COUNCIL OF EUROPE CONSEIL DE L'EUROPE

December 2004 **DGIV/EDU/LANG (2004) 13**

Reference Supplement

to the

Preliminary Pilot version of the Manual for

Relating Language examinations to the
Common European Framework of Reference for Languages:
learning, teaching, assessment

Section C: Classical Test Theory

Language Policy Division, Strasbourg

Section C

Classical Test Theory

N.D. Verhelst
National Institute for Educational Measurement (Cito)
Arnhem, The Netherlands

In this section an overview is given of the main concepts and theoretical results of Classical Test Theory (CTT). The text has been written to be as accessible as possible for the non-technical reader. The first two sections (Basic Concepts and Procedures) do not contain any formulae. They are meant to be read as a whole, since concepts introduced at the start are used in later parts. As CTT is a statistical theory, it is not possible to present and discuss it in great depth without having recourse to the exact and compact way of expression provided by mathematical formulae. Where it is felt that some deeper understanding of the theory might be wished, reference is made to a more technical section. These technical sections are stand-alone sections, and are added to the main text in the order they are referred to.

Classical Test Theory has been used for more than fifty years as a guide for test constructors to understand the statistical properties of test scores, and to use these properties to optimise the test under construction in a number of ways. The main purpose of this appendix is to review the main issues of Classical Test Theory, and to emphasise what can be expected from Classical Test Theory and what not. We will first present some basic concepts and then go on to procedures which are used in the framework of Classical Test Theory.

C.1. Basic Concepts

Items. In many cases a test is composed of a number of elementary parts, for example, twenty questions. A generic name for such a part is: ‘item’. There is, however, no stringent rule of identifying items with questions. Suppose a reading test consists of five text passages, and four questions are to be answered about each passage. One might conceive the twenty questions as twenty items, but one might also consider the four questions associated with each text as a single item. In the latter case, one sometimes refers to those composite items as super items, testlets or item bundles.

Observed score. When a test is administered, the result is summarized by a **number** (for example, the number of correct item responses). This number is called the (observed) **test score.** Usually the test score is the sum of the **item scores.** In all analyses to be carried out in CTT, the item scores are usually the basic quantities that enter such analyses. But it should be kept in mind that these scores are not given as such; they come about through a decision by the test constructor, and CTT does not provide any rules for taking such a decision. It is customary to grant one point for the ‘correct answer’ in a multiple choice item, and zero points for any other choice. In some cases, however, it might be more informative to grant 2 points for a particular choice, 1 point for another (not optimal) choice and zero points for the remaining choices. The actual choice the test taker makes is the basic observation; the granting of points is a decision to be taken a priori, sometimes on intuitive grounds, sometimes on the basis of extended qualitative studies and quantitative analyses of the set of observations. Therefore it is wise to keep as detailed records of the observations as possible: for multiple choice questions, the option actually chosen; for open ended questions, it is advisable to develop a quite detailed categorizing system, and to keep records (in a data base) of as much detail as possible. To the data stored in this manner, different scoring rules may then applied, yielding in each case a file with (numerical) item scores which may then be submitted to quantitative analyses.

True score. The basic assumption of CTT is that in a second administration of the same test to the same person under similar circumstances as the first time, we will probably not observe the same score as the first time. This reasoning can be generalized to an arbitrary number of similar test administrations, giving rise to the idea of a **distribution of (possible) test scores.** This distribution is associated with a single person, and hence could be characterized as his or her ‘private’ distribution. In CTT the average of this private distribution is called the person’s true score. True score is a statistical concept, and has nothing to do with conceptions like ‘ideal score’ or ‘the score a person really deserves’. The observed score actually obtained is conceived as a sample (of size 1) from the ‘private’ distribution. True scores are not observed. The true score of a person is symbolized (in this appendix) with the Greek letter tau (τ). Notice that it is a number.

Measurement error. In CTT the measurement error is defined as the difference between the observed score and the true score. If the observed score is greater than the true score, we say that the measurement error is positive; if it is smaller, the measurement error is negative. Since the true score of a person is not known, the measurement error (in a particular case) is not known either. It is possible, however, to say something more concrete of measurement errors in a population. The symbol used for the measurement error is E.

Variability: standard deviation and variance. Phenomena showing no variability are not very informative. If everybody (from a certain population) gets the maximum score on a test, all one can say is that the test is apparently too easy for this population. Things are becoming interesting if they show variability, as test scores in a calibration sample usually do. In statistics one needs a **measure of variability**. A well known measure is the standard deviation. The variance is the square of the standard deviation. Although the standard deviation is usually easier to interpret, the variance is a more useful concept in statistics (e.g., in such techniques as analysis of variance.)

Sources of variance. Suppose John’s observed score is 18 and Mary’s is 20. One could ask why these observed scores differ. CTT distinguishes two sources of variability: the scores may differ because John’s and Mary’s true scores differ or because the two measurement errors differ; or both. These two sources cannot be disentangled at the individual level, i.e., we cannot know the answer in the concrete case of John and Mary; but they can be distinguished at the level of the population. In the population the true score is not a number, but a variable (which can assume different values for different persons). To indicate the true score as a variable we use the symbol T. The important result is that (in the population) the variance of the observed scores is the sum of the variance of the true scores and the variance of the measurement errors. (Notice that this decomposition rule does not hold for standard deviations.) Shorthand names are sometimes used: observed variance for the variance of observed scores, true and error variance for variance of true scores and measurement errors, respectively.

Reliability of test scores. The reliability of test scores is defined as the ratio of the true variance to the observed variance. Multiplied by 100, it can be interpreted as a percentage: it is the percentage of the observed variance which is true variance. The minimum value of the reliability is zero, meaning that all variation in the observed scores is due to measurement error. The maximum value is one, meaning that there is no measurement error. A reliability coefficient of 0.8 means that 80% of the observed score variance is due to variation in the true scores and 20% to measurement error. Reliability is a key concept in CTT, but from the definition it is not clear how it can be determined. Further down, this problem will be discussed, together with some examples of the importance of the concept. The expression ‘reliability of a test’ is often used, but it is not correct; it should be understood as ‘reliability of test scores’.

C.2 Procedures

***P*-values**.

In the process of constructing test items it is important to have a rather precise idea of the target population. Administration of items that are too easy or too hard is not adequate for several reasons. It may lead to boredom or frustration, which in turn will almost invariably cause loss of motivation for the test taker. Moreover, in this case, the item responses will give very little information about the proficiency level of the test takers. Therefore, it is important to have a rather precise idea about the degree of difficulty of the items; decisions on inclusion or exclusion of items are often based on information about their degree of difficulty, usually called *p*-values. (The '*p*' refers to proportion or probability.) For binary (scored 0/1) items, the *p*-value of an item is the proportion of correct responses in the population. Usually, a *p*-value is considered as a property of an item, which is correct, as long as one realises that this property is valid only with respect to a certain population. A common way of expressing this relativity is to say that *p*-values are **population dependent**. This can easily be understood with the following simple example. Suppose an item is developed for a test to be applied in the fourth year of English learning. With respect to this population, let us assume that the item is rather easy, and has a *p*-value of 0.8. It will easily be understood that such an item may be much harder in the population of second year students, yielding a *p*-value of 0.25 or even lower in this population. Thus speaking of the *p*-value of an item has no meaning; implicitly or explicitly there is always a reference to some population. This population dependency has immediate implications when one tries to establish the psychometric properties of a test from the sample observations. The sample must be representative for the population.

