})}} = \frac{4V_{t_1}}{2V_{t_1} + 2V_{x_1}} = \frac{4r_1}{2r_1 + 2}$$

and thus

$$r_{xx'} = \frac{2r_1}{1 + r_1} \qquad (12)$$

6,435 possible eight item splits of the 16 ability items


Split Half reliabilities of a test with 16 ability items

Domain sampling

Other techniques to estimate the reliability of a single test are based on the *domain sampling* model in which tests are seen as being made up of items randomly sampled from a domain of items. Analogous to the notion of estimating characteristics of a population of people by taking a sample of people is the idea of sampling items from a universe of items.
Consider a test meant to assess English vocabulary. A person's vocabulary could be defined as the number of words in an unabridged dictionary that he or she recognizes. But since the total set of possible words can exceed 500,000, it is clearly not feasible to ask someone all of these words. Rather, consider a test of k words sampled from the larger domain of n words. What is the correlation of this test with the domain? That is, what is the correlation across subjects of test scores with their domain scores.?

Correlation of an item with the domain

First consider the correlation of a single (randomly chosen) item with the domain. Let the domain score for an individual be D_i and the score on a particular item, j, be X_{ij}. For ease of calculation, convert both of these to deviation scores. $d_i = D_i - \bar{D}$ and $x_{ij} = X_{ij} - \bar{X}_j$. Then

$$r_{x_j d} = \frac{cov_{x_j d}}{\sqrt{\sigma^2_{x_j} \sigma^2_d}}.$$

Now, because the domain is just the sum of all the items, the domain variance σ^2_d is just the sum of all the item variances and all the item covariances

$$\sigma^2_d = \sum_{j=1}^{n} \sum_{k=1}^{n} cov_{x_{jk}} = \sum_{j=1}^{n} \sigma^2_{x_j} + \sum_{j=1}^{n} \sum_{k \neq j} cov_{x_{jk}}.$$

Correlation of an item with the domain

Then letting $\bar{c} = \frac{\sum_{j=1}^{j=n} \sum_{k \neq j} cov_{x_{jk}}}{n(n-1)}$ be the average covariance and $\bar{v} = \frac{\sum_{j=1}^{j=n} \sigma^2_{x_j}}{n}$ the average item variance, the correlation of a randomly chosen item with the domain is

$$r_{x_j d} = \frac{\bar{v} + (n-1)\bar{c}}{\sqrt{\bar{v}(n\bar{v} + n(n-1)\bar{c})}} = \frac{\bar{v} + (n-1)\bar{c}}{\sqrt{n\bar{v}(\bar{v} + (n-1)\bar{c}))}}.$$

Squaring this to find the squared correlation with the domain and factoring out the common elements leads to

$$r^2_{x_j d} = \frac{(\bar{v} + (n-1)\bar{c})}{n\bar{v}}.$$

and then taking the limit as the size of the domain gets large is

$$\lim_{n \to \infty} r^2_{x_j d} = \frac{\bar{c}}{\bar{v}}. \tag{13}$$

That is, the squared correlation of an average item with the domain is the ratio of the average interitem covariance to the average item variance. Compare the correlation of a test with true

Domain sampling – correlation of an item with the domain

$$\lim_{n \to \infty} r^2_{x_j d} = \frac{\bar{c}}{\bar{v}}. \tag{14}$$

That is, the squared correlation of an average item with the domain is the ratio of the average interitem covariance to the average item variance. Compare the correlation of a test with true score (Eq 4) with the correlation of an item to the domain score (Eq 14). Although identical in form, the former makes assumptions about true score and error, the latter merely describes the domain as a large set of similar items.

Correlation of a test with the domain

A similar analysis can be done for a test of length k with a large domain of n items. A k-item test will have total variance, V_k, equal to the sum of the k item variances and the k(k-1) item covariances:

$$V_k = \sum_{i=1}^{k} v_i + \sum_{i=1}^{k} \sum_{j \neq i}^{k} c_{ij} = k\bar{v} + k(k-1)\bar{c}.$$

The correlation with the domain will be

$$r_{kd} = \frac{cov_k d}{\sqrt{V_k V_d}} = \frac{k\bar{v} + k(n-1)\bar{c}}{\sqrt{(k\bar{v} + k(k-1)\bar{c})(n\bar{v} + n(n-1)\bar{c})}} = \frac{k(\bar{v} + (n-1)\bar{c})}{\sqrt{nk(\bar{v} + (k-1)\bar{c})(\bar{v} + (n-1)\bar{c})}}$$

Correlation of a test with the domain

Then the squared correlation of a k item test with the n item domain is

$$r_{kd}^2 = \frac{k(\bar{v} + (n-1)\bar{c})}{n(\bar{v} + (k-1)\bar{c})}$$

and the limit as n gets very large becomes

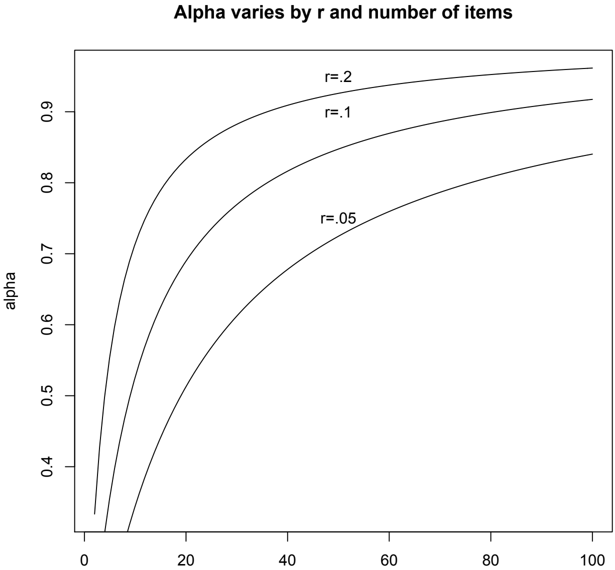
$$\lim_{n \to \infty} r_{kd}^2 = \frac{k\bar{c}}{\bar{v} + (k-1)\bar{c}}. \qquad (15)$$

Coefficient α

Find the correlation of a test with a test just like it based upon the internal structure of the first test. Basically, we are just estimating the error variance of the individual items.

$$\alpha = r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{k^2 \frac{\sigma_x^2 - \sum \sigma_i^2}{k(k-1)}}{\sigma_x^2} = \frac{k}{k-1} \frac{\sigma_x^2 - \sum \sigma_i^2}{\sigma_x^2} \tag{16}$$

Alpha varies by the number of items and the inter item correlation



Signal to Noise Ratio

The ratio of reliable variance to unreliable variance is known as the Signal/Noise ratio and is just

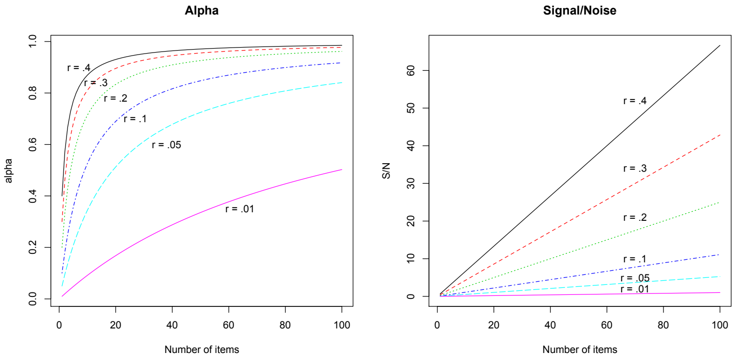
$$\frac{S}{N} = \frac{\rho^2}{1 - \rho^2}$$

, which for the same assumptions as for α, will be

$$\frac{S}{N} = \frac{n\bar{r}}{1 - \bar{r}}. \tag{17}$$

That is, the S/N ratio increases linearly with the number of items as well as with the average intercorrelation

Alpha vs signal/noise: and r and n



Find alpha using the alpha function

```
> alpha(bfi[16:20])
Reliability analysis
Call: alpha(x = bfi[16:20])

  raw_alpha std.alpha G6(smc) average_r mean  sd
      0.81      0.81     0.8      0.46   15 5.8

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r
N1      0.75      0.75    0.70      0.42
N2      0.76      0.76    0.71      0.44
N3      0.75      0.76    0.74      0.44
N4      0.79      0.79    0.76      0.48
N5      0.81      0.81    0.79      0.51

 Item statistics
      n    r r.cor mean  sd
N1  990 0.81  0.78  2.8 1.5
N2  990 0.79  0.75  3.5 1.5
N3  997 0.79  0.72  3.2 1.5
N4  996 0.71  0.60  3.1 1.5
N5  992 0.67  0.52  2.9 1.6
```

