Chapter 7
Classical Test Theory and the Measurement of Reliability

Whether discussing ability, affect, or climate change, as scientists we are interested in the relationships between our theoretical constructs. We recognize, however, that our measurements are not perfect and that any particular observation has some unknown amount of error associated with that measurement for “all measurement is befuddled by error” (McNemar, 1946, p 294). When estimating central tendencies, the confidence interval of the mean may be estimated by the standard error of the sample observations, and may be calculated from the observed standard deviation and the number of observations. This is, of course, the basic concept of Gossett and Fisher.

Error may be both random as well systematic. Random error reflects trial by trial variability due to unknown sources while systematic error may reflect situational or individual effects that may be specified. Perhaps the classic example of systematic error (known as the *personal equation*) is the analysis of individual differences in reaction time in making astronomical observations. “The personal equation of an observer is the interval of time which habitually intervenes between the actual and the observed transit of a star...” (Rogers, 1869). Before systematic individual differences were analyzed, the British astronomer Maskelyn fired his assistant Kennebrook for making measurements that did not agree with his own (Stigler, 1986). Subsequent investigations of the systematic bias (Safford, 1898; Sanford, 1889) showed consistent individual differences as well as the effect of situational manipulations such as hunger and sleep deprivation (Rogers, 1869).

Systematic error may be removed. What about the effect of random error? When estimating individual scores we want to find the standard error for each individual as welll as the central tendency for that individual. More importantly, if we want to find the relationship between two variables, the errors of observation will affect the strength of the correlation between them. Charles Spearman (1904b) was the first psychologist to recognize that observed correlations are attenuated from the true correlation if the observations contain 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, 1904b, p 90).

This *attenuation* of the relationship between p and q may be seen graphically in Figure 7.1 panel A. An alternative way of considering this is examine the effect of combining true scores (dashed line) with error scores (light solid lines) to produce observed score (heavy solid line) as shown in Figure 7.2.

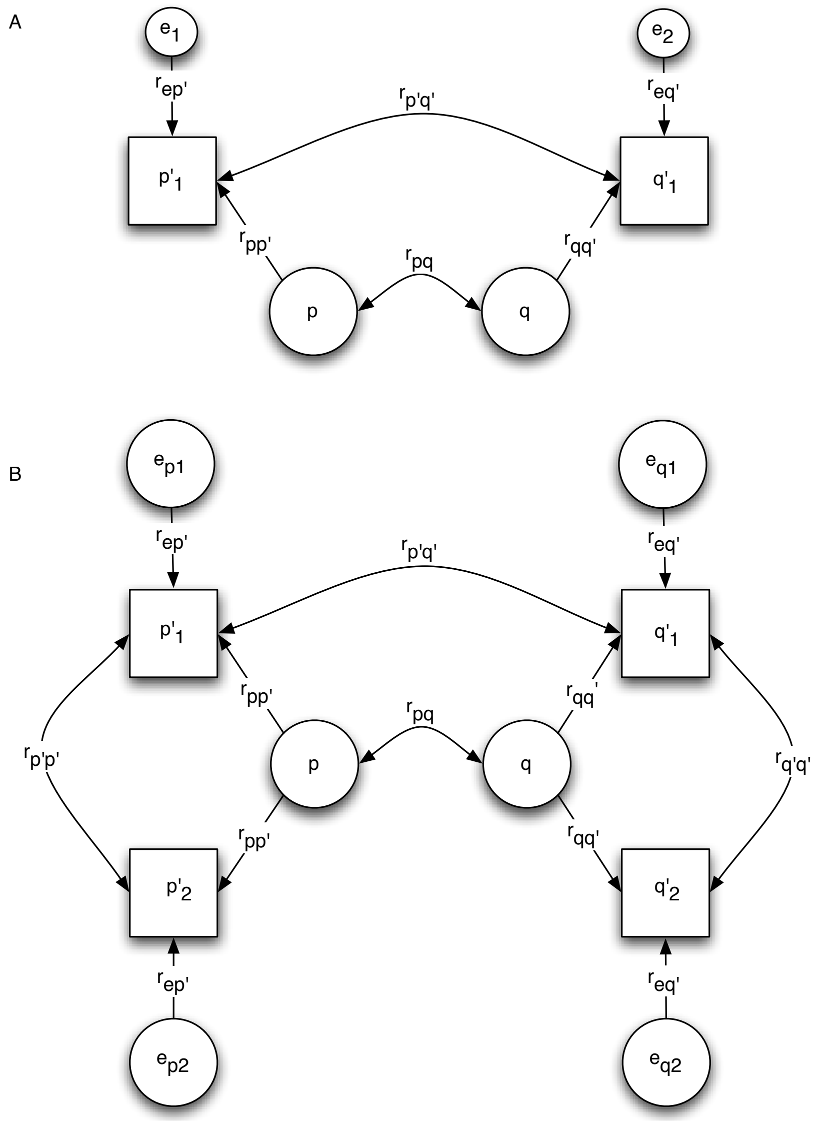


Fig. 7.1 Spearman‘s model of attenuation and reliability. Panel A: The true relationship between p and q is attenuated by the error in p’ and q’. Panel B: the correlation between the latent variable p and the observed variable p’ may be estimated from the correlation of p’ with a parallel test.

7.1 Reliability and True Scores

The classic model of *reliability* treats an *observed score*, p', as made up of two independent components: the latent *true score*, p, and a latent *error score*, e (Figure 7.1 panel A). Errors are "accidental deviations [that] are different in every individual case (hence are often called the 'variable errors') and occur quite impartially in every direction according to the known laws of probability" (Spearman, 1904b, p 76), and may be seen as randomly "augmenting and diminishing" observed values, and "tending in a prolonged series to always more and more perfectly counterbalance one another" (Spearman, 1904b, p 89).

Consider a simple case of asking students their ages but to insure privacy, asking them to flip a coin before giving their answer. If the coin comes up heads, they should add 1 to their real age, if it comes up tails, they should subtract 1. Clearly no observed score corresponds to the true scores. But if we repeat this exercise 10 times, then the mean for each student will be quite close to their true age. Indeed, as the number of observations per student increases, the mean score will tend towards their true score with a precision based upon the inverse of the square root of the number of observations. True score can then be defined as the expected value, over multiple trials, of the observed score (Figure 7.2). Unfortunately, if errors are systematically biased in one direction or another, this definition of true score will not produce Platonic Truth. (The classic example is if one is attempting to determine the sex of young chickens, errors will not have an expected value of zero, but rather will be slightly biased towards the other sex Lord and Novick (1968)).

Using more modern notation by replacing p' with x (for *observed score*) and p with t (for *true score*), then each individual score, x, reflects 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^2_x, is just

$$\sigma^2_x = \sigma^2_t + \sigma^2_e + 2\sigma_{te} = \sigma^2_t + \sigma^2_e.$$

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

$$\sigma_{xt} = \sigma^2_t + \sigma_{te} = \sigma^2_t$$

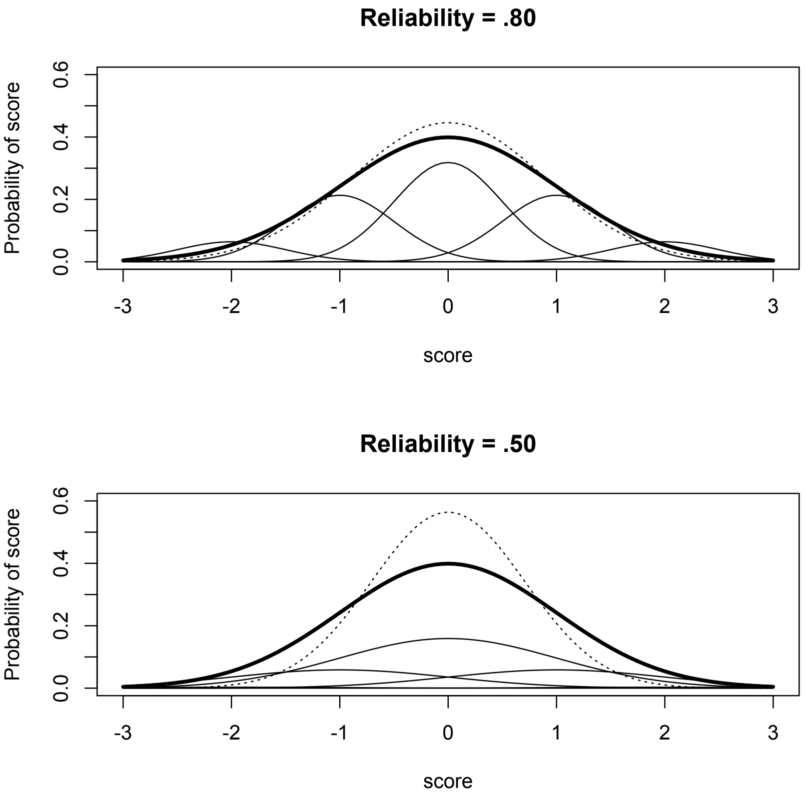
and the correlation of observed score with true score is

$$\rho_{xt} = \frac{\sigma_{xt}}{\sqrt{(\sigma^2_t + \sigma^2_e)(\sigma^2_t)}} = \frac{\sigma^2_t}{\sqrt{\sigma^2_x \sigma^2_t}} = \frac{\sigma_t}{\sigma_x}. \tag{7.1}$$

By knowing the correlation between observed score and true score, ρ_{xt}, and from the definition of linear regression (Eqs. 4.2,4.6) predicted true score, $\hat{t}$, for an observed x may be found from

$$\hat{t} = b_{t.x}x = \frac{\sigma^2_t}{\sigma^2_x}x = \rho^2_{xt}x. \tag{7.2}$$

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



Reliability = .80

Reliability = .50

Fig. 7.2 Observed score (heavy solid line) is made up of true score (dotted line) and error scores at each level of true score (light solid lines). The variance of true score more closely approximates that of observed score when the error variances are small and the reliability is greater.

