Psychology 371: 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 $\alpha$

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

\[ X \]

\[ X_1 \]
\[ X_2 \]
\[ X_3 \]
\[ X_4 \]
\[ X_5 \]
\[ X_6 \]

\[ Y \]

\[ Y_1 \]
\[ Y_2 \]
\[ Y_3 \]
\[ Y_4 \]
\[ Y_5 \]
\[ Y_6 \]
Latent Variables

$\xi_1$  $\eta_1$

$\xi_2$  $\eta_2$
Theory: A regression model of latent variables

\[ \xi \rightarrow \eta \]

\[ \xi_1 \rightarrow \eta_1 \]

\[ \xi_2 \rightarrow \eta_2 \]

\[ \xi_1 \rightarrow \xi_2 \]

\[ \eta_1 \rightarrow \eta_2 \]

\[ \zeta_1 \rightarrow \zeta_2 \]

\[ \mu \]

\[ \nu \]

\[ \psi \]

\[ \lambda \]

\[ \theta \]

\[ \alpha \]

\[ \beta \]

\[ \gamma \]

\[ \delta \]

\[ \epsilon \]

\[ \zeta \]

\[ \eta \]

\[ \xi \]

\[ \sigma \]

\[ \tau \]

\[ \omega \]

\[ \chi \]

\[ \psi \]

\[ \varphi \]

\[ \delta \]

\[ \epsilon \]

\[ \theta \]

\[ \phi \]

\[ \chi \]

\[ \psi \]

\[ \varphi \]

\[ \delta \]

\[ \epsilon \]

\[ \theta \]

\[ \phi \]

\[ \chi \]

\[ \psi \]

\[ \varphi \]

\[ \delta \]

\[ \epsilon \]

\[ \theta \]

\[ \phi \]

\[ \chi \]

\[ \psi \]

\[ \varphi \]

\[ \delta \]

\[ \epsilon \]

\[ \theta \]

\[ \phi \]

\[ \chi \]

\[ \psi \]

\[ \varphi \]

\[ \delta \]

\[ \epsilon \]

\[ \theta \]

\[ \phi \]

\[ \chi \]

\[ \psi \]

\[ \varphi \]

\[ \delta \]

\[ \epsilon \]

\[ \theta \]

\[ \phi \]

\[ \chi \]

\[ \psi \]

\[ \varphi \]
A measurement model for $X$ – Correlated factors

\[ \delta \rightarrow X \rightarrow \xi \]

\[ \delta_1 \rightarrow X_1 \]
\[ \delta_2 \rightarrow X_2 \rightarrow \xi_1 \]
\[ \delta_3 \rightarrow X_3 \rightarrow \xi_1 \]
\[ \delta_4 \rightarrow X_4 \rightarrow \xi_2 \]
\[ \delta_5 \rightarrow X_5 \rightarrow \xi_2 \]
\[ \delta_6 \rightarrow X_6 \]
A measurement model for Y - uncorrelated factors

\[ \eta \rightarrow Y_1, Y_2, Y_3, Y_4, Y_5, Y_6 \rightarrow \epsilon \]
A complete structural model

\[ \delta \] \quad \[ X \] \quad \[ \xi \] \quad \[ \eta \] \quad \[ Y \] \quad \[ \epsilon \]

- \( \delta_1 \) \rightarrow \( X_1 \)
- \( \delta_2 \) \rightarrow \( X_2 \)
- \( \delta_3 \) \rightarrow \( X_3 \)
- \( \delta_4 \) \rightarrow \( X_4 \)
- \( \delta_5 \) \rightarrow \( X_5 \)
- \( \delta_6 \) \rightarrow \( X_6 \)

\( \xi_1 \) \quad \( \eta_1 \) \quad \( \zeta_1 \) \quad \( Y_1 \)

\( \xi_2 \) \quad \( \eta_2 \) \quad \( \zeta_2 \) \quad \( Y_2 \)

\( \epsilon_1 \) \quad \( \epsilon_2 \) \quad \( \epsilon_3 \) \quad \( \epsilon_4 \)

\( \epsilon_5 \) \quad \( \epsilon_6 \)
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

Reliability = .80

Reliability = .50
Spearman’s parallell test theory

A

B
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 ($\sigma_{te}$) is zero, and the variance of $x$, $\sigma_x^2$, 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

$$\rho_{xt} = \frac{\sigma_{xt}}{\sqrt{(\sigma_t^2 + \sigma_e^2)(\sigma_t^2)}} = \frac{\sigma_t^2}{\sqrt{\sigma_x^2 \sigma_t^2}} = \frac{\sigma_t}{\sigma_x}. \quad (1)$$
By knowing the correlation between observed score and true score, $\rho_{xt}$, and from the definition of linear regression predicted true score, $\hat{t}$, for an observed $x$ may be found from

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

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

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

$$\rho_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \rho_{xt}^2. \quad (4)$$

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.

$$r_{pq} = \frac{\sigma_{pq}}{\sqrt{\sigma_p^2 \sigma_q^2}} \quad (5)$$

and

$$r_{p'q'} = \frac{\sigma_{p'q'}}{\sqrt{\sigma_{p'}^2 \sigma_{q'}^2}} = \frac{\sigma_{p+e'_1 \sigma q+e'_2}}{\sqrt{\sigma_{p'}^2 \sigma_{q'}^2}} = \frac{\sigma_{pq}}{\sqrt{\sigma_{p'}^2 \sigma_{q'}^2}} \quad (6)$$
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 ($\sigma^2_{p1} = \sigma^2_{p2} = \sigma^2_{p'}$), and the error variances across subjects for each test are the same ($\sigma^2_{e1} = \sigma^2_{e2} = \sigma^2_{e'}$) (see Figure 19), (Lord & Novick, 1968; McDonald, 1999). The correlation between two parallel tests will be

$$\rho_{p1p2}' = \rho_{p'p'} = \frac{\sigma_{p1p2}'}{\sqrt{\sigma^2_{p1}' \sigma^2_{p2}'}} = \frac{\sigma^2_p + \sigma_{pe1} + \sigma_{pe2} + \sigma_{e1e2}}{\sigma^2_{p'}} = \frac{\sigma^2_p}{\sigma^2_{p'}}. \quad (7)$$
but from Eq 4,

\[ \sigma_p^2 = \rho_{p'p'} \sigma_{p'}^2 \]  \hspace{1cm} (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

\[ r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} r_{q'q'}}} \]  \hspace{1cm} (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, $r_{p'_1p'_2}$, represents the two known variances $s^2_{p'_1}$ and $s^2_{p'_2}$ and their covariance $s_{p'_1p'_2}$. 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.
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