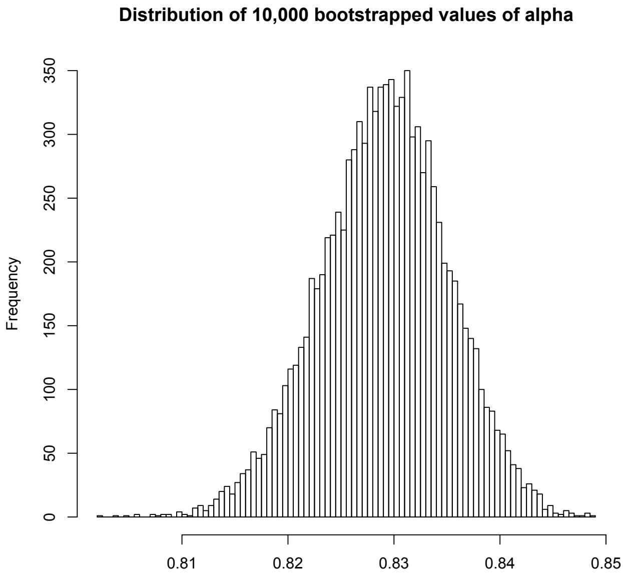
What if items differ in their direction?

```
> alpha(bfi[6:10],check.keys=FALSE)


Reliability analysis
Call: alpha(x = bfi[6:10], check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r mean   sd
     -0.28     -0.22    0.13    -0.038  3.8 0.58

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r
C1    -0.430    -0.472  -0.020   -0.0871
C2    -0.367    -0.423  -0.017   -0.0803
C3    -0.263    -0.295   0.094   -0.0604
C4    -0.022     0.123   0.283    0.0338
C5    -0.028     0.022   0.242    0.0057

 Item statistics
      n    r r.cor  r.drop mean  sd
C1 2779 0.56  0.51  0.0354  4.5 1.2
C2 2776 0.54  0.51 -0.0076  4.4 1.3
C3 2780 0.48  0.27 -0.0655  4.3 1.3
C4 2774 0.20 -0.34 -0.2122  2.6 1.4
C5 2784 0.29 -0.19 -0.1875  3.3 1.6
```

But what if some items are reversed keyed?

```
 alpha(bfi[6:10])
Reliability analysis
Call: alpha(x = bfi[6:10])

  raw_alpha std.alpha G6(smc) average_r mean   sd
      0.73      0.73    0.69      0.35  3.8 0.58
 Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r
C1       0.69      0.70    0.64      0.36
C2       0.67      0.67    0.62      0.34
C3       0.69      0.69    0.64      0.36
C4-      0.65      0.66    0.60      0.33
C5-      0.69      0.69    0.63      0.36
 Item statistics
       n    r r.cor r.drop mean  sd
C1  2779 0.67  0.54   0.45  4.5 1.2
C2  2776 0.71  0.60   0.50  4.4 1.3
C3  2780 0.67  0.54   0.46  4.3 1.3
C4- 2774 0.73  0.64   0.55  2.6 1.4
C5- 2784 0.68  0.57   0.48  3.3 1.6
Warning message:  In alpha(bfi[6:10]) :
  Some items were negatively correlated with total scale and were automatically
```

Bootstrapped confidence intervals for α



Guttman's alternative estimates of reliability

Reliability is amount of test variance that is not error variance. But what is the error variance?

$$r_{xx} = \frac{V_x - V_e}{V_x} = 1 - \frac{V_e}{V_x}. \tag{18}$$

$$\lambda_1 = 1 - \frac{tr(\mathbf{V_x})}{V_x} = \frac{V_x - tr(\mathbf{V}_x)}{V_x}. \tag{19}$$

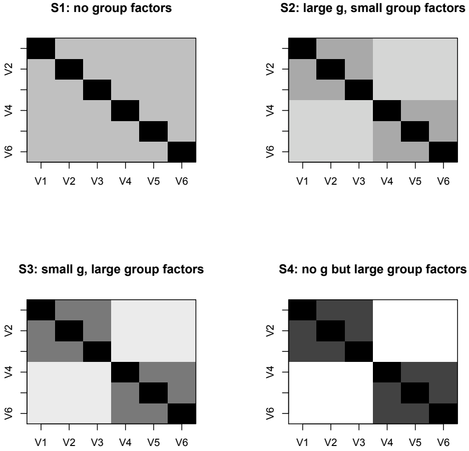
$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{n}{n-1}C_2}}{V_x} = \frac{V_x - tr(\mathbf{V}_x) + \sqrt{\frac{n}{n-1}C_2}}{V_x}. \tag{20}$$

$$\lambda_3 = \lambda_1 + \frac{\frac{V_X - tr(\mathbf{V}_X)}{n(n-1)}}{V_X} = \frac{n\lambda_1}{n-1} = \frac{n}{n-1}\left(1 - \frac{tr(\mathbf{V})_x}{V_x}\right) = \frac{n}{n-1}\frac{V_x - tr(\mathbf{V}_x)}{V_x} = \alpha \tag{21}$$

$$\lambda_4 = 2\left(1 - \frac{V_{X_a} + V_{X_b}}{V_X}\right) = \frac{4c_{ab}}{V_x} = \frac{4c_{ab}}{V_{X_a} + V_{X_b} + 2c_{ab}V_{X_a}V_{X_b}}. \tag{22}$$

$$\lambda_6 = 1 - \frac{\sum e_j^2}{V_x} = 1 - \frac{\sum(1 - r_{smc}^2)}{V_x} \tag{23}$$

Four different correlation matrices, one value of α


S1: no group factors

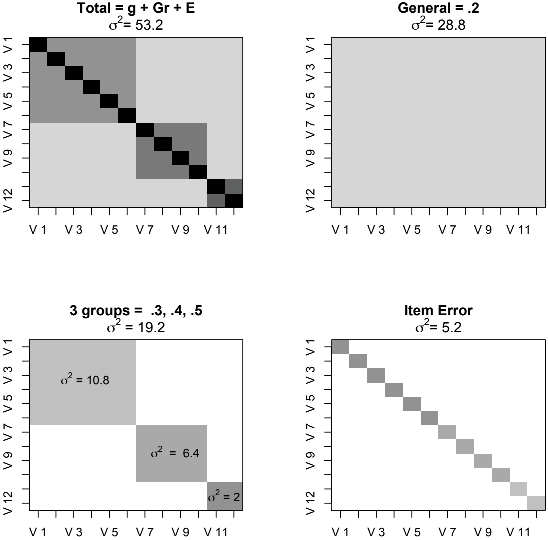
S2: large g, small group factors

S3: small g, large group factors

S4: no g but large group factors

1. The problem of group factors
2. If no groups, or many groups, α is ok

Decomposing a test into general, Group, and Error variance



1. Decompose total variance into general, group, specific, and error
2. $\alpha <$ total
3. $\alpha >$ general

Two additional alternatives to α: $\omega_{hierarchical}$ and $omega_{total}$

If a test is made up of a general, a set of group factors, and specific as well as error:

$$\mathbf{x} = \mathbf{cg} + \mathbf{Af} + \mathbf{Ds} + \mathbf{e} \tag{24}$$

then the communality of item$_j$, based upon general as well as group factors,

$$h_j^2 = c_j^2 + \sum f_{ij}^2 \tag{25}$$

and the unique variance for the item

$$u_j^2 = \sigma_j^2(1 - h_j^2) \tag{26}$$

may be used to estimate the test reliability.

$$\omega_t = \frac{\mathbf{1cc'1'} + \mathbf{1AA'1'}}{V_x} = 1 - \frac{\sum(1 - h_j^2)}{V_x} = 1 - \frac{\sum u^2}{V_x} \tag{27}$$

McDonald (1999) introduced two different forms for ω

$$\omega_t = \frac{\mathbf{1cc'1'} + \mathbf{1AA'1'}}{V_x} = 1 - \frac{\sum(1 - h_j^2)}{V_x} = 1 - \frac{\sum u^2}{V_x} \qquad (28)$$

and

$$\omega_h = \frac{\mathbf{1cc'1}}{V_x} = \frac{(\sum \Lambda_i)^2}{\sum \sum R_{ij}}. \qquad (29)$$

These may both be find by factoring the correlation matrix and finding the g and group factor loadings using the `omega` function.