7.1.1 Parallel Tests, Reliability, and Corrections 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, 1904b, 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

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} r_{q'q'}}}. \tag{7.3}$$

Rephrasing Spearman (1904b, 1910) in more current terminology (Lord and Novick, 1968; McDonald, 1999), *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 7.1). The correlation between two parallel tests will be

$$\rho_{p'_1 p'_2} = \rho_{p'p'} = \frac{\sigma_{p'_1 p'_2}}{\sqrt{\sigma^2_{p'_1}\sigma^2_{p'_2}}} = \frac{\sigma^2_p + \sigma_{pe_1} + \sigma_{pe_2} + \sigma_{e_1 e_2}}{\sigma^2_{p'}} = \frac{\sigma^2_p}{\sigma^2_{p'}}. \tag{7.4}$$

Expressing Equation 7.4 in terms of observed and true scores and comparing it to Equation 7.1 we see that 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

$$\rho_{xx} = \frac{\sigma^2_t}{\sigma^2_x} = \rho^2_{xt}. \tag{7.5}$$

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.

$$r_{pq} = \frac{\sigma_{pq}}{\sqrt{\sigma^2_p \sigma^2_q}} \tag{7.6}$$

and

$$r_{p'q'} = \frac{\sigma_{p'q'}}{\sqrt{\sigma^2_{p'}\sigma^2_{q'}}} = \frac{\sigma_{p+e'_1}\sigma_{q+e'_2}}{\sqrt{\sigma^2_{p'}\sigma^2_{q'}}} = \frac{\sigma_{pq}}{\sqrt{\sigma^2_{p'}\sigma^2_{q'}}} \tag{7.7}$$

but from Eq 7.5,

$$\sigma^2_p = \rho_{p'p'}\sigma^2_{p'} \tag{7.8}$$

and thus, by combining equation 7.6 with 7.7 and 7.8 the *unattenuated correlation* between p and q corrected for reliability is Spearman's equation 7.3

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'}r_{q'q'}}}. \tag{7.9}$$

As Spearman recognized, *correcting for attenuation* could show structures that otherwise, because of unreliability, would be hard to detect. A very thoughtful discussion of the necessity of correcting measures for attenuation has been offered by Schmidt and Hunter (1999) who suggest that all measures should be so corrected. Borsboom and Mellenbergh (2002) disagree and suggests that rather than apply *corrections for attenuation* from classical test theory, it is more appropriate to think in a structural modeling context. But as will be discussed in Chapter 10, this will lead to almost the same conclusion. An example of the power of correcting for attenuation may be seen in Table 7.1. The correlations below the diagonal represent observed correlations, the entries on the diagonal the reliabilities, and the entries above the diagonal, the correlations corrected for reliability using equation 7.9. The data were generated using the `sim.structural` function to represent three different latent variables with a particular structure and then the corrections for attenuation were made using the `correct.cor` function. Note how the structural relationships are much clearer when correcting for attenuation. That is, variables loading on the same factor have dis-attenuated correlations of unity, and the dis-attenuated correlations between variables loading on different factors reflects the correlations between the factors.

Table 7.1 Correlations can be corrected for attenuation using Equation 7.9. The raw correlations in the last matrix were created from the factor (fx) and structure matrices (Phi) shown at the top of the table using the `sim.structural` function. In the last matrix, raw correlations are shown below the diagonal, reliabilities on the diagonal, and disattenuated correlations above the diagonal. Note how the structural relationships are much clearer when correcting for attenuation. That is, variables loading on the same factor have dis-attenuated correlations of unity, and the dis-attenuated correlations between variables loading on different factors reflects the correlations between the factors.

```
  #define the observed variable factor loadings
> fx <- matrix(c(.9,.8,.6,rep(0,7),.6,.8,-.7,rep(0,8),.6,.5,.4),ncol=3)
> colnames(fx) <- colnames(Phi)
> rownames(fx) <- paste("V",1:8,sep="")
> fx

    F1   F2  F3
V1 0.9  0.0 0.0
V2 0.8  0.0 0.0
V3 0.6  0.6 0.0
V4 0.0  0.8 0.0
V5 0.0 -0.7 0.0
V6 0.0  0.0 0.6
V7 0.0  0.0 0.5
V8 0.0  0.0 0.4

#define the structural relationships
> Phi <- matrix(c(1,0,.707,0,1,rep(.707,3),1),ncol=3)
> colnames(Phi) <- rownames(Phi) <- paste("F",1:3,sep="")
> print(Phi,2)

    F1  F2  F3
F1 1.0 0.0 0.7
F2 0.0 1.0 0.7
F3 0.7 0.7 1.0

> r <- sim.structural(fx,Phi) #create a correlation matrix with known structure
> print (correct.cor(r$model,r$reliability),2) #correct for reliability

     V1   V2    V3    V4    V5    V6    V7    V8
V1 0.81 1.00  0.71  0.00  0.00  0.71  0.71  0.71
V2 0.72 0.64  0.71  0.00  0.00  0.71  0.71  0.71
V3 0.54 0.48  0.72  0.71 -0.71  1.00  1.00  1.00
V4 0.00 0.00  0.48  0.64 -1.00  0.71  0.71  0.71
V5 0.00 0.00 -0.42 -0.56  0.49 -0.71 -0.71 -0.71
V6 0.38 0.34  0.51  0.34 -0.30  0.36  1.00  1.00
V7 0.32 0.28  0.42  0.28 -0.25  0.30  0.25  1.00
V8 0.25 0.23  0.34  0.23 -0.20  0.24  0.20  0.16
```

However, defining reliability as the correlation between parallel tests still requires finding 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.

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).

7.1.2 Tau equivalent and congeneric tests

With three tests, the number of assumptions may be reduced, and if the tests are *tau equivalent* (individuals differ from each other in their true scores but each person has the same true score on each test) or *essentially tau equivalent* (tests differ in their true score means but not true score variance) then each test has the same covariance with true score), reliability for each of the three tests may be found Novick and Lewis (1967). *Tau equivalence* or at least essential tau equivalence is a basic (if unrealized) assumption of internal consistency estimates of reliability (see 7.2.3). Using the notation of Table 7.2, for τ equivalence, $\lambda_1 = \lambda_2 = \lambda_3$, but the σ_i^2 need not be the same.

With four tests, to find the reliability of each test, we need only assume that the tests all measure the same construct (to be "*congeneric*"), although possibly with different true score saturations ($\lambda_1 ... \lambda_4$) and error score variances (Lord and Novick, 1968). The set of observed variables and unknown parameters for each of four tests are shown in Table 7.2. When four variables are all measures (of perhaps different quality) of one common factor, the variables are said to be *congeneric*. The parameters may be estimated by exploratory or confirmatory factor analysis. The reliabilities are the communalities of each variable (the squared factor loadings) or 1- the uniqueness for each variable (Table 7.3). With three tests, the parameters can be estimated, but the model is said to be fully saturated, in that there are no extra degrees of freedom (six parameters are estimated by six observed variances and covariances. With four tests, there are two degrees of freedom (eight parameters are estimated from 10 observed variances and covariances).

There are multiple ways of finding the parameters of a set of congeneric tests. Table 7.3 shows the results of an exploratory factor analysis using the `factanal` function. In Chapter 10 this same model is fit using a *structural equation model* by the **sem** package. Given the loadings (λ_i) on the single latent factor, the reliability of each test is 1 - the uniqueness (or error variance) for that test.

$$r_{xx} = 1 - u^2 = \lambda_i^2$$

.

Table 7.2 Two parallel tests have two variances and one covariance. These allow us to estimate $\lambda_1 = \lambda_2$ and $\sigma^2_{e_1} = \sigma^2_{e_2}$ and the true score variance. The parameters of τ equivalent tests can be estimated if $\lambda_1 = \lambda_2 = \lambda_3$. For four congeneric tests, all parameters are free to vary.

	V1	V2	V3	V4	V1	V2	V3	V4
V1	s^2_1				$\lambda_1\sigma^2_t + \sigma^2_{e_1}$			
V2	s_{12}	s^2_2			$\lambda_1\lambda_2\sigma^2_t$	$\lambda_2\sigma^2_t + \sigma^2_{e_2}$		
V3	s_{13}	s_{23}	s^2_3		$\lambda_1\lambda_3\sigma^2_t$	$\lambda_2\lambda_3\sigma^2_t$	$\lambda_3\sigma^2_t + \sigma^2_{e_3}$	
V4	s_{14}	s_{24}	s_{34}	s^2_4	$\lambda_1\lambda_4\sigma^2_t$	$\lambda_2\lambda_3\sigma^2_t$	$\lambda_3\lambda_4\sigma^2_t$	$\lambda_4\sigma^2_t + \sigma^2_{e_4}$

Table 7.3 The congeneric model is a one factor model of the observed covariance or correlation matrix. The test reliabilities will be 1- the uniquenesses of each test. The correlation matrix was generated from a factor model with loadings of .9, .8, .7, and .6. Not surprisingly, a factor model correctly estimates these parameters. If this analysis is redone with sample values for three variables, the one factor model still fits perfectly (with 0 df), but the one factor model will not necessarily fit perfectly for four variables. The reliability of each test is then $\rho_{ii} = \lambda^2_i = 1 - u^2_i$. Thus, the reliabilities are .81, .64, .49, and .36 for $V_1 \dots V_4$ respectively. Although the `factanal` function reports the uniquenesses (u^2), the `fa` function in the *psych* package reports $h^2 = 1 - u^2$ as well.

```
> f <- c(.9,.8,.7,.6)
> r <- sim.structural(f)
> r

Call: sim.structural(fx = f)

 $model (Population correlation matrix)
     V1   V2   V3   V4
V1 1.00 0.72 0.63 0.54
V2 0.72 1.00 0.56 0.48
V3 0.63 0.56 1.00 0.42
V4 0.54 0.48 0.42 1.00

$reliability (population reliability)
[1] 0.81 0.64 0.49 0.36

> factanal(covmat=r$model,factors=1)

Call:
factanal(factors = 1, covmat = r$model)

Uniquenesses:
  V1   V2   V3   V4
0.19 0.36 0.51 0.64

Loadings:
   Factor1
V1 0.9
V2 0.8
V3 0.7
V4 0.6

               Factor1
SS loadings      2.300
Proportion Var   0.575

The degrees of freedom for the model is 2 and the fit was 0
```

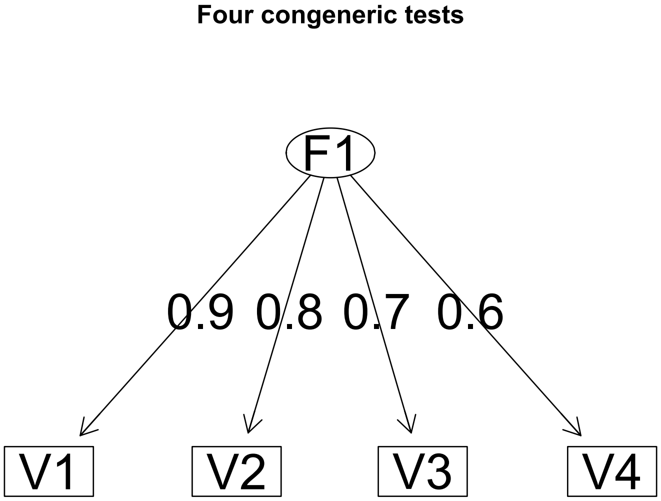


Fig. 7.3 Fitting four congeneric measures by one factor. There are four observed variances and six observed covariances. There are eight unknowns to estimate. This model was fit by a one factor exploratory factor model, although a one factor confirmatory model would work as well. The confirmatory solution using the **sem** package is discussed in Chapter 10. The graph was done using the `fa.diagram` function.

7.2 Reliability and internal structure

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 7.1), 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.