Four congeneric tests

\[ F1 \]

\[ 0.9 \quad 0.8 \quad 0.7 \quad 0.6 \]

\[ V1 \quad V2 \quad V3 \quad V4 \]
### Observed variables and estimated parameters of a congeneric test

<table>
<thead>
<tr>
<th>Variable</th>
<th>Test₁</th>
<th>Test₂</th>
<th>Test₃</th>
<th>Test₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test₁</td>
<td>( \sigma_{x_1}^2 = \lambda_1 \sigma_\theta^2 + \epsilon_1^2 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test₂</td>
<td>( \sigma_{x_1x_2} = \lambda_1 \sigma_\theta \lambda_2 \sigma_\theta )</td>
<td>( \sigma_{x_2}^2 = \lambda_2 \sigma_\theta^2 + \epsilon_2^2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test₃</td>
<td>( \sigma_{x_1x_3} = \lambda_1 \sigma_\theta \lambda_3 \sigma_\theta )</td>
<td>( \sigma_{x_2x_3} = \lambda_2 \sigma_\theta \lambda_3 \sigma_\theta )</td>
<td>( \sigma_{x_3}^2 = \lambda_3 \sigma_\theta^2 + \epsilon_3^2 )</td>
<td></td>
</tr>
<tr>
<td>Test₄</td>
<td>( \sigma_{x_1x_4} = \lambda_1 \sigma_\theta \lambda_4 \sigma_\theta )</td>
<td>( \sigma_{x_2x_4} = \lambda_2 \sigma_\theta \lambda_4 \sigma_\theta )</td>
<td>( \sigma_{x_3x_4} = \lambda_3 \sigma_\theta \lambda_4 \sigma_\theta )</td>
<td>( \sigma_{x_4}^2 = \lambda_4 \sigma_\theta^2 + \epsilon_4^2 )</td>
</tr>
</tbody>
</table>
## Observed variables and estimated parameters of a congeneric test

<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1^2$</td>
<td>$s_2^2$</td>
<td>$s_3^2$</td>
<td>$s_4^2$</td>
<td>$\lambda_1 \sigma_t^2 + \sigma_{e1}^2$</td>
<td>$\lambda_1 \lambda_2 \sigma_t^2$</td>
<td>$\lambda_2 \sigma_t^2 + \sigma_{e2}^2$</td>
<td>$\lambda_3 \sigma_t^2 + \sigma_{e3}^2$</td>
</tr>
<tr>
<td>$s_{12}$</td>
<td>$s_{23}$</td>
<td>$s_{34}$</td>
<td></td>
<td>$\lambda_1 \sigma_t^2$</td>
<td>$\lambda_2 \lambda_3 \sigma_t^2$</td>
<td>$\lambda_3 \lambda_4 \sigma_t^2$</td>
<td>$\lambda_4 \sigma_t^2$</td>
</tr>
</tbody>
</table>
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

\[ \Sigma_{XX'} = \begin{pmatrix} V_x & C_{xx'} \\ \vdots & \ddots & \ddots \\ C_{xx'} & \vdots & V_{x'} \end{pmatrix} \]  \hspace{1cm} (10)

and letting \( V_x = 1V_x1' \) and \( C_{XX'} = 1C_{XX'}1' \) 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:

\[ V_x = 1V_x1' = 1C_t1' + 1V_e1' = V_t + V_e \]
Split halves

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

\[
\Sigma_{XX'} = \begin{pmatrix}
V_X = V_t + V_e & : & C_{xx'} = V_t \\
dots & & \\
V_t = C_{xx'} & : & V_{t'} + V_{e'} = V_{X'}
\end{pmatrix}
\]

and because \( V_t = V_{t'} \) and \( V_e = V_{e'} \) the correlation between each half, (their reliability) is

\[
\rho = \frac{C_{xx'}}{V_X} = \frac{V_t}{V_X} = 1 - \frac{V_e}{V_t}.
\]
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:

$$\sum_{XX'} = \begin{pmatrix} V_{x_1} & C_{x_1x_2} & C_{x_1'x_1} & C_{x_1'x_2} \\ C_{x_1x_2} & V_{x_2} & C_{x_2x_1'} & C_{x_2x_2'} \\ C_{x_1'x_1} & C_{x_2x_1'} & V_{x_1'} & C_{x_1'x_1'} \\ C_{x_1'x_2} & C_{x_2x_2'} & C_{x_1'x_2'} & V_{x_2'} \end{pmatrix}$$
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 ($V_{t1}$), and the variance of a split is the sum of true score and error variances:

$$\Sigma_{XX'} = \begin{pmatrix}
V_{t1} + V_{e1} & \cdots & V_{t1} \\
\cdots & \cdots & \cdots \\
V_{t1} & \cdots & V_{t1} + V_{e1} \\
\end{pmatrix}
\begin{pmatrix}
V_{t1} \\
\cdots \\
V_{t1} \\
\end{pmatrix} = \begin{pmatrix}
V_{t1} \\
\cdots \\
V_{t1} \\
\end{pmatrix}
\begin{pmatrix}
V_{t1} \\
\cdots \\
V_{t1} \\
\end{pmatrix} + \begin{pmatrix}
V_{t1}' + V_{e1}' \\
\cdots \\
V_{t1}' \\
\end{pmatrix}
\begin{pmatrix}
V_{t1}' \\
\cdots \\
V_{t1}' \\
\end{pmatrix}
$$

