Chapter 6

Psychometric Foundations of Neuropsychological Assessment

John R Crawford

Department of Psychology

King's College

University of Aberdeen

To appear as:

Crawford, J. R. (2003). Psychometric foundations of neuropsychological assessment. In L. H. Goldstein & J. McNeil (Eds.), Clinical Neuropsychology: A Practical Guide to Assessment and Management for Clinicians. Chichester: Wiley, in press.

Word Count: 8274 excluding references (9805 including references)

Introduction

Clinical neuropsychologists make a unique contribution to the assessment of clients with neurological or psychiatric disorders. Firstly, they can call on their specialised knowledge of sophisticated models of the human cognitive architecture when arriving at a formulation. Secondly, they have expertise in quantitative methods for both the measurement of cognitive and behavioural functioning and for the interpretation of resultant findings. This chapter will set out the basics of these quantitative methods.

Among the topics covered will be the pros and cons of the various metrics for expressing test scores, the use of reliability information in assessment, the distinction between the reliability and the abnormality of test scores and test score differences, and the measurement of change in the individual case. Finally, the role of measures of intelligence in neuropsychological assessment will be considered, as will methods for estimating *premorbid* intelligence.

Metrics for expressing test scores

In constructing a neuropsychological profile of a client's strengths and weaknesses most clinicians use instruments drawn from diverse sources. These instruments will differ from each other in the metric used to express test scores (in extreme cases no formal metric will have been applied so that clinicians will be working from the means and *SD*s of the raw scores from normative samples). The process of assimilating the information from these tests is greatly eased if the scores are all converted to a common metric (Crawford et al., 1998c; Lezak, 1995).

Converting all scores to percentiles has the important advantage that percentiles directly express the rarity or abnormality of an individual's score. In addition, percentiles are easily comprehended by other health workers. However, because such a conversion involves an area (i.e. non-linear) transformation they are not ideally suited for the rapid and accurate assimilation of information from a client's profile. For example, as scores on most standardized tests are normally distributed, the difference between a percentile score of 10 and 20 does not reflect the same underlying raw score (or standard score) difference as that between 40 and 50. In addition, percentiles are not a suitable metric for use with most inferential statistical methods. Expressing scores as percentiles can however be a useful fall-back option when raw scores depart markedly from a normal distribution and a normalising transformation cannot be found (such as when skew is acute and there is a limited number of scale points).

Z scores are a simple method of expressing scores and do not suffer from the limitations outlined above. However, they have the disadvantage of including negative values and decimal places which makes them awkward to work with and can cause problems in communication. It is also important to be aware that simply converting raw scores to standard or *z* scores has no effect on the shape of the distribution. If the raw scores are normally distributed (as will normally be the case with standardized tests) then so will the resultant *z* scores. However, if raw scores are skewed (i.e., the distribution is asymmetric), as may be the case when working with raw scores from a non-standardized measure, then the *z* scores will be equally skewed. In contrast, *normalized* z scores, as the term suggests, *are* normally distributed. There are a number of methods employed to normalize distributions; a common method is to convert raw scores to percentiles then convert the percentiles to normalized *z* scores by referring to a table of the areas under the normal curve (or using a computer package to the same effect). For example, if a raw score corresponds to the 5th percentile then the corresponding normalized *z* score is -1.64.

McKinlay (1992) suggested converting all scores to have a mean of 100 and *SD* of 15 as tests commonly forming a part of the neuropsychologist's armamentarium are already expressed on this metric; e.g., IQs and Indexes on the Wechsler Adult Intelligence Scale-3rd Edition (WAIS-III; Wechsler, 1997a), memory indices from the

Wechsler Memory Scale-3rd Edition (WMS-III; Wechsler, 1997b) and estimates of premorbid ability such as the National Adult Reading Test (NART; Nelson & Willison, 1991). A common alternative is to use *T* scores (mean=50, *SD*=10) which have much to recommend them. The gradation between *T* scores is neither too coarse, so that potentially meaningful differences between raw scores are obscured (such as would commonly be the case with sten scores in which a difference of one unit corresponds to 0.5 of a *SD*), nor too finely graded, so as to lend a spurious air of precision (for *T* scores, a difference of one unit corresponds to 0.1 of a *SD*). The meaning of *T* scores is also easy to communicate and are free of the conceptual baggage associated with IQs (Lezak, 1995).

With the exception of percentiles (which, as noted, involves a non-linear transformation) conversion of scores expressed on any of these different metrics can be achieved using a simple formula (the formula is generic in that it can be used to convert scores having *any* particular mean and *SD* to scores having any other desired mean and *SD*):

$$X_{\text{new}} = \frac{s_{\text{new}}}{s_{\text{old}}}\left(X_{\text{old}} - \overline{X}_{\text{old}}\right) + \overline{X}_{\text{new}}, \qquad (1)$$

where X_{new} = the transformed score, X_{old} = the original score, s_{old} = the standard deviation of the original scale, s_{new} = the standard deviation of the metric you wish to convert to, $\overline{X}_{\text{old}}$ = the mean of the original scale, and $\overline{X}_{new}$ = the mean of the metric you wish to convert to. Much of the time this formula is superfluous as the mapping of one metric on to another is straightforward (e.g., no thought is required to transform an IQ of 115 to a *T* score of 60). However, if a clinician is regularly converting large numbers of test scores, then entering the formula into a spreadsheet can save time and reduce the chance of clerical errors.

Regardless of which method is used to express score on a common metric, the clinician must be aware that the validity of any inferences regarding relative strengths and weaknesses in the resultant profile of scores is heavily dependent on the degree of equivalence of the normative samples involved. Although the quality of normative data for neuropsychological tests has improved markedly, there are still tests used in clinical practice that are normed on small samples of convenience. Thus, discrepancies in an individual's profile may in some cases be more a reflection of differences between normative samples than differences in the individual's relative level of functioning in the domains covered by the tests.

Reliability

Adequate reliability is a fundamental requirement for any instrument used in neuropsychology regardless of purpose. However, when the concern is with assessing the cognitive status of an *individual* its importance is magnified; particularly as clinicians frequently need to arrive at a formulation based on information from single administrations of each instrument (Crawford et al., 1998c).