Note 1. *P*-values are values which pertain to items in some population, but they are computed on a sample. Representativeness of the sample does not mean that the value computed will be equal to the value in the population. If we compute the *p*-value of an item in two independent samples, we will usually find two different values. The *p*-value found in the sample is to be considered as an **estimate** of the *p*-value in the population. The accuracy of the estimate depends mainly on the sample size. Details and examples are given in Section C.3

Note 2. Items where one can get 0, 1 or 2 points, or 0, 1, 2 or 3 points, etc., are called partial credit items. *P*-values of partial credit items are defined as the average relative score. See Section C.4 for details.

Note 3. It is common to interpret *p*-values as measures of difficulty, but notice that the higher the *p*-value, the easier the item is. Some authors use 1 – *p* as the measure of difficulty. Both measures are acceptable, as long as it is clearly indicated which one is used.

Item discrimination

Simply stated, the discriminating power of an item is to extent to which it is possible to separate high proficiency levels from low levels on the basis of the responses to the item. Or, stated otherwise: what is the psychometric quality of a test which consists of this particular item? Suppose that a quite difficult binary item is used as a test. We will say that the item discriminates well if the very best students have the item correct, and the others not, but since a binary item has only two categories (right or wrong), if the item separates the very best from the others, it cannot separate the students of medium proficiency from the weak ones. That is, discrimination is a local property, and it is fairly difficult to catch (and describe) the discriminating power of an item in a single number. Yet, there exist several indices of discrimination which are used within CTT. We list some of them:

- the correlation between item score and test score (item-test correlation);
- the correlation between item score and the score on the test with that item excluded (item-rest correlation);
- in particular for multiple choice items: the correlation between test score and each of the distractors.

Item-test and item-rest correlations should be positive; correlations between the test score and the distractors should be negative. (See Section C5 for the exact meaning of this notion) Rules of thumb for a minimum value of item-test or item-rest correlations may be misleading, because the correlation is strongly influenced by the *p*-value of the item.

Graphical Item Analysis

The usual output from software for item analysis consists of a number of tables containing *p*-values, discrimination indices like item-test and item-rest correlations, and other indices usually interpreted also as indices of discrimination. There exists, however, a simple and powerful tool to judge the quality of the items. Each item is represented by one or more curves as exemplified in Figure C.1

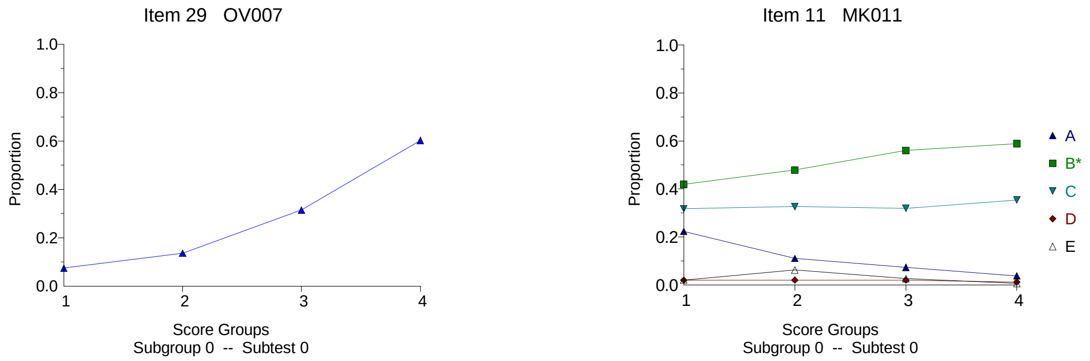

Figure C.1. Graphical item analysis

The figures are constructed using the same principle: the total sample is split into a small number of homogeneous groups (four in the examples; ‘1’ denotes the groups with the lowest scores, ‘4’ the group with the highest scores, and ‘2’ and ‘3’ intermediate groups) on the basis of the test scores. In each group the proportion of correct responses is computed and plotted against the group number, as is shown in the left-hand panel of the figure (item 29). One sees immediately that the item is relatively difficult: even in the highest group (4) only about 60% of the test takers gave a correct answer. We also see that the item-test correlation will be positive: the higher the group number, the higher the proportion of correct responses. In the right-hand panel a similar figure is drawn for a multiple choice item with five alternatives (where B is the correct alternative). Here we see immediately that the item is not of a very high quality: the discriminating power is low (the curve for alternative B increases very slowly); the distractors D and E are almost never chosen (and so prove to be useless as distractors), and distractor C remains attractive at a constant and quite elevated level (more than 30%), which may suggest that this item is a catch item. In summary, the figure suggests clearly that the item deserves revision, and cannot function as a ‘model item’ to help train item writers. More examples are given and discussed in Section C.6.

The figures displayed above are standard output of the computer program TiaPlus. To obtain this program, a request should be sent to Ton.Heuvelmans@Citogroep.nl.

Estimation of Reliability

From the definition of reliability, it is clearly not possible to compute the reliability coefficient directly, because of the presence of a quantity which is not observable: the variance of the true scores. In order to compute the reliability, a new concept has to be introduced, the concept of a **parallel test**. Two tests are parallel if the following two conditions hold: the true scores on both tests are equal for all persons in the population, and the variance of the measurement error is equal in both tests.

An important and reassuring result of CTT is that the reliability of a test equals the correlation between the test and a parallel test. Two parallel tests have the same reliability.

There are two problems associated with this finding: (1) how do we know that two tests are parallel, and (2) in order to compute the correlation, we need test scores on the same sample of test takers on the two tests, i.e., two test administrations are required. We comment on both problems.

1 The construction of parallel tests
   a Two parallel tests have the same average observed score and the same observed variance. Moreover, their correlation with all other tests, whatever these measure, should be the same. But this holds in the population; we cannot expect that these equalities will also hold in a sample. In practice, significance testing can be used, but one should be careful: if the differ-

ence between two sample averages does not differ significantly from zero, this does not imply that the population differences do not differ. The risk that a real difference in the population is not detected by a significance test is larger when the sample size is small.

b Two methods are commonly used in applications of CTT, parallel form and retesting. In the retesting method, the same test is applied at two different points in time. The main threat to parallelism is the memory effect. Here we have to distinguish between two cases:

i) In general memory effects are beneficial to the test performance, yielding a higher test score on the second administration than on the first one. If memory effects are uniform, i.e., if the increase from the first to the second administration (in true score) is the same for every person, the two series of test scores are not parallel, but their correlation nevertheless is the reliability of the test. If the increase is uniform, the two (population) means may differ, but the variances will not differ.

ii) If memory effects are not the same for every person, the retesting will not yield a parallel form. This may occur when there are ceiling effects: low scores in the first administration may increase considerably by memory effects, but high scores may probably not increase by the same amount, because they are already close to the maximum score of the test. If this is the case, the correlation between the two series of test scores is not the reliability.