7.2.1 Split half reliability

If a test is split into two random halves, then the correlation between these two halves can be used to estimate the *split half reliability* of the total test. That is, two tests, $\mathbf{X}$, and a test just like it, $\mathbf{X'}$, with covariance, $\mathbf{C_{xx'}}$ can be represented as

$$\Sigma_{XX'} = \left(\begin{array}{ccc} \mathbf{V_x} & \vdots & \mathbf{C_{xx'}} \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots} \\ \mathbf{C_{xx'}} & \vdots & \mathbf{V_{x'}} \end{array}\right) \tag{7.10}$$

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:

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

and the structure of the two tests seen in Equation 7.10 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

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

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:

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc} \mathbf{V_{x_1}} & \vdots & \mathbf{C_{x_1x_2}} & \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_1x'_2}} \\ \multicolumn{3}{c|}{\cdots\cdots\cdots\cdots} & \multicolumn{3}{c}{\cdots\cdots\cdots\cdots} \\ \mathbf{C_{x_1x_2}} & \vdots & \mathbf{V_{x_2}} & \mathbf{C_{x_2x'_1}} & \vdots & \mathbf{C_{x_2x'_1}} \\ \hline \mathbf{C_{x_1x'_1}} & \vdots & \mathbf{C_{x_2x'_1}} & \mathbf{V_{x'_1}} & \vdots & \mathbf{C_{x'_1x'_2}} \\ \mathbf{C_{x_1x'_2}} & \vdots & \mathbf{C_{x_2x'_2}} & \mathbf{C_{x'_1x'_2}} & \vdots & \mathbf{V_{x'_2}} \end{array}\right)$$

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:

$$\Sigma_{XX'} = \left(\begin{array}{ccc|ccc}
\mathbf{V_{t_1}+V_{e_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} \\
\multicolumn{3}{c|}{\cdots\cdots\cdots\cdots\cdots\cdots} & \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\
\mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}+V_{e_1}} & \mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} \\
\hline
\mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}+V_{e'_1}} & \vdots & \mathbf{V_{t'_1}} \\
\mathbf{V_{t_1}} & \vdots & \mathbf{V_{t_1}} & \mathbf{V_{t'_1}} & \vdots & \mathbf{V_{t'_1}+V_{e'_1}}
\end{array}\right)$$

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

$$r_{xx'} = \frac{2r_1}{1+r_1} \tag{7.11}$$

This way of estimating the correlation of a test with a parallel test based upon the correlation of two split halves correcting for the fact they were half tests rather than full tests (the *split half reliability*)is a special case (n=2) of the the more general Spearman-Brown correction (Brown, 1910; Spearman, 1910).

$$r_{xx} = \frac{nr_1}{1+(n-1)r_1}. \tag{7.12}$$

It is important to remember that when finding the *split half reliability* that the observed correlation between the two halves needs to be adjusted by the *Spearman-Brown* prophecy formula (Equation 7.12) which for $n = 2$ is is just Equation 7.11.

7.2.2 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. (Lord (1955), made the distinction between “Type 1” sampling of people, “Type 2” sampling of items and “Type12” sampling of persons and 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.?

7.2.2.1 Correlation of an item with a 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

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

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

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

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

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

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

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

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

$$\lim_{n\to\infty} r^2_{x_jd} = \frac{\bar{c}}{\bar{v}}. \tag{7.13}$$

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 score (Eq 7.5) with the correlation of an item to the domain score (Eq 7.13). 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.

7.2.2.2 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

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

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

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

and the limit as n gets very large becomes

$$\lim_{n \to \infty} r^2_{kd} = \frac{k\bar{c}}{\bar{v} + (k-1)\bar{c}}. \tag{7.14}$$

This is an important conclusion: the squared correlation of a k item test with a very large domain will be a function of the number of items in the test (k) and the average covariance of the items within the test (and by assumption, the domain). Compare Eq 7.12 to Eq 7.14. The first, the *Spearman-Brown* prophecy formula estimates the reliability of a n-part test based upon the average correlation between the n parts. The second, the squared correlation of a test with the domain, estimates the fraction of test variance that is domain variance based upon the average item variance, the average item covariance, and the number of items. For standardized items, $\bar{v} = 1$ and $\bar{c} = \bar{r}$ the two equations are identical.

7.2.3 The internal structure of a test. Part 1: coefficient α

Although defined in terms of the correlation of a test with a test just like it, reliability can be estimated by the characteristics of the items within the test. The desire for an easy to use "magic bullet" based upon the domain sampling model has led to a number of solutions for estimating the reliability of a test based upon characteristics of the covariances of the items. All of these estimates are based upon classical test theory and assume that the covariances between items represents true covariance, but that the variances of the items reflect an unknown sum of true and unique variance. From the variance of a composite (Eq 5.1), it is known that the variance of a total test, σ^2_x made up of a sum of individual items, x_i is

$$\sigma^2_x = \sum_{i \neq j} \sigma_{x_i x_j} + \sum \sigma^2_{x_i}. \tag{7.15}$$

After earlier work introduced various shortcuts (Kuder and Richardson, 1937) that did not require finding the covariances, Guttman (1945), in an attempt to formalize the estimation of reliability, proposed six lower bounds for reliability, ρ_{xx}, that took advantage of the internal structure of the test

$$\rho_{xx} = \frac{\sigma^2_t}{\sigma^2_x} = \frac{\sigma^2_x - \sigma^2_e}{\sigma^2_x}.$$

Each one successively modifies the way that the error variance of the items are estimated. Unfortunately, although many psychometricians deplore its use, one of these estimates, λ_3 (Guttman, 1945), also known as coefficient *alpha* (Cronbach, 1951), is by far and away the most common estimate of reliability. The appeal of α is both that it is easy to compute, is easy to understand, and is available in all statistical programs (including the **psych** package in R). To understand the appeal of α, as well as the reasons not to rely solely on it, it is necessary to consider a number of alternatives.

Although splitting a test into two and calculating the reliability based upon the correlation between the two halves corrected by the *Spearman-Brown* formula was one way of estimating reliability, the result would depend upon the particular split. When considering the number of possible split halves of an k-item test ($\frac{k!}{2(\frac{k}{2}!)^2}$), Kuder and Richardson (1937) introduced a short cut formula for reliability in terms of the total test variance, σ_x^2, and the average item variance, $\overline{pq}$ of a true-false test where p_i and q_i represent the percentage passing and failing any one item. The *Kuder-Richardson 20* formula could be calculated without finding the average covariance or correlation required by Eq 7.12 and Eq 7.14 by taking advantage of the identity of Eq 7.15

$$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{k}{k-1}\frac{\sigma_x^2 - k\overline{pq}}{\sigma_x^2}. \tag{7.16}$$

Functionally, this is finding the total true score variance of an n item test as k^2 times the average covariance, $\sigma_t^2 = k^2\bar{c}$, and recognizing that the total test variance represents the sum of the item covariances plus the sum of the item error variances $\sigma_x^2 = k^2\bar{c} + k\bar{\sigma}_e^2$. Taking the total test variance, subtracting the sum of the item variances (k times the average variance) and dividing by $k(k-1)$ (the number of covariance terms) gives an average covariance. Multiplying this by k^2 gives the total true score variance which when divided by the total test variance is the test reliability. The clear advantage of *KR20* was that it could be calculated without finding the inter-item covariances or correlations, but just the total test variance and the average item variance.

An even easier shortcut was to estimate the average variance by finding the variance of the average item (the 21st formula of Kuder and Richardson, 1937) now known as *KR21*. That is, by finding the percent passing the average item, $\bar{p}$, it is possible to find the variance of the average item, $\bar{p}\bar{q}$, which will be a positively biased estimate of the average item variance and thus a negatively biased estimate of reliablity. (Unless all items have equal probabilities of success, the variance of the average item will be greater than the average of the variance of the items).

$$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{k}{k-1}\frac{\sigma_x^2 - k\bar{p}\bar{q}}{\sigma_x^2}.$$

Coefficient alpha (Cronbach, 1951) is a straight forward generalization of *KR20* to the case of non-dichotomous as well as dichotomous items. Letting σ_i^2 represent the variance of $item_i$, and σ_x^2 the variance of the total total test, then the average covariance of an item with any other item is

$$\bar{c} = \frac{\sigma_x^2 - \sum\sigma_i^2}{k(k-1)}$$

and thus the ratio of the total covariance in the test to the total variance in the test is

$$\alpha = r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{k^2\frac{\sigma_x^2 - \sum\sigma_i^2}{k(k-1)}}{\sigma_x^2} = \frac{k}{k-1}\frac{\sigma_x^2 - \sum\sigma_i^2}{\sigma_x^2} \tag{7.17}$$

which is just KR20 but using the sum of the item variances rather than n times the average variance and allows for non-dichotomous items.

An alternate way of finding *coefficient alpha* based upon finding the average covariance between items is to consider the ratio of the total covariance of the test to the total variance

$$\alpha = \frac{k^2\bar{c}}{k\bar{v}+k(k-1)\bar{c}} = \frac{k\bar{c}}{\bar{v}+(k-1)\bar{c}} \tag{7.18}$$

which for standardized items is just

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}}. \tag{7.19}$$

In the preceding few pages, six important equations have been introduced. All six of these equations reflect characteristics of the reliability of a test and how observed scores relate to true score. Reliability, defined as the correlation between two parallel forms of a test, is the same as the squared correlation of either test with true score and is the amount of true score variance in the test. Reliability is an increasing function of the correlation between random split halves of either test. Coefficient α, derived from Eq 7.19, is the same as the reliability as estimated by the Spearman-Brown prophecy formula (Eq 7.12), but is derived from domain sampling principles, and as is seen in Eq 7.18 is the same as the squared correlation of a k-item test with the domain (Eq 7.14). As seen in Eq 7.17, coefficient α is the same as the the reliability of test found for dichotomous items using formula KR20, Eq 7.16. That is, all six of these equations, although derived in different ways by different people have identical meanings.

As an example of finding coefficient α, consider the five neuroticism items from the `bfi` data set. This data set contains 25 items organized in five sets of five items to measure each of the so-called “Big Five” dimensions of personality (Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness). All five of these are scored in the same direction and can be analyzed without reverse keying any particular item. The analysis in Table 7.7 reports both the values of α based upon the covariances (raw) as well as the correlations (standardized). In addition, Guttman's coefficient λ_6 (discussed below) of the reliability based upon the squared multiple correlation (smc) for each items as well as the average intercorrelation of the items are reported. The correlation of each item with the total scale is reported in two forms, the first is just the raw correlation which reflects item overlap, the second corrects for item overlap by replacing each item's correlation with itself (1.0) with the estimated item reliability based upon the smc.