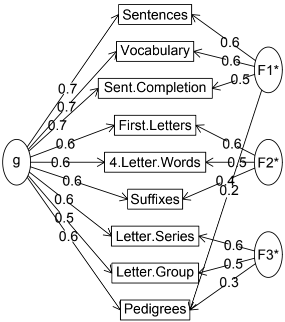
Using omega on the Thurstone data set to find alternative reliability estimates

```
> lower.mat(Thurstone)
> omega(Thurstone)
```

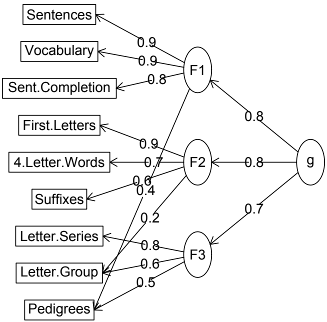
```
                Sntnc Vcblr Snt.C Frs.L 4.L.W Sffxs Ltt.S Pdgrs Ltt.G
Sentences       1.00
Vocabulary      0.83  1.00
Sent.Completion 0.78  0.78  1.00
First.Letters   0.44  0.49  0.46  1.00
4.Letter.Words  0.43  0.46  0.42  0.67  1.00
Suffixes        0.45  0.49  0.44  0.59  0.54  1.00
Letter.Series   0.45  0.43  0.40  0.38  0.40  0.29  1.00
Pedigrees       0.54  0.54  0.53  0.35  0.37  0.32  0.56  1.00
Letter.Group    0.38  0.36  0.36  0.42  0.45  0.32  0.60  0.45  1.00

Omega
Call: omega(m = Thurstone)
Alpha:                 0.89
G.6:                   0.91
Omega Hierarchical:    0.74
Omega H asymptotic:    0.79
Omega Total            0.93
```

Two ways of showing a general factor

Omega



Hierarchical (multilevel) Structure



omega function does a Schmid Leiman transformation

```
> omega(Thurstone,sl=FALSE)
Omega
Call: omega(m = Thurstone, sl = FALSE)
Alpha:                 0.89
G.6:                   0.91
Omega Hierarchical:    0.74
Omega H asymptotic:    0.79
Omega Total            0.93
Schmid Leiman Factor loadings greater than  0.2
                   g   F1*   F2*   F3*   h2   u2   p2
Sentences       0.71  0.57             0.82 0.18 0.61
Vocabulary      0.73  0.55             0.84 0.16 0.63
Sent.Completion 0.68  0.52             0.73 0.27 0.63
First.Letters   0.65        0.56       0.73 0.27 0.57
4.Letter.Words  0.62        0.49       0.63 0.37 0.61
Suffixes        0.56        0.41       0.50 0.50 0.63
Letter.Series   0.59              0.61 0.72 0.28 0.48
Pedigrees       0.58  0.23        0.34 0.50 0.50 0.66
Letter.Group    0.54              0.46 0.53 0.47 0.56
With eigenvalues of:
   g  F1*  F2*  F3*
3.58 0.96 0.74 0.71
```

Types of reliability

- Internal consistency
  - α
  - $\omega_{hierarchical}$
  - ω_{total}
  - β
- Intraclass
- Agreement
- Test-retest, alternate form
- Generalizability

- Internal consistency
  - `alpha`, `score.items`
  - `omega`
  - `iclust`
- `icc`
- `wkappa`, `cohen.kappa`
- `cor`
- `aov`

Alpha and its alternatives

- Reliability $= \frac{\sigma_t^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$
- If there is another test, then $\sigma_t = \sigma_{t_1 t_2}$ (covariance of test X_1 with test $X_2 = C_{xx}$)
- But, if there is only one test, we can *estimate* σ_t^2 based upon the observed covariances within test 1
- How do we find σ_e^2 ?
- The worst case, (Guttman case 1) all of an item's variance is error and thus the error variance of a test X with variance-covariance C_x
  - $C_x = \sigma_e^2 = diag(C_x)$
  - $\lambda_1 = \frac{C_x - diag(C_x)}{C_x}$
- A better case (Guttman case 3, α) is that that the average covariance between the items on the test is the same as the average true score variance for each item.
  - $C_x = \sigma_e^2 = diag(C_x)$
  - $\lambda_3 = \alpha = \lambda_1 * \frac{n}{n-1} = \frac{(C_x - diag(C_x)) * n/(n-1)}{C_x}$

Guttman 6: estimating using the Squared Multiple Correlation

- Reliability $= \frac{\sigma_t^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$
- Estimate true item variance as squared multiple correlation with other items
- $\lambda_6 = \frac{(C_x - diag(C_x) + \Sigma(smc_i)}{C_x}$
  - This takes observed covariance, subtracts the diagonal, and replaces with the squared multiple correlation
  - Similar to α which replaces with average inter-item covariance
- Squared Multiple Correlation is found by `smc` and is just $smc_i = 1 - 1/R_{ii}^{-1}$

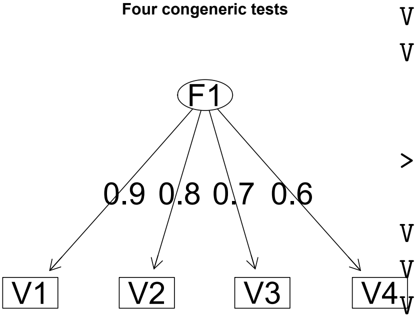
Alpha and its alternatives: Case 1: congeneric measures

First, create some simulated data with a known structure

```
> set.seed(42)
> v4 <- sim.congeneric(N=200,short=FALSE)
> str(v4)  #show the structure of the resulting object
List of 6
 $ model   : num [1:4, 1:4] 1 0.56 0.48 0.4 0.56 1 0.42 0.35 0.48 0.42 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "V1" "V2" "V3" "V4"
  .. ..$ : chr [1:4] "V1" "V2" "V3" "V4"
 $ pattern : num [1:4, 1:5] 0.8 0.7 0.6 0.5 0.6 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "V1" "V2" "V3" "V4"
  .. ..$ : chr [1:5] "theta" "e1" "e2" "e3" ...
 $ r       : num [1:4, 1:4] 1 0.546 0.466 0.341 0.546 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "V1" "V2" "V3" "V4"
  .. ..$ : chr [1:4] "V1" "V2" "V3" "V4"
 $ latent  : num [1:200, 1:5] 1.371 -0.565 0.363 0.633 0.404 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:5] "theta" "e1" "e2" "e3" ...
 $ observed: num [1:200, 1:4] -0.104 -0.251 0.993 1.742 -0.503 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:4] "V1" "V2" "V3" "V4"
 $ N       : num 200
 - attr(*, "class")= chr [1:2] "psych" "sim"
```

A congeneric model

```
> f1 <- fa(v4\$model)
> fa.diagram(f1)
```



```
> v4$model
     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
```

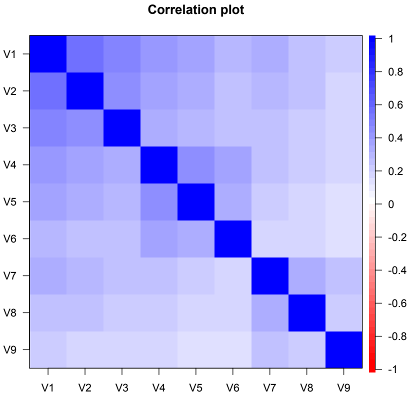
```
> round(cor(v4$observed),2)
     V1   V2   V3   V4
V1 1.00 0.55 0.47 0.34
V2 0.55 1.00 0.38 0.30
V3 0.47 0.38 1.00 0.31
V4 0.34 0.30 0.31 1.00
```

Find α and related stats for the simulated data

```
> alpha(v4$observed)

Reliability analysis
Call: alpha(x = v4$observed)

  raw_alpha std.alpha G6(smc) average_r   mean   sd
      0.71      0.72    0.67      0.39 -0.036 0.72

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r
V1      0.59      0.60    0.50      0.33
V2      0.63      0.64    0.55      0.37
V3      0.65      0.66    0.59      0.40
V4      0.72      0.72    0.64      0.46

 Item statistics
     n    r r.cor r.drop   mean   sd
V1 200 0.80  0.72   0.60 -0.015 0.93
V2 200 0.76  0.64   0.53 -0.060 0.98
V3 200 0.73  0.59   0.50 -0.119 0.92
V4 200 0.66  0.46   0.40  0.049 1.09
```