The construction of parallel forms is not easy either. A necessary condition for parallelism is that the contents of both forms should be comparable, which may be hard to accomplish in cases where complex items are constructed (e.g., a text passage with four or five associated questions). There exists a rather simple method to use psychometric indices to aid in constructing parallel forms. This method is discussed in section C.7.

c Sometimes, only one test is available, but for the sake of estimating the reliability it is split into two halves which are meant to be parallel. Notice that the correlation between the two halves – if they are really parallel - is not the reliability of the test, but of the half tests. To obtain the reliability of the test, the Spearman-Brown formula has to be applied (see below).This method is known as the split-half method. If the two halves are not parallel, the resulting coefficient underestimates the reliability.

2 Reliability estimation from a single test administration

a In principle it is impossible to determine the reliability of a test from a single test administration. All that can be reached is a so called lower bound to the reliability; this is a number such that one can be certain that the reliability is not lower than that number. If for a given test this lower bound is 0.7, all one can be sure of is that the reliability is at least 0.7. If the lower bound is high (more than 0.95, for example) this will not be a big problem. If it is low, however, 0.30 say, it does not follow that the reliability is that low.

b The best known lower bound is Cronbach's coefficient alpha. It can be used for any mixture of binary and partial credit items.

c The KR20-coefficient is the same as Cronbach's alpha, but it is defined only for binary items.

d Cronbach's alpha is sometimes labelled as an index of internal consistency, i.e., an index that shows the extent to which all items in the test measure the same concept. If the test is really one-dimensional, the index will be close to the reliability; if the test is heterogeneous, alpha can be substantially lower than the reliability.

e There exist more lower bounds. In fact there exists a **greatest lower bound** (GLB). It is at least as large as all possible lower bounds. The computation of the GLB is not easy (there does not exist a closed formula), but it is available in published software; the program TiaPlus does compute it.

f Lower bounds such as Cronbach's alpha, the KR20 and the GLB are quantities which apply to the population. They are estimated from the calibration sample and contain an estimation error. The estimate of the GLB from small samples tends to be a serious overestimate of the population GLB. In the program TiaPlus, a correction to this bias is applied if the sample size is not too small.

The Spearman-Brown formula. Tests are administered to collect information on a person's proficiency. The information is conveyed through the scores obtained on the items, but we have to admit that these scores contain errors, some positive, others negative. By summing the item scores, positive and negative errors will tend to cancel each other, the more so if the test gets longer. It follows that we can trust the result of a long test generally more than the result of a short test, or, what is the same, the reliability of a long test is higher than that of a short test. The Spearman-Brown formula expresses the relation between test length and reliability. It can be used in two ways, which we illustrate with an example:

1. A test consisting of 25 items has a reliability of 0.7. What will the reliability be if 10 items are added? (The answer is 0.766; see Section C.8)
2. A test consisting of 25 items has a reliability of 0.7. How many items must the test contain to have a reliability of 0.8? (The answer is 43; see Section C.8.)

The second example shows how the Spearman-Brown formula can be used to plan work on extra item writing. It should be noticed that it is more expensive (in terms of the number of items) to raise the reliability from 0.8 to 0.9 than from 0.7 to 0.8. The increase from 0.7 to 0.8 requires 43 – 25 = 18 extra items; to reach 0.9, another 54 items are required.

The Spearman-Brown formula must be used very cautiously: it only applies if the added items are of the same quality as the items already present. The standard expression is that the test must be lengthened homogeneously.

The formula can also be used in the reverse sense: if a planned test with a known reliability happens to be too long to be useful in practice, the formula can be used to compute the reliability of a shortened version of the test. Taking the example above: if the test with 43 items and a reliability of 0.8 is shortened (homogeneously) to 25 items, the shorter version will have a reliability of 0.7.

Finally, it can be used to compute the reliability in case of the split-half method. If the correlation between the two test halves is symbolized as r, the reliability of the full test is $2r/(1+r)$.

The Standard Error of Measurement. Although we can never know in a particular case what the measurement error is, we can have a quite precise idea of the magnitude of the measurement error 'on the average'. Recall the 'private' distributions of the observed scores. If in such a private distribution of possible observed scores all (or most) of the values are very near to the average (the true score), this distribution will have a small standard deviation; if on the contrary, many values are far away from the average, the standard deviation will become large. So the standard deviation of the private distribution gives an indication of a typical error. This standard deviation is called the standard error of measurement.

There is a strong relation between the standard error of measurement and the reliability of the test: the standard error of measurement is the standard deviation of the observed scores (in the population) multiplied by the square root of one minus the reliability.

The standard error of measurement can be used to define confidence intervals for the true score. It is instructive to look at examples of such confidence intervals to learn about the relative merits of testing. Even with a reliability as high as 0.96, the 90% confidence interval for the true score is larger than half a standard deviation. Details are discussed in Section C.9.

Decisions on individuals are sometimes based on a test score, for instance an examination score. One should realize that such decisions are of necessity based on observed test scores, which contain an unknown measurement error. This implies that able candidates may fail on an examination because of a negative measurement error, and weak candidates may succeed because of a positive error. This leads to wrong (unintended) classifications. The percentage of such erroneous classifications depends strongly on the reliability of the test. Even if it is as high as 0.9, the percentage of wrong classifications can be substantial.

Kelley's formula. Sometimes an estimate of the true score is needed. The best known estimate is computed using the famous formula by Kelley. The result of this formula is a compromise between the observed score and the population mean of the scores. A compromise means a weighted sum; the weight of the observed score is the reliability of the test, the weight for the population mean is one minus the reliability. Suppose X = 112 and the population mean is 100; the reliability equals 0.88. Kelley's estimate of the true score is $112 \times 0.88 + 100 \times (1 - 0.88) = 110.56$. Notice that the estimate is closer to the population mean than is the observed score. This is known as 'shrinkage'. This estimate can be interpreted as follows: it is the average true score of all people in the population having an observed score of 112. If John's observed score happens to be 112, we cannot infer from this that his true score is exactly 110.56, i.e., the estimate also contains an error. This error is called the estimation error, and its standard deviation is called the standard error of estimation. It is smaller than the standard error of measurement.

Theoretical results.

There are three important results which are useful in the discussion of external validation. One can conceive test results as measurements that are polluted in some way by measurement error. It may be interesting to know as precisely as possible what the results would be if one could measure without measurement errors, i.e., the results in the ideal case where the observed scores are equal to the true scores. These are the results: (details can be found in Section C.10)

1. The correlation between observed scores and true scores is the square root of the reliability.
2. The correlation between the observed scores on two tests is 'attenuated' (lowered) by the unreliability of the two tests. The correlation between the true scores on both tests equals the correlation between the observed scores divided by the square root of the product of their reliabilities. The corresponding formula is called the correction for attenuation.
3. If two tests really measure the same concept, the correlation between the true scores of both tests should equal one. If this is the case, the tests are called **congeneric**. But the correlation between the observed scores will be attenuated by their unreliability. If two tests are congeneric, the correlation between the observed scores is equal to the square root of the product of their reliabilities.