The reliability coefficient represents the proportion of variance in test scores that is true variance. Thus, if a test has a reliability of 0.90, 90% of the variance reflects real differences between individuals and 10% reflects measurement error. Information on test reliability is used to quantify the degree of confidence that can be placed in test scores e.g., when comparing an individual's scores with appropriate normative data, or assessing whether discrepancies between scores on different tests represent genuine differences in the functioning of the underlying components of the cognitive system, as opposed to simply reflecting measurement error in the tests employed to measure the functioning of these components. In the latter case, i.e. where evidence for a dissociation or differential deficit is being evaluated, it is important to consider the extent to which the tests are matched for reliability; an apparent deficit in function A with relative sparing of function B may simply reflect the fact that the measure of function B is less reliable.

This point was well made in a classic paper by Chapman & Chapman (1973) in which the performance of a schizophrenic sample on two *parallel* reasoning tests was examined. By manipulating the number of test items, and hence the reliability of the tests, the schizophrenic sample could be made to appear to have a large differential deficit on either of the tests. Particular care should be taken in comparing test scores when one of the measures is not a simple score but a difference (or ratio score). Such measures will typically have modest reliability (the measurement error in the individual components that are used to form the difference score is additive).

The standard error of measurement (SEM) is the vehicle used to convert a test's reliability coefficient into information that is directly relevant to the assessment of individuals. The SEM can be conceived of as the standard deviation of obtained scores around an individual's hypothetical true score that would result from administering an infinite number of parallel tests. The formula for the SEM is

$$\text{SEM} = s_x\sqrt{1-r_{xx}}\ , \qquad (2)$$

where s_x = the standard deviation of scores on test *X*, and r_{xx} is the test's reliability coefficient. As noted, the reliability coefficient is the proportion of variance that is true variance; therefore subtracting this from unity gives us the proportion of variance that is error variance (i.e., measurement error). But we want to obtain the standard deviation of errors (rather than the variance) on the metric used to express the obtained scores. Therefore we take the square root of this quantity and multiply it by the *SD* of obtained scores.

The SEM allows us to form a confidence interval (CI) on a score. Most authorities on psychological measurement stress the use of these intervals (e.g., Nunnally &

Bernstein, 1994); they serve the general purpose of reminding us that all test scores are fallible and serve the specific purpose of allowing us to quantify the effects of this fallibility. Confidence intervals are formed by multiplying the SEM by the standard normal deviate corresponding to the desired level of confidence. Therefore, for a 95% CI, the SEM is multiplied by 1.96. To illustrate, if an individual obtained a score of 80 on a test and the SEM was 5.0, then the (rounded) confidence interval would be 80±10; i.e., the interval would range from 70 to 90.

There is however a slight complication: many authorities on measurement have argued that the confidence interval should be centred round the individual's estimated true score rather than their obtained score (e.g., Nunnally & Bernstein, 1994; Stanley, 1971). The estimated true score is obtained by multiplying the obtained score, in deviation form, by the reliability of the test,

$$\text{Estimated true score} = r_{xx}\left(X - \overline{X}\right) + \overline{X}\ , \tag{3}$$

where X is the obtained score and $\overline{X}$ is the mean for the test. The estimated true score represents a compromise between plumping for an individual being at the mean (which is our best guess if we had no information) and plumping for them being as extreme as the score they obtained on the particular version of the test on the particular occasion on which they were tested. The more reliable the test, the more we can trust the score and therefore the less the estimated true score is regressed to the mean.

To extend the previous example, suppose that the mean of the test in question was 100 and the reliability coefficient was 0.7. Therefore the estimated true score is 84 and, using this to centre the CI, we find that it ranges from 74 to 94. Before leaving this topic it can be noted that this confidence interval does not encompass the mean of the test; therefore it can be concluded that the individual's level of ability on the test is reliably below the mean level of ability.

As noted, a central aim in neuropsychological assessment is to identify relative strengths and weaknesses in a client's cognitive profile; as a result clinicians will commonly be concerned with evaluating test score *differences*. One question that can be asked of any difference is whether it is reliable, i.e., whether it is unlikely to simply reflect measurement error. To answer this question requires the standard error of measurement of the difference. When we are only concerned with comparing a pair of test scores then one formula for this quantity is as follows

$$\mathrm{SEM}_{X-Y} = \sqrt{\mathrm{SEM}_X^2 + \mathrm{SEM}_Y^2}\ . \qquad (4)$$

To use this formula the two scores must already be expressed on the same metric or transformed so that they are. The SEM_{X-Y} can be multiplied by a standard normal deviate corresponding to the required level of significance to obtain a critical value (i.e., multiplying by 1.96 gives the critical value for a reliable difference at the 0.05 level, two-tailed). If the difference between a client's scores exceeds this critical value it can be concluded that the scores are reliably different. This is the method usually employed in test manuals. Alternatively, the difference between the client's scores can be divided by the SEM_{X-Y} to yield a standard normal deviate and the precise probability determined using a table of areas under the normal curve. To illustrate both methods, suppose that the SEM for Tests *X* and *Y* are 3.0 and 4.0 respectively; therefore the SEM_{X-Y} is 5.0 and the critical value is 9.8. Further suppose that a client's scores on Tests *X* and *Y* were 104 and 92 respectively. The difference (12) exceeds the critical value; therefore the scores are reliably different (p< .05). Alternatively, dividing the difference by the SEM_{X-Y} yields a *z* of 2.4 and reference to a table of the normal curve reveals that the precise (two-tailed) probability that this difference occurred by chance is 0.016.

The method outlined is concerned with testing for a difference between a client's obtained scores. An alternative (but less common) method is to test for a reliable difference between estimated true scores; see Silverstein (1989) and Crawford et al. (2003) for examples of this latter approach.

Formula (4) is for comparing a client's scores on a single pair of tests. However, in a typical neuropsychological assessment many tests will have been administered. This leads to a large number of potential pairwise comparisons. For example, if 12 tests have been administered then there are 66 potential pairwise comparisons. Even with a relatively modest number of tests the process of assimilating this information on differences is formidable (particularly when it has to be integrated with all the other data available to the clinician). It can be also be readily appreciated that, when a large number of comparisons are involved, there will be an increase in the probability of making Type I errors (in this context a Type I error would occur if we concluded that there was a difference between a client's scores when there is not). Limiting the number of pairwise comparisons does not get round this problem, unless the decision as to which tests will be compared is made *prior* to obtaining the test results; if the clinician selects the comparisons to be made post-hoc on the basis of the magnitude of the observed differences then this is equivalent to having conducted all possible comparisons.