7.2.4 *The internal structure of a test. Part 2: Guttman's lower bounds of reliability*

Although arguing that reliability was only meaningful in the case of test-retest, Guttman (1945) may be credited with introducing a series of lower bounds for reliability, $\lambda_1 \dots \lambda_6$, each based upon the item characteristics of a single test. These six have formed the base for most of the subsequent estimates of reliability based upon the item characteristics of a single test. Of the six, λ_3 is most well known and was called *coefficient alpha* or α by Cronbach (1951). All of these measures decompose the total test variance, V_x, into two parts, that associated with error, V_e and what ever is left over, $V_x - V_e$. (although not using the term true score, this is implied). Reliability is then just

$$r_{xx} = \frac{V_x - V_e}{V_x} = 1 - \frac{V_e}{V_x}. \tag{7.20}$$

Table 7.4 Coefficient α may be found using the **alpha** function in *psych*. The analysis is done for 5 neuroticism items taken from the **bfi** data set.

```
> 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
```

The problem then becomes how to estimate the error variance, V_e.

Consider the case of a test made up of 12 items, all of which share 20% of their variance with a general factor, but form three subgroups of (6, 4 and 2) items which share 30%, 40% or 50% of their respective variance with some independent group factors. The remaining item variance is specific or unique variance (Table 7.5, Figure 7.4). An example of this kind of a test might be a measure of Neuroticism, with a broad general factor measuring general distress, group factors representing anger, depression, and anxiety, and specific item variance. For standardized items this means that the correlations between items in different groups are .2, those within groups are .5, .6 or .7 for groups 1, 2 and 3 respectively. The total test variance is thus

$$V_t = 12^2 * .2 + 6^2 * .3 + 4^2 * .4 + 2^2 * .5 + 6 * .5 + 4 * .4 + 2 * .3 = 53.2$$

and the error variance V_e is

$$V_e = 6 * .5 + 4 * .4 + 2 * .3 = 5.2$$

for a reliability of

$$r_{xx} = 1 - \frac{5.2}{53.2} = .90.$$

Using the data matrix formed in Table 7.5 and shown in Figure 7.4, we can see how various estimates of reliability perform.

The first Guttman lowest bound, λ_1 considers that all of an item variance is error and that only the interitem covariances reflect true variability. Thus, λ_1 subtracts the sum of the

Table 7.5 A hypothetical 12 item test can be thought of as the sum of the general variances of all items, group variances for some items, and specific variance for each item. (See Figure 7.4). An example of this kind of a test might be a measure of Neuroticism, with a broad general factor, group factors representing anger, depression, and anxiety, and specific item variance.

```
> general  <- matrix(.2,12,12)
> group <- super.matrix( super.matrix(matrix(.3,6,6),matrix(.4,4,4)),matrix(.5,2,2))
> error <- diag(c(rep(.5,6),rep(.4,4),rep(.3,2)),12,12)
> Test <- general + group + error
> colnames(Test ) <- rownames(Test) <-  paste("V",1:12,sep="")
> round(Test,2)
```

```
     V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12
V1  1.0 0.5 0.5 0.5 0.5 0.5 0.2 0.2 0.2 0.2 0.2 0.2
V2  0.5 1.0 0.5 0.5 0.5 0.5 0.2 0.2 0.2 0.2 0.2 0.2
V3  0.5 0.5 1.0 0.5 0.5 0.5 0.2 0.2 0.2 0.2 0.2 0.2
V4  0.5 0.5 0.5 1.0 0.5 0.5 0.2 0.2 0.2 0.2 0.2 0.2
V5  0.5 0.5 0.5 0.5 1.0 0.5 0.2 0.2 0.2 0.2 0.2 0.2
V6  0.5 0.5 0.5 0.5 0.5 1.0 0.2 0.2 0.2 0.2 0.2 0.2
V7  0.2 0.2 0.2 0.2 0.2 0.2 1.0 0.6 0.6 0.6 0.2 0.2
V8  0.2 0.2 0.2 0.2 0.2 0.2 0.6 1.0 0.6 0.6 0.2 0.2
V9  0.2 0.2 0.2 0.2 0.2 0.2 0.6 0.6 1.0 0.6 0.2 0.2
V10 0.2 0.2 0.2 0.2 0.2 0.2 0.6 0.6 0.6 1.0 0.2 0.2
V11 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 1.0 0.7
V12 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.7 1.0

> sum(Test)
[1] 53.2
> sum(general)
[1] 28.8
> sum(group)
[1] 19.2
> sum(error)
[1] 5.2
```

diagonal of the observed item covariance matrix from the total test variance:

$$\lambda_1 = 1 - \frac{tr(\mathbf{V_x})}{V_x} = \frac{V_x - tr(\mathbf{V}_x)}{V_x}. \tag{7.21}$$

This leads to an estimate of

$$\lambda_1 = 1 - \frac{12}{53.2} = \frac{41.2}{53.2} = .774.$$

The second bound, λ_2 replaces the diagonal with a function of the square root of the sums of squares of the off diagonal elements. Let $C_2 = \mathbf{1}(\mathbf{V} - diag(\mathbf{V}))^2\mathbf{1}'$, then

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{n}{n-1}C_2}}{V_x} = \frac{V_x - tr(\mathbf{V}_x) + \sqrt{\frac{n}{n-1}C_2}}{V_x}. \tag{7.22}$$

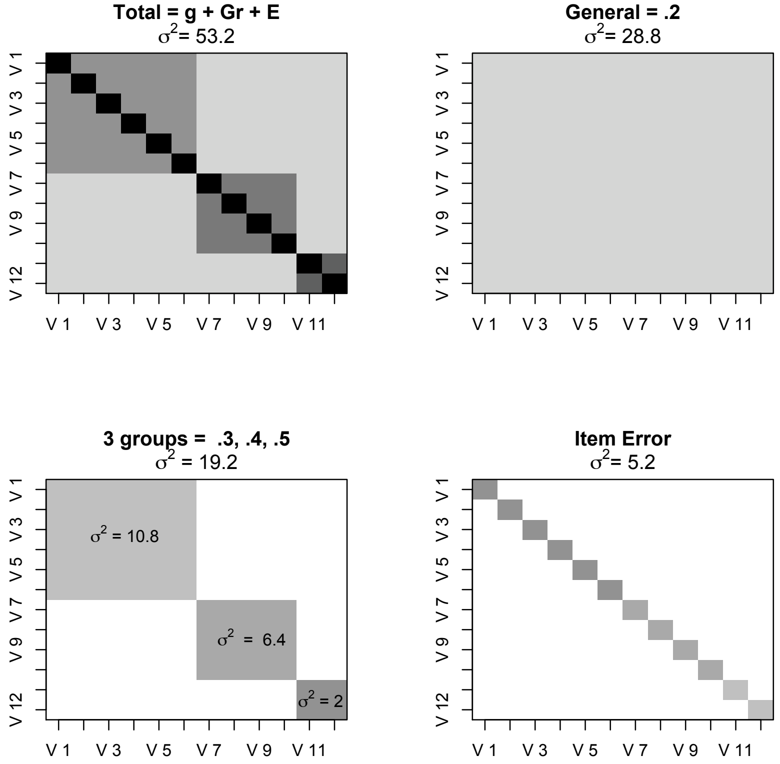


Fig. 7.4 The total variance of a test may be thought of as composed of a general factor, several group factors, and specific item variance. The problem in estimating total reliability is to determine how much of the total variance is due to specific or unique item variance. Shading represent the magnitude of the correlation. All items are composed of 20% general factor variance, the first six items have a group factor accounting for 30% of their variance, the next four items have a stronger group factor accounting for 40% of their variance, and the last two items define a very powerful group factor, accounting for 50% of their variance. The items are standardized, and thus have total variance of 1.

Effectively, this is replacing the diagonal with n * the square root of the average squared off diagonal element.

$$\lambda_2 = .774 + \frac{\sqrt{\frac{12}{11}16.32}}{53.2} = .85$$

Guttman's 3rd lower bound, λ_3, also modifies λ_1 and estimates the true variance of each item as the average covariance between items and is, of course, the same as Cronbach's α.

$$\lambda_3 = \lambda_1 + \frac{\frac{V_X - tr(\mathbf{V}_X)}{n(n-1)}}{V_X} = \frac{n\lambda_1}{n-1} = \frac{n}{n-1}\left(1 - \frac{tr(\mathbf{V})_x}{V_x}\right) = \frac{n}{n-1}\frac{V_x - tr(\mathbf{V}_x)}{V_x} = \alpha \qquad (7.23)$$

This is just replacing the diagonal elements with the average off diagonal elements. $\lambda_2 \geq \lambda_3$ with $\lambda_2 > \lambda_3$ if the covariances are not identical.

$$\lambda_3 = \alpha = \frac{12}{11}\left(1 - \frac{12}{53.2}\right) = \frac{12}{11}\frac{(53.2 - 12)}{53.2} = .84$$

As pointed out by Ten Berge and Zegers (1978), λ_3 and λ_2 are both corrections to λ_1 and this correction may be generalized as an infinite set of successive improvements.

$$\mu_r = \frac{1}{V_x}\left(p_o + (p_1 + (p_2 + \dots (p_{r-1} + (p_r)^{1/2})^{1/2} \dots)^{1/2})^{1/2}\right), r = 0, 1, 2, \dots \tag{7.24}$$

where

$$p_h = \sum_{i \neq j} \sigma_{ij}^{2h}, h = 0, 1, 2, \dots r - 1$$

and

$$p_h = \frac{n}{n-1}\sigma_{ij}^{2h}, h = r.$$

Clearly $\mu_0 = \lambda_3 = \alpha$ and $\mu_1 = \lambda_2$. $\mu_r \geq \mu_{r-1} \geq \dots \mu_1 \geq \mu_0$, although the series does not improve much after the first two steps (Ten Berge and Zegers, 1978).

Guttman's fourth lower bound, λ_4 was originally proposed as any spit half reliability (Guttman, 1945) but has been interpreted as the greatest split half reliability (Jackson and Agunwamba, 1977). If $\mathbf{X}$ is split into two parts, $\mathbf{X}_a$ and $\mathbf{X}_b$, with covariance c_{ab} then

$$\lambda_4 = 2\left(1 - \frac{V_{X_a} + V_{X_b}}{V_X}\right) = \frac{4c_{ab}}{V_x} = \frac{4c_{ab}}{V_{X_a} + V_{X_b} + 2c_{ab}V_{X_a}V_{X_b}}. \tag{7.25}$$

If the two parts are standardized, the covariance becomes a correlation, the two variances are equal and λ_4 is just the normal split half reliability, but in this case, of the most similar splits. In the case of the example, there are several ways that lead to a "best split", but any scale made up 3 items from the first six, two from the second four and one from the last two will correlate .82 with the corresponding other scale. Correcting this for test length using the *Spearman-Brown* correction leads to

$$\lambda_4 = \frac{2 * .82}{1 + .82} = .90.$$

In the general case of splits with unequal variances, it is better to use Equation 7.25 rather than 7.12.