A hierarchical structure

```
cor.plot(r9)
```

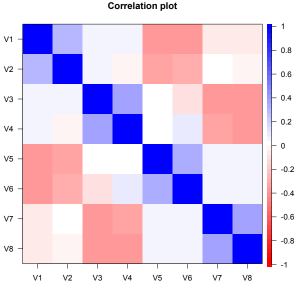


```
> set.seed(42)
> r9 <- sim.hierarchical()
> lower.mat(r9)

   V1   V2   V3   V4   V5   V6   V7   V8   V9
V1 1.00
V2 0.56 1.00
V3 0.48 0.42 1.00
V4 0.40 0.35 0.30 1.00
V5 0.35 0.30 0.26 0.42 1.00
V6 0.29 0.25 0.22 0.35 0.30 1.00
V7 0.30 0.26 0.23 0.24 0.20 0.17 1.00
V8 0.25 0.22 0.19 0.20 0.17 0.14 0.30 1.00
V9 0.20 0.18 0.15 0.16 0.13 0.11 0.24 0.20 1.00
```

α of the 9 hierarchical variables

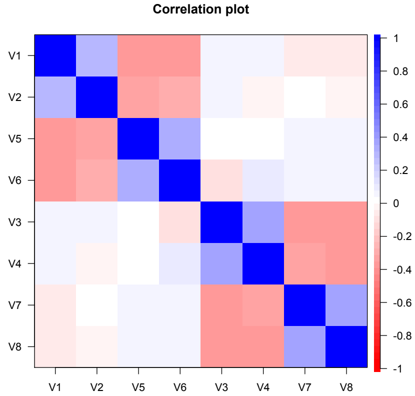
```
> alpha(r9)

Reliability analysis
Call: alpha(x = r9)

  raw_alpha std.alpha G6(smc) average_r
      0.76      0.76    0.76      0.26

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r
V1      0.71      0.71    0.70      0.24
V2      0.72      0.72    0.71      0.25
V3      0.74      0.74    0.73      0.26
V4      0.73      0.73    0.72      0.25
V5      0.74      0.74    0.73      0.26
V6      0.75      0.75    0.74      0.27
V7      0.75      0.75    0.74      0.27
V8      0.76      0.76    0.75      0.28
V9      0.77      0.77    0.76      0.29

 Item statistics
      r r.cor
V1 0.72  0.71
```

An example of two different scales confused as one



```
> set.seed(17)
> two.f <- sim.item(8)
> lower.mat(cor(two.f))
cor.plot(cor(two.f))
   V1    V2    V3    V4    V5    V6    V7    V8
V1  1.00
V2  0.29  1.00
V3  0.05  0.03  1.00
V4  0.03 -0.02  0.34  1.00
V5 -0.38 -0.35 -0.02 -0.01  1.00
V6 -0.38 -0.33 -0.10  0.06  0.33  1.00
V7 -0.06  0.02 -0.40 -0.36  0.03  0.04  1.00
V8 -0.08 -0.04 -0.39 -0.37  0.05  0.03  0.37  1.00
```

Rearrange the items to show it more clearly



```
> cor.2f <- cor(two.f)
> cor.2f  <- cor.2f[c(1:2,5:6,3:4,7:8),
             c(1:2,5:6,3:4,7:8)]
> lower.mat(cor.2f)
>cor.plot(cor.2f)
```

```
     V1    V2    V5    V6    V3    V4    V7    V8
V1  1.00
V2  0.29  1.00
V5 -0.38 -0.35  1.00
V6 -0.38 -0.33  0.33  1.00
V3  0.05  0.03 -0.02 -0.10  1.00
V4  0.03 -0.02 -0.01  0.06  0.34  1.00
V7 -0.06  0.02  0.03  0.04 -0.40 -0.36  1.00
V8 -0.08 -0.04  0.05  0.03 -0.39 -0.37  0.37  1.00
```

α of two scales confused as one

Note the use of the keys parameter to specify how some items should be reversed.

```
> alpha(two.f,keys=c(rep(1,4),rep(-1,4)))

Reliability analysis
Call: alpha(x = two.f, keys = c(rep(1, 4), rep(-1, 4)))

  raw_alpha std.alpha G6(smc) average_r    mean   sd
      0.62      0.62    0.65      0.17 -0.0051 0.27

 Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r
V1      0.59      0.58    0.61      0.17
V2      0.61      0.60    0.63      0.18
V3      0.58      0.58    0.60      0.16
V4      0.60      0.60    0.62      0.18
V5      0.59      0.59    0.61      0.17
V6      0.59      0.59    0.61      0.17
V7      0.58      0.58    0.61      0.17
V8      0.58      0.58    0.60      0.16

 Item statistics
     n    r r.cor r.drop   mean   sd
V1 500 0.54  0.44   0.33  0.063 1.01
V2 500 0.48  0.35   0.26  0.070 0.95
V3 500 0.56  0.47   0.36 -0.030 1.01
V4 500 0.48  0.37   0.28 -0.130 0.97
V5 500 0.52  0.42   0.31 -0.073 0.97
V6 500 0.52  0.41   0.31 -0.071 0.95
V7 500 0.53  0.44   0.34  0.035 1.00
V8 500 0.56  0.47   0.36  0.097 1.02
```

Score as two different scales

First, make up a keys matrix to specify which items should be scored, and in which way

```
> keys <- make.keys(nvars=8,keys.list=list(one=c(1,2,-5,-6),two=c(3,4,-7,-8)))
> keys
     one two
[1,]   1   0
[2,]   1   0
[3,]   0   1
[4,]   0   1
[5,]  -1   0
[6,]  -1   0
[7,]   0  -1
[8,]   0  -1
```

Now score the two scales and find α and other reliability estimates

```
> score.items(keys,two.f)
Call: score.items(keys = keys, items = two.f)
(Unstandardized) Alpha:
       one two
alpha 0.68 0.7
Average item correlation:
           one  two
average.r 0.34 0.37
 Guttman 6* reliability:
          one  two
Lambda.6 0.62 0.64
Scale intercorrelations corrected for attenuation
 raw correlations below the diagonal, alpha on the diagonal
 corrected correlations above the diagonal:
     one  two
one 0.68 0.08
two 0.06 0.70
Item by scale correlations:
 corrected for item overlap and scale reliability
     one   two
V1  0.57  0.09
V2  0.52  0.01
V3  0.09  0.59
V4 -0.02  0.56
V5 -0.58 -0.05
V6 -0.57 -0.05
V7 -0.05 -0.58
V8 -0.09 -0.59
```

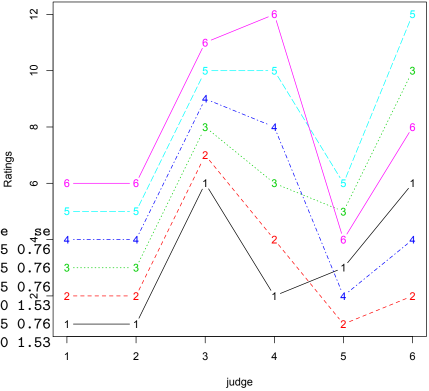
Reliability of judges

- When raters (judges) rate targets, there are multiple sources of variance
  - Between targets
  - Between judges
  - Interaction of judges and targets
- The intraclass correlation is an analysis of variance decomposition of these components
- Different ICC's depending upon what is important to consider
  - Absolute scores: each target gets just one judge, and judges differ
  - Relative scores: each judge rates multiple targets, and the mean for the judge is removed
  - Each judge rates multiple targets, judge and target effects removed

Ratings of judges

What is the reliability of ratings of different judges across ratees? It depends. Depends upon the pairing of judges, depends upon the targets. ICC does an Anova decomposition.

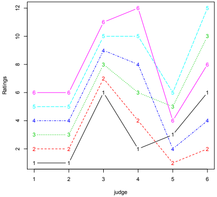
```
> Ratings
  J1 J2 J3 J4 J5 J6
1  1  1  6  2  3  6
2  2  2  7  4  1  2
3  3  3  8  6  5 10
4  4  4  9  8  2  4
5  5  5 10 10  6 12
6  6  6 11 12  4  8

> describe(Ratings,skew=FALSE)

   var n mean   sd median trimmed  mad min max range   se
J1   1 6  3.5 1.87    3.5     3.5 2.22   1   6     5 0.76
J2   2 6  3.5 1.87    3.5     3.5 2.22   1   6     5 0.76
J3   3 6  8.5 1.87    8.5     8.5 2.22   6  11     5 0.76
J4   4 6  7.0 3.74    7.0     7.0 4.45   2  12    10 1.53
J5   5 6  3.5 1.87    3.5     3.5 2.22   1   6     5 0.76
J6   6 6  7.0 3.74    7.0     7.0 4.45   2  12    10 1.53
```



Sources of variances and the Intraclass Correlation Coefficient

Table : Sources of variances and the Intraclass Correlation Coefficient.