A useful solution to these problems was proposed independently by Silverstein (1982) and Knight and Godfrey (1984). In their approach a patient's score on each of *k* individual tests is compared with the patient's mean score on the *k* tests (just as is the case when comparing a pair of tests, all the tests must be expressed on the same metric or transformed so that they are). It can be seen that with 12 tests there are 12 comparisons rather than the 66 involved in a full pairwise comparison. Another feature of this approach is that a Bonferroni correction is applied to maintain the overall Type I error rate at the desired level. This approach has been applied to the analysis of strengths and weaknesses on the subtests of the Wechsler intelligence scales, including the WAIS-III (see Table B.3 of the WAIS-III manual). It has also been applied to various other tests;

e.g., Crawford et al. (1997b) have applied it to the Test of Everyday Attention (Robertson et al., 1994).

Reliability versus abnormality of test scores and test score differences

The distinction between the reliability and the abnormality of test scores and test score differences is an important one in clinical neuropsychology. As noted, if the confidence interval on a client's score does not encompass the mean of the test then we can consider it to be reliably different from the mean (i.e., a difference of this magnitude is unlikely to have arisen from measurement error). However, it does not follow from this that the score is necessarily unusually low (i.e., rare or abnormal), nor that the score reflects an acquired impairment.

Provided that the normative sample for a test is large (see next section), estimating the *abnormality* of a test score is straightforward. If scores are expressed as percentiles then we immediately have the required information; e.g., if a client's score is at the 5th percentile then we know that 5% of the population would be expected to obtain lower scores. If scores are expressed on other metrics we need only refer to a table of the normal curve. For example, a *T* score of 30 or an IQ score of 70 are exactly 2 *SD*s below the mean (i.e., z = -2.0) and therefore only 2.3% of the population would be expected to obtain lower scores (experienced clinicians will have internalised such information and so will rarely need to consult a table).

Most of the confusion around the distinction between the reliability and abnormality of test scores seems to arise when the focus is on differences between an individual's scores. Methods of testing for reliable differences between test scores were covered in the previous section. However, establishing if a difference is reliable is only the first step in neuropsychological profile analysis. There is considerable *intra*-individual variability in cognitive abilities in the general population such that reliable differences between tests of different abilities are common; indeed, if the reliabilities of the tests involved are very high, then such differences may be very common. Therefore, when evaluating the possibility that a difference between scores reflects acquired impairment, evidence on the reliability of the difference should be supplemented with information on the abnormality or rarity of the difference. That is, we need to ask the question "what percentage of the healthy population would be expected to exhibit a discrepancy larger than that exhibited by my client?"

To highlight the distinction between the reliability and abnormality of a difference take the example of a discrepancy between the Verbal and Perceptual Organization Indexes of the WAIS-III. Consulting Table B.1 of the WAIS-III manual it can be seen that a discrepancy of 10 points would be necessary for a reliable difference (p<0.05). However, such a discrepancy is by no means unusual; from Table B.2 we can see that 42% of the general population would be expected to exhibit a discrepancy of this magnitude. If we define an abnormal discrepancy as one that would occur in less than 5% of the general population, then a 26 point discrepancy would be required to fulfill this criterion.

Base rate data on differences between test scores such as that contained in Table B.2 of the WAIS-III manual are available for a number of tests used in neuropsychology. An alternative to this empirical approach is to estimate the degree of abnormality of a discrepancy using a formula provided by Payne and Jones (1957). This formula will be described briefly below so that clinicians understand it when they encounter it in the literature and can use it themselves when the necessary summary statistics are available for a healthy sample. The method can be employed when it is reasonable to assume that the scores are normally distributed and requires only the means and *SD*s of the two tests plus their intercorrelation (r_{xy}). The first step is to convert the individual's scores on the two tasks to z scores and then enter them into the formula,

$$z_D = \frac{z_X - z_Y}{\sqrt{2 - 2r_{xy}}} \tag{5}$$

This formula is very straightforward. The denominator is the standard deviation of the difference between scores when the scores are expressed as *z* scores. The numerator instructs us to subtract the individual's *z* score on Test *X* from their *z* score on Test *Y*. In the numerator the difference between the individual's *z* scores are subtracted from the mean difference in controls. However, the mean difference between *z* scores in the controls is necessarily zero and therefore need not appear. In summary, the difference between *z* scores is divided by the standard deviation of the difference to obtain a *z* score *for the difference.*

This *z* score (z_D) can then be referred to a table of the areas under the normal curve to provide an estimate of the proportion or percentage of the population that would exhibit a difference more extreme than the patient. For example, suppose scores on tests of verbal and spatial short-term memory were expressed as *T* scores, further suppose that the correlation between the tasks is 0.6 and that a patient obtained scores of 55 and 36 respectively. Therefore the patient's *z* scores on the tasks are 0.50 and –1.40, the difference is 1.90, the *SD* for the difference is 0.894, and so z_D = 2.13. Referring to a table of the normal curve reveals that only 3.32% of the population would be expected to exhibit a difference larger than that exhibited by the patient (1.66% if we concern ourselves only with a difference in the same direction as the patient’s).

It is also possible to assess the abnormality of discrepancies between a client's mean score on *k* tests and her/his scores on each of the tests contributing to that mean (Silverstein, 1984). This method complements the method discussed in the previous section that was concerned with the *reliability* of such discrepancies; it has the same advantage of reducing the comparisons to a manageable proportion. Silverstein's formula estimates the degree of abnormality from the statistics of a normative sample, an alternative is to generate the base rate data empirically (this latter approach is used for the WAIS-III).

The importance of evaluating the abnormality of discrepancies through the use of base rate data or methods such as the Payne and Jones formula cannot be overstressed. Most clinical neuropsychologists have not had the opportunity to administer neuropsychological measures to significant numbers of individuals drawn from the general population. It is possible therefore to form a distorted impression of the degree of intra-individual variability found in the general population; the indications are that clinicians commonly underestimate the degree of normal variability leading to a danger of over-interference when working with clinical populations (Crawford et al., 1998c).

Assessing the abnormality of test scores and test score differences when normative or control samples are small.

In the procedures just described for assessing the abnormality of scores and score differences the normative sample against which a patient is compared is treated as if it were a population; i.e., the means and *SD*s are used as if they were population parameters rather than sample statistics. When the normative sample is large (e.g., such as when a patient's score or score difference is compared against normative data from the WAIS-III or WMS-III) this is not a problem as the sample provides very good estimates of these parameters.