λ_5, Guttman's fifth lower bound, replaces the diagonal values with twice the square root of the maximum (across items) of the sums of squared interitem covariances

$$\lambda_5 = \lambda_1 + \frac{2\sqrt{\bar{C}_2}}{V_X}. \tag{7.26}$$

Although superior to λ_1, λ_5 underestimates the correction to the diagonal. A better estimate would be analogous to the correction used in λ_3:

$$\lambda_{5+} = \lambda_1 + \frac{n}{n-1}\frac{2\sqrt{\bar{C}_2}}{V_X}. \tag{7.27}$$

Guttman's final bound considers the amount of variance in each item that can be accounted for the linear regression of all of the other items (the squared multiple correlation or smc), or more precisely, the variance of the errors, e_j^2, and is

$$\lambda_6 = 1 - \frac{\sum e_j^2}{V_x} = 1 - \frac{\sum(1 - r_{smc}^2)}{V_x} \tag{7.28}$$

Although the smc used in finding Guttman's λ_6 is normally found by using just the other items in the particular scale, in a multiple scale inventory, the concept can be generalized to consider the smc based upon all the other items. In the *psych* package this is labeled as λ_{6+}. Using the `smc` function to find the smcs for each item and them summing them across all items

$$\lambda_6 = 1 - \frac{6.51}{53.2} = .878.$$

Yet another estimate that has been proposed for the reliability of a principal component (Ten Berge and Hofstee, 1999) unfortunately also uses λ_1 as a symbol, but this time as the magnitude of the the eigenvalue of the principal component

$$\alpha_{pc} = \frac{n}{n-1}(1 - \frac{1}{\lambda_1}). \tag{7.29}$$

$$\alpha_{pc} = \frac{12}{11}\left(1 - \frac{1}{4.48}\right) = .847.$$

The discussion of various lower bounds of reliability seemed finished when Jackson and Agunwamba (1977) and Bentler and Woodward (1980) introduced their "greatest lower bound", or glb. Woodhouse and Jackson (1977) organized Guttman's six bounds into a series of partial orders, and provided an algorithm for estimating the glb of Jackson and Agunwamba (1977). An alternative algorithm was proposed by Bentler and Woodward (1980) and discussed by Sijtsma (2008). Unfortunately, none of these authors considered ω_t (see below), which tends to exceed the glbs reported in the various discussions of the utility of the glb (Revelle and Zinbarg, 2009).

The Guttman statistics as well as those discussed Ten Berge and Zegers (1978) by may be found using the `guttman` function in *psych*.

7.2.5 The internal structure of a test. Part 3: coefficients α, β, ω_h and ω_t

Two additional coefficients, ω_h and ω_t, that were not considered by either Guttman (1945) or Cronbach (1951) were introduced by McDonald (1978, 1999). These two coefficients require factor analysis to estimate, but are particularly useful measures of the structure of a test. McDonald's ω_t is similar to Guttman's λ_6, but uses the estimates of uniqueness (u_j^2) for each item from factor analysis to find e_j^2. This is based on a decomposition of the variance of a test score, V_x, into four parts: that due to a general factor, $\mathbf{g}$, that due to a set of group factors, $\mathbf{f}$, (factors common to some but not all of the items), specific factors, $\mathbf{s}$ unique to each item, and $\mathbf{e}$, random error. (Because specific variance can not be distinguished from random error unless the test is given at least twice, McDonald (1999) combines these both into error).

Letting

$$\mathbf{x} = \mathbf{cg} + \mathbf{Af} + \mathbf{Ds} + \mathbf{e} \tag{7.30}$$

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

$$h_j^2 = c_j^2 + \sum f_{ij}^2 \tag{7.31}$$

and the unique variance for the item

$$u_j^2 = \sigma_j^2(1 - h_j^2) \tag{7.32}$$

may be used to estimate the test reliability. That is, if h_j^2 is the communality of item$_j$, based upon general as well as group factors, then for standardized items, $e_j^2 = 1 - h_j^2$ and

$$\omega_t = \frac{\mathbf{1cc'1'} + \mathbf{1AA'1'}}{V_x} = 1 - \frac{\sum(1-h_j^2)}{V_x} = 1 - \frac{\sum u^2}{V_x} \tag{7.33}$$

Because $h_j^2 \geq r_{smc}^2$, $\omega_t \geq \lambda_6$. For the example data set, the uniquenesses may be found by factor analysis (Table 7.6) and their sum is 5.2 (compare with Figure 7.4). Thus,

$$\omega_t = 1 - \frac{5.2}{53.2} = .90.$$

McDonald introduced another reliability coefficient, also labeled ω, based upon the saturation of the general factor. It is important to distinguish here between the two ω coefficients of McDonald (1978) and (McDonald, 1999, Equation 6.20a), ω_t and ω_h. While the former is based upon the sum of squared loadings on all the factors, the latter is based upon the sum of the squared loadings on the general factor, **g**. For a correlation matrix, **R** with general factor **g** with loadings **c**

$$\omega_h = \frac{\mathbf{1cc'1}}{V_x} = \frac{(\sum \Lambda_i)^2}{\sum\sum R_{ij}}. \tag{7.34}$$

That is, ω_h is the ratio of the sum of correlations reproduced by the general factor to the sum of all correlations. It is the percentage of a correlation matrix associated with the general factor. For the example,

$$\omega_h = \frac{5.366563^2}{53.2} = \frac{28.8}{53.2} = .54.$$

As is true for the other estimates of reliability, because the variance associated with the uniquenesses of each item becomes a smaller and smaller fraction of the test as the test becomes longer, ω_t will increase as a function of the number of variables and tend asymptotically towards 1.0. However, ω_h will not, and rather will tend towards a limit of

$$\omega_{h_\infty} = \frac{\mathbf{1cc'1}}{\mathbf{1cc'1} + \mathbf{1AA'1'}}. \tag{7.35}$$

ω_h is particularly important when evaluating the importance and reliability of the *general factor* of a test, while ω_t is an estimate of the total reliable variance in a test. As was discussed earlier (6.3.4) measures of cognitive ability have long been analyzed in terms of lower order factors (group factors) as well as a higher order, general factor **?**Horn and Cattell (1966, 1982). More recently, this approach has also been applied to the measurement of personality

Table 7.6 The `omega` function does a factor analysis followed by an oblique rotation and extraction of a general factor using the *schmid-leiman* transformation Schmid and Leiman (1957). The sum of the uniquenesses is used to find ω_t and the squared sum of the g loadings to find ω_h).

```
> omega(Test)

Omega
Call: omega(m = Test)
Alpha:                 0.84
G.6:                   0.88
Omega Hierarchical:    0.54
Omega H asymptotic:    0.6
Omega Total            0.9

Schmid Leiman Factor loadings greater than  0.2
         g  F1*  F2*  F3*  h2  u2
V1    0.45 0.55           0.5 0.5
V2    0.45 0.55           0.5 0.5
V3    0.45 0.55           0.5 0.5
V4    0.45 0.55           0.5 0.5
V5    0.45 0.55           0.5 0.5
V6    0.45 0.55           0.5 0.5
V7    0.45      0.63      0.6 0.4
V8    0.45      0.63      0.6 0.4
V9    0.45      0.63      0.6 0.4
V10   0.45      0.63      0.6 0.4
V11   0.45           0.71 0.7 0.3
V12   0.45           0.71 0.7 0.3

With eigenvalues of:
  g F1* F2* F3*
2.4 1.8 1.6 1.0

general/max  1.33   max/min =   1.8
The degrees of freedom for the model is 33  and the fit was  0

Measures of factor score adequacy
                                                 g  F1*  F2*  F3*
Correlation of scores with factors            0.74 0.77 0.81 0.81
Multiple R square of scores with factors      0.55 0.60 0.66 0.66
Minimum correlation of factor score estimates 0.10 0.20 0.31 0.33
```

traits such as anxiety which shows lower order factors as well as higher order one (Chen et al., 2006; Zinbarg and Barlow, 1996; Zinbarg et al., 1997). For tests that are thought to have a higher order structure, measures based upon just the average interitem correlation, α or λ_6, are not appropriate. Coefficients that reflect the structure such as ω_h and ω_t are more appropriate. If a test is composed of relatively homogeneous items then α and λ_6 will provide very similar estimates to ω_h and ω_t. $\omega_h, \omega_{h_\infty}, \omega_t, \alpha$ and λ_6 may all be found using the `omega` function (Table 7.6).

ω_h is an estimate of the general factor saturation of a test based upon a factor analytic model. An alternative estimate, coefficient β Revelle (1979), uses hierarchical cluster analysis to find the two most unrelated split halves of the test and then uses the implied inter-group itemcorrelation to estimate the total variance accounted for by a general factor. This is based

Table 7.7 Four data sets with equal α reliability estimates but drastically different structures. Each data set is assumed to represent two correlated clusters. The between cluster correlations are .45, .32, .14, and 0. For all four data sets, $\alpha = .72$. Because the within cluster factor loadings are identical within each set, the two estimates of general factor saturation, β and ω_h, are equal. They are .72, .48, .25, and 0 for sets S1, S2, S3 and S4 respectively. Figure 7.5 displays these matrices graphically using `cor.plot`.

S1

	V1	V2	V3	V4	V5	V6
V1	1.0	0.3	0.3	0.3	0.3	0.3
V2	0.3	1.0	0.3	0.3	0.3	0.3
V3	0.3	0.3	1.0	0.3	0.3	0.3
V4	0.3	0.3	0.3	1.0	0.3	0.3
V5	0.3	0.3	0.3	0.3	1.0	0.3
V6	0.3	0.3	0.3	0.3	0.3	1.0

S2

V1	V2	V3	V4	V5	V6
1.00	0.45	0.45	0.20	0.20	0.20
0.45	1.00	0.45	0.20	0.20	0.20
0.45	0.45	1.00	0.20	0.20	0.20
0.20	0.20	0.20	1.00	0.45	0.45
0.20	0.20	0.20	0.45	1.00	0.45
0.20	0.20	0.20	0.45	0.45	1.00

S3

	V1	V2	V3	V4	V5	V6
V1	1.0	0.6	0.6	0.1	0.1	0.1
V2	0.6	1.0	0.6	0.1	0.1	0.1
V3	0.6	0.6	1.0	0.1	0.1	0.1
V4	0.1	0.1	0.1	1.0	0.6	0.6
V5	0.1	0.1	0.1	0.6	1.0	0.6
V6	0.1	0.1	0.1	0.6	0.6	1.0

S4

V1	V2	V3	V4	V5	V6
1.00	0.75	0.75	0.00	0.00	0.00
0.75	1.00	0.75	0.00	0.00	0.00
0.75	0.75	1.00	0.00	0.00	0.00
0.00	0.00	0.00	1.00	0.75	0.75
0.00	0.00	0.00	0.75	1.00	0.75
0.00	0.00	0.00	0.75	0.75	1.00

upon the observation that the correlation between the two worst splits reflects the covariances of items that have nothing in common other that what is common to all the items in the test. For the example data set, the two most unrelated parts are formed from the first 10 items and the last two items. The correlation between these two splits is .3355 which implies an average correlation of the items between these two halves of .20. These correlations reflect the general factor saturation and when corrected for test length implies that the general saturation of this 12 item test is 144* .2 = 28.8. The total test variance is 53.2 and thus

$$\beta = \frac{12 * 12 * .2}{53.2} = .54.$$

Although in the case of equal correlations within groups and identical correlations between groups, ω_h and β are identical, this is not the case for group factors with unequal general factor loadings. Whether β or ω_h will be greater depends upon the specific pattern of the general factor loadings (Zinbarg et al., 2005).