Population dependency

In the discussion on the *p*-values, it was stressed that it is meaningless to speak about the *p*-value of an item, because there is always a reference (explicitly or implicitly) to a certain population. The same argument applies to all item- and test-indices that are used in Classical Test Theory. In particular it applies to the concept of reliability. The reliability of a test is a characteristic of the test scores in some population. The same test can have a high reliability in some population and a very low one in another population. Here is an example. Suppose a test is used as an entrance test to the university, and assume it has a reliability of .85 in the population of candidates. This very same test will have a lower reliability in the population of first year students at the university, because this population is more homogeneous with respect to true score than the population of candidates, i.e. the variance of the true scores at the university will be smaller than in the population of candidates. Or more generally, the more homogeneous the population (with respect to true score), the lower the reliability will be. But, of course, this is not the only reason why the reliability of a test can be low. Sloppy items with ambiguous scoring rules will usually lead to low reliability, and one cannot use the homogeneity of the population as an excuse for the bad quality of the test.

C.3. The accuracy of *p*-values

A good method of getting an impression of a *p*-value of an item is to construct **confidence intervals**. A *p*-value is a theoretical quantity which applies to the population, and which one usually estimates by a corresponding quantity in the sample. If the *p*-value of an item in the population is 0.75, say, it is almost sure that one will not find a proportion correct of 0.75 in the sample. But in general, we do not know the population value, we only observe a proportion correct in the sample. The problem of **inferential statistics** is to make clear what one can say about the population value on the basis of a sample value. To this end, one usually constructs **confidence intervals**. In what follows, the theory of confidence intervals is summarized and a practical formula for constructing intervals is given.

We represent the unknown *p*-value in the population by the Greek letter π; the proportion one can observe in a sample is denoted as *p*. The observed proportion is called a **random variable**, because it can assume different values in different samples.

1. Assume we could draw a very great number of samples, all independent of each other, and all of the same size, *n*. In each sample we can compute the observed *p*-value, and we can construct a histogram with these *p*-values. From theoretical statistics we can tell interesting things about this histogram:
    a. Its average equals the unknown value π;
    b. Its standard deviation equals $\sqrt{\pi(1-\pi)/n}$; this standard deviation is called the standard error of the random variable *p*;
    c. The form of the histogram looks very much like the graph of the normal distribution, and the similarity is more striking for large *n* than for small *n*.
2. Of course, we do not draw many samples, we usually draw a single one, but from the theoretical results we can say that the *p*-value we will observe will, with a probability of 90% lie in an interval from the mean (π) minus 1.645 times the standard deviation to the mean plus 1.645 times the standard deviation. The value of 1.645 is to be found in published tables of the normal distribution. If we want a 95% interval, we have to replace 1.645 by 1.96, and for a 99% interval, we use 2.58.
3. We express the preceding paragraph by means of a formula:

$$P\left(\pi-1.645\sqrt{\frac{\pi(1-\pi)}{n}}\le p\le\pi+1.645\sqrt{\frac{\pi(1-\pi)}{n}}\right)=0.9 \tag{c1}$$

4. The expression between parentheses in the preceding formula is called an **event** (*p* lies in some interval). The whole formula reads as: the probability of this event is 0.9 But we can replace this event by an equivalent event. We do this in two steps: the first step concentrates on the first inequality, where we move the term with the square root to the other side of the inequality sign:

$$\pi-1.645\sqrt{\frac{\pi(1-\pi)}{n}}\le p \;\Leftrightarrow\; \pi\;\le p+1.645\sqrt{\frac{\pi(1-\pi)}{n}}$$

and, in the second step (concentrating on the second inequality in formula (c1)) by a similar move we get:

$$p\le\pi+1.645\sqrt{\frac{\pi(1-\pi)}{n}} \;\Leftrightarrow\; p-1.645\sqrt{\frac{\pi(1-\pi)}{n}}\;\le\pi$$

and combining the two right-hand sides gives

$$p-1.645\sqrt{\frac{\pi(1-\pi)}{n}}\le\pi\le p+1.645\sqrt{\frac{\pi(1-\pi)}{n}}$$

and this event reads as: the population value π is embraced by two values which will vary from sample to sample, because the observed *p*-value is a random variable. And since we work with equivalent events, we can say that

$$P\left(p-1.645\sqrt{\frac{\pi(1-\pi)}{n}}\le\pi\le p+1.645\sqrt{\frac{\pi(1-\pi)}{n}}\right)=0.9 \tag{c2}$$

5. It deserves some attention to understand well the equivalence of (c1) and (c2) and the difference in wording of the two statements. In (c1) we say that the event is that the value of a random variable

(p) will lie between two fixed values; in (c2) we say (equivalently) that the fixed population value (π) will be embraced by two variable bounds.

6. There is, however, a further problem with formula (c2): the two bounds depend on the variable p, but also on the unknown value of π. In practice, then, one replaces π by the observed value p under the square root sign, giving a practical formula:

$$P\left(p-1.645\sqrt{\frac{p(1-p)}{n}} \le \pi \le p+1.645\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.9 \qquad \text{(c3)}$$

7. Here is a simple example. Suppose p = 0.51 and n = 100. Then, $\sqrt{0.51(1-0.51)/100}$ = 0.04999 (≈ 0.05), and using these values in (c3), we find that

$$\begin{aligned}&P(0.51-1.645\times 0.05 \le \pi \le 0.51+1.645\times 0.05)\\ &= P(0.428 \le \pi \le 0.592) = 0.9\end{aligned}$$

8. Notice that the observed p-value (0.51) lies precisely in the middle of the defined interval, or, as one says, the confidence interval is symmetric around the observed p-value. If the observed p-value is around 0.5, this is reasonable. But now, suppose that the observed p-value is as high as 0.95, n=100 and we want a 99% confidence interval. The standard error of p is now approximated by $\sqrt{0.95(1-0.95)/100} \approx 0.0218$ and $2.58\times 0.0218 = 0.056$ so that we find

$$P(0.894 \le \pi \le 1.006) = 0.99$$

but the upper bound of the confidence interval is larger than 1, while we know that π can not be larger than one. Moreover, with very high observed p-values, we would rather believe that the population value is smaller than that it is larger than the observed value. But this asks for an **asymmetric** interval, for which we need another formula. Here is one which looks complicated but which gives nice results in many cases (Hays, 1977, p. 379[1]):

$$\frac{n}{n+z^2}\left[p+\frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n}+\frac{z^2}{4n^2}}\right]$$

In the formula, z stands for the value from the tables of the normal distribution: 1.645 for a 90% interval; 1.96 for a 95% and 2.58 for a 99% interval. The sign '±' must be replaced by a '+' to yield the upper bound and by a '-' to find the lower bound of the interval. We apply this to the preceding example (2.58^2 = 6.656), finding

$$\begin{aligned}&\frac{100}{100+6.656}\left[0.95+\frac{6.656}{200} \pm 2.58\sqrt{\frac{0.95\times 0.05}{100}+\frac{6.656}{4\times 100^2}}\right]\\ &= \frac{100}{106.656}[0.983 \pm 2.58\times 0.0253]\end{aligned}$$

which gives 0.860 as lower bound and 0.983 as upper bound. Notice that the observed p-value of 0.95 is much closer to the upper bound than to the lower bound.

