CHAPTER 4

RELIABILITY STATISTICAL ANALYSIS

The purpose of this chapter is to analyse data (scores) obtained from the tests written by students (see 3.5). Specific statistical analysis used for establishing the reliability of a measurement instrument is applied to the test. It should be emphasised that the focus is not on the reliability or not of the specific test under discussion, but on how a reliability calculator (See Annexure B) and the results obtained from the reliability calculator be used to establish the reliability of scores obtained from that measurement. Scores obtained from a measurement instrument must be reliable because any measurement must have reliability as a prerequisite to validity (Oosterhof 1994:74). Therefore the enhancement of test items in order to improve validity, as will be discussed in Chapter 5, can only take place after reliability has been established.

In this chapter each statistical measurement is first described and then an example from the data is provided in the subsequent section. This was done with the aim of improving the readability and interpretation of results. In some cases the results obtained from the reliability calculator were compared to those obtained from the statistician using the SPSS program, to verify the values obtained with the reliability calculator.

Sax (1997) pointed out that the reliability refers to test scores or measurements, not the tests themselves. Tests consist of items that by themselves provide no estimate of reliability. The test must be administered and scored before reliability can be estimated. It is important to note that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Wilkenson & Taskforce on statistical inference 1999:596). Vacha-Haase, Kogan and Thompson (2000) support the view that reliability is a property of scores and not of tests.

The tendency toward consistency from one set of measurements to another is called reliability (Stanley & Hopkins 1972:357). The fact that repeated sets of measurements never exactly duplicate one another is what is meant by unreliability. At the same time however, repeated measurements of a series of objectives or individuals will ordinarily show some consistency. It is highly possible to determine the reliability of the scores obtained from a measurement instrument. Several empirical procedures have been devised to estimate reliability, as indicated in Figure 22.

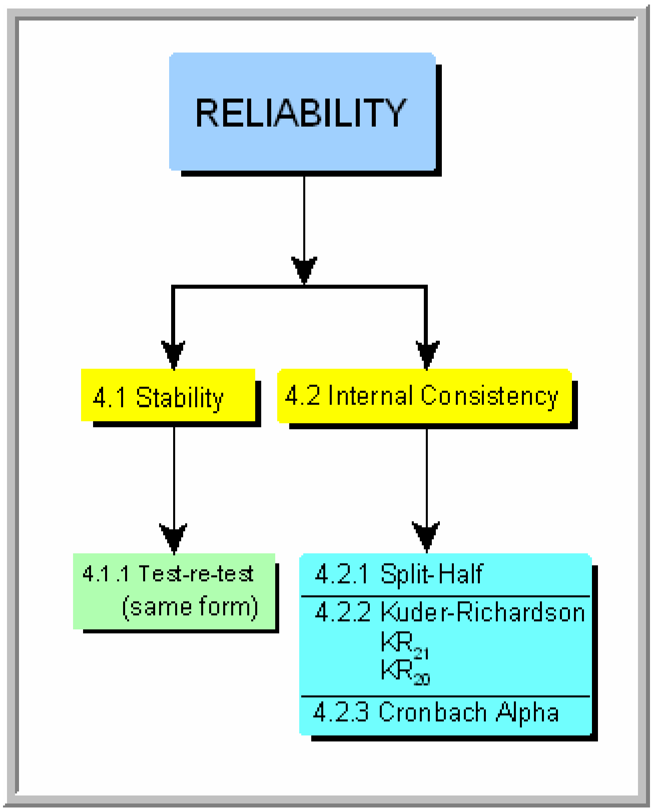


Figure 22 Estimating reliability

4.1. STABILITY

Stability is measured by correlating test scores obtained from the same students over a period of time. If individuals respond consistently from one test to another, the correla tion between the scores will be high (Sax 1997:275). Factors that affect stability, namely time and objectivity, will be elaborated on in the next paragraph.

In the first place, stability is strongly affected by the amount of time that elapses between successive administrations of the same test. If a second test is administered immediately after the first test, chances are good that the students will mark the same answer twice. This can be contributed to students using their short-term memory, which may produce highly consistent test scores However, the usefulness of high reliabilities over very short periods of time is questionable (Sax 1997:275). In general, educators want to know how stable measurements are over periods of time. For the purpose of the study, a time interval of two weeks was used as suggested by Metsämuuronen (2002:51) to be an acceptable period. This time period was taken into consideration for the research design as described in 3.6.