The correlation between a test made of up two halves with intercorrelation ($r_1 = V_{t1}/V_{x1}$) with another such test is

$$r_{xx'} = \frac{4V_{t1}}{\sqrt{(4V_{t1} + 2V_{e1})(4V_{t1} + 2V_{e1})}} = \frac{4V_{t1}}{2V_{t1} + 2V_{x1}} = \frac{4r_1}{2r_1 + 2}$$

and thus...
The Spearman Brown Prophecy Formula

The correlation between a test made of up two halves with intercorrelation \( r_1 = V_{t1}/V_{x1} \) with another such test is

\[
r_{xx'} = \frac{4V_{t1}}{\sqrt{(4V_{t1} + 2V_{e1})(4V_{t1} + 2V_{e1})}} = \frac{4V_{t1}}{2V_{t1} + 2V_{x1}} = \frac{4r_1}{2r_1 + 2}
\]

and thus

\[
r_{xx'} = \frac{2r_1}{1 + r_1}
\]  \hspace{1cm} (12)
6,435 possible eight item splits of the 16 ability items
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

$$r_{xjd} = \frac{\text{COV}_{xjd}}{\sqrt{\sigma^2_{xj} \sigma^2_d}}.$$

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

$$\sigma^2_d = \sum_{j=1}^{n} \sum_{k=1}^{n} \text{COV}_{xjk} = \sum_{j=1}^{n} \sigma^2_{xj} + \sum_{j=1}^{n} \sum_{k \neq j} \text{COV}_{xjk}.$$
Correlation of an item with the domain

Then letting $\bar{c} = \frac{\sum_{j=1}^{n} \sum_{k \neq j} \text{cov}_{xjk}}{n(n-1)}$ be the average covariance and $\bar{v} = \frac{\sum_{j=1}^{n} \sigma_{xj}^2}{n}$ the average item variance, the correlation of a randomly chosen item with the domain is

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

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

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

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

$$\lim_{n \to \infty} r_{xjd}^2 = \frac{\bar{c}}{\bar{v}}. \quad (13)$$

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 Theory

Domain sampling – correlation of an item with the domain

\[
\lim_{n \to \infty} r_{xjd}^2 = \frac{\bar{c}}{\bar{v}}.
\]  (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.
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:

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

The correlation with the domain will be

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

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

\[ \alpha = r_{xx} = \frac{\sigma^2_t}{\sigma^2_x} = \frac{k^2 \sigma^2_x - \sum \sigma^2_i}{k(k-1)} = \frac{k}{k-1} \frac{\sigma^2_x - \sum \sigma^2_i}{\sigma^2_x} \]  

(16)
Coefficients based upon the internal structure of a test

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

Alpha varies by r and number of items

\[ \alpha \]

\[ \text{Number of items} \]

\[ \alpha \]

\[ r = 0.2 \]

\[ r = 0.1 \]

\[ r = 0.05 \]
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 \( \alpha \), will be

\[ \frac{S}{N} = \frac{n\bar{r}}{1 - \bar{r}}. \quad (17) \]

That is, the S/N ratio increases linearly with the number of items as well as with the average intercorrelation.
Coefficients based upon the internal structure of a test

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])

<table>
<thead>
<tr>
<th></th>
<th>raw_alpha</th>
<th>std.alpha</th>
<th>G6(smc)</th>
<th>average_r</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.81</td>
<td>0.81</td>
<td>0.8</td>
<td>0.46</td>
<td>15</td>
<td>5.8</td>
</tr>
</tbody>
</table>

Reliability if an item is dropped:

<table>
<thead>
<tr>
<th></th>
<th>raw_alpha</th>
<th>std.alpha</th>
<th>G6(smc)</th>
<th>average_r</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>0.75</td>
<td>0.75</td>
<td>0.70</td>
<td>0.42</td>
</tr>
<tr>
<td>N2</td>
<td>0.76</td>
<td>0.76</td>
<td>0.71</td>
<td>0.44</td>
</tr>
<tr>
<td>N3</td>
<td>0.75</td>
<td>0.76</td>
<td>0.74</td>
<td>0.44</td>
</tr>
<tr>
<td>N4</td>
<td>0.79</td>
<td>0.79</td>
<td>0.76</td>
<td>0.48</td>
</tr>
<tr>
<td>N5</td>
<td>0.81</td>
<td>0.81</td>
<td>0.79</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Item statistics

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>r</th>
<th>r.cor</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>990</td>
<td>0.81</td>
<td>0.78</td>
<td>2.8</td>
<td>1.5</td>
</tr>
<tr>
<td>N2</td>
<td>990</td>
<td>0.79</td>
<td>0.75</td>
<td>3.5</td>
<td>1.5</td>
</tr>
<tr>
<td>N3</td>
<td>997</td>
<td>0.79</td>
<td>0.72</td>
<td>3.2</td>
<td>1.5</td>
</tr>
<tr>
<td>N4</td>
<td>996</td>
<td>0.71</td>
<td>0.60</td>
<td>3.1</td>
<td>1.5</td>
</tr>
<tr>
<td>N5</td>
<td>992</td>
<td>0.67</td>
<td>0.52</td>
<td>2.9</td>
<td>1.6</td>
</tr>
</tbody>
</table>
What if items differ in their direction?