[1] Hays, W.L., *Statistics for the social sciences*. London: Holt, Rinehart and Winston, 1977^2.

C.4. Partial credit items and *p*-values

A binary item is an item where the score can assume only two values: zero for an incorrect and one for a correct response. A partial credit item is an item where the score can range from zero to a certain maximum that is larger than one, and where all intermediate (whole numbered) scores can be obtained as ‘partial credits’. The simplest form is where one gets two points for a perfect response, zero points for a totally wrong answer and one point for an answer that is neither totally wrong nor totally correct.

The (observed) *p*-value of a binary item is the proportion of test takers in the sample having the item correct. When one tries to generalize the definition of the *p*-value for binary items to partial credit items, one runs into trouble, because the notion of ‘correct’ becomes ambiguous in this case. There is, however, a convenient way to look at *p*-values which easily generalizes to partial credit items, namely, the notion of average relative (item) score. For binary items this is illustrated in Table C.1 with a numerical example and symbolically.

Table C.1 The observed *p*-value as average score

	example		symbolically	
score	frequency	proportion	frequency	proportion
0	189	0.30	N_{i0}	$1-p_i$
1	441	0.70	N_{i1}	p_i
total	630	1	N_i	1

The average score on this item is computed as

$$\frac{189\times 0+441\times 1}{630}=\frac{189}{630}\times 0+\frac{441}{630}\times 1=\frac{441}{630}=0.7=p_i$$

So, in the case of a binary item, we see that the proportion correct or the average score mean the same thing. Now we apply the same procedure to a partial credit item with a maximum score of 3. (See Table C.2.)

Table C.2 The average item score for a partial credit item

	example		symbolically	
score	frequency	proportion	frequency	proportion
0	126	0.20	N_{i0}	p_{i0}
1	189	0.30	N_{i1}	p_{i1}
2	252	0.40	N_{i2}	p_{i2}
3	63	0.10	N_{i3}	p_{i3}
total	630	1	N_i	1

It is easily checked that the average score in this case is

$$\frac{126\times 0+189\times 1+252\times 2+63\times 3}{630}=1.4$$

As an index of difficulty this average is not very useful, because we have to remember that the maximum score for this item is 3. Therefore, the average score is divided by the maximum score (yielding a relative average score) of 1.4/3 = 0.467, i.e. 46.7% of the maximum score. The relative average score is (by definition) a number between zero and one. Notice that with binary items, average score and relative average score coincide, because the maximum score is one. If the term *p*-value is used with partial credit items, it refers to the average relative score.

C.5. Correlations between distractors and test score

To compute a correlation, one needs two series of scores. To compute the item test correlation, for example, one score is the test score, and the other score is the item score. The latter equals one if the answer is correct and zero if the answer is incorrect. The correlation is computed using the usual formula for a product-moment correlation (Pearson correlation). The computation will only fail if the observed *p*-value of the item is either zero or one, because in these cases the variance of the item score is zero.

To compute the correlation between a distractor and the test score, one must **recode** the answers given by the test takers. Suppose the item under study is a multiple choice item with four alternatives (A, B, C and D), alternative B being the correct answer: this means that an item score of one is given to every test taker who chose B, and a zero to the others. To compute the correlation between test score and distractor A, one has to create a new binary variable, giving a ‘score’ of one to every test taker who chose A, and zero to the others. The correlation looked for is the correlation between this new variable and the test score. To compute the correlation between test score and distractors C and D, one should proceed in a similar way. When using multiple choice items, it is good practice to compute the correlations between distractors and test score. In well constructed items, these correlations should be negative.

This application also illustrates the need of storing in some way the original observations. If one stores only the item scores (zeros and ones), it is not possible to compute the correlation between distractor and item score, because it is impossible to know which one of the distractors has been chosen from the mere knowledge that the answer was not correct.

C.6. More on graphical item analysis: DIF

The discussion on graphical item analysis is a good opportunity to introduce a concept that has received a lot of attention in the last two decennia, the so-called Differential Item Functioning (DIF). The ideal of fair testing requires that an item ‘behaves similarly’ in distinct populations, for example in the populations of boys and girls. It is, however, not so easy to state what is meant or should be meant by ‘similar behaviour’. One could claim, for example, that an item should be equally difficult in the populations of boys and girls, but using such a definition will cause serious trouble. It is a well established fact that at the age of 12, girls tend to be less proficient in arithmetic than boys. If the difficulty of the item is operationalised by its *p*-value, it is to be expected that the *p*-value of a typical arithmetic item will be lower in the girls’ population than in the boys’ population. This illustrates nicely the population dependence of the *p*-value. Usually this will hold for most or all items in an arithmetic test. But if we stick to the requirement that to be fair each item should be equally difficult in both populations, (and suppose an admissible test is required to have this property, and that only items with this property are included in the test), then by necessity we will find that on a ‘fair’ test, the average score of boys and girls is the same. But this approach implies that all differences are unfair, because it can be applied to any pair of populations, including the populations consisting of myself and my neighbour respectively.

So we need a more qualified definition of DIF, one that leaves room for differences between populations. Such a definition is formulated as a conditional statement. We apply it to the example of boys and girls. An item shows no DIF if in the (conceptual) population of boys with an arbitrary but fixed level of proficiency and the (conceptual) population of girls with the same level of proficiency, the *p*-values of the item are identical. Notice that this identity of the two *p*-values must hold at each level of proficiency. Stated more simply: absence of DIF means that the item should be equally difficult for boys and girls with the same level of proficiency.

In practice of course, we do not know the exact proficiency level of any test taker, but we can use the test score as a proxy. If, as before, test takers are grouped in a number of groups (of reasonable size), we can plot the observed *p*-values in each group for boys and girls separately. In Figure C.2, two examples are given from a mathematics examination. The legend refers to girls (Sg = 1; Sg stands for subgroup) and boys (Sg = 2).