	(J1, J2)	(J3, J4)	(J5, J6)	(J1, J3)	(J1, J5)	(J1 ... J3)	(J1 ... J4)	(J1 ...
Variance estimates								
MS_b	7	15.75	15.75	7.0	5.2	10.50	21.88	2
MS_w	0	2.58	7.58	12.5	1.5	8.33	7.12	
MS_j	0	6.75	36.75	75.0	0.0	50.00	38.38	3
MS_e	0	1.75	1.75	0.0	1.8	0.00	.88	
Intraclass correlations								
ICC(1,1)	1.00	.72	.35	-.28	.55	.08	.34	
ICC(2,1)	1.00	.73	.48	.22	.53	.30	.42	
ICC(3,1)	1.00	.80	.80	1.00	.49	1.00	.86	
ICC(1,k)	1.00	.84	.52	-.79	.71	.21	.67	
ICC(2,k)	1.00	.85	.65	.36	.69	.56	.75	
ICC(3,k)	1.00	.89	.89	1.00	.65	1.00	.96	



ICC is done by calling anova

```
aov.x <- aov(values ~ subs + ind, data = x.df)
   s.aov <- summary(aov.x)
   stats <- matrix(unlist(s.aov), ncol = 3, byrow = TRUE)
   MSB <- stats[3, 1]
   MSW <- (stats[2, 2] + stats[2, 3])/(stats[1, 2] + stats[1,
       3])
   MSJ <- stats[3, 2]
   MSE <- stats[3, 3]
   ICC1 <- (MSB - MSW)/(MSB + (nj - 1) * MSW)
   ICC2 <- (MSB - MSE)/(MSB + (nj - 1) * MSE + nj * (MSJ - MSE)/n.obs)
   ICC3 <- (MSB - MSE)/(MSB + (nj - 1) * MSE)
   ICC12 <- (MSB - MSW)/(MSB)
   ICC22 <- (MSB - MSE)/(MSB + (MSJ - MSE)/n.obs)
   ICC32 <- (MSB - MSE)/MSB
```

Intraclass Correlations using the ICC function

```
> print(ICC(Ratings),all=TRUE)   #get more output than normal
$results
                         type  ICC     F df1 df2    p lower bound upper bound
Single_raters_absolute   ICC1 0.32  3.84   5  30 0.01        0.04        0.79
Single_random_raters     ICC2 0.37 10.37   5  25 0.00        0.09        0.80
Single_fixed_raters      ICC3 0.61 10.37   5  25 0.00        0.28        0.91
Average_raters_absolute ICC1k 0.74  3.84   5  30 0.01        0.21        0.96
Average_random_raters   ICC2k 0.78 10.37   5  25 0.00        0.38        0.96
Average_fixed_raters    ICC3k 0.90 10.37   5  25 0.00        0.70        0.98

$summary
            Df  Sum Sq Mean Sq F value    Pr(>F)
subs         5 141.667 28.3333  10.366 1.801e-05 ***
ind          5 153.000 30.6000  11.195 9.644e-06 ***
Residuals   25  68.333  2.7333
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

$stats
             [,1]         [,2]      [,3]
[1,] 5.000000e+00 5.000000e+00 25.000000
[2,] 1.416667e+02 1.530000e+02 68.333333
[3,] 2.833333e+01 3.060000e+01  2.733333
[4,] 1.036585e+01 1.119512e+01        NA
[5,] 1.800581e-05 9.644359e-06        NA

$MSW
[1] 7.377778

$Call
ICC(x = Ratings)
```

Cohen's kappa and weighted kappa

- When considering agreement in diagnostic categories, without numerical values, it is useful to consider the kappa coefficient.
  - Emphasizes matches of ratings
  - Doesn't consider how far off disagreements are.
- Weighted kappa weights the off diagonal distance.
- Diagnostic categories: normal, neurotic, psychotic

Cohen kappa and weighted kappa

```
> cohen
     [,1] [,2] [,3]
[1,] 0.44 0.07 0.09
[2,] 0.05 0.20 0.05
[3,] 0.01 0.03 0.06
> cohen.weights
     [,1] [,2] [,3]
[1,]    0    1    3
[2,]    1    0    6
[3,]    3    6    0
> cohen.kappa(cohen,cohen.weights)
Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha)

Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundari
                  lower estimate upper
unweighted kappa  -0.92     0.49   1.9
weighted kappa   -10.04     0.35  10.7
```

see the other examples in ?cohen.kappa

Outline of Part II: the New Psychometrics

7. Two approaches
8. Various IRT models
9. Polytomous items
   - Ordered response categories
   - Differential Item Functioning
10. Factor analysis & IRT
    - Non-monotone Trace lines
11. (C) A T

Classical Reliability

1. Classical model of reliability
   - Observed = True + Error
   - Reliability $= 1 - \frac{\sigma^2_{error}}{\sigma^2_{observed}}$
   - Reliability $= r_{xx} = r^2_{x_{domain}}$
   - Reliability as correlation of a test with a test just like it
2. Reliability requires variance in observed score
   - As σ^2_x decreases so will $r_{xx} = 1 - \frac{\sigma^2_{error}}{\sigma^2_{observed}}$
3. Alternate estimates of reliability all share this need for variance
   1. Internal Consistency
   2. Alternate Form
   3. Test-retest
   4. Between rater
4. Item difficulty is ignored, items assumed to be sampled at random

The “new psychometrics”

1. Model the person as well as the item
   - People differ in some latent score
   - Items differ in difficulty and discriminability
2. Original model is a model of ability tests
   - $p(correct|ability, difficulty, ...) = f(ability - difficulty)$
   - What is the appropriate function?
3. Extensions to polytomous items, particularly rating scale models

Classic Test Theory as 0 parameter IRT

Classic Test Theory considers all items to be random replicates of each other and total (or average) score to be the appropriate measure of the underlying attribute. Items are thought to be endorsed (passed) with an increasing probability as a function of the underlying trait. But if the trait is unbounded (just as there is always the possibility of someone being higher than the highest observed score, so is there a chance of someone being lower than the lowest observed score), and the score is bounded (from p=0 to p=1), then the relationship between the latent score and the observed score must be non-linear. This leads to the most simple of all models, one that has no parameters to estimate but is just a non-linear mapping of latent to observed:

$$p(correct_{ij}|\theta_i) = \frac{1}{1+e^{-\theta_i}}. \tag{30}$$

Rasch model – All items equally discriminating, differ in difficulty

Slightly more complicated than the zero parameter model is to assume that all items are equally good measures of the trait, but differ only in their difficulty/location. The *one parameter logistic* (*1PL*) *Rasch model* (Rasch, 1960) is the easiest to understand:

$$p(correct_{ij}|\theta_i, \delta_j) = \frac{1}{1 + e^{\delta_j - \theta_i}}. \tag{31}$$

That is, the probability of the i^{th} person being correct on (or endorsing) the j^{th} item is a logistic function of the difference between the person's ability (latent trait) (θ_i) and the item difficulty (or location) (δ_j). The more the person's ability is greater than the item difficulty, the more likely the person is to get the item correct.