The second condition that affects stability coefficients is the objectivity of the measurements (Sax 1997:276). Stability coefficients may be low because raters or scorers used different scoring criteria at different times. This effect can be reduced by developing more objective scoring systems. This is more applicable to measurement instruments consisting of essay-type (open-ended) items. For the purpose of the research, the measurement instrument that was used consisted of objective items only. It is important to note that tests cannot be stable or unstable, but observations (measurements) based on tests can (Sax 1997:276). The technique to determine stability is discussed next.

4.1.1 Test-re-test with same forms technique

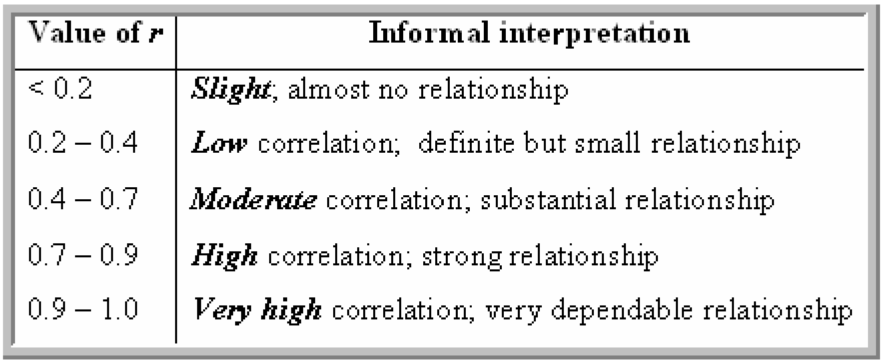
Discussion

The test-re-test approach is intended to determine stability as defined in the definition of reliability: If the measurement is reliable, the same students will give the same answers with the same measurement instrument. The reliability is the correlation between the scores on the two instruments. If the results are consistent (stable) over time, the scores should be similar.

To determine stability, the relationship between the two scores obtained from the *test* and the *re-test* must be considered. This is referred to as the correlation. When one works with relationships a formal method based on calculations can give a numerical value for the degree of correlation between the two sets of scores. This is done using the Pearson product-moment correlation coefficient (*r*).

The value of *r* will always fall within the range -1 to +1. An *r* of -1 means a perfect negative correlation and an *r* of +1 means a perfect positive correlation. An *r* of 0 means zero correlation. These values are easy to interpret but values that fall between 0 and +1, or between 0 and -1, are more complex. Guilford (1982) offers an informal interpretation of the value *r*, as shown in Table 1.

Table 1 Interpretation of Pearson product-moment correlation coefficient (*r*)



Value of *r*	Informal interpretation
< 0.2	***Slight***; almost no relationship
0.2 – 0.4	***Low*** correlation; definite but small relationship
0.4 – 0.7	***Moderate*** correlation; substantial relationship
0.7 – 0.9	***High*** correlation; strong relationship
0.9 – 1.0	***Very high*** correlation; very dependable relationship

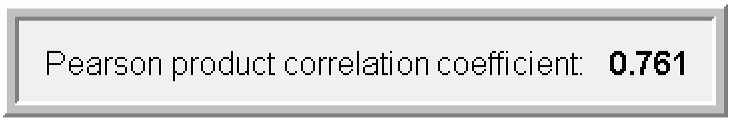
Another way of establishing a relationship between two sets of scores is by examining a scatter plot drawn from the data. Results from each of these sets are shown respectively:

Results obtained from the Pearson product-moment correlation coefficient

Two sets of data (*test* and *re-test*) were exported from the CCAT into the reliability calculator where the Pearson product-moment correlation coefficient was calculated. The following formula was used.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

where:
x = the sample mean AVERAGE(array *test*) and
y = sample mean AVERAGE(array *re-test*).