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

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

<table>
<thead>
<tr>
<th>raw_alpha</th>
<th>std.alpha</th>
<th>G6(smc)</th>
<th>average_r</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.28</td>
<td>-0.22</td>
<td>0.13</td>
<td>-0.038</td>
<td>3.8</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Reliability if an item is dropped:

<table>
<thead>
<tr>
<th>raw_alpha</th>
<th>std.alpha</th>
<th>G6(smc)</th>
<th>average_r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>-0.430</td>
<td>-0.472</td>
<td>-0.020</td>
</tr>
<tr>
<td>C2</td>
<td>-0.367</td>
<td>-0.423</td>
<td>-0.017</td>
</tr>
<tr>
<td>C3</td>
<td>-0.263</td>
<td>-0.295</td>
<td>0.094</td>
</tr>
<tr>
<td>C4</td>
<td>-0.022</td>
<td>0.123</td>
<td>0.283</td>
</tr>
<tr>
<td>C5</td>
<td>-0.028</td>
<td>0.022</td>
<td>0.242</td>
</tr>
</tbody>
</table>

Item statistics

<table>
<thead>
<tr>
<th>n</th>
<th>r</th>
<th>r.cor</th>
<th>r.drop</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>2779</td>
<td>0.56</td>
<td>0.51</td>
<td>0.0354</td>
<td>4.5</td>
</tr>
<tr>
<td>C2</td>
<td>2776</td>
<td>0.54</td>
<td>0.51</td>
<td>-0.0076</td>
<td>4.4</td>
</tr>
<tr>
<td>C3</td>
<td>2780</td>
<td>0.48</td>
<td>0.27</td>
<td>-0.0655</td>
<td>4.3</td>
</tr>
<tr>
<td>C4</td>
<td>2774</td>
<td>0.20</td>
<td>-0.34</td>
<td>-0.2122</td>
<td>2.6</td>
</tr>
<tr>
<td>C5</td>
<td>2784</td>
<td>0.29</td>
<td>-0.19</td>
<td>-0.1875</td>
<td>3.3</td>
</tr>
</tbody>
</table>
But what if some items are reversed keyed?

\[
\text{alpha(bfi[6:10])}
\]

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

\[
\begin{array}{cccccc}
\text{raw_alpha} & \text{std.alpha} & \text{G6(smc)} & \text{average_r} & \text{mean} & \text{sd} \\
0.73 & 0.73 & 0.69 & 0.35 & 3.8 & 0.58 \\
\end{array}
\]

Reliability if an item is dropped:

\[
\begin{array}{cccccc}
\text{raw_alpha} & \text{std.alpha} & \text{G6(smc)} & \text{average_r} \\
C1 & 0.69 & 0.70 & 0.64 & 0.36 \\
C2 & 0.67 & 0.67 & 0.62 & 0.34 \\
C3 & 0.69 & 0.69 & 0.64 & 0.36 \\
C4- & 0.65 & 0.66 & 0.60 & 0.33 \\
C5- & 0.69 & 0.69 & 0.63 & 0.36 \\
\end{array}
\]

Item statistics

\[
\begin{array}{ccccccc}
\text{n} & \text{r} & \text{r.cor} & \text{r.drop} & \text{mean} & \text{sd} \\
C1 & 2779 & 0.67 & 0.54 & 0.45 & 4.5 & 1.2 \\
C2 & 2776 & 0.71 & 0.60 & 0.50 & 4.4 & 1.3 \\
C3 & 2780 & 0.67 & 0.54 & 0.46 & 4.3 & 1.3 \\
C4- & 2774 & 0.73 & 0.64 & 0.55 & 2.6 & 1.4 \\
C5- & 2784 & 0.68 & 0.57 & 0.48 & 3.3 & 1.6 \\
\end{array}
\]

Warning message: In alpha(bfi[6:10]) : Some items were negatively correlated with total scale and were automatically
Bootstrapped confidence intervals for $\alpha$

Distribution of 10,000 bootstrapped values of alpha

![Histogram showing the distribution of 10,000 bootstrapped values of alpha with a peak around 0.83]
Reliability is about the amount of test variance that is not error variance. But what is the error variance?

\[ r_{xx} = \frac{V_x - V_e}{V_x} = 1 - \frac{V_e}{V_x} . \]  

(18)

\[ \lambda_1 = 1 - \frac{\text{tr}(V_x)}{V_x} = \frac{V_x - \text{tr}(V_x)}{V_x} . \]  

(19)

\[ \lambda_2 = \lambda_1 + \frac{\sqrt{n - 1} C_2}{V_x} = \frac{V_x - \text{tr}(V_x) + \sqrt{n - 1} C_2}{V_x} . \]  

(20)

\[ \lambda_3 = \lambda_1 + \frac{\frac{V_x - \text{tr}(V_x)}{n(n-1)}}{V_x} = \frac{n\lambda_1}{n-1} = \frac{n}{n-1} \left( 1 - \frac{\text{tr}(V)}{V_x} \right) = \frac{n}{n-1} \frac{V_x - \text{tr}(V_x)}{V_x} = \alpha \]  

(21)

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

(22)

\[ \lambda_6 = 1 - \frac{\sum e_j^2}{V_x} = 1 - \frac{\sum (1 - r_{smc}^2)}{V_x} \]  

(23)
Four different correlation matrices, one value of $\alpha$

1. The problem of group factors
2. If no groups, or many groups, $\alpha$ is ok
Decomposing a test into general, Group, and Error variance

1. Decompose total variance into general, group, specific, and error

2. $\alpha < \text{total}$

3. $\alpha > \text{general}$
Two additional alternatives to $\alpha$: $\omega_{hierarchical}$ and $\omega_{total}$

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

$$x = cg + Af + Ds + e \quad (24)$$

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

$$h_j^2 = c_j^2 + \sum f_{ij}^2 \quad (25)$$

and the unique variance for the item

$$u_j^2 = \sigma_j^2 (1 - h_j^2) \quad (26)$$

may be used to estimate the test reliability.