7.3 A comparison of internal consistency estimates of reliability

If there are so many different measures of reliability, the question to ask is which reliability estimate should be used, and why. Consider the four example data sets in Table 7.7 (shown graphically in Figure 7.5). All four of these data sets (S1 ... S4) have equal average correlations (.3) and thus identical values of coefficient α (.72). However, by looking at the correlations, it is clear that the items in S1 represent a single construct, with all items having equal

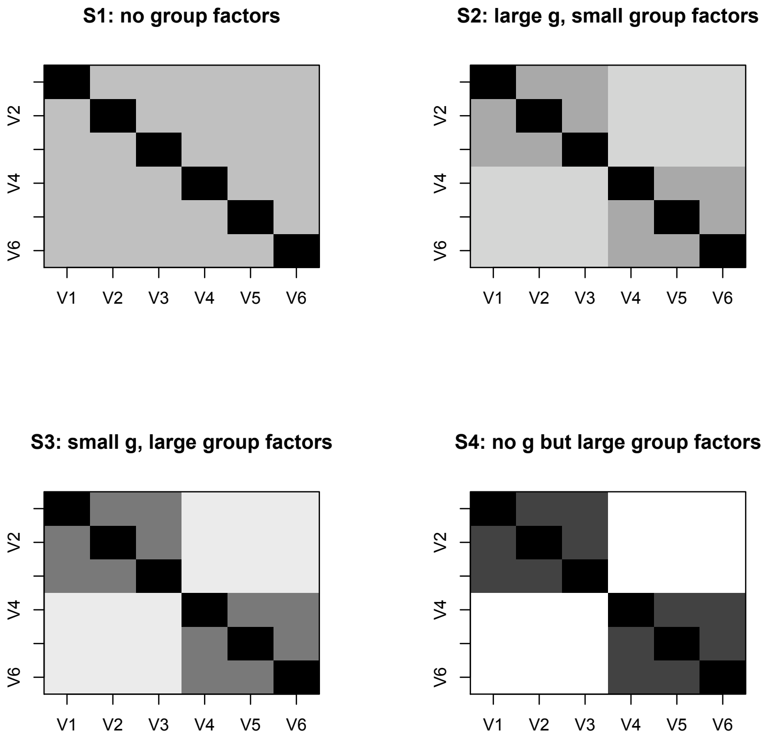


Fig. 7.5 The correlation matrices of Table 7.7 may be represented graphically using the `cor.plot` function. Although all four matrices have equal α reliabilities (.72), they differ drastically in the saturation of a general factor ($omega_h = \beta =$ .72, .48, .25 and .00.)

correlations of .3. The items in set S-4, on the other hand, form two distinct groups, even though the average intercorrelation of all the items remains .3 and α remains .72. Some of the other estimates of reliability discussed above, on the other hand, vary from sets S1 to S4 (Table 7.8). In particular, the two ways of estimating the *general factor saturation*, ω_h and β go from .72 when the test is unifactorial (S-1) to 0 when it contains two independent factors (S-4). ω_t, on the other hand, increases from .72 in the case of a unifactorial test (S-1) to .90 in the case of a test containing two independent factors (S-4).

To understand how Guttman's bounds relate to each other and to ω_h, ω_t, and β, it is useful to consider the "Test" data from Table 7.5 as well as the four sample correlation matrices from Table 7.7 as well as three demonstrations of a higher order structure, one simulated and two from the `bifactor` data set in the *psych* package (see Table 7.8). The simulated data set (S.9) was created using the `sim.hierarchical` function to demonstrate a hierarchical factor model as discussed by Jensen and Weng (1994) and shown earlier (see Table 6.13). The two real data sets are the *Thurstone* example from McDonald (1999) of 9 cognitive variables used to show a clear bifactor (hierarchical) structure. The second example of a bifactor structure

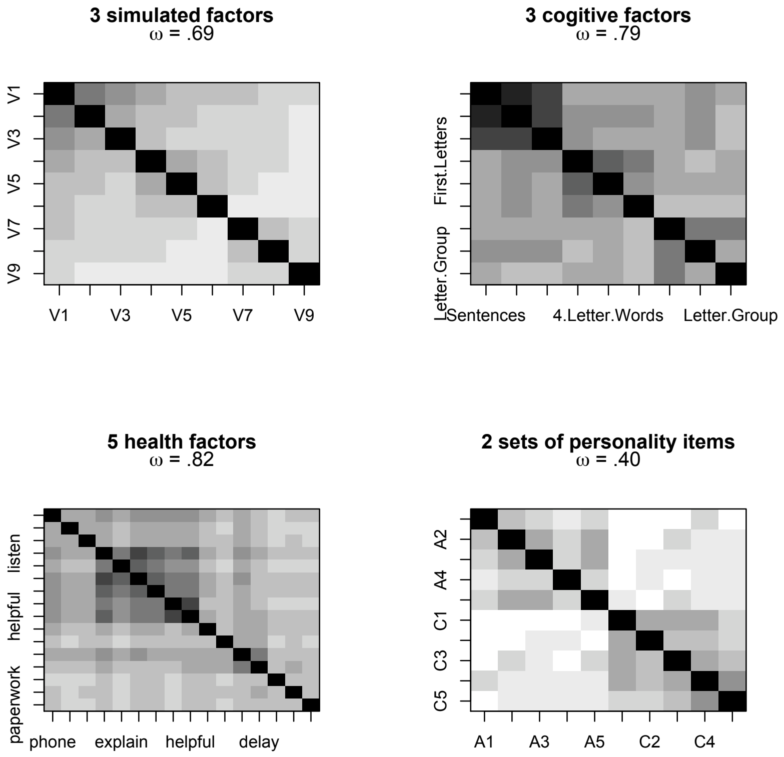


Fig. 7.6 One simulated and three real data sets. The first is an example of hierarchical structure created by the `sim.hierachical` function based upon an article by Jensen and Weng (1994). The second and third data sets, *Thurstone* and *Reise* are from the `bifactor` data set and are nine cognitive variables adapted by Bechtoldt (1961) from Thurstone and Thurstone (1941), the last is a set of 10 items thought to measure two different traits (Agreeableness and Conscientiousness) taken from the `bfi` data set.

are 14 health related items from Reise et al. (2007). The last real example, the BFI, uses 10 items from the `bfi` dataset (examples of the "Big Five Inventory") and is an example of two distinct constructs incorrectly combined into one scale. The 10 items from the BFI represent five items measuring "Agreeableness" and five measuring "Conscientiousness". Normally seen as separate traits, they are included as example of how a large α is not a sign of a single dimension.

Comparing the S1 $\dots$ S4 data sets, alpha is equal to all other measures and superior to λ_4 and λ_6 if there is exactly one factor in the data (S1). But as the general factor becomes less important, and the group factors more important (S2 ... S4), α does not change, but the other Guttman coefficients do. λ_6, based upon the *smc* as an estimate of item reliability, underestimates the reliability for the completely homogeneous set, but exceeds α as the test becomes more multifaceted.

Table 7.8 Comparison of 13 estimates of reliability. Test is the data set from Table 7.5. The next four data sets are S1-S4 from Table 7.7 and Figure 7.5. S.9 is a simulated hierarchical structure using the `sim.hierarchical` function based upon Jensen and Weng (1994), T.9 is the 9 cognitive variables used by McDonald (1999), R14 is 14 health related items from Reise et al. (2007), the BFI is an example of two distinct constructs "incorrectly" combined into one scale. (See Figure 7.6). $\lambda_1 \dots \lambda_6$ are the Guttman (1945) bounds found by the `guttman` function as are $\mu_0 \dots \mu_3$ from Ten Berge and Zegers (1978), ω_h and ω_t are from McDonald (1999) and found by the `omega` function, β is from Revelle (1979) and found by the `ICLUST` function. Because β and ω_h reflect the general factor saturation they vary across the S-1 ... S-4 data sets and are much lower than α or ω_t for incorrectly specified scales such as the BFI.

Estimate	Test	S-1	S-2	S-3	S-4	S.9	T.9	R14	BFI
β (min)	.54	.72	.48	.25	.00	.57	.76	.79	.40
ω_h	.54	.72	.48	.25	.00	.69	.74	.78	.36
ω_{h_∞}	.60	1.00	.62	.29	.00	.86	.79	.85	.47
λ_1	.77	.60	.60	.60	.60	.68	.79	.85	.65
$\lambda_3(\alpha, \mu_0)$	.84	.72	.72	.72	.72	.76	.89	.91	.72
α_{pc}	.85	.72	.72	.72	.72	.77	.89	.91	.73
$\lambda_2(\mu_1)$	.85	.72	.73	.75	.79	.77	.89	.91	.74
μ_2	.86	.72	.73	.76	.80	.77	.90	.91	.74
μ_3	.86	.72	.73	.76	.80	.77	.90	.91	.74
λ_5	.82	.69	.70	.72	.74	.75	.87	.89	.71
λ_6 (smc)	.88	.68	.72	.78	.86	.76	.91	.92	.75
λ_4 (max)	.90	.72	.76	.83	.89	.76	.93	.93	.82
glb	.90	.72	.76	.83	.89	.76	.93	.93	.82
ω_t	.90	.72	.78	.84	.90	.86	.93	.92	.77

In that reliability is used to correct for attenuation (equation 7.3), underestimating the reliability will lead to an over estimate of the unattenuated correlation and overestimating the reliability will lead to an under estimate of the unattenuated correlation. Choosing the proper reliability coefficient is therefore very important and should be guided by careful thought and strong theory. In the case in which our test is multidimensional and several of the dimensions contribute to the prediction of the criterion of interest, α will underestimate the reliabilty, and thus lead to an overcorrection, but unfortunately, so will most of the estimates. ω_t will lead to a more accurate correction. In the case in which the test is multidimensional but only the test's general factor contributes to the prediction of the criterion of interest, α will over estimate the reliability associated with the general factor and lead to an undercorrection. ω_h would lead to a more accurate correction in this case.

7.4 Estimation of reliability

As initially introduced by Spearman, reliability was used to correct for the attenuation of relationships due to error in measurement. The initial concept of reliability, r_{xx}, was the correlation with a parallel test. This correlation allowed for an estimate of the percent of error variance in the test. Congeneric test theory elaborated this concept such that test reliability was the test's communality (the squared factor loading) on a latent factor common to multiple measures of the construct. Further refinements in domain sampling theory led to ways of estimating the percentage of reliable variance associated with the general factor of the test, ω_h, or the entire test, ω_t. But all of these estimates are based upon the idea that there is one source of true variance to be estimated. An alternative approach recognizes that scores have multiple sources of reliable variance and that for different questions we want to generalize across different sources of variance. That is, do not say that a measure has a reliability, but rather that it has different reliabilities, depending upon what aspects of the test are being considered.