Pearson product correlation coefficient: **0.761**

When measuring stability, in this case using the test-re-test (same form), there is no real agreement on when the result is considered adequate due to various factors that can influence the result. It is therefore not surprising that different textbooks provide different suggestions on what value for correlation is acceptable. The interpretation of the Pearson product-moment correlation coefficient largely depends on the purpose of the test. If the purpose is to make serious decisions regarding students, then the correlation should be high, confirming the stability of the measurement instrument used. Considering Guilford's (1982) values in Table 1, the correlation between *test* and *re-test* the value of 0.761 can be considered as high, indicating a strong relationship.

Results obtained from the scatter plot graph

A scatter plot is also needed when evaluating the correlation between two sets of data. The purpose of the scatter plot is to show whether there is a linear relationship amongst the two sets of data. The scatter plot graph in Figure 23 was generated with the SPSS statistical analysis program for the scores obtained from the *test* and *re-test*.

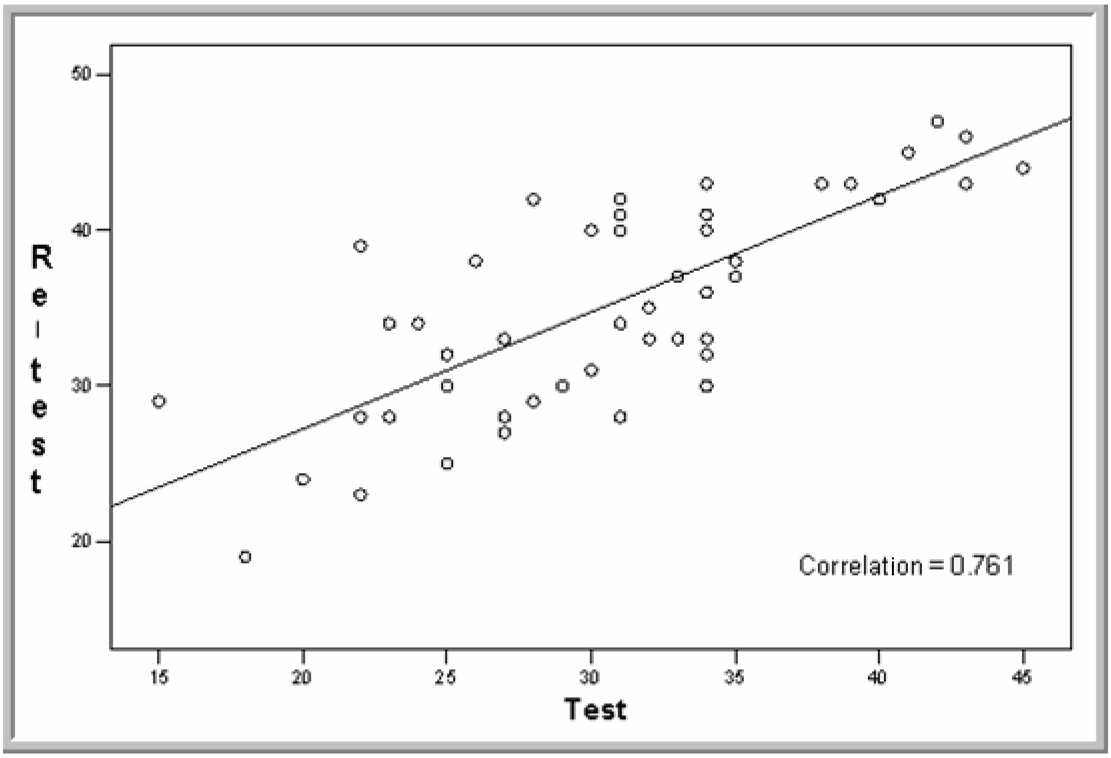


Figure 23 Scatter plot of *test* versus *re-test*

An analysis of the scatter plot shows a definite tendency towards linearity, as most of the scores are fairly close to the regression line (as in Figure 23). Thus, from the correlation value of 0.761 and the scatter plot it can be concluded that the measuring instrument was stable over the time span for which it was administered.

4.2. INTERNAL CONSISTENCY

The techniques discussed above have the disadvantage that it is time-consuming to administer the same test or parallel forms test twice. This also implies that two forms need to be constructed. When two parallel forms are constructed, the question will always arise whether the items are really the 'same'. In most cases educators want to estimate reliability from a single administration of a test (Sax 1997:277). This requirement has led to the measuring of internal consistency, or homogeneity. Internal-consistency measures consistency within the instrument (consistency among the items). Several internal-consistency methods exist.