$$\omega_t = \frac{1cc'1' + 1AA'1'}{V_x} = 1 - \frac{\sum(1 - h_j^2)}{V_x} = 1 - \frac{\sum u^2}{V_x} \quad (27)$$
Problems with \( \alpha \)

McDonald (1999) introduced two different forms for \( \omega \)

\[
\omega_t = \frac{1cc'1' + 1AA'1'}{V_x} = 1 - \frac{\sum(1 - h_j^2)}{V_x} = 1 - \frac{\sum u^2}{V_x} \tag{28}
\]

and

\[
\omega_h = \frac{1cc'1}{V_x} = \frac{(\sum \Lambda_i)^2}{\sum \sum R_{ij}}. \tag{29}
\]

These may both be find by factoring the correlation matrix and finding the g and group factor loadings using the \( \omega \)ma function.
Using **omega** on the Thurstone data set to find alternative reliability estimates

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.83</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sent.Completion</td>
<td>0.78</td>
<td>0.78</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First.Letters</td>
<td>0.44</td>
<td>0.49</td>
<td>0.46</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.Letter.Words</td>
<td>0.43</td>
<td>0.46</td>
<td>0.42</td>
<td>0.67</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suffixes</td>
<td>0.45</td>
<td>0.49</td>
<td>0.44</td>
<td>0.59</td>
<td>0.54</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Letter.Series</td>
<td>0.45</td>
<td>0.43</td>
<td>0.40</td>
<td>0.38</td>
<td>0.40</td>
<td>0.29</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedigrees</td>
<td>0.54</td>
<td>0.54</td>
<td>0.53</td>
<td>0.35</td>
<td>0.37</td>
<td>0.32</td>
<td>0.56</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Letter.Group</td>
<td>0.38</td>
<td>0.36</td>
<td>0.36</td>
<td>0.42</td>
<td>0.45</td>
<td>0.32</td>
<td>0.60</td>
<td>0.45</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>g</th>
<th>F1*</th>
<th>F2*</th>
<th>F3*</th>
<th>h2</th>
<th>u2</th>
<th>p2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>0.71</td>
<td>0.57</td>
<td></td>
<td></td>
<td>0.82</td>
<td>0.18</td>
<td>0.61</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.73</td>
<td>0.55</td>
<td></td>
<td></td>
<td>0.84</td>
<td>0.16</td>
<td>0.63</td>
</tr>
<tr>
<td>Sent.Completion</td>
<td>0.68</td>
<td>0.52</td>
<td></td>
<td></td>
<td>0.73</td>
<td>0.27</td>
<td>0.63</td>
</tr>
<tr>
<td>First.Letters</td>
<td>0.65</td>
<td>0.56</td>
<td></td>
<td></td>
<td>0.73</td>
<td>0.27</td>
<td>0.57</td>
</tr>
<tr>
<td>4.Letter.Words</td>
<td>0.62</td>
<td>0.49</td>
<td></td>
<td></td>
<td>0.63</td>
<td>0.37</td>
<td>0.61</td>
</tr>
<tr>
<td>Suffixes</td>
<td>0.56</td>
<td>0.41</td>
<td></td>
<td></td>
<td>0.50</td>
<td>0.50</td>
<td>0.63</td>
</tr>
<tr>
<td>Letter.Series</td>
<td>0.59</td>
<td>0.56</td>
<td></td>
<td></td>
<td>0.61</td>
<td>0.72</td>
<td>0.28</td>
</tr>
<tr>
<td>Pedigrees</td>
<td>0.58</td>
<td>0.23</td>
<td></td>
<td></td>
<td>0.34</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Letter.Group</td>
<td>0.54</td>
<td>0.41</td>
<td></td>
<td></td>
<td>0.46</td>
<td>0.53</td>
<td>0.47</td>
</tr>
</tbody>
</table>

With eigenvalues of:

<table>
<thead>
<tr>
<th></th>
<th>g</th>
<th>F1*</th>
<th>F2*</th>
<th>F3*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.58</td>
<td>0.96</td>
<td>0.74</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Types of reliability

- **Internal consistency**
  - $\alpha$
  - $\omega_{hierarchical}$
  - $\omega_{total}$
  - $\beta$

- **Intra-class**

- **Agreement**

- **Test-retest, alternate form**

- **Generalizability**

- **Internal consistency**
  - alpha,
  - score.items

- **Intraclass Agreement**
  - icc

- **Kappa**
  - wkappa,
  - cohen.kappa

- **Generalizability**
  - cor
  - aov
Alpha and its alternatives

Reliability = \( \frac{\sigma^2_x}{\sigma^2_x} = 1 - \frac{\sigma^2_e}{\sigma^2_x} \)

If there is another test, then \( \sigma_t = \sigma_{t1t2} \) (covariance of test \( X_1 \) with test \( X_2 = C_{xx} \))

But, if there is only one test, we can estimate \( \sigma^2_t \) based upon the observed covariances within test 1

How do we find \( \sigma^2_e \)?

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 \( C_x \)

- \( C_x = \sigma^2_e = \text{diag}(C_x) \)
- \( \lambda_1 = \frac{C_x - \text{diag}(C_x)}{C_x} \)

A better case (Guttman case 3, \( \alpha \)) is that that the average covariance between the items on the test is the same as the average true score variance for each item.