However, there are a number of reasons why the neuropsychologists may wish to compare the test scores of an individual with norms derived from a small sample. For example, although the quality of normative data has improved in recent years, there are still many useful neuropsychological instruments that have modest normative data. This is not surprising; advances in neuropsychological theory occur at a rapid rate whereas the process of devising practical measures of new constructs and obtaining norms from a large and representative sample is a time-consuming and arduous process. Furthermore, even when the overall *N* for a normative sample is reasonably large, the size of the actual sample (*n*) against which an individual's score is compared can be small if the normative data have been broken down by demographic characteristics (e.g., age and/or gender).

Therefore, there is a need for methods that are suitable for use with small normative samples. This need is perhaps most apparent when one considers single-case research in neuropsychology. The resurgence of interest in single-case studies has led to significant advances in our understanding of normal and pathological cognitive function (e.g., see Ellis & Young, 1996). However, in the typical single-case study the sample size of the control or normative sample recruited for comparison purposes is often < 10 or even < 5; it is inadvisable to treat samples of this size as though they were a population yet this is what is commonly done (Crawford & Howell, 1998a). The most common method of comparing a patient against such samples is to convert the patient's score to a *z* score and refer it to a table of the normal curve. This gives an estimate of the abnormality of the score but is also used as an inferential statistical method, i.e., if the *z* score exceeds a specified critical value then the patient's score is considered to be significantly different from controls. However, because this approach treats the control sample statistics as parameters, it overestimates the abnormality of the score and increases Type I errors.

The solution is to use a modified *t*-test to conduct such comparisons as this method treats the control sample *as* a sample (Crawford & Howell, 1998a). Crawford et al. (1998a) extended this approach to allow neuropsychologists to estimate the abnormality of the difference between a patient's scores on a pair of tests and to test whether the patient's difference is significantly different from the differences observed in a control or normative sample (this method is the equivalent of the Payne and Jones method but is not constrained by the need for a large normative sample).

This test can be used to provide an operational definition of a dissociation. The typical existing definitions of a classical dissociation are not stringent as they require only that a patient is "impaired" on Task *X* and "normal" or "within normal limits" on Task *Y*. One half of this typical definition requires us to prove the null hypothesis (we must demonstrate that a patient is not different from the controls), whereas, as is well known, we can only fail to reject it. Further, a patient's score on the impaired task could lie just below the critical value for defining impairment and the performance on the other test lie just above it. That is, the difference between the patient's relative standing on the two tasks of interest could be very trivial. A test on the difference between scores overcomes both these problems.

Crawford and Garthwaite (2002) have further extended these methods by providing a means of establishing confidence intervals on the abnormality of a score or score difference. These intervals quantify the uncertainty arising from using sample statistics to estimate population parameters and, in combination with the methods discussed above, provide clinicians or researchers with information of the form "the estimated percentage of the healthy population that would obtain a score (or score difference) lower than the patient is 2.1% and the 95% CI on the percentage is from 0.2% to 6.7%". Computer programs that implement all of the methods covered in this section are available; see Crawford & Garthwaite (2002) for details.

Detecting change in neuropsychological functioning in the individual case

There are many situations in which the neuropsychologist needs to measure potential changes in cognitive functioning. Common examples would be to determine whether cognitive decline has occurred in an individual in whom a degenerative neurological process is suspected, or to determine the extent of recovery of function following a stroke or traumatic brain injury. In both these cases neuropsychological assessment will provide useful information to assist clients, relatives and other health professionals to plan for the future. Monitoring the cognitive effects of surgical, pharmacological or cognitive interventions in the individual case is also an important role for the neuropsychologist. Although the aim here is most commonly to determine if there has been any improvement, the possibility of detrimental effects can also often be an issue. For example, many drugs can potentially impair cognitive functioning, particularly in the elderly.

In assessing the effectiveness of a rehabilitation effort it is often possible to obtain *multiple* repeated measures of an individual's performance before, during and after intervention (see Wilson, 1987). A number of inferential statistical techniques can be used in this situation because there are multiple data points for the different phases. However, in general clinical practice the neuropsychologist often must come to conclusions about change from only a *single* retesting. This situation will also arise in rehabilitation settings; although multiple measures may have been obtained on the training task(s), the issue of the generalisability of any improvement is often addressed by comparing single before-and-after scores on related but separate tests.

Monitoring change on the basis of a single retesting is a formidable task except in cases where the level of change has been dramatic. It is rendered more formidable by the fact that many of the standard instruments currently used in clinical neuropsychology do not have parallel or analogue versions. This is a particular problem for tests of executive function, which often depend on novelty for their effectiveness.

The clinician must differentiate changes resulting from systematic practice effects and random measurement error from change reflecting genuine improvement or deterioration. Amongst other complications are the fact that (1) the magnitude of practice effects vary with the nature of the task; for example tasks with a psychomotor component tend to have larger practice effects than tasks that do not, (2) the length of time that has elapsed between test and retest is liable to influence the magnitude of effects, and (3) a diminution of practice effects is to be expected in neurological populations (given the high prevalence of memory and learning deficits) but the expected diminution is difficult to estimate for individuals.

One approach to dealing with some, although not all, of these considerable interpretive problems is to use regression to predict scores at retest from scores on initial testing. An individual's predicted score is compared with the retest score actually obtained to assess whether the observed gain or decline significantly exceeds that expected by chance. This can be achieved by dividing the difference between predicted and obtained scores by the equation's standard error of estimate to obtain a *z* for the difference[1].

The utility of this approach is determined by the extent and nature of retest studies available for a particular test. For example, the test-retest scores on memory tasks for an elderly client with suspected dementia could be compared with estimated retest scores derived from a healthy, elderly sample retested after a similar period to gauge how atypical any decline may be. For some questions equations derived from a clinical sample may be used. For example, a head-injured client's scores on measures of attention or speed of processing could be compared with estimated retest scores from a head-injured sample if the clinician suspects that the extent of recovery is atypical.

An excellent example of the regression-based approach is provided by Temkin et al. (1999). Using a healthy sample (*N*=384), these authors built regression equations to predict performance at retest for a large number of tests commonly used in clinical neuropsychology. The regression-based approach compared favourably with alternative approaches (e.g., Jacobson, 1991). This study also serves to illustrate a further point; if other variables (e.g., age or years of education) moderate the relationship between initial performance and performance at retest, these variables can simply be incorporated into the equations to improve the accuracy of prediction.