All internal consistency measurements have one thing in common, namely that the measurement is based on the results of a single measurement (a test is written only once). For this reason, internal consistency is considered to be the easiest form of reliability to investigate. This method measures consistency within the instrument using three different techniques: the split-half, the Kuder-Richardson and the Cronbach alpha. A discussion follows of the techniques mostly used to determine internal consistency without re-administering a test. In the study all three techniques were used to obtain and to analyse the data.

4.2.1 Split-Half Technique

Discussion

A total score for the odd number questions is correlated with a total score for the even number questions. This is often used with dichotomous variables that are scored 0 for incorrect and 1 for correct. When the test items are split, the assumption is made that they are homogeneous, meaning that they are measuring the same content. Tests can be split in other ways as well, but those are generally not recommended (Sax 1997:278).

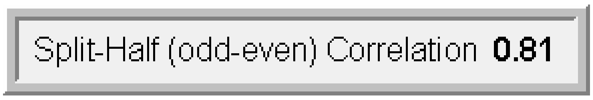
Another method entails that the test is split into a first half and a last half, and then correlated. The disadvantage of doing this is that as with most achievement tests, items are arranged in ascending order of difficulty level. Therefore, the domain coverage might differ, and the scores on the first half of the test may not correlate well with the scores on the second half. Tredoux and Durrheim (2002:213) are also concerned with the way the test is split in half. He says that ideally, the scale should be split in such a way that the halves are roughly equivalent in items in terms of difficulty and domain coverage. The statistical program SPSS used for comparing the values on the reliability calculator (See Annexure B) uses the latter split-half methods.

Results obtained

The reliability calculator uses the odd-even split, which is a more acceptable method as will be seen further on.

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

where:
x = means AVERAGE(array 1) and
y = sample and AVERAGE(array 2).


Split-Half (odd-even) Correlation **0.81**

As the split-half technique correlation provides a measure of reliability of measurements for half a test but not for a test as a whole, the split-half technique has the following two limitations, according to Sax (1997:278):

Firstly, the major source of error is the items themselves; any changes in students over time do not affect this type of reliability. Unreliability therefore results from the differences in item-content quality on the two halves of the test rather than from the students.

Secondly, splitting the test in two effectively halves the number of items, thus decreasing reliability, as we know that more items in a test provide a more adequate sample of whatever trait or ability is being measured and therefore a longer test tends to be more reliable than a shorter one.

To estimate the reliability of the whole test from knowledge of the correlation between the halves, the Spearman-Brown formula must be used to compensate for the reduction in items (Sax 1997:278). The Spearman-Brown prophecy formula is applied to the correlation to determine the reliability. The Spearman-Brown formula compensates for the fact that the test is split in two halves, decreasing the number of items, which decreases the reliability. As always: the more items, the higher the reliability.

Spearman-Brown formula from Tredoux and Durrheim (2002:213):

$$r_{sb} = \frac{2r_{hh}}{1 + r_{hh}}$$

where: r_{sb} = Spearman-Brown reliability
r_{hh} = the correlation coefficient between the two halves

Spearman-Brown Prophecy: **0.90**

Using the Spearman-Brown prophecy formula, the test internal consistency value increased from 0.81 to 0.9. This yields a very good value for reliability.

4.2.2 Kuder-Richardson Techniques

Discussion

Two persons, Kuder and Richardson, devised several methods for estimating the reliability of scores from a single administration of a test. Their research paper, dated 1937, included numerous derivations of which the 20th and 21st formulas have become the most widely used of their methods for estimating reliability (Oosterhof 1994:84). The formulas have become known as the KR_{20} and KR_{21} formulas and will be referred to as such in this dissertation.

There are two alternative formulas for calculating how consistent student responses are among the questions on a measurement instrument. Items on the measurement instrument must be dichotomously scored (0 for incorrect and 1 for correct). Both the Kuder-Richardson KR_{20} and KR_{21} formulas provide an estimate of the average reliability found by taking all possible splits without actually having to do so (Sax 1997: 279).