Estimating the model

The probability of missing an item, q, is just 1 - p(correct) and thus the *odds ratio* of being correct for a person with ability, θ_i, on an item with difficulty, δ_j is

$$OR_{ij} = \frac{p}{1-p} = \frac{p}{q} = \frac{\frac{1}{1+e^{\delta_j-\theta_i}}}{1-\frac{1}{1+e^{\delta_j-\theta_i}}} = \frac{\frac{1}{1+e^{\delta_j-\theta_i}}}{\frac{e^{\delta_j-\theta_i}}{1+e^{\delta_j-\theta_i}}} = \frac{1}{e^{\delta_j-\theta_i}} = e^{\theta_i-\delta_j}. \tag{32}$$

That is, the odds ratio will be a exponential function of the difference between a person's ability and the task difficulty. The odds of a particular pattern of rights and wrongs over n items will be the product of n odds ratios

$$OR_{i1}OR_{i2}\dots OR_{in} = \prod_{j=1}^{n} e^{\theta_i-\delta_j} = e^{n\theta_i}e^{-\sum_{j=1}^{n}\delta_j}. \tag{33}$$

Estimating parameters

Substituting P for the pattern of correct responses and Q for the pattern of incorrect responses, and taking the logarithm of both sides of equation 33 leads to a much simpler form:

$$ln\frac{P}{Q} = n\theta_i + \sum_{j=1}^{n}\delta_j = n(\theta_i + \bar{\delta}). \tag{34}$$

That is, the log of the pattern of correct/incorrect for the i^{th} individual is a function of the number of items * (θ_i - the average difficulty). Specifying the average difficulty of an item as $\bar{\delta} = 0$ to set the scale, then θ_i is just the logarithm of P/Q divided by n or, conceptually, the average logarithm of the p/q

$$\theta_i = \frac{ln\frac{P}{Q}}{n}. \tag{35}$$

Difficulty is just a function of probability correct

Similarly, the pattern of the odds of correct and incorrect responses across people for a particular item with difficulty δ_j will be

$$OR_{1j}OR_{2j}\ldots OR_{nj} = \frac{P}{Q} = \prod_{i=1}^{N} e^{\theta_i - \delta_j} = e^{\sum_{i=1}^{N}(\theta_i) - N\delta_j} \tag{36}$$

and taking logs of both sides leads to

$$ln\frac{P}{Q} = \sum_{i=1}^{N}(\theta_i) - N\delta_j. \tag{37}$$

Letting the average ability $\bar{\theta} = 0$ leads to the conclusion that the difficulty of an item for all subjects, δ_j, is the logarithm of Q/P divided by the number of subjects, N,

$$\delta_j = \frac{ln\frac{Q}{P}}{N}. \tag{38}$$

Rasch model in words

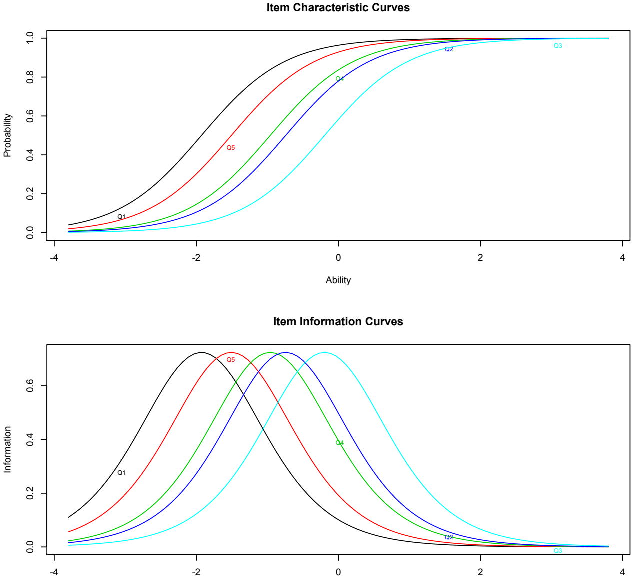
That is, the estimate of ability (Equation 35) for items with an average difficulty of 0 does not require knowing the difficulty of any particular item, but is just a function of the pattern of corrects and incorrects for a subject across all items.

Similarly, the estimate of item difficulty across people ranging in ability, but with an average ability of 0 (Equation 38) is a function of the response pattern of all the subjects on that one item and does not depend upon knowing any one person's ability. The assumptions that average difficulty and average ability are 0 are merely to fix the scales. Replacing the average values with a non-zero value just adds a constant to the estimates.

Rasch as a high jump

The independence of ability from difficulty implied in equations 35 and 38 makes estimation of both values very straightforward. These two equations also have the important implication that the number correct ($n\bar{p}$ for a subject, $N\bar{p}$ for an item) is monotonically, but not linearly related to ability or to difficulty. That the estimated ability is independent of the pattern of rights and wrongs but just depends upon the total number correct is seen as both a strength and a weakness of the Rasch model. From the perspective of *fundamental measurement*, Rasch scoring provides an additive interval scale: for all people and items, if $\theta_i < \theta_j$ and $\delta_k < \delta_l$ then $p(x|\theta_i, \delta_k) < p(x|\theta_j, \delta_l)$. But this very additivity treats all patterns of scores with the same number correct as equal and ignores potential information in the pattern of responses.

Rasch estimates from ltm



The LSAT example from ltm

```
data(bock)
> ord <- order(colMeans(lsat6),decreasing=TRUE)
> lsat6.sorted <- lsat6[,ord]
> describe(lsat6.sorted)
> Tau <- round(-qnorm(colMeans(lsat6.sorted)),2)  #tau = estimates of threshold
> rasch(lsat6.sorted,constraint=cbind(ncol(lsat6.sorted)+1,1.702))

   var    n mean   sd median trimmed mad min max range  skew kurtosis   se
Q1   1 1000 0.92 0.27      1    1.00   0   0   1     1 -3.20     8.22 0.01
Q5   2 1000 0.87 0.34      1    0.96   0   0   1     1 -2.20     2.83 0.01
Q4   3 1000 0.76 0.43      1    0.83   0   0   1     1 -1.24    -0.48 0.01
Q2   4 1000 0.71 0.45      1    0.76   0   0   1     1 -0.92    -1.16 0.01
Q3   5 1000 0.55 0.50      1    0.57   0   0   1     1 -0.21    -1.96 0.02

> Tau
   Q1    Q5    Q4    Q2    Q3
-1.43 -1.13 -0.72 -0.55 -0.13

Call:
rasch(data = lsat6.sorted, constraint = cbind(ncol(lsat6.sorted) +
    1, 1.702))

Coefficients:
Dffclt.Q1  Dffclt.Q5  Dffclt.Q4  Dffclt.Q2  Dffclt.Q3     Dscrmn
   -1.927     -1.507     -0.960     -0.742     -0.195      1.702
```

Item information

When forming a test and evaluating the items within a test, the most useful items are the ones that give the most information about a person's score. In classic test theory, *item information* is the reciprocal of the squared *standard error* for the item or for a one factor test, the ratio of the item communality to its uniqueness:

$$I_j = \frac{1}{\sigma^2_{e_j}} = \frac{h^2_j}{1 - h^2_j}.$$

When estimating ability using IRT, the information for an item is a function of the first derivative of the likelihood function and is maximized at the inflection point of the *icc*.

Estimating item information

The information function for an item is

$$I(f, x_j) = \frac{[P_j'(f)]^2}{P_j(f)Q_j(f)} \tag{39}$$

For the 1PL model, P', the first derivative of the probability function $P_j(f) = \frac{1}{1+e^{\delta-\theta}}$ is

$$P' = \frac{e^{\delta-\theta}}{(1+e^{\delta-\theta})^2} \tag{40}$$

which is just $P_j Q_j$ and thus the information for an item is

$$I_j = P_j Q_j. \tag{41}$$

That is, information is maximized when the probability of getting an item correct is the same as getting it wrong, or, in other words, the best estimate for an item's difficulty is that value where half of the subjects pass the item.

Elaborations of Rasch

1. Logistic or cumulative normal function
    - Logistic treats any pattern of responses the same
    - Cumulative normal weights extreme scores more
2. Rasch and 1PN models treat all items as equally discriminating
    - But some items are better than others
    - Thus, the two parameter model

$$p(correct_{ij}|\theta_i, \alpha_j, \delta_j) = \frac{1}{1 + e^{\alpha_i(\delta_j - \theta_i)}} \tag{42}$$

2PL and 2PN models

$$p(correct_{ij}|\theta_i, \alpha_j, \delta_j) = \frac{1}{1 + e^{\alpha_i(\delta_j - \theta_i)}} \tag{43}$$

while in the *two parameter normal* ogive (*2PN*) model this is

$$p(correct|\theta, \alpha_j, \delta) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{\alpha(\theta-\delta)} e^{-\frac{u^2}{2}} du \tag{44}$$

where $u = \alpha(\theta - \delta)$.