It is not necessary for the clinician to have access to existing equations to make use of regression to assess change. If basic summary data are available from a sample retested on a neuropsychological measure then an equation can be built with relative ease; all that is required is the correlation between scores at test and retest and the means and *SD*s of these scores. The relevant formulae for obtaining a regression equation and its associated standard error of estimate can be found in most statistics textbooks (e.g., Howell, 2002).

The role of measures of intelligence in neuropsychological assessment

Lezak (1988) has described the Wechsler intelligence scales as the "the workhorse of neuropsychological assessment" and noted that it is the single most utilized component of the neuropsychological repertory" (p. 53). In the UK, the British Psychological Society proposed that Wechsler IQ cut-off scores should be used as the sole formal criteria for the legal definition of mental impairment (Alves et al., 1991).

The most recent incarnation of these scales is the WAIS-III and its UK edition (WAIS-IIIUK; Wechsler et al., 1998). The WAIS-III has many of the strengths of its

[1] This method is used widely in clinical neuropsychology but provides only an approximation to the technically correct method. However, in practice, the approximation is adequate unless the sample used to

predecessors, including a large and highly representative standardisation sample, and very impressive levels of reliability. For example, the reliability of the WAIS-III Full Scale IQ is 0.98 (therefore only 2% of the variance in scores is error variance); this compares very favourably with the reliability of most medical testing procedures.

The WAIS-III also has a number of advantages over earlier editions. These include an upward extension of the age range (norms are available to cover ages 16 to 89) and an increased ability to differentiate at the lower end of ability (i.e., there has been a downward extension of measurable IQ to 54). In addition, the inclusion of the Matrix Reasoning subtest has bolstered the measurement of fluid ability and contributes to the lesser emphasis on timed performance in determining scores (there are also lesser time bonuses on some existing subtests). A minor but welcome change is that subtest scaled scores are age-corrected (previously the clinician had to convert to scaled scores to derive IQs and age-graded scaled scores if they wanted to examine strengths and weaknesses at the subtest level).

Perhaps the most significant improvement is the inclusion of factor-based composite measures (i.e., the Indexes) as alternatives to IQs. Two of these indexes also represent attempts to measure constructs that are of great theoretical importance in neuropsychology and cognitive psychology, namely working memory (Baddeley & Hitch, 1994) and speed-of-processing (e.g. Salthouse, 1991).

One limitation of earlier versions of the Wechsler scales was that they did not have parallel versions. Although a full parallel version is still not available for the WAIS-III, the development of the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) means that there are now parallel versions for four of the subtests (Vocabulary, Similarities, Block Design, and Matrix Reasoning). The WASI is also

generate the equation is very small; see Crawford & Howell (1998b) for details.

useful in its own right as a short-form when clinical considerations rule out use of a full WAIS-III.

Given that the WAIS-III performance can be analysed at the level of IQs, Indexes and subtests, it is appropriate to consider what the *primary* level of analysis should be. Lezak (1988) argued that IQs obscure clinically important strengths and weaknesses and suggested that the solution was to focus on the subtest profile. Few clinical neuropsychologists would disagree with the first half of this position but issue could be taken with the proposed solution. A number of authors have argued that the primary level of analysis should be at the level of factorially derived indexes (e.g., Atkinson, 1991; Crawford et al., 1997a).

In the present author's view the arguments in favour of Indexes are compelling. The indexes are empirically derived, whereas the subtests comprising the Verbal and Performance IQs were allocated to these scales on intuitive grounds. Factor analysis of the WAIS-III and its predecessors have consistently found that the ability dimensions underlying the Wechsler fail to map on to the VIQ and PIQ scales; three factors emerge on the WAIS-R (subtests from both the VIQ and PIQ scales load on the third factor). Four factors emerge on the WAIS-III, reflecting the addition of an additional speeded measure (Symbol Search) to accompany Digit Symbol. Furthermore, the factor structure has proven itself to be surprisingly robust; the same factors emerge across cultures and across healthy versus clinical populations (Atkinson, 1991).

Evidence of the superior clinical utility of factor-based indexes over IQs is provided by Crawford et al's. (1997a) study of WAIS-R performance following head-injury. Their head-injured (HI) sample (*N*=233) exhibited highly significant and broadly equivalent deficits on VIQ and PIQ. In contrast to this uninformative pattern, examination of the indexes revealed that the HI sample did not differ significantly from controls on the Verbal Index (despite high statistical power) but exhibited a very substantial deficit on the attention/concentration index (now termed Working Memory in the WAIS-III). This pattern of spared and severely compromised abilities was hidden within the VIQ scale because it represents an amalgam of variance reflecting well consolidated verbal abilities (as indexed by the Verbal index), and attention/concentration or working memory (Arithmetic and Digit Span contribute strongly to this index). Furthermore, discriminant function analysis revealed that VIQ and PIQ did not improve differentiation of the HI sample from healthy controls (*N*=356) over that achieved by FSIQ alone (i.e., the VIQ/PIQ discrepancy was not useful). In contrast, the factor-based indexes achieved significantly better discrimination than the IQ scales.

The primary advantage of the Indexes over individual subtests is that they have superior reliability. The reliability of a composite will normally substantially exceed that of its individual components (i.e., the subtests) if these components are correlated with each other. Substantial difference among individual subtests is not uncommon in the healthy unimpaired population because of measurement error and also because each subtest has a proportion of unique (but not necessarily clinically important) variance associated with it in addition to variance that reflects the common underlying factor. Such extreme differences in normals are less likely at the level of Indexes (because they are composites); therefore, when large differences are observed, they are more likely to be of clinical significance. Empirical evidence of this was provided by Crawford et al's (1997a) study of HI; the factor-based indexes achieved vastly superior discrimination between healthy and HI cases than indexes of subtest scatter (these latter indexes quantified the overall level of intra-individual variability in subtest scores). Indeed, counter to expectations, the head-injured sample exhibited no more subtest scatter than healthy controls.

The arguments and evidence reviewed above suggest that clinical interpretation of strengths and weaknesses on the WAIS-III should primarily be conducted at the level of the Indexes. As previously noted, Table B.1. of the manual allows clinicians to examine the reliability of differences between Indexes and Table B.2 allows them to assess the abnormality of the differences. In addition, a fundamental role for the WAIS-III in neuropsychological assessment is to provide context for interpretation of *other* tests results. As Crawford et al. (1997a) note, the Wechsler provides "broad indicators of current functioning against which more specific neuropsychological measures are compared" (p. 352). The evidence and arguments reviewed above also suggests that Indexes, rather than the IQs, should fulfill this role.

