Chapter 9
Test-Retest

Case Example 9

You administer the DASH Copy Best subtest to Jimmy, aged 15 years, who obtains a scaled score of 7. You then repeat this subtest after 30 weeks and he then obtains a scaled score of 11. You want to know if the increase of 4 scaled scores is significant or not.

Points to Consider

A scaled score is a form of standard score with a mean of 10 and a standard deviation of 3. Chapter 3 stated that for scores on two different tests that were significantly different from one another, we used the SE_{mdiff} formula, which for two tests with the same standard deviation is:

$$SE_{mdiff} = SD \times \sqrt{2 - r_a - r_b}$$

If we wish to compare two scores on the same test, then the formula can be amended to:

$$SE_{mdiff} = SD \times \sqrt{2 - (2 \times r_a)}$$

So, for our case example we have the standard deviation as 3 and the test-retest reliability coefficient from Table 6.2, page 80 of the DASH manual for this age as 0.72, making the formula:

$$SE_{mdiff} = 3 \times \sqrt{2 - (2 \times 0.72)} = 2.45$$

If we then multiply 2.45 by 1.96 that is derived from the *z* score amount of scores that lie outside 95% of the total number of scores within the normal distribution, we obtain the amount 4.8 in units of scaled scores. This is the amount that needs to be exceeded for you to conclude that the second test score is significantly different to the first score at the 95% level of confidence. The value of 4.8 scaled score points is higher than the observed difference of 4 scaled score points, so you cannot conclude that the difference is significant at the selected level of confidence in this case.

Clearly, the choice of *z* is crucial for your investigation. At the 90% level of confidence, *z* would be 1.645.

Applying this value to the SE_{mdiff} would give 2.45 x 1.645 = 4.03, which you would accept as being a significant difference for this level of confidence.

Note that the above formula does not consider the influence of practice effects as a result of a client taking the same test again. This is another reason why using parallel test forms is recommended.

The formula for the SE_{mdiff} is:

$$SE_{mdiff} = SD \times \sqrt{2 - r_a - r_b}$$

As explained in Chapter 3, this enables you to determine if the difference between the obtained two scores is reliably different from one another or not. An additional formula is available (Reynolds, 1990) to help you decide if the difference can be regarded to be unusual. This formula can be used when you have no direct

source of information as to how many children actually display a range of differences across the two relevant tests. The formula is:

$$\textit{Severe Discrepancy} = SD \times z_a \times \sqrt{2 - 2r_{xy}}$$

where SD = the standard deviation of both tests (recalculated into a common metric, if necessary), z_a = the z score corresponding to the point on the normal curve that you decide designates the frequency of occurrence of a '*severe discrepancy*', and r_{xy} = the correlation coefficient between the two tests.

Note that the above formulae, both for the confirmation of a reliable difference between scores and the degree of the difference indicating a severe discrepancy, make no allowance for regression affects.

The major weakness of comparing standardised scores over time periods is the impact of practice/learning effects on the repeated measures where the client profits from knowledge of the test items and procedures, and the related problem of deciding on the time period(s) for repeat measurements. Teachers often prefer to have knowledge about their students' progress over relatively short time periods in order to evaluate efficiently and quickly the impact of their teaching inputs. It is unsafe to repeat the same psychometric test across a short time period. In addition, changes in standardised scores over time are difficult to interpret because they are comparative measures where the comparison group is the test's standardisation sample. Over time, this group of students is also progressing. Thus a student may progress over time in relation to his own baseline score on a test, but in standard score terms may be seen as deteriorating in relation to his peers as they also continue to learn. Chapter 10 highlights why Growth Scale Value (GSV) scores have therefore been created in an attempt to overcome these problems.