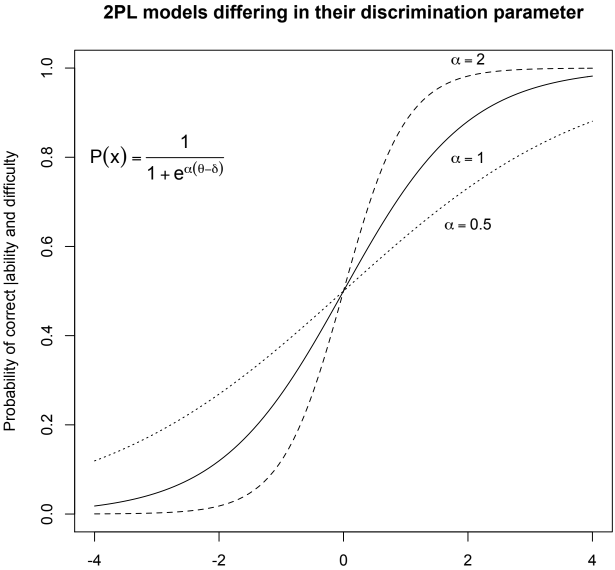
The information function for a two parameter model reflects the item discrimination parameter, α,

$$I_j = \alpha^2 P_j Q_j \tag{45}$$

which, for a 2PL model is

$$I_j = \alpha_j^2 P_j Q_j = \frac{\alpha_j^2}{(1 + e^{\alpha_j(\delta_j - \theta_j)})^2}. \tag{46}$$

The problem of non-parallel trace lines



Parameter explosion – better fit but at what cost

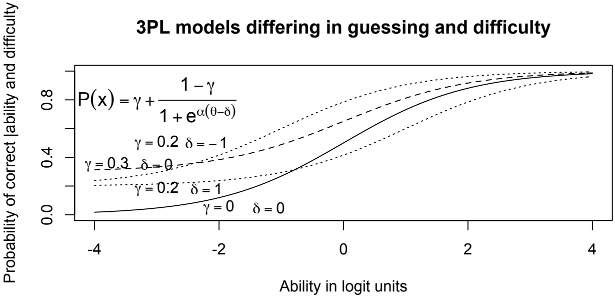
The 3 parameter model adds a guessing parameter.

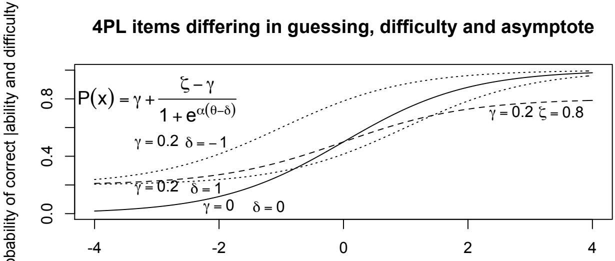
$$p(correct_{ij}|\theta_i, \alpha_j, \delta_j, \gamma_j) = \gamma_j + \frac{1-\gamma_j}{1+e^{\alpha_i(\delta_j-\theta_i)}} \tag{47}$$

And the four parameter model adds an asymtotic parameter

$$P(x|\theta_i, \alpha, \delta_j, \gamma_j, \zeta_j) = \gamma_j + \frac{\zeta_j-\gamma_j}{1+e^{\alpha_j(\delta_j-\theta_i)}}. \tag{48}$$

frame


3PL models differing in guessing and difficulty
Probability of correct |ability and difficulty
0.0
0.4
0.8
$P(x) = \gamma + \frac{1-\gamma}{1+e^{\alpha(\theta-\delta)}}$
γ = 0.2 δ = −1
γ = 0.3 δ = 0
γ = 0.2 δ = 1
γ = 0 δ = 0
-4
-2
0
2
4
Ability in logit units


4PL items differing in guessing, difficulty and asymptote
Probability of correct |ability and difficulty
0.0
0.4
0.8
$P(x) = \gamma + \frac{\zeta-\gamma}{1+e^{\alpha(\theta-\delta)}}$
γ = 0.2 δ = −1
γ = 0.2 ζ = 0.8
γ = 0.2 δ = 1
γ = 0 δ = 0
-4
-2
0
2
4

Personality items with monotone trace lines

A typical personality item might ask "How much do you enjoy a lively party" with a five point response scale ranging from "1: not at all" to "5: a great deal" with a neutral category at 3. An alternative response scale for this kind of item is to not have a neutral category but rather have an even number of responses. Thus a six point scale could range from "1: very inaccurate" to "6: very accurate" with no neutral category

The assumption is that the more sociable one is, the higher the response alternative chosen. The probability of endorsing a 1 will increase monotonically the less sociable one is, the probability of endorsing a 5 will increase monotonically the more sociable one is.

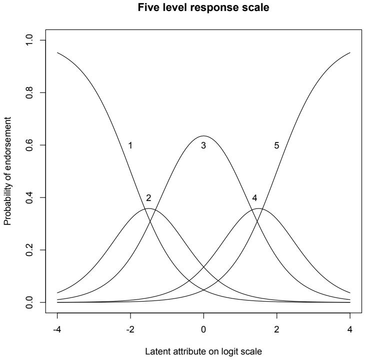
Threshold models

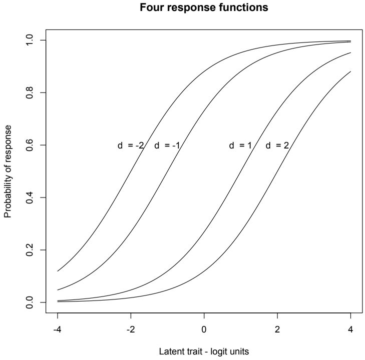
For the 1PL or 2PL logistic model the probability of endorsing the k^{th} response is a function of ability, item thresholds, and the discrimination parameter and is

$$P(r=k|\theta_i,\delta_k,\delta_{k-1},\alpha_k)=P(r|\theta_i,\delta_{k-1},\alpha_k)-P(r|\theta_i,\delta_k,\alpha_k)=\frac{1}{1+e^{\alpha_k(\delta_{k-1}-\theta_i)}}-\frac{1}{1+e^{\alpha s_k(\delta_k-\theta_i)}} \tag{49}$$

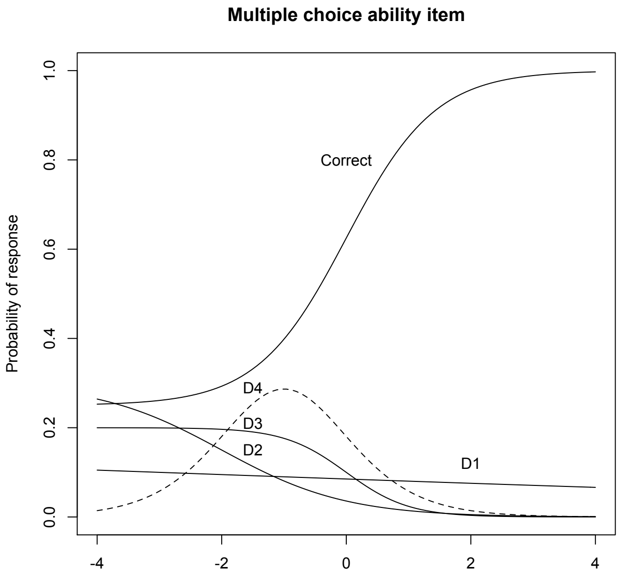
where all b_k are set to $b_k=1$ in the 1PL Rasch case.

Responses to a multiple choice polytomous item



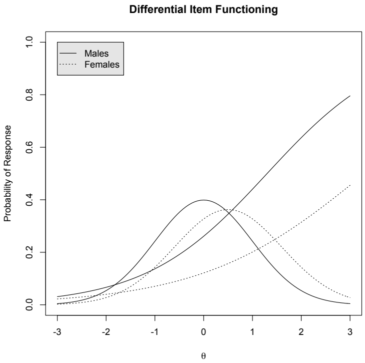


Differences in the response shape of mulitple choice items



Differential Item Functioning

1. Use of IRT to analyze item quality
   - Find IRT difficulty and discrimination parameters for different groups
   - Compare response patterns



FA and IRT

If the correlations of all of the items reflect one underlying latent variable, then factor analysis of the matrix of tetrachoric correlations should allow for the identification of the regression slopes (α) of the items on the latent variable. These regressions are, of course just the factor loadings. Item difficulty, δ_j and item discrimination, α_j may be found from factor analysis of the tetrachoric correlations where λ_j is just the factor loading on the first factor and τ_j is the normal threshold reported by the `tetrachoric` function (McDonald, 1999; Lord & Novick, 1968; Takane & de Leeuw, 1987).

$$\delta_j = \frac{D\tau}{\sqrt{1-\lambda_j^2}}, \qquad \alpha_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}} \tag{50}$$

where D is a scaling factor used when converting to the parameterization of *logistic* model and is 1.702 in that case and 1 in the case of the normal ogive model.