Methods for estimating premorbid ability

The detection and quantification of cognitive impairment in the individual case is problematic because of the wide variability in cognitive abilities in the general adult population. Scores on cognitive measures that are average, or even above average, can still represent a significant impairment for an individual of high premorbid ability. Conversely, for individuals with modest premorbid resources, test scores that fall well below the mean may be entirely consistent with their prior level of functioning. Because of this, simple normative comparison standards are of limited utility in the detection of impairment, and are supplemented with *individual* comparison standards when attempting to assess acquired deficits (Crawford et al., 1998c; Lezak, 1995; O'Carroll, 1995). Ideally, these individual standards would be obtained from cognitive test scores obtained in the premorbid period. However, this is rarely a viable option as most individuals have had no prior formal testing and even where they have, it is difficult or impossible to obtain the results. Because of these difficulties, clinicians and researchers normally have to settle for some means of estimating an individual's premorbid ability.

The most common means of obtaining such an estimate is to use a measure of present ability that is relatively unaffected by neurological or psychiatric disorder.

Currently, the test most widely used for this purpose is the National Adult Reading Test (NART; Nelson & Willison, 1991). The NART is a 50 item single word reading test of graded difficulty. It is argued that the NART makes minimal demands on current cognitive capacity because it requires only oral reading of short, single words (Nelson & O'Connell, 1978). In addition, all the words are irregular; i.e. they violate grapheme–phoneme correspondence rules (e.g., *ache*, *thyme*, *topiary*). The supposition is that, as a result, the test depends on prior or premorbid ability because a testee must have prior knowledge of a word's pronunciation; deployment of *current* cognitive resources (e.g., the intelligent application of grapheme-phoneme correspondence rules) will not result in a correct pronunciation.

To be considered valid, any putative measure of premorbid intelligence must fulfil three criteria (Crawford, 1992). Firstly, like any psychological test, it must possess adequate reliability. Secondly, it must have adequate criterion validity, that is, it must correlate highly with measures of psychometric intelligence. Thirdly, performance on the measure must be largely impervious to the effects of neurological or psychiatric disorder. The NART fulfils the first criterion in that it possesses high internal consistency, test-retest reliability, and inter-rater reliability (Crawford, 1992; O'Carroll, 1995).

The results from studies that have addressed the criterion validity of the NART have also been generally positive. For example, in UK studies the NART predicted between 55% and 72% of the variance in FSIQ as measured by the WAIS and WAIS-R (see Crawford, 1992; O'Carroll, 1995). The NART also predicts a substantial proportion of the variance in most of the IQs and Indexes of the WAIS-III (Crawford et al., submitted).

A feature of all these studies is that, although the NART is intended to measure *prior* or *premorbid* intelligence, the criterion validity information they provide is based solely on *concurrent* administration of the NART and an IQ measure. Therefore, it is important to know whether the criterion validity of the NART is equally impressive when it is compared against a criterion that genuinely represents prior ability. Crawford et al. (2001a) administered the NART to a sample of 77 year olds (*N* = 197) and reported that it retrospectively accounted for 53% of the variance in the IQ scores obtained by this sample at age 11 (i.e., 66 years previously). Furthermore, the NART's correlation with prior IQ was higher than its correlation with IQ measured concurrently and was also higher than the correlation between the two IQ tests. This remarkable pattern of results lends support to the validity of the notion that the NART primarily indexes prior rather than current ability.

Using a North American variant of the NART, Smith et al. (1997) conducted an important longitudinal study, that simultaneously assessed the issue of criterion validity and the question of the robustness of NART performance (further evidence on this latter issue is reviewed below). Smith et al. reported that, in a sample of participants with cognitive impairment, the NART provided accurate predictions of Verbal IQs obtained 5 years earlier when the sample had been cognitively normal.

The final criterion for a putative measure of premorbid ability is that test performance be resistant to neurological or psychiatric disorder. NART performance appears to be largely resistant to the effects of many neurological and psychiatric disorder (e.g., depression, acute schizophrenia, alcoholic dementia, and Parkinson's disease (see Crawford, 1992; Franzen et al., 1997 for reviews; O'Carroll, 1995). The results for head-injury have generally been positive (e.g., Watt & O'Carroll, 1999) although there are indications of impaired performance in a minority of cases (Freeman et al., 2001); the available evidence also suggests that the NART "holds" following focal frontal lesions (Bright et al., 2002; Crawford & Warrington, 2002).

The findings from early studies have been most mixed in the case of dementia of the Alzheimer type (DAT); see aforementioned reviews. Subsequent studies have continued to produce conflicting results in DAT, some reporting clear evidence of impairment (e.g., Cockburn et al., 2000; Patterson et al., 1994), others finding that NART performance "held" (Bright et al., 2002; Sharpe & O'Carroll, 1991). However, it has become increasingly clear that NART performance is impaired in many cases of severe and even moderate dementia.

Any findings of impaired NART performance pose a threat to the validity of this approach. However, the practical implications of impaired performance in cases of *severe* neurological disorder are not as serious as they may appear; in such cases the presence of deficits is unfortunately only too obvious, thereby largely obviating the need for the NART or similar instrument to assist in its detection and quantification (Crawford et al., 1998c). Nevertheless, given the accumulating evidence that NART performance can be impaired in some conditions, it would be useful to have a means of evaluating the likely validity of the premorbid estimate provided by the NART in the individual case.

To address this need, Crawford, et al. (1990a) built a regression equation in a healthy sample (N =659) to predict NART scores from demographic variables (e.g., years of education and occupational classification); an obtained NART score that was significantly lower than the demographically predicted NART score would indicate that the NART is unlikely to provide a valid estimate of premorbid ability for the individual concerned.

It has been suggested (Crawford et al., 1990b) that the NART should fulfill a fourth criterion that is more concerned with its clinical utility than its validity. It could be that, contrary to the assumptions underlying the use of the NART, simply using obtained scores alone, rather than the discrepancy between NART estimates of premorbid scores and obtained scores, would be just as effective a means of detecting impairment.