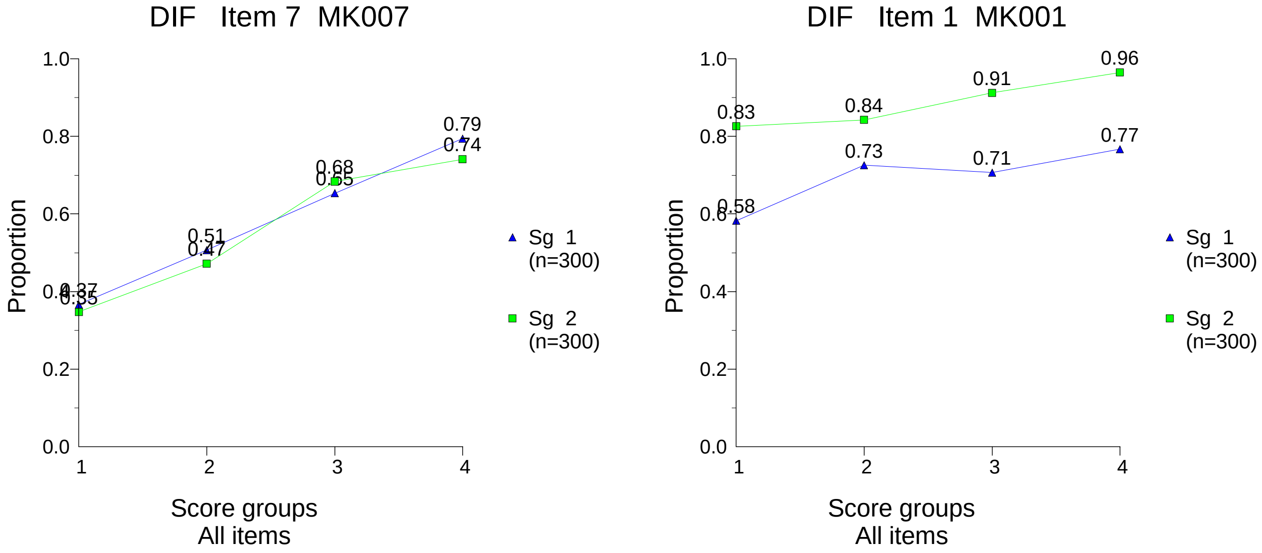

Figure C.2. Examples of DIF analysis

For item 7, there is no evidence of DIF: the *p*-values for boys and girls are very similar in each group (remember that these *p*-values contain an estimation error; so we cannot expect them to be identical in a sample). For item 1, on the other hand, there is clear evidence of DIF: the item is substantially harder in each girls' group than in the corresponding boys' group. Although there exist techniques for testing these differences statistically, in a clear-cut case as this, a plot is convincing enough. Scanning similar plots for all items in the test will reveal immediately important DIF as with item 1.

Although gender is commonly used as an example to explain and illustrate DIF, it is by no means the only variable where DIF can be investigated. In the United States of America cultural fairness of tests is often a strong requirement, and ethnical and racial background is often used as the contrasting variable in DIF-studies. In the general domain of achievement tests, an important variable to be used in DIF studies is the method of instruction used: it may be the case that some items turn out to be easier when the content matter of the test has been taught by method A, say, rather than by method B. A detailed DIF analysis may be revealing in such a context. Another highly relevant example is the use of mother tongue as the DIF-variable in case a test is administered to groups with different linguistic backgrounds, like the TOEFL.

C.7. A graphical aid in constructing parallel forms

The construction of parallel forms can occur in different situations:

- A parallel form for an existing (and already used) test has to be constructed;
- Two (or even more) parallel forms are to be constructed from scratch;
- An existing test has to be split in two halves which are parallel (to use the split half method for estimating the reliability).

In all these cases a simple method can be used to construct the parallel forms in a graphical way. The idea is to construct two test forms which are approximately **strictly parallel**. This means that each item in one form has a twin in the other form with (approximately) the same psychometric qualities. In the framework of CTT one tries to have a match on two qualities: the difficulty and the discrimination, which are usually operationalised by the *p*-value and the item-test (or item-rest) correlation.

The starting point of the method is to construct a scatter diagram where each item is represented by a point in the plane. The *x*-coordinate is the *p*-value of the item, the *y*-coordinate the item-test correlation. The position of the item is symbolized by a (short) item label, such that items can easily be identified. An example is given in Figure C.3. Two items with graphical representation near each other have approximately the same *p*-value and the same discrimination. In Figure C.3 pairs are represented by lines connecting two item points. Pairs are formed such that the distance between the two item points in each pair is a small as possible.

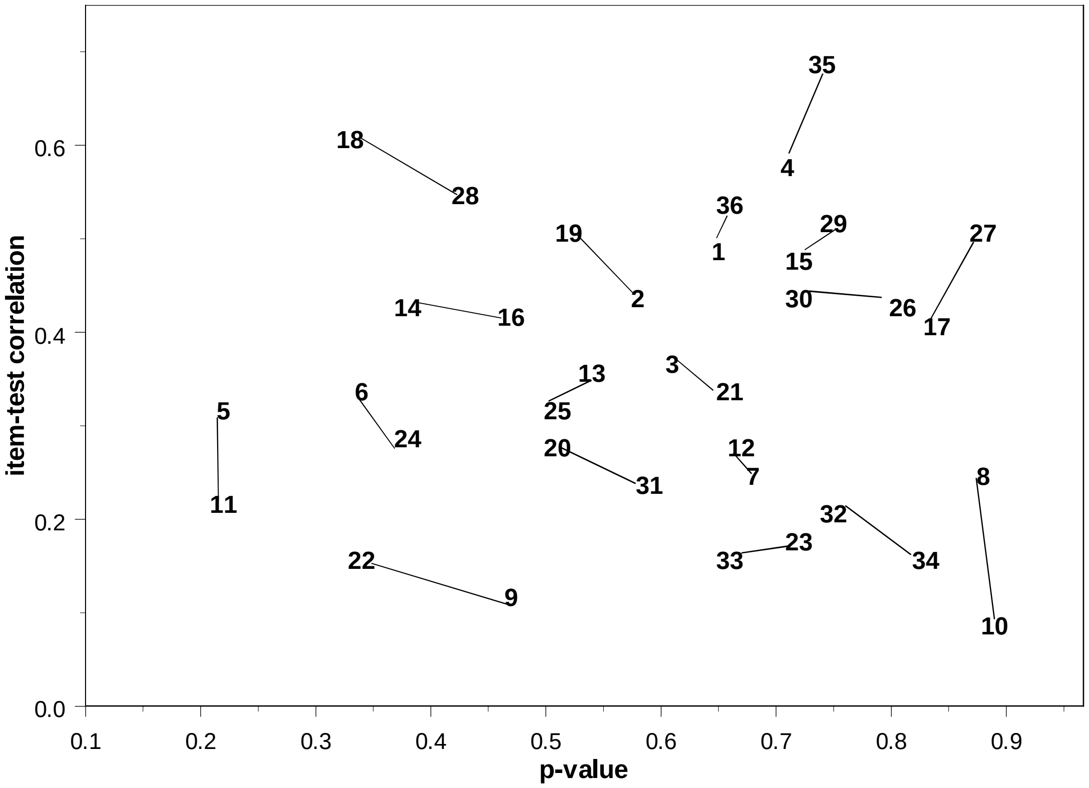

Figure C.3 Graphical construction of parallel forms

To construct the approximate parallel forms, the two items belonging to a pair should be assigned to the forms at random. There are a number of remarks to be made at this point:

1 If data are available on all the items from the same sample (and this will be the case when splitting an existing test in parallel halves, or in the construction of two parallel forms from scratch), it is always wise to check the extent to which the formation of parallel forms has been successful. In the two parallel forms the *p*-values of the items will not be different from their values when considering all items as belonging to a single test, but usually the item-test correlations will change.

2 If data are collected on two different samples (which may be the case if a new parallel form to an existing test has to be constructed), one should be very careful in using statistically equivalent samples. Both samples should be representative for the same target population.

3 If a parallel form for an existing test has to be constructed, it is wise to have more items to select from than what is strictly needed in the test. If the existing test consists of 35 items, it is advisable to have at least 50 items for the new test, such that 35 pairs can be formed, leaving 15 or more items unused. If one does not have such a provision, it may appear that it is not possible to construct a parallel form, because, for example, the new items are on average easier than the old ones.