FA and IRT

IRT parameters from FA

$$\delta_j = \frac{D\tau}{\sqrt{1-\lambda_j^2}}, \qquad \alpha_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}} \tag{51}$$

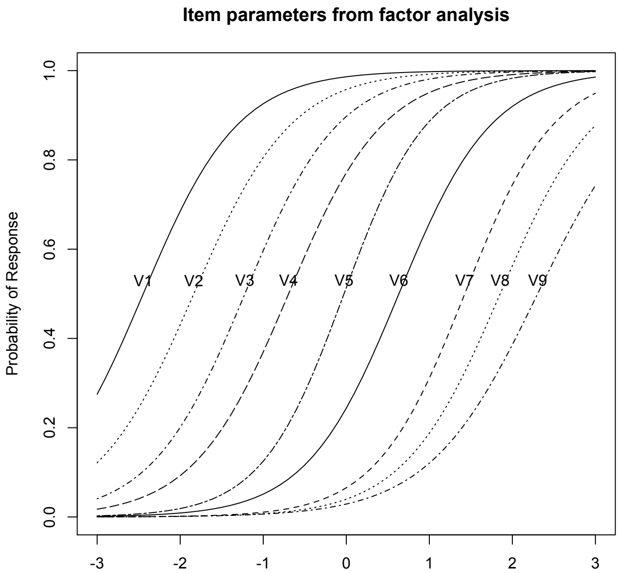
FA parameters from IRT

$$\lambda_j = \frac{\alpha_j}{\sqrt{1+\alpha_j^2}}, \qquad \tau_j = \frac{\delta_j}{\sqrt{1+\alpha_j^2}}.$$

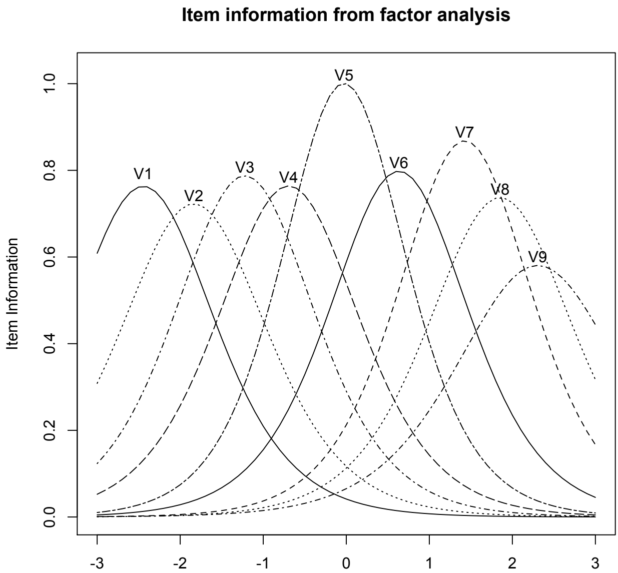
the irt.fa function

```
> set.seed(17)
> items <- sim.npn(9,1000,low=-2.5,high=2.5)$items
> p.fa <-irt.fa(items)

Summary information by factor and item
 Factor =  1
              -3   -2   -1    0    1    2     3
V1          0.61 0.66 0.21 0.04 0.01 0.00  0.00
V2          0.31 0.71 0.45 0.12 0.02 0.00  0.00
V3          0.12 0.51 0.76 0.29 0.06 0.01  0.00
V4          0.05 0.26 0.71 0.54 0.14 0.03  0.00
V5          0.01 0.07 0.44 1.00 0.40 0.07  0.01
V6          0.00 0.03 0.16 0.59 0.72 0.24  0.05
V7          0.00 0.01 0.04 0.21 0.74 0.66  0.17
V8          0.00 0.00 0.02 0.11 0.45 0.73  0.32
V9          0.00 0.00 0.01 0.07 0.25 0.55  0.44
Test Info   1.11 2.25 2.80 2.97 2.79 2.28  0.99
SEM         0.95 0.67 0.60 0.58 0.60 0.66  1.01
Reliability 0.10 0.55 0.64 0.66 0.64 0.56 -0.01
```

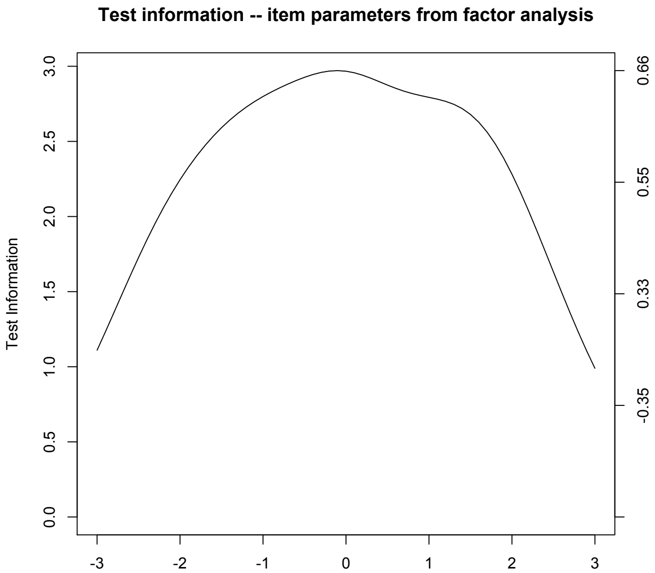
Item Characteristic Curves from FA



Item information from FA



Test Information Curve



Comparing three ways of estimating the parameters

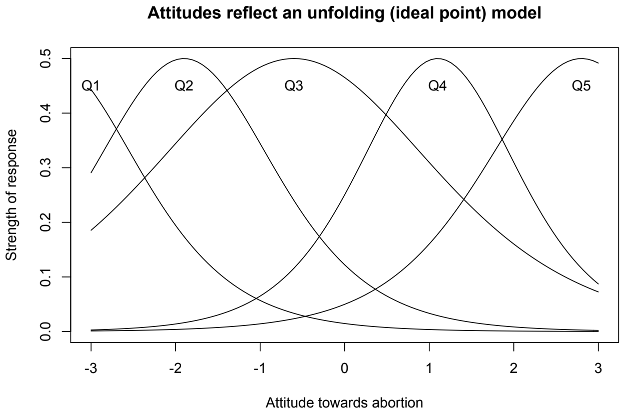
```
set.seed(17)
items <- sim.npn(9,1000,low=-2.5,high=2.5)$items
p.fa <- irt.fa(items)$coefficients[1:2]
p.ltm <- ltm(items~z1)$coefficients
p.ra <- rasch(items, constraint = cbind(ncol(items) + 1, 1))$coefficients
a <- seq(-2.5,2.5,5/8)
p.df <- data.frame(a,p.fa,p.ltm,p.ra)
round(p.df,2)
```

```
           a Difficulty Discrimination X.Intercept.   z1 beta.i beta
Item 1 -2.50      -2.45           1.03         5.42 2.61   3.64    1
Item 2 -1.88      -1.84           1.00         3.35 1.88   2.70    1
Item 3 -1.25      -1.22           1.04         2.09 1.77   1.73    1
Item 4 -0.62      -0.69           1.03         1.17 1.71   0.98    1
Item 5  0.00      -0.03           1.18         0.04 1.94   0.03    1
Item 6  0.62       0.63           1.05        -1.05 1.68  -0.88    1
Item 7  1.25       1.43           1.10        -2.47 1.90  -1.97    1
Item 8  1.88       1.85           1.01        -3.75 2.27  -2.71    1
Item 9  2.50       2.31           0.90        -5.03 2.31  -3.66    1
```

Attitudes might not have monotone trace lines

1. *Abortion is unacceptable under any circumstances.*
2. *Even if one believes that there may be some exceptions, abortions is still generally wrong.*
3. *There are some clear situations where abortion should be legal, but it should not be permitted in all situations.*
4. *Although abortion on demand seems quite extreme, I generally favor a woman's right to choose.*
5. *Abortion should be legal under any circumstances.*

Ideal point models of attitutude



IRT and CTT don't really differ except

1. Correlation of classic test scores and IRT scores $> .98$.
2. Test information for the person doesnt't require people to vary
3. Possible to item bank with IRT
    - Make up tests with parallel items based upon difficulty and discrimination
    - Detect poor items
4. Adaptive testing
    - No need to give a person an item that they will almost certainly pass (or fail)
    - Can tailor the test to the person
    - (Problem with anxiety and item failure)

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. The Addison-Wesley series in behavioral science: quantitative methods. Reading, Mass.: Addison-Wesley Pub. Co.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: reprinted in 1980 by The University of Chicago Press /Paedagogike Institut, Copenhagen.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. 10.1007/BF02294363.