This basic issue has received little empirical scrutiny. Crawford et al. (1990b) used hierarchical discriminant function analysis to examine the ability of the NART in combination with WAIS IQ to correctly classify a sample consisting of healthy participants and patients with Alzheimer’s disease (AD). The inclusion of the NART significantly improved the accuracy of classification over that achieved by WAIS IQs alone; 85% of cases were correctly classified by IQ scores and this rose to 96% for the combination of IQs and NART scores. An analogous result was obtained by Crawford and Warrington (2002) who reported that the use of the NART significantly improved the ability of a verbal fluency task (homophone meaning generation) to discriminate between the performance of healthy participants and patients with focal anterior lesions.

Most research on the NART's ability to estimate premorbid ability has used scores on IQ tests as the criterion variable. Although this is in keeping with the notion of obtaining an estimate of an individual's general level of premorbid functioning, the NART also has the potential to provide estimates of premorbid functioning for more specific neuropsychological tests. For example, Crawford et al. (1992) built a regression equation which can be used to estimate premorbid performance on the FAS verbal fluency test; an equation is also available for the Homophone Meaning Generation Test (Crawford & Warrington, 2002). Similarly, Crawford et al. (1998b) have provided an equation which uses NART and age to estimate premorbid scores on the PASAT (Gronwall, 1977), a measure that is sensitive to the presence of attentional dysfunction following head injury (McMillan & Glucksman, 1987).

A proposed modification to the NART is the Cambridge Contextual Reading Test (Beardsall & Huppert, 1994). As its name suggests, the NART words are embedded in sentences to provide context for the examinee. Baddeley et al. (1993) developed another alternative reading test which they termed the Spot TheWord Test (STW). STW is a lexical decision task in which the examinee has to identify the legitimate words from a series word / pseudo-word pairs (e.g. stamen / floxid); see Law and O'Carroll (1998) for an evaluation. Finally, the Wechsler Test of Adult Reading (WTAR; Holdnack, 2001) is based on the same rationale as the NART (it requires the reading of single, irregular words). All of these tests have considerable potential as alternatives to the NART but have a relatively modest research base at present. Space limitations means that they cannot be considered in any further detail in the present work.

An entirely different approach uses regression equations to estimate premorbid ability from demographic variables (e.g. years of education, occupational classification etc). This approach can be seen as a formalisation of clinical "guesstimates" based on the same information, but one that is more accurate and free from demonstrable clinical biases (Crawford et al., 2001b). The advantage of the demographic approach over the use of tests such as the NART is that the estimate it provides is entirely independent of a client's current level of functioning. As evidence of impaired NART performance in various clinical conditions accumulates, the demographic approach offers an alternative method in cases where use of the NART would be inappropriate. However, the obvious disadvantage of the demographic approach is its modest criterion validity. Where Wechsler IQs have been used as the criterion variables, regression equations based solely on demographic variables account for between only 36% and 54% of FSIQ variance (e.g., Barona et al., 1984; Crawford & Allan, 1997). This compares unfavorably with the corresponding figures for the NART and its variants. Finally, regression equations have been developed that use both the NART *and* demographic variables as predictors; see O'Carroll (1995) for an evaluation.

Conclusion

This chapter has covered basic quantitative methods for analyzing test scores. A solid grounding in these methods and associated issues (i.e. the distinction between the reliability and abnormality of differences) is fundamental to the practice of clinical neuropsychology and, as noted, is an important part of what makes the clinical neuropsychologist unique among health professionals.

References

Alves, E., Williams, C., Stephen, I., & Prosser, G. (1991). *Mental impairment and severe mental impairment*. Leicester: British Psychological Society.

Atkinson, L. (1991). Some tables for statistically based interpretation of WAIS-R factor scores. *Psychological Assessment, 3*, 288-291.

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word Test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology, 32*, 55-65.

Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology, 8*, 485-493.

Barona, A., Reynolds, C. R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 52*, 885-887.

Beardsall, L., & Huppert, F. A. (1994). Improvement in NART word reading in demented and normal older persons using the Cambridge Contextual Reading Test. *Journal of Clinical and Experimental Neuropsychology, 16*, 232-242.

Bright, P., Jaldow, E., & Kopelman, M. D. (2002). The National Adult Reading Test as a measure of premorbid intelligence: A comparison with estimates derived from demographic variables. *Journal of the International Neuropsychological Society, 8*, 847-854.

Chapman, L. J., & Chapman, J. P. (1973). Problems in the measurement of cognitive deficit. *Psychological Bulletin, 79*, 380-385.

Cockburn, J., Keene, J., Hope, T., & Smith, P. (2000). Progressive decline in NART score with increasing dementia severity. *Journal of the International Neuropsychological Society, 22*, 508-517.

Crawford, J. R. (1992). Current and premorbid intelligence measures in neuropsychological assessment. In J. R. Crawford & D. M. Parker & W. W. McKinlay (Eds.), *A handbook of neuropsychological assessment* (pp. 21-49). London: Erlbaum.

Crawford, J. R., & Allan, K. M. (1997). Estimating premorbid IQ with demographic variables: Regression equations derived from a U.K. sample. *The Clinical Neuropsychologist, 11*, 192-197.

Crawford, J. R., Allan, K. M., Cochrane, R. H. B., & Parker, D. M. (1990a). Assessing the validity of NART estimated premorbid IQs in the individual case. *British Journal of Clinical Psychology, 29*, 435-436.

Crawford, J. R., Deary, I. J., Starr, J. M., & Whalley, L. J. (2001a). The NART as an index of prior intellectual functioning: A retrospective validity study covering a 66 year interval. *Psychological Medicine, 31*, 451-458.

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia, 40*, 1196-1208.

Crawford, J. R., Hart, S., & Nelson, H. E. (1990b). Improved detection of cognitive impairment with the NART: An investigation employing hierarchical discriminant function analysis. *British Journal of Clinical Psychology, 29*, 239-241.

Crawford, J. R., & Howell, D. C. (1998a). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12*, 482-486.

Crawford, J. R., & Howell, D. C. (1998b). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology, 20*, 755-762.

Crawford, J. R., Howell, D. C., & Garthwaite, P. H. (1998a). Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples t-test. *Journal of Clinical and Experimental Neuropsychology, 20*, 898-905.

Crawford, J. R., Johnson, D. A., Mychalkiw, B., & Moore, J. W. (1997a). WAIS-R performance following closed head injury: A comparison of the clinical utility of summary IQs, factor scores and subtest scatter indices. *The Clinical Neuropsychologist, 11*, 345-355.