- \( C_x = \sigma^2_e = \text{diag}(C_x) \)
- \( \lambda_3 = \alpha = \lambda_1 \times \frac{n}{n-1} = \frac{(C_x - \text{diag}(C_x)) \times n/(n-1)}{C_x} \)
Guttman 6: estimating using the Squared Multiple Correlation

- Reliability = \( \frac{\sigma^2_t}{\sigma^2_x} = 1 - \frac{\sigma^2_e}{\sigma^2_x} \)
- Estimate true item variance as squared multiple correlation with other items
- \( \lambda_6 = \frac{(C_x - \text{diag}(C_x) + \Sigma(smci))}{C_x} \)
  - This takes observed covariance, subtracts the diagonal, and replaces with the squared multiple correlation
  - Similar to \( \alpha \) which replaces with average inter-item covariance
- Squared Multiple Correlation is found by \( smc \) and is just
  \( smci = 1 - 1/R_{ii}^{-1} \)
First, create some simulated data with a known structure

```r
> 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 0.42 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)

Four congeneric tests

F1

0.9 0.8 0.7 0.6

V1 V2 V3 V4

> 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 $\alpha$ and related stats for the simulated data

```r
> alpha(v4$observed)

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

        raw_alpha std.alpha  G6(smc) average_r mean    sd
V1          0.59      0.60   0.50      0.33 0.50 0.50
V2          0.63      0.64   0.55      0.37 0.64 0.64
V3          0.65      0.66   0.59      0.40 0.66 0.66
V4          0.72      0.72   0.64      0.46 0.72 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
```
Hierarchical structures

**A hierarchical structure**

cor.plot(r9)

```r
> 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
```
Hierarchical structures

\( \alpha \) of the 9 hierarchical variables

```R
> alpha(r9)

Reliability analysis
Call: alpha(x = r9)

        raw_alpha std.alpha  G6(smc)  average_r
V1       0.76      0.76    0.76    0.26
V2       0.72      0.72    0.71    0.25
V3       0.74      0.74    0.73    0.26
V4       0.74      0.74    0.73    0.26
V5       0.73      0.73    0.72    0.25
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
V2  0.67  0.63
V3  0.61  0.55
V4  0.59  0.52
V5  0.53  0.43
V6  0.56  0.46
V7  0.50  0.39
V8  0.45  0.32
V9  0.76  0.76
```
An example of two different scales confused as one

\[
\begin{array}{cccccccc}
V1 & V2 & V3 & V4 & V5 & V6 & V7 & V8 \\
V1 & 1.00 & & & & & & \\
V2 & 0.29 & 1.00 & & & & & & \\
V3 & 0.05 & 0.03 & 1.00 & & & & & \\
V4 & 0.03 & -0.02 & 0.34 & 1.00 & & & & \\
V5 & -0.38 & -0.35 & -0.02 & -0.01 & 1.00 & & & \\
V6 & -0.38 & -0.33 & -0.10 & 0.06 & 0.33 & 1.00 & & \\
V7 & -0.06 & 0.02 & -0.40 & -0.36 & 0.03 & 0.04 & 1.00 & \\
V8 & -0.08 & -0.04 & -0.39 & -0.37 & 0.05 & 0.03 & 0.37 & 1.00 \\
\end{array}
\]

> set.seed(17)
> two.f <- sim.item(8)
> lower.mat(cor(two.f))
cor.plot(cor(two.f))
Rearrange the items to show it more clearly
Note the use of the keys parameter to specify how some items should be reversed.

```r
> 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)))

<table>
<thead>
<tr>
<th>raw_alpha</th>
<th>std.alpha</th>
<th>G6(smc)</th>
<th>average_r</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.62</td>
<td>0.62</td>
<td>0.65</td>
<td>0.17</td>
<td>-0.0051</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Reliability if an item is dropped:

```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

<table>
<thead>
<tr>
<th>n</th>
<th>r</th>
<th>r.cor</th>
<th>r.drop</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>500</td>
<td>0.54</td>
<td>0.44</td>
<td>0.33</td>
<td>0.063</td>
</tr>
<tr>
<td>V2</td>
<td>500</td>
<td>0.48</td>
<td>0.35</td>
<td>0.26</td>
<td>0.070</td>
</tr>
<tr>
<td>V3</td>
<td>500</td>
<td>0.56</td>
<td>0.47</td>
<td>0.36</td>
<td>-0.030</td>
</tr>
<tr>
<td>V4</td>
<td>500</td>
<td>0.48</td>
<td>0.37</td>
<td>0.28</td>
<td>-0.130</td>
</tr>
<tr>
<td>V5</td>
<td>500</td>
<td>0.52</td>
<td>0.42</td>
<td>0.31</td>
<td>-0.073</td>
</tr>
<tr>
<td>V6</td>
<td>500</td>
<td>0.52</td>
<td>0.41</td>
<td>0.31</td>
<td>-0.071</td>
</tr>
<tr>
<td>V7</td>
<td>500</td>
<td>0.53</td>
<td>0.44</td>
<td>0.34</td>
<td>0.035</td>
</tr>
<tr>
<td>V8</td>
<td>500</td>
<td>0.56</td>
<td>0.47</td>
<td>0.36</td>
<td>0.097</td>
</tr>
</tbody>
</table>
Score as two different scales

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

```r
> 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 $\alpha$ and other reliability estimates

```r
> 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
```
Intraclass correlations

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
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.

```r
> Ratings
   J1 J2 J3 J4 J5 J6
1  1  1  1  6  2  3  6
2  2  2  2  4  1  2
3  3  3  3  8  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.

<table>
<thead>
<tr>
<th>Variance estimates</th>
<th>(J1, J2)</th>
<th>(J3, J4)</th>
<th>(J5, J6)</th>
<th>(J1, J3)</th>
<th>(J1, J5)</th>
<th>(J1 ... J3)</th>
<th>(J1 ... J4)</th>
<th>(J1 ... J6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MS_b$</td>
<td>7</td>
<td>15.75</td>
<td>15.75</td>
<td>7.0</td>
<td>5.2</td>
<td>10.50</td>
<td>21.88</td>
<td>22.00</td>
</tr>
<tr>
<td>$MS_w$</td>
<td>0</td>
<td>2.58</td>
<td>7.58</td>
<td>12.5</td>
<td>1.5</td>
<td>8.33</td>
<td>7.12</td>
<td>3.00</td>
</tr>
<tr>
<td>$MS_j$</td>
<td>0</td>
<td>6.75</td>
<td>36.75</td>
<td>75.0</td>
<td>0.0</td>
<td>50.00</td>
<td>38.38</td>
<td>3.00</td>
</tr>
<tr>
<td>$MS_e$</td>
<td>0</td>
<td>1.75</td>
<td>1.75</td>
<td>0.0</td>
<td>1.8</td>
<td>0.00</td>
<td>.88</td>
<td>3.00</td>
</tr>
</tbody>
</table>

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`