In addition to having true score variance, tests are seen as having additional sources of variance, some of which are relevant and some of which are irrelevant when making decisions using the test. The test variance thus needs to be decomposed into variance associated with item, form, time, and source of information. To make the problem more complex, all of these components can interact with each other and produce their own components of variance. Thus, rather than use correlations as the index of reliability, *generalizability theory* introduced by Cronbach et al. (1972) used an *analysis of variance* or *ANOVA* approach to decompose the test variance. Only some of these sources of variance are relevant when making decisions using the test.

Table 7.9 Reliability is the ability to generalize about individual differences across alternative sources of variation. Generalizations within a domain of items use internal consistency estimates. If the items are not necessarily internally consistent reliability can be estimated based upon the worst split half, β, the average split (corrected for test length) or the best split, λ_4. Reliability across forms or across time is just the Pearson correlation. Reliability across raters depends upon the particular rating design and is one of the family of Intraclass correlations.

Generalization over	Type of reliability	Name
Unspecified	Parallel tests	r_{xx}
Items	Internal consistency	
	general factor (g)	ω_h
	$> g < h^2$	α
	all common (h^2)	ω_t
Split halves	random split half	$\frac{2r_{12}}{1+r_{12}}$
	worst split half	β
	best split half	λ_4
Form	Alternative form	r_{xx}
Time	Test-retest	r_{xx}
Raters	Single rater	ICC_2
	Average rater	ICC_{2k}

7.4.1 Test-retest reliability: Stability across time

Perhaps the most simple example of the different components of variance associated with reliability is to consider the reliabilities of a test of an emotional *state* with a test of a personality or ability *trait*. For both tests, we would expect that items within each test given at the same time should correlate with each other. That is, the tests should be internally consistent. But if a test of mood state shows reliability over time (stability), then we question whether it is in fact a test of mood. Similarly, a test of intellectual ability should be internally consistent at any one time, but should also show stability across time. More formally, consider the score for a particular person, i, on a particular test, j, at a particular time, o_k, with a particular random error, e_{ijk}.

$$X_{ijk} = t_{ij} + o_k + e_{ijk}.$$

For two parallel tests at the same time, the time component drops out and the expected score is just

$$X_{ijk} = t_{ij} + e_{ijk} = t_i + e_i$$

and reliability will be

$$r_{xx} = \frac{\sigma_t^2}{\sigma_X^2} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}. \tag{7.36}$$

But, if the tests are given at different times, there might be an effect of time (practice, learning, maturation) as well as an interaction of true score by time (people respond to the different occasions differently.) Thus, the variance of an observation (not specifying time) will be

$$\sigma_X^2 = \sigma_t^2 + \sigma_o^2 + \sigma_{to}^2 + \sigma_e^2.$$

The correlation of the test at time 1 with that at time 2 standardizes the observations at both times and thus removes any mean change across time. However, the interaction of time with true score remains and thus:

$$r_{xx} = \frac{\sigma_t^2}{\sigma_X^2} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{to}^2 + \sigma_e^2}. \tag{7.37}$$

In that the test-restest correlation reflects an additional variance component in the denominator (the trait by time interaction), it can normally be expected to be less than the reliability at one time. Two different examples of this effect emphasize the difference between measures of emotional *states* versus cognitive *traits*. In a short term study of the effect of a movie manipulation on mood and emotional state, tense arousal showed a momentary reliability of .65 with a 30 minute temporal stability of .28 (Rafaeli and Revelle, 2006). Indeed, four mood measures with an average internal consistency reliability of .84 (ranging from .65 to .92) had an average temporal stability over 30 minutes of just .50 (ranging from .28 to .68). Indeed, if a mood measure shows any stability across intervals as short as one or two days it is probably no longer a measure of temporary mood! However, we would expect stability in traits such as intellectual ability. In a longitudinal study with a 66 year interval, ability test scores show an amazing amount of *temporal stability* of .66 with an estimated short term test-retest reliability of .90 (Deary et al., 2004).

7.4.2 Intraclass correlations and the reliability of ratings across judges

The components of variance approach associated with *generalizability theory* is particularly appropriate when considering the reliability of multiple raters or judges. By forming appropriate ratios of variances, various *intraclass correlation coefficients* may be found (Shrout and Fleiss, 1979). The term *intraclass* is used because judges are seen as indistinguishable members of a “class”. That is, there is no logical way of distinguishing them.

Consider the problem of a study using coders (judges) to assess some construct (e.g., the amount of future orientation that each person reports in a set of essays, or the amount of anxiety a judge rates in subject based upon a 30 second video clip). Suppose that there are 100 subjects to be rated but a very limited number of judges. The rating task is very difficult and it is much too much work for any one person to judge all the subjects. Rather, some judges will rate some essays and other judges will rate other essays. What is the amount of reliable subject variance estimated by the set of available judges? To answer this question, one can take a subset of the subjects and have all of the judges rate just those targets. From the analysis of the components of variance in those ratings (the *generalizability study*), it is possible to estimate the amount of reliable variance in the overall set of ratings. The `ICC` function calculates intraclass correlations by taking advantage of the power of `R` to use the output from one function as input to another function. That is, `ICC` calls the `aov` function to do an analysis of variance and then just organizes the mean square estimates from that function to calculate the appropriate intraclass correlations and their confidence intervals (Shrout and Fleiss, 1979).

For example, six subjects are given some score by six different judges (Table 7.10 and Figure 7.7. Judges 1 and 2 give identical ratings, Judges 3 and 4 agree with Judges 1 and 2 in the relative ratings, but disagree in terms of level (Judge 3) or variance (Judge 4). Finally, Judges 5 and 6 differ from the first four judges in their rank orders and differ from each other in terms of their mean and variance. `ICC` reports the variance between subjects (MS_b), the variance within subjects (MS_w), the variances due to the judges (MS_j), and the variance due to the interaction of judge by subject (MS_e). The variance within subjects is based upon the pooled The reliability estimates from this *generalizability analysis* will depend upon how the scores from the judges are to be used in the *decision analysis*.

The next three equations are adapted from Shrout and Fleiss (1979) who give In a very thorough discussion of the ICC as it is used in ratings and discuss six different ICCs and formulas for their confidence intervals. Another useful discussion is by McGraw and Wong (1996) and an errata published six months later.

$ICC_{(1,1)}$: Each target is rated by a different judge and the judges are selected at random. This is a one-way ANOVA fixed effects model where the judge effect is part of the error term and is found by

$$ICC_{(1,1)} = \frac{MS_b - MS_w}{MS_b + (n_j - 1)MS_w}.$$

$ICC_{(1,1)}$ is sensitive to differences in means and variances between raters and is a measure of absolute agreement. The interaction of rater by judge is included in the error term. Compare the results for Judges 1 and 2 versus 1 and 3. Although the variances are identical, because the mean for Judge 3 is 5 points higher the Judge 1, $ICC_{(1,1)}$ for these two judges is actually negative.

$ICC_{(2,1)}$: A random sample of k judges rate the targets. The measure is one of absolute agreement in the ratings. Mean differences in judges as well as the judge by target interaction will affect the scores. Defined as

$$ICC_{(2,1)} = \frac{MS_b - MS_e}{MS_b + (n_j - 1)MS_e + n_j(MS_j - MS_e)/n}.$$

Because $ICC_{(2,1)}$ has a smaller residual error term (MS_e) it will usually, but not always be greater than $ICC_{(1,1)}$ (but see the analysis for J1 and J5).

Table 7.10 The Intraclass Correlation Coefficient (ICC) measures the correlation between multiple observers when the observations are all of the same class. It is found by doing an analysis of variance to identify the effects due to subjects, judges, and their interaction. These are combined to form the appropriate ICC. There are at least six different ICCs, depending upon the type of generalization that is to be made. See Table 7.11 for results taken from these data.

```
> 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,ranges=FALSE,skew=FALSE)
   var n mean   sd   se
J1   1 6  3.5 1.87 0.76
J2   2 6  3.5 1.87 0.76
J3   3 6  8.5 1.87 0.76
J4   4 6  7.0 3.74 1.53
J5   5 6  3.5 1.87 0.76
J6   6 6  7.0 3.74 1.53

> print(ICC(Ratings),all=TRUE)
$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

Number of subjects = 6     Number of Judges =  6

$summary
           Df  Sum Sq Mean Sq F value    Pr(>F)
subs        5 141.667  28.333  10.366 1.801e-05 ***
ind         5 153.000  30.600  11.195 9.644e-06 ***
Residuals  25  68.333   2.733
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1
```

$ICC_{(3,1)}$: A fixed set of k judges rate each target. Mean differences between judges are removed. There is no generalization to a larger population of judges.

$$ICC_{(3,1)} = \frac{MS_b - MS_e}{MS_b + (n_j - 1)MS_e}$$

By removing the mean for each judge, $ICC_{(3,1)}$ is sensitive to variance differences between judges (e.g., Judges 4 and 6 have four times the variance of Judges 1...3 and 5).

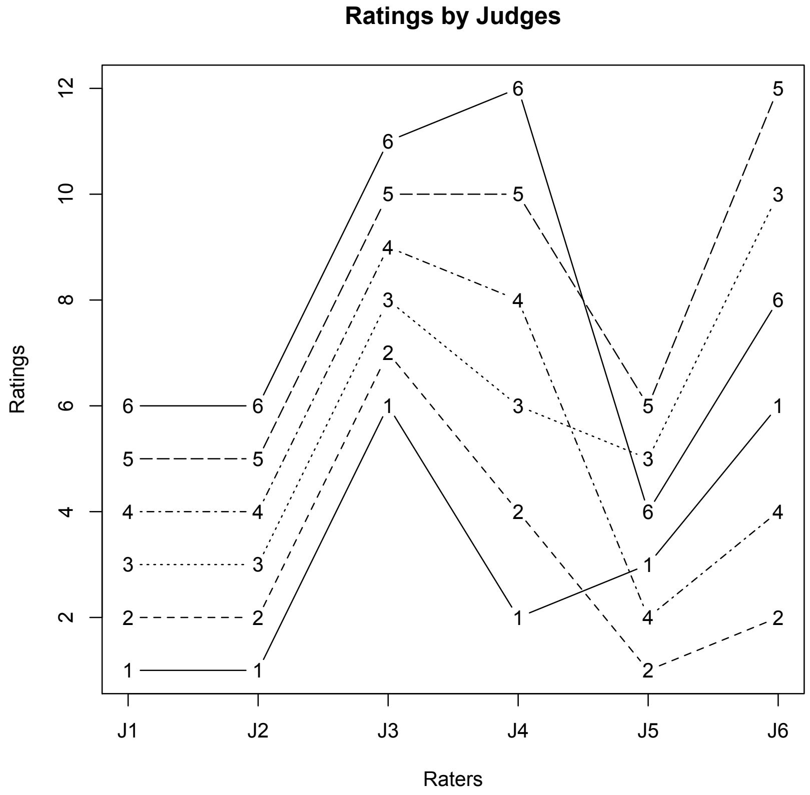


Fig. 7.7 When estimating the reliability of raters, it is important to consider what kind of reliability is relevant. Although clearly the correlation between raters J1 ... J4 is 1 between all 4 raters, the raters differ in their leniency as well as their variability. The Intraclass correlation considers different types of reliability.