Crawford, J. R., Martin, D., Mockler, D., Allner, K., & Cipolotti, L. (submitted). Estimation of premorbid performance on WAIS-III IQs and indexes using the National Adult Reading Test (NART).

Crawford, J. R., Miller, J., & Milne, A. B. (2001b). Estimating premorbid IQ from demographic variables: A comparison of a regression equation versus clinical judgement. *British Journal of Clinical Psychology, 40*, 97-105.

Crawford, J. R., Moore, J. W., & Cameron, I. M. (1992). Verbal fluency: A NART-based equation for the estimation of premorbid performance. *British Journal of Clinical Psychology, 31*, 327-329.

Crawford, J. R., Obonsawin, M. C., & Allan, K. M. (1998b). PASAT and components of WAIS-R performance: Convergent and discriminant validity. *Neuropsychological Rehabilitation, 8*, 255-272.

Crawford, J. R., Smith, G. V., Maylor, E. A. M., Della Sala, S., & Logie, R. H. (2003). The Prospective and Retrospective Memory Questionnaire (PRMQ): Normative data and latent structure in a large non-clinical sample. *Memory, in press*.

Crawford, J. R., Sommerville, J., & Robertson, I. H. (1997b). Assessing the reliability and abnormality of subtest differences on the Test of Everyday Attention. *British Journal of Clinical Psychology, 36*, 609-617.

Crawford, J. R., Venneri, A., & O'Carroll, R. E. (1998c). Neuropsychological assessment of the elderly. In A. S. Bellack & M. Hersen (Eds.), *Comprehensive clinical psychology, vol. 7: Clinical geropsychology* (pp. 133-169). Oxford, UK: Pergamon.

Crawford, J. R., & Warrington, E. K. (2002). The Homophone Meaning Generation Test: Psychometric properties and a method for estimating premorbid performance. *Journal of The International Neuropsychological Society, 8*, 547-554.

Ellis, A. W., & Young, A. W. (1996). *Human cognitive neuropsychology: A textbook with readings*. Hove, UK: Psychology Press.

Franzen, M. D., Burgess, E. J., & Smith-Seemiller, L. (1997). Methods of estimating premorbid functioning. *Archives of Clinical Neuropsychology, 12*, 711-738.

Freeman, J., Godfrey, H. P. D., Harris, K. J. H., & Partridge, F. M. (2001). Utility of a demographic equation in detecting impaired NART performance. *British Journal of Clinical Psychology, 40*, 221-224.

Gronwall, D. (1977). Paced Auditory Serial Addition Task: A measure of recovery from concussion. *Perceptual and Motor Skills, 44*, 367-373.

Holdnack, J. A. (2001). *WTAR. Wechsler Test of Adult Reading manual*. San Antonio, TX: Psychological Corporation.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Duxbury Press.

Jacobson, N. S. T., Paula. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12-19.

Knight, R. G., & Godfrey, H. P. D. (1984). Assessing the significance of differences between subtests on the Wechsler Adult Intelligence Scale - Revised. *Journal of Clinical Psychology, 40*, 808-810.

Law, R., & O'Carroll, R. E. (1998). A comparison of three measures of estimating premorbid intellectual level in dementia of the Alzheimer type. *International Journal of Geriatric Psychiatry, 13*, 727-730.

Lezak, M. D. (1988). IQ: R.I.P. *Journal of Clinical & Experimental Neuropsychology, 10*, 351-361.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

McKinlay, W. W. (1992). Assessment of the head-injured for compensation. In J. R. Crawford & D. M. Parker & W. W. McKinlay (Eds.), *A handbook of neuropsychological assessment* (pp. 381-392). Hove: Lawrence Erlbaum.

McMillan, T. M., & Glucksman, E. E. (1987). The neuropsychology of moderate head injury. *Journal of Neurology, Neurosurgery and Psychiatry, 50*, 393-397.

Nelson, H. E., & O'Connell, A. (1978). Dementia: The estimation of premorbid intelligence levels using the new adult reading test. *Cortex, 14*, 234-244.

Nelson, H. E., & Willison, J. (1991). *National Adult Reading Test manual* (2nd ed.). Windsor: NFER-Nelson.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Carroll, R. (1995). The assessment of premorbid ability: A critical review. *Neurocase, 1*, 83-89.

Patterson, K., Graham, N., & Hodges, J. R. (1994). Reading in dementia: A preserved ability? *Neuropsychology, 8*, 395-407.

Payne, R. W., & Jones, G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology, 13*, 115-121.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). *The Test of Everyday Attention*. Bury St Edmunds: Thames Valley Test Company.

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging*. Hillsdale NJ: Erlbaum.

Sharpe, K., & O'Carroll, R. (1991). Estimating premorbid intellectual level in dementia using the National Adult Reading Test: A Canadian study. *British Journal of Clinical Psychology, 30*, 381-384.

Silverstein, A. B. (1982). Pattern analysis as simultaneous statistical inference. *Journal of Consulting and Clinical Psychology, 50*, 234-240.

Silverstein, A. B. (1984). Pattern analysis: The question of abnormality. *Journal of Consulting and Clinical Psychology, 52*, 936-939.

Silverstein, A. B. (1989). Confidence intervals for test scores and significance tests for test score differences: A comparison of methods. *Journal of Clinical Psychology, 45*, 828-832.

Smith, G. E., Bohac, D. L., Ivnik, R. J., & Malec, J. F. (1997). Using word recognition scores to predict premorbid IQ in dementia patients. *Journal of the International Neuropsychological Society, 3*, 528-533.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington D.C.: American Council on Education.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5*, 357-369.

Watt, K. J., & O'Carroll, R. E. (1999). Evaluating methods for estimating premorbid intellectual ability in closed head injury. *Journal of Neurology, Neurosurgery and Psychiatry, 66*, 474-479.

Wechsler, D. (1997a). *Manual for the Wechsler Adult Intelligence Scale -Third Edition*. San Antonio TX: The Psychological Corporation.

Wechsler, D. (1997b). *Manual for the Wechsler Memory Scale -Third Edition*. San Antonio TX: The Psychological Corporation.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence. Manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D., Wycherley, R. J., Benjamin, L., Crawford, J. R., & Mockler, D. (1998). *Manual for the Wechsler Adult Intelligence Scale -Third Edition (U.K.)*. London: The Psychological Corporation.

Wilson, B. (1987). Single-case experimental designs in neuropsychological rehabilitation. *Journal of Clinical and Experimental Neuropsychology, 9*, 527-544.