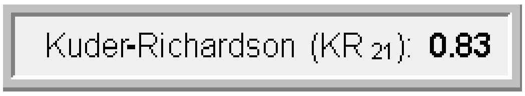
Results obtained from Kuder-Richardson (KR_{21})

The KR_{21} is a shortcut method that will yield reliability coefficients identical to $KR_{20,}$ but only when all items are equally difficult (if this assumption is violated, KR_{21} will always underestimate KR_{20}). For the test on which the formula was applied the items are definitely not equally difficult. When comparing the results of KR_{21} with those of $KR_{20,}$ one sees that it is lower, proving the fact that the items are not equally difficult.

The formula for KR_{21} is:

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{M - \frac{(M)^2}{n}}{SD^2}\right)$$

where n = number of items on the test
M = mean score on the test
SD^2 = variance of scores (the standard deviation squared)



Kuder-Richardson (KR 21): **0.83**

Results obtained from Kuder-Richardson (KR_{20})

All items are compared with each other, rather than one half of the items with the other half of the items. It can be shown mathematically that the Kuder-Richardson reliability coefficient is actually the mean of all split-half coefficients (provided the Rulon formula is used) resulting from different splitting of a test. KR_{21} assumes that all of the questions are equally difficult. KR_{20} does not assume that. It is particularly simple to use if the difficulty level of each item has been determined by means of *an item analysis* which is described in the item analysis section.

The formula for KR_{20} is:

$$\mathrm{KR}_{20} = \frac{n}{n-1}\left(\frac{\mathrm{SD}^2 - \Sigma pq}{\mathrm{SD}^2}\right)$$

where
n = number of items on the test
SD^2 = variance of scores (the standard deviation squared)
p = difficulty level of each item (the proportion of the group that responded correctly)
q = proportion responded incorrectly to each item, or $1 - p$

A value as low as 0.5 is satisfactory for short tests (10 – 15 items), while tests with over 50 items should yield KR_{20} values of 0.8 or higher, with 1.00 the maximum. The KR_{20} value obtained from the test adhered to the aforementioned criteria.

Kuder-Richardson (KR_{20}): **0.86**

An important aspect of this value is that; when important decisions concerning an individual student are to be made, they should not be based on a test where the KR_{20} of the particular test is lower than 0.8. A low value in KR_{20} is usually due to an excess of very easy or difficult items, poorly written items that do not discriminate, or violation of the precondition that the items test a unified (homogeneous) domain.

4.2.3 Cronbach's Alpha

Discussion

A statistical analysis computer program such as SPSS can also be used to calculate Cronbach's alpha (α). Although Cronbach's alpha is usually used for scores which fall along a continuum, it will produce the same results as KR_{20} with dichotomous data (0 or 1). The Cronbach's alpha value for the test was calculated using SPSS, and the results obtained yielded the same value as that calculated by the reliability calculator.

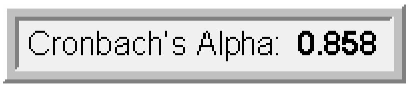
Results obtained

Both values for Cronbach's alpha and KR_{20} for the test are **0.858**.

When a high reliability coefficient is obtained it is no guarantee that the test is well matched with the outcomes. It is only an indication that the items in the test are strongly or weakly related with regard to student performance (Tredoux & Durrheim 2002:213).

$$r_\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_j^2}{\sigma^2}\right)$$

where: $\Sigma\sigma_j^2$ = sum of the item variances
σ^2 = variance of the total score on the scale
n = number of items



Cronbach's Alpha: **0.858**

A general guideline for rejecting a measurement instrument (test) is that in which the alpha value is less than 0.6 (Nunally & Bernstein 1994). It can be seen that the alpha value obtained from the test is well within this limit.

4.3 CONCLUSION

Internal consistency should be high if all items are a variation of the same knowledge base or skill that is being measured by the test. If one test is used to measure multiple outcomes, the reliability coefficient value might be lower due to the fact that a student who knows the content of one outcome may not be as proficient in relation to another outcome. Thus low reliability coefficient is in this case not an indication that some of the items used in the test need to be re-evaluated, and the only way to know is to do an evaluation of the test to confirm validity and reliability.

Results in this chapter prove that objective items administered online can yield a high reliability. The functionality provided by CCAT supports the analysis and subsequent improvement of test items.