4 The construction of the two parallel forms, as exemplified in Figure C.3 is done ‘by hand’, and it is not guaranteed that the proposed solution in the figure is the best possible. This is not a big problem, however: the aim is to construct two forms which are reasonably in balance with respect to the two psychometric qualities of the items. But it may appear that by proceeding in this way the two test forms show a quite strong unbalance in other respects, for example, with respect to

content. It is **not** the case that psychometric balance has priority to content. The ultimate decision is in the hands of the test constructor, and the method exemplified in Figure C.3 is only meant as a convenient tool in the construction of the parallel forms. One can extend control by very simple means, just as using a different colour of the item labels to distinguish between open ended and multiple choice items, or underlining and italicising to distinguish different content categories, and try to form pairs where content category, item format, *p*-value and discrimination are as similar as possible.

C.8 The Spearman-Brown formula

There exists a powerful formula to control the test reliability, known as the Spearman-Brown formula. It says how the reliability changes as the test is lengthened (or shortened). Suppose a prototype of a test has been constructed which contains twenty items; this number of items is in some way considered as a standard length. So, we could say that it has the length of 1. The reliability of this test will be denoted by $\rho(1)$ for short. The Spearman-Brown formula can tell us what the reliability of the test would be if it contained forty items, that is, if it had the length of 2. And more generally, it tells us what the relation is between the reliabilities of a test of length 1 and a test of length k, where k is an arbitrary positive number. Here is the formula:

$$\rho(k) = \frac{k\rho(1)}{1+(k-1)\rho(1)}$$

and here is an example. Suppose the test of 20 items has a reliability of 0.63, but the possibility exists to extend the test to 30 items, i.e. to make the test 1.5 times as long is it actually is. So, we have to apply the formula with $k = 1.5$ and $\rho(1) = 0.63$, yielding

$$\rho(1.5) = \frac{1.5 \times 0.63}{1+(1.5-1)\times 0.63} = 0.719$$

The formula can be applied also to see the effect of shortening the test. Suppose we can apply only a test of 10 items instead of 20, then $k = 10/20 = 0.5$ and applying the formula gives

$$\rho(0.5) = \frac{0.5 \times 0.63}{1+(0.5-1)\times 0.63} = 0.460$$

Some users do not understand fully the meaning of ‘*k*’ in formula (10). It definitely does not denote the number of items; it denotes the ratio of a new number of items to some reference number, usually the number of items in an existing test. This latter number is then considered as the standard length (length of 1). The effect of test lengthening (or shortening) can be displayed graphically by a number of curves, as in Figure C.4.

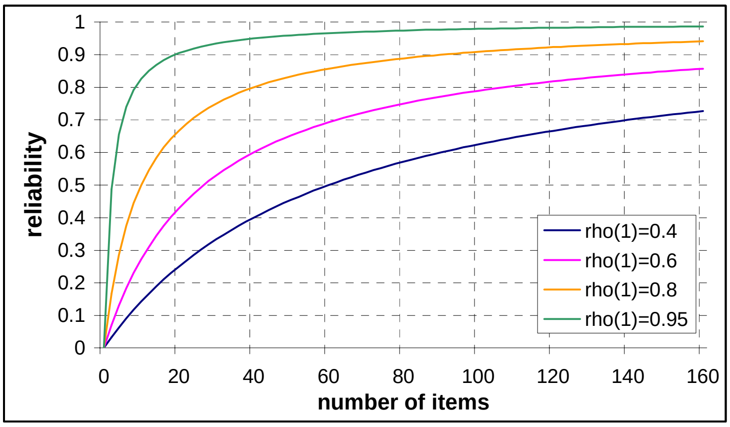

Figure C.4 Graphs of the Spearman-Brown formula

These graphs display a number of interesting characteristics:

1. All curves will eventually go to 1 if the number of items is large enough.
2. Of course, many more curves can be produced. The curves in Figure C.4 are just a few examples, and were produced with 40 items as standard length.
3. All curves have the same feature: starting with a small number of items, and then adding progressively more items, makes the curves grow rapidly at the start and more and more slowly as the number of items increases. A nice example is offered by the second curve from below. With 20 items, the reliability is (about) 0.40; adding 20 items causes an increase to 0.60, but adding another 20 items is not sufficient to reach a reliability of 0.70. Or, in short, adding items leads to a modest gain but removing items causes a great loss in reliability.

The Spearman-Brown formula is the most important practical tool to control the reliability of a test under construction. Sometimes a certain reliability is set as a minimum requirement for a test (in a certain population). One starts with the construction of the test, and the first analysis reveals that the target is not reached. Then one can use the Spearman-Brown formula to estimate the number of items that must be added to reach the target. Here is an example. Assume that the target reliability of a test is 0.85. Assume that a first analysis is done with a provisional test of 25 items, which yields an (estimated) reliability of 0.77. A very practical question then is to know how many items should be added to reach the target. If we take 25 items as the standard length, then it must hold (by applying the Spearman-Brown formula) that

$$0.85 = \frac{k \times 0.77}{1 + (k-1) \times 0.77}$$

and this equation (with *k* unknown) can be solved to find *k*:

$$k = \frac{0.85 \times (1 - 0.77)}{0.77 \times (1 - 0.85)} = 1.693$$

meaning that the test should have 1.693 times its present length, that is, contain 25× 1.693 =42.3 items. As fractions of items do not exist, this means that we will need at least 43 items to reach the target (42 will not be enough.). The preceding calculation leads to a very useful and practical formula:

$$k = \frac{\rho_{\text{target}}(1 - \rho_{\text{obs}})}{\rho_{\text{obs}}(1 - \rho_{\text{target}})}$$

where ρ_{obs} is the reliability one actually has reached, and ρ_{target} is the target reliability. (But again, remember that the result *k* of the formula is not the number of items, but the factor with which the actual number has to be multiplied.)

We will end this section with an example of the popular saying: the sting is in the tail. There is a big risk in applying the Spearman-Brown formula purely mechanically. The Spearman-Brown formula is only valid under quite strict conditions (which can not be discussed in detail in this appendix). Suppose one has to double the actual test length to reach the target reliability. If the provisional test contains 25 items that are constructed in a careful and professional way, one cannot hope to reach the target by adding 25 sloppy items, constructed in a hurry on a Sunday afternoon. More generally, one can express the requirement for the validity of the formula by saying that the test should be lengthened homogeneously. This means the added items should be very comparable (as a whole) to the items already present in many respects: the content coverage should be the same, the general level of difficulty and discrimination, perhaps also the format (a test consisting of 25 essay questions is not doubled homogeneously by adding 25 multiple choice questions.) All this of course cannot be controlled in full detail, and that is why the Spearman-Brown formula, beautiful as it is, will only yield approximations in practice.