```r
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

```r
> print(ICC(Ratings), all=TRUE) # get more output than normal

$results

<table>
<thead>
<tr>
<th>type</th>
<th>ICC</th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>p</th>
<th>lower bound</th>
<th>upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single_raters_absolute</td>
<td>ICC1</td>
<td>0.32</td>
<td>3.84</td>
<td>5</td>
<td>0.01</td>
<td>0.04</td>
<td>0.79</td>
</tr>
<tr>
<td>Single_random_raters</td>
<td>ICC2</td>
<td>0.37</td>
<td>10.37</td>
<td>5</td>
<td>0.00</td>
<td>0.09</td>
<td>0.80</td>
</tr>
<tr>
<td>Single_fixed_raters</td>
<td>ICC3</td>
<td>0.61</td>
<td>10.37</td>
<td>5</td>
<td>0.00</td>
<td>0.28</td>
<td>0.91</td>
</tr>
<tr>
<td>Average_raters_absolute</td>
<td>ICC1k</td>
<td>0.74</td>
<td>3.84</td>
<td>5</td>
<td>0.01</td>
<td>0.21</td>
<td>0.96</td>
</tr>
<tr>
<td>Average_random_raters</td>
<td>ICC2k</td>
<td>0.78</td>
<td>10.37</td>
<td>5</td>
<td>0.00</td>
<td>0.38</td>
<td>0.96</td>
</tr>
<tr>
<td>Average_fixed_raters</td>
<td>ICC3k</td>
<td>0.90</td>
<td>10.37</td>
<td>5</td>
<td>0.00</td>
<td>0.70</td>
<td>0.98</td>
</tr>
</tbody>
</table>

$summary

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>subs</td>
<td>5</td>
<td>141.667</td>
<td>28.3333</td>
<td>10.366</td>
</tr>
<tr>
<td>ind</td>
<td>5</td>
<td>153.000</td>
<td>30.6000</td>
<td>11.195</td>
</tr>
<tr>
<td>Residuals</td>
<td>25</td>
<td>68.333</td>
<td>2.7333</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ^O***~O 0.001 ^O**~O 0.01 ^O*~O 0.05 ^O.~O 0.1 ^O ~O 1

$stats

<table>
<thead>
<tr>
<th>[,1]</th>
<th>[,2]</th>
<th>[,3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>5.000000e+00</td>
<td>5.000000e+00</td>
</tr>
<tr>
<td>[2,]</td>
<td>1.416667e+02</td>
<td>1.530000e+02</td>
</tr>
<tr>
<td>[3,]</td>
<td>2.833333e+01</td>
<td>3.060000e+01</td>
</tr>
<tr>
<td>[4,]</td>
<td>1.036585e+01</td>
<td>1.119512e+01</td>
</tr>
<tr>
<td>[5,]</td>
<td>1.800581e-05</td>
<td>9.644359e-06</td>
</tr>
</tbody>
</table>

$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 boundaries

  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

<table>
<thead>
<tr>
<th>7</th>
<th>Two approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Various IRT models</td>
</tr>
<tr>
<td>9</td>
<td>Polytomous items</td>
</tr>
<tr>
<td></td>
<td>Ordered response categories</td>
</tr>
<tr>
<td></td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td>10</td>
<td>Factor analysis &amp; IRT</td>
</tr>
<tr>
<td></td>
<td>Non-monotone Trace lines</td>
</tr>
<tr>
<td>11</td>
<td>(C) A T</td>
</tr>
</tbody>
</table>
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 $\sigma^2_X$ 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
   - \( p(\text{correct}|\text{ability}, \text{difficulty}, \ldots) = f(\text{ability} - \text{difficulty}) \)
   - 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:

\[ p(\text{correct}_{ij}|\theta_i) = \frac{1}{1 + e^{-\theta_i}}. \]  

(30)
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:

\[ p(\text{correct}_{ij} | \theta_i, \delta_j) = \frac{1}{1 + e^{\delta_j - \theta_i}}. \]  

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) \( (\theta_i) \) and the item difficulty (or location) \( (\delta_j) \). 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 \(1 - p(\text{correct})\) and thus the odds ratio of being correct for a person with ability, \(\theta_i\), on an item with difficulty, \(\delta_j\) is

\[
OR_{ij} = \frac{p}{1-p} = \frac{p}{q} = \frac{1}{1 + e^{\delta_j - \theta_i}} = \frac{1}{1 + e^{\delta_j - \theta_i}} \cdot \frac{1}{1 + e^{\delta_j - \theta_i}} = \frac{1}{e^{\delta_j - \theta_i}} = e^{\theta_i - \delta_j}.
\]  

(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

\[
OR_{i1} OR_{i2} \ldots OR_{in} = \prod_{j=1}^{n} e^{\theta_i - \delta_j} = e^{n \theta_i} e^{-\sum_{j=1}^{n} \delta_j}.
\]  

(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:

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

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

\[
\theta_i = \frac{\ln P}{n}.
\] (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 \( \delta_j \) will be

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

(36)

and taking logs of both sides leads to

\[
ln\frac{P}{Q} = \sum_{i=1}^{N} (\theta_i) - N\delta_j.
\]  

(37)

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

\[
\delta_j = \frac{ln \frac{Q}{P}}{N}.
\]  

(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

**Item Characteristic Curves**

![Item Characteristic Curves](image)

**Item Information Curves**

![Item Information Curves](image)
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))

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

> 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  Dscrnm
-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:

\[ I_j = \frac{1}{\sigma_{e_j}^2} = \frac{h_j^2}{1 - h_j^2}. \]

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