Table 7.11 Sources of variances and the Intraclass Correlation Coefficient.

	(J1, J2)	(J3, J4)	(J5, J6)	(J1, J3)	(J1, J5)	(J1 ... J3)	(J1 ... J4)	(J1 ... J6)
Variance estimates								
MS_b	7	15.75	15.75	7.0	5.2	10.50	21.88	28.33
MS_w	0	2.58	7.58	12.5	1.5	8.33	7.12	7.38
MS_j	0	6.75	36.75	75.0	0.0	50.00	38.38	30.60
MS_e	0	1.75	1.75	0.0	1.8	0.00	.88	2.73
Intraclass correlations								
ICC(1,1)	1.00	.72	.35	-.28	.55	.08	.34	.32
ICC(2,1)	1.00	.73	.48	.22	.53	.30	.42	.37
ICC(3,1)	1.00	.80	.80	1.00	.49	1.00	.86	.61
ICC(1,k)	1.00	.84	.52	-.79	.71	.21	.67	.74
ICC(2,k)	1.00	.85	.65	.36	.69	.56	.75	.78
ICC(3,k)	1.00	.89	.89	1.00	.65	1.00	.96	.90

Then, for each of these three cases, if reliability is to be estimated for the average rating of multiple judges? In that case, each target gets the the average of k ratings and the reliability are increased by the Spearman Brown adjusted reliability.

7.4.3 Generalizability theory: reliability over facets

The intraclass correlation analysis of the reliability of ratings in terms of components of variance associated with raters, targets, and their interactions, can be extended to other domains. That is, the analysis of variance approach to the measurement of reliability focuses on the relevant facets in an experimental design. If ratings are nested within teachers whom are nested within schools, and are given at different times, then all of these terms and their interactions are sources of variance in the ratings. First do an analysis of variance in the *generalizability study* to identify the variance components. Then determine which variance components are relevant for the application in the *decision study* in which one is trying to use the measure (Cronbach et al., 1972). Similarly, the components of variance associated with parts of a test can be analyzed in terms of the generalizability of the entire test.

7.4.4 Reliability of a composite test

If a test is made up of subtests with known reliability, it is possible to find the reliability of the composite in terms of the known reliabilities and the observed correlations. Early discussions of this by Cronbach et al. (1965) considered the composite reliability of a test to be a function of the reliabilities of each subtest, ρ_{xx_i} (for which Cronbach used α_i), the subtest variance, σ_i^2, and the total test variance, σ_X^2,

$$\alpha_s = 1 - \frac{\Sigma(1-\rho_{xx_i})\sigma_i^2}{\sigma_X^2}. \tag{7.38}$$

The example in Table 7.5 had three groups with reliabilities of .857, .857 and .823, total variance of 53.2, and variance/covariances of

```
    G1   G2  G3
G1 21.0  4.8 2.4
G2  4.8 11.2 1.6
G3  2.4  1.6 3.4
```

The composite α is therefore

$$1 - \frac{(1-.857)21.0+(1-.857)11.2+(1-.823)3.4}{53.2} = .90$$

which is the same value as found for ω_t and is the correct value given the known structure of this problem. However, the items in the example all have equal correlations within groups. For the same reason that α underestimates reliability, α_s will also underestimate the reliability if the items within groups do not have identical correlations. α_s is preferred to α for estimating the reliability of a composite, but is still not as accurate as ω_t. Although Cronbach et al.

(1965) used α_i as an estimate for the subtest reliability, ρ_{xx_i}, it is possible to use a better estimate of reliability for the subtests. ω_t can, of course, be found using the `omega` function or can be found by using a "phantom variable" approach (Raykov, 1997) in a structural equation solution using `sem` (Chapter 10).

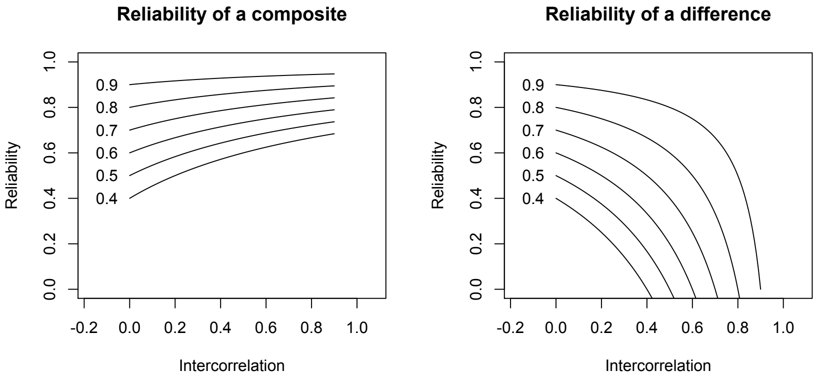
7.4.5 Reliability of a difference score

It is sometimes useful to create a score made up of the difference between two tests. Just as the variance of a composite of X_1 and X_2 is the sum of the variances and twice the covariance of X_1 and X_2 (Equation 7.39, so is the variance of a difference, except in this case the covariance is negative. The reliability of this difference score, $r_{\Delta\Delta}$, may be found by the ratio of the reliable variance to the total variance and is a function of the reliable variances for the two components as well as their intercorrelation:

$$r_{\Delta\Delta} = \frac{\sigma_1^2 r_{xx_1} + \sigma_2^2 r_{xx_2} - 2\sigma_1^2\sigma_2^2 r_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1^2\sigma_2^2 r_{12}}. \tag{7.39}$$

This is equivalent to finding the reliability of a sum of two tests (Equation 7.38 but now the tests are negatively rather than positively correlated. As the correlation between the two tests increases, the reliability of their differences decreases.

Fig. 7.8 The reliability of composites or differences of two tests depends upon the reliability of the tests as well as their intercorrelation. The reliability of the composite increases as the tests are more correlated (left hand panel). However, the reliability of a difference decreases as the tests are more correlated (right hand panel).



7.5 Using reliability to estimate true scores

Although developed to correct for *attenuation* of correlations due to poor measurement, another important use of reliability is to estimate a person's true or domain score given their observed score. From Equation 7.40 it is clear that

$$\hat{t} = b_{t.x}x = \frac{\sigma_t^2}{\sigma_x^2}x = \rho_{xt}^2 x = r_{xx}x. \tag{7.40}$$

That is, expected true scores will regress towards the mean observed score as a function of 1 - r_{xx}. This *regression to the mean* is the source of great confusion for many people because it implies that the best estimate for a person's retest score is closer to the mean of the population than is the observed score. Real life examples of this are seen in sports and finance where a baseball player who does well one year can be expected to do less well the next year just as a financial advisor who does well one year will probably not do as well the next year (Bernstein, 1996; Stigler, 1999). Perhaps the classic example is that of flight instructors who observe that praising good performance results in decreases in performance while punishing poor performance results in improved performance (Tversky and Kahneman, 1974).

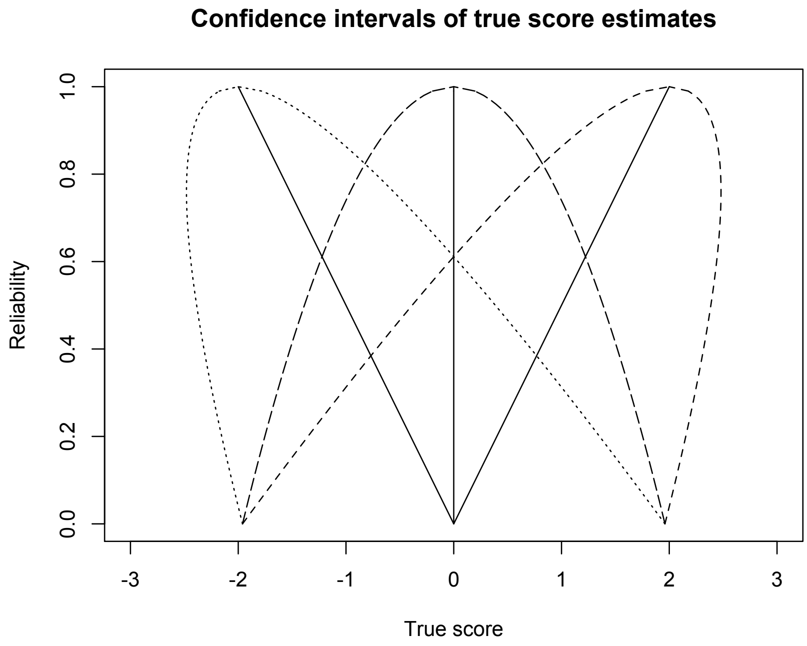
Knowing the reliability also allows for confidence intervals around the estimated true score. For the proportion of error variance in an estimate is just 1- rxx, and thus the *standard error of measurement* is

$$\sigma_e = \sigma_x\sqrt{r_{xt}^2} = \sigma_x\sqrt{1 - r_{xx}} \tag{7.41}$$

Because the estimated true score is closer to the mean than the observed score, the confidence intervals of true score will be assymetric around the observed score. Consider three observed scores of +2, 0, and -2 and their estimated true scores and confidence intervals as a function of the test reliability.

All of the approaches discussed in this chapter have considered the reliability of a measure in terms of the variance of observed scores and the variance of latent scores. Although the effect of items is considered in terms of how the items intercorrelate, items are assumed to sampled at random from some universe of items. Reliability is a characteristic of the test and of a sample of people. A test's reliability will be increased if the true score variance is increased, but this can be done in a somewhat artificial manner. Consider a class of first year graduate students in psychometrics. An exam testing their knowledge will probably not be very reliable in that the variance of the students' knowledge is not very great. But if several first year undergraduates who have not had statistics and several psychometricians are added to the group of test takers, suddenly the reliability of the test is very high because the between person variance has increased. But the precision for evaluating individual differences within the set of graduate students has not changed. What is needed is a model of how an individual student responds to an individual item rather than how a group of students responds to a group of items (a test). In the next chapter we consider how these ideas can be expanded to include a consideration of how individual items behave, and how it is possible to get an estimate of the error associated with the measure for a single individual.

Fig. 7.9 Confidence intervals vary with reliability, For observed scores of -2, 0, and 2, the estimated true scores and confidence intervals vary by reliability. The confidence range is symmetric about the estimated true score. Even with reliability = .90, the confidence intervals for a true score with an observed score of 2 range from 1.18 to 2.42. $(1.8-\sqrt{1-.9}$ to $1.8+\sqrt{(1-.9)})$


Confidence intervals of true score estimates
Reliability
True score