C.9 Confidence intervals for the true score

We need some mathematical notation to express the relation between the standard error of measurement and the reliability. The symbol X will be used to represent the **observed** test score, and the reliability of X will be symbolized as $\text{Rel}(X)$. The standard deviation of the observed test scores is denoted as $\text{SD}(X)$, and the standard error of measurement as $\text{SE}(X)$. The relation between the standard error of measurement and reliability is given by the following formula:

$$\text{SE}(X) = \text{SD}(X)\sqrt{1-\text{Rel}(X)}$$

The important fact about this formula is that we can compute the standard error of measurement from observable quantities: the standard deviation of the observed scores and the reliability. We use a well-known case as an example. In the use of intelligence tests, the scores (IQ) are expressed on a scale such that (in a well defined population) the mean IQ is 100 and the standard deviation is 15. Notice, that these quantities refer to observed scores, not to true scores, and that the reliability of many intelligence tests is well above 0.9, but certainly not equal to one. In Table C.3, the standard error of measurement is given for a number of cases.

Table C.3. Standard error of measurement with $\text{SD}(X) = 15$

Reliability	$\text{SE}(X)$
0.85	5.81
0.88	5.20
0.91	4.50
0.94	3.67
0.97	2.60

These figures may come as a surprise, yet they are the result of a simple calculation. The table is important, as it should dissuade us from statements like “the reliability is as high as 0.97, which is virtually one” and then proceed as if it is really equal to one. Let us see what we can say about John’s IQ, if we have found that his observed IQ equals 112, and the reliability of the IQ-test is indeed as high as 0.97.

Since our measurement is not perfect, but contains a measurement error, the best we can hope is to define an interval that contains John’s real IQ (to be understood as his true score). But here a new problem crops up: Classical Test Theory does not say anything about the shape of John’s private error distribution. We cannot say that it is symmetric, and a fortiori we cannot be sure that it has the form of a normal distribution. Although it is possible in statistics to define confidence intervals without any additional assumption about the shape of the distribution, these intervals are usually disappointingly large. We can narrow these, but at the price of extra assumptions. Commonly, it is assumed that the error distribution is normal. If we buy this assumption, we can define a confidence interval in the usual way (see Section C.3), which as a mathematical expression looks like this:

$$\text{Prob}(X_{\text{John}} - 1.645 \times \text{SE}(X) \leq \tau_{\text{John}} \leq X_{\text{John}} + 1.645 \times \text{SE}(X)) = 0.90$$

or, in words, there is a probability of 90% that the constructed symmetric interval true score will contain the true score; the lower bound of the interval is the observed score minus 1.645 times the standard error of measurement and the upper bound is the observed score plus 1.645 times SD(E). Replacing the symbols by the numbers we know, we find

$$\text{Prob}(112 - 1.645 \times 2.6 \leq \tau_{\text{John}} \leq 112 + 1.645 \times 2.6) =$$
$$\text{Prob}(107.7 \leq \tau_{\text{John}} \leq 116.3) = 0.90$$

This means that the 90% confidence interval is 116.3 – 107.7 = 8.6 IQ points, which is more than half a standard deviation of the observed scores. Of course, we can apply a similar procedure not only to John but to an arbitrary member of the population. But if we do so, we have to remember that in 10% of the cases, the true score will lie outside the thus defined interval. So we see clearly that we cannot treat a reliability of 0.97 as being 'virtually one'.

C.10 Important theoretical results

The theoretical definition of reliability (see Section C.1) is the ratio of true score and observed score variance. This ratio cannot be computed in practice, because the true score variance is not known. If, we have a test which is parallel to a certain X (and which is commonly denoted as X'), then the reliability can be computed because it is theoretically shown that the correlation between two parallel tests equals the reliability of the test (and of its parallel form as well). There is, however, another important theoretical concept which is closely related to the reliability, namely, the correlation (in the target population) between observed and true scores. This relation is presented together with the earlier results in the following composite equation:

$$\text{Rel}(X) = \frac{\text{Var}(T)}{\text{Var}(X)} = \rho(X, X') = \rho^2(X, T)$$

Notice that the reliability is the squared correlation between observed and true score, and it follows immediately that

$$\rho(X,T) = \sqrt{\text{Rel}(X)} \tag{C.1}$$

This is an important theoretical result. One might wish to be able to measure without measurement error, but in language testing, as in many other areas, this is practically not possible, and all one can obtain is fallible results: the observed outcomes of a measurement procedure are in error. The above formula expresses directly the correlation between observed values and the theoretical construct of interest.

Since the reliability of a test is a number between zero and one, the correlation between observed and true score is larger than the reliability (it is equal only in case the reliability is zero or one). In Table C.4, some examples are displayed.

Table C.4. The relation between reliability and $\rho(X,T)$

$\text{Rel}(X)$	$\rho(X,T)$
0.2	0.45
0.4	0.63
0.6	0.77
0.8	0.89
0.9	0.95

This relation has important implications for the discussion on validity. An important aspect of validity concerns the relation between the test scores and some other variable, which in many cases is also a test score. But both test scores are in error, and these measurement errors will tend to attenuate (i.e., lower) the correlation. Ideally one would like to know the correlation between the true scores on both tests. There exists a famous formula for this correlation, but we need some extension of the notation to write it down compactly. The two observed test scores will be denoted by X and Y and their corresponding true scores are denoted by T_X and T_Y respectively. The formula is:

$$\rho(T_X, T_Y) = \frac{\rho(X,Y)}{\sqrt{\mathrm{Rel}(X)\ \mathrm{Rel}(Y)}} \qquad \text{(C.2)}$$

or in words, the correlation between the true scores is the correlation between the observed scores divided by the square root of the product of the reliabilities. Since reliabilities are generally smaller than one, the denominator of the fraction will also be smaller than one, whence it follows that the correlation between true scores is larger than the correlations between observed values, or, as one usually says, the correlation between observed scores is attenuated by measurement error. The formula is also called ‘the correction for attenuation’. (Notice that the formula does not apply when one or both reliabilities are zero, but in such a case the correlation between the true scores is also zero.)

This formula plays an important role in discussions about the construct validity of a test. If two tests measure the same concept, one usually finds that they correlate less than one, and this can be explained by the attenuation formula: the correlation is lowered by the fact that both test scores contain measurement error. But if X and Y really measure the same concept, then the correlation between their true scores should be equal to one, i.e., they should be **congeneric**. Replacing the left hand side of the attenuation formula by 1, we find immediately that

$$X \text{ and } Y \text{ are congeneric } \Leftrightarrow \rho(X,Y) = \sqrt{\mathrm{Rel}(X)\mathrm{Rel}(Y)}$$

i.e., if X and Y are congeneric then their correlation should be equal to the square root of the product of their reliabilities.

In practice, one cannot use formula (C.2) as its stands, because this formula refers to population values, and in practical situations one has to use sample estimates for the correlation and the two reliabilities, and because of the fraction in the formula, the result can be a number that is larger than one, which of course cannot be a correlation. The most notorious pitfall, however, with this formula is when one uses a lower bound to the reliabities, such as Cronbach's alpha. If tests are heterogeneous, this coefficient can be substantially lower than the reliability, and using these as estimates of the reliability in the formula, will make its denominator too small, and as a result the result of the fraction too high, giving in some cases results far exceeding one, or results near one, even if the two tests are not congeneric at all.