\[ I(f, x_j) = \frac{[P'_j(f)]^2}{P_j(f)Q_j(f)} \]  \hspace{1cm} (39)

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

\[ P' = \frac{e^{\delta-\theta}}{(1 + e^{\delta-\theta})^2} \]  \hspace{1cm} (40)

which is just \( P_j Q_j \) and thus the information for an item is

\[ I_j = P_j Q_j. \]  \hspace{1cm} (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

\[ p(\text{correct}_{ij} | \theta_i, \alpha_j, \delta_j) = \frac{1}{1 + e^{\alpha_i(\delta_j - \theta_i)}} \] (42)
2PL and 2PN models

\[ p(\text{correct}_{ij} | \theta_i, \alpha_j, \delta_j) = \frac{1}{1 + e^{\alpha_i(\delta_j - \theta_i)}} \]  (43)

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

\[ p(\text{correct} | \theta, \alpha_j, \delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha(\theta - \delta)} e^{-\frac{u^2}{2}} du \]  (44)

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

The information function for a two parameter model reflects the item discrimination parameter, \( \alpha \),

\[ I_j = \alpha^2 P_j Q_j \]  (45)

which, for a 2PL model is

\[ I_j = \alpha_j^2 P_j Q_j = \frac{\alpha_j^2}{(1 + e^{\alpha_j(\delta_j - \theta_j)})^2}. \]  (46)
The problem of non-parallel trace lines

2PL models differing in their discrimination parameter

\[ P(x) = \frac{1}{1 + e^{\alpha(\theta-\delta)}} \]
Parameter explosion – better fit but at what cost

The 3 parameter model adds a guessing parameter.

\[
p(\text{correct}_{ij}|\theta_i, \alpha_j, \delta_j, \gamma_j) = \gamma_j + \frac{1 - \gamma_j}{1 + e^{\alpha_i(\delta_j-\theta_i)}} \quad (47)
\]

And the four parameter model adds an asymptotic parameter

\[
P(x|\theta_i, \alpha, \delta_j, \gamma_j, \zeta_j) = \gamma_j + \frac{\zeta_j - \gamma_j}{1 + e^{\alpha_j(\delta_j-\theta_i)}}. \quad (48)
\]
Two approaches to various IRT models

Polytomous items

Factor analysis & IRT

3PL models differing in guessing and difficulty

\[ P(x) = \gamma + \frac{1 - \gamma}{1 + e^{\alpha(\theta - \delta)}} \]

\[ \gamma = 0 \quad \delta = 0 \]

\[ \gamma = 0.2 \quad \delta = -1 \]

\[ \gamma = 0.2 \quad \delta = 1 \]

\[ \gamma = 0.3 \quad \delta = 0 \]

\[ \gamma = 0.2 \quad \delta = -1 \]

4PL items differing in guessing, difficulty and asymptote

\[ P(x) = \gamma + \frac{\zeta - \gamma}{1 + e^{\alpha(\theta - \delta)}} \]

\[ \gamma = 0.2 \quad \delta = -1 \]

\[ \gamma = 0.2 \quad \delta = 1 \]

\[ \gamma = 0 \quad \delta = 0 \]

\[ \gamma = 0.2 \quad \zeta = 0.8 \]
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

$$P(r = k|\theta_i, \delta_k, \delta_{k-1}, \alpha_k) = P(r|\theta_i, \delta_{k-1}, \alpha_k) - P(r|\theta_i, \delta_k, \alpha_k) = \frac{1}{1 + e^{-\alpha_k(\delta_{k-1} - \theta_i)}} - \frac{1}{1 + e^{-\alpha_k(\delta_k - \theta_i)}}$$  \hspace{1cm} (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 multiple 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

![Differential Item Functioning](image)
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 ($\alpha$) of the items on the latent variable. These regressions are, of course, just the factor loadings. Item difficulty, $\delta_j$ and item discrimination, $\alpha_j$ may be found from factor analysis of the tetrachoric correlations where $\lambda_j$ is just the factor loading on the first factor and $\tau_j$ is the normal threshold reported by the tetrachoric function (McDonald, 1999; Lord & Novick, 1968; Takane & de Leeuw, 1987).

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

(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

\[ \delta_j = \frac{D \tau}{\sqrt{1 - \lambda_j^2}}, \]

FA parameters from IRT

\[ \lambda_j = \frac{\alpha_j}{\sqrt{1 + \alpha_j^2}}, \]

\[ \alpha_j = \frac{\lambda_j}{\sqrt{1 - \lambda_j^2}} \] (51)

\[ \tau_j = \frac{\delta_j}{\sqrt{1 + \alpha_j^2}}. \]
the \texttt{irt.fa} function

\begin{verbatim}
> 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
\end{verbatim}
Item Characteristic Curves from FA

Item parameters from factor analysis

V1, V2, V3, V4, V5, V6, V7, V8, V9
Item information from factor analysis

The graph shows the item information from factor analysis for variables V1 to V9. The y-axis represents item information, while the x-axis represents the latent trait on a normal scale. Each variable has a distinct bell-shaped curve indicating the distribution of item information across different levels of the latent trait.
Test Information Curve

Test information -- item parameters from factor analysis
Comparing three ways of estimating the parameters

```r
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)
```

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

Attitudes reflect an unfolding (ideal point) model
Two approaches Various IRT models Polytomous items Factor analysis & IRT (C) A T References

IRT and CTT don’t really differ except

1. Correlation of classic test scores and IRT scores $> .98$.
2. Test information for the person doesn’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)


