PERGAMON

Neuropsychologia 40 (2002) 1196–1208

NEUROPSYCHOLOGIA

www.elsevier.com/locate/neuropsychologia

Investigation of the single case in neuropsychology: confidence limits on the abnormality of test scores and test score differences

J.R. Crawford [a,*], Paul H. Garthwaite [b]

[a] *Department of Psychology, King's College, University of Aberdeen, Aberdeen AB24 2UB, UK*
[b] *Department of Statistics, The Open University, Milton Keynes, UK*

Received 4 June 2001; received in revised form 20 November 2001; accepted 23 November 2001

Abstract

Neuropsychologists often need to estimate the abnormality of an individual patient's test score, or test score discrepancies, when the normative or control sample against which the patient is compared is modest in size. Crawford and Howell [The Clinical Neuropsychologist 12 (1998) 482] and Crawford et al. [Journal of Clinical and Experimental Neuropsychology 20 (1998) 898] presented methods for obtaining *point* estimates of the abnormality of test scores and test score discrepancies in this situation. In the present study, we extend this work by developing methods of setting *confidence limits* on the estimates of abnormality. Although these limits can be used with data from normative or control samples of any size, they will be most useful when the sample sizes are modest. We also develop a method for obtaining point estimates and confidence limits on the abnormality of a discrepancy between a patient's mean score on *k*-tests and a test entering into that mean. Computer programs that implement the formulae for the confidence limits (and point estimates) are described and made available.

Keywords: Neuropsychology; Abnormality; Confidence limits; Single case studies

1. Introduction

Estimating the rarity or abnormality of an individual's test score is a fundamental part of the assessment process in neuropsychology. The procedure for statistical inference in this situation is well known. When it is reasonable to assume that scores from a normative sample are normally distributed, the individual's score is converted to a *z* score and evaluated using tables of the area under the normal curve [20,23]. Thus, if a neuropsychologist has formed a directional hypothesis concerning the individual's score prior to testing (e.g. that the score will be below the normative mean), then a *z* score which fell below −1.64 would be considered statistically significant (using the conventional 0.05 level). More generally, and it could be argued more usefully (given that any significance level is an arbitrary convention that does not address the issue of severity), the probability for *z* provides the neuropsychologist with information on the rarity or abnormality of the individual's score. Thus, for example, if a patient obtained a *z* score of −1.28 on a given test, then a table of the normal curve will tell us that approximately 10% of the population would be expected to obtain a score lower than this.

Many tests used in neuropsychology are expressed on a conventional metric such as an IQ scale (mean = 100, S.D. = 15), or test score (mean = 50, S.D. = 10). In such cases it is clearly often not necessary to convert the score to *z* to arrive at the estimate of abnormality. For example, a patient obtaining a score of 85 on the Working Memory Index of the WAIS-III [41,42] is exactly 1 S.D. below the mean. Most neuropsychologists will know that therefore approximately 16% of the population would be expected to obtain a score as low or lower than this. However, the principle in this latter example is identical, i.e. the score is referred to the normal curve.

In the standard procedure just described the normative or control sample is treated as if it was a population; i.e. the mean and standard deviation are used as if they were *parameters* rather than *sample statistics*. When the normative sample is reasonably large this is justifiable. However, Crawford and Howell [10] point out that there are a number of reasons why neuropsychologists may wish to compare the test scores of an individual with norms derived from a small sample. For example, although there has been a marked improvement in the quality of normative data in recent years, there are still many useful neuropsychological instruments

* Corresponding author. Tel.: +44-1224-272-231.
E-mail address: j.crawford@abdn.ac.uk (J.R. Crawford).

0028-3932/02/$ – see front matter
PII: S0028-3932(01)00224-X

that have modest normative data. Even when the overall N for a normative sample is reasonably large, the actual sample size (n) against which an individual's score is compared can be small when the sample is broken down by demographic characteristics. Secondly, many clinical neuropsychologists have gathered local norms for neuropsychological instruments, but because of the time and expense involved, the size of the normative samples are often modest.

Finally, in recent years there has been an enormous resurgence of interest within academic neuropsychology in single case studies [3,4,16,21,25,33]. In many of these studies the theoretical questions posed cannot be addressed using existing instruments and therefore novel instruments are designed specifically for the study. The sample size of the control or normative group recruited for comparison purposes in such studies is typically <10 and often <5.

Crawford and Howell [10] have described and illustrated the use of a method that can be used to compare an individual with normative or control samples that have modest N. Their approach uses a formula given by Sokal and Rohlf [37] that treats the statistics of the normative or control sample as statistics rather than as population parameters and uses the t-distribution (with $N-1$ degrees of freedom (d.f.)), rather than the standard normal distribution, to evaluate the abnormality of the individual's scores. Essentially, this method is a modified independent samples t-test in which the individual is treated as a sample of $M=1$, and therefore does not contribute to the estimate of the within group variance. The formula for this test is presented in Appendix A.1.

The disadvantage of the standard (z score) method is that, with small samples, it exaggerates the rarity/abnormality of an individual's score. This is because the normal distribution has "thinner tails" than t-distributions. Intuitively, the less that is known, the less extreme should be statements about abnormality/rarity. The z score method treats the variance as being known, when it is not, and consequently makes statements that are too extreme. A fuller illustration of this will be provided in a worked example, but in the interim, suppose that an individual obtains a score of 20 on a test and that the mean and S.D. for this test in a control sample are 40 and 10, respectively. If the N of the control sample was 10, then the estimate provided by the modified t-test procedure is that approximately 4.4% of the population would obtain a score lower than the individual's score. The z score method exaggerates the rarity of the individual's score as the estimate it provides is that approximately 2.3% of the population would obtain a lower score.

Up to this point we have been concerned with the simple case of comparing a single test score obtained from an individual with a normative or control sample. However, in the assessment of acquired neuropsychological deficits, simple normative comparison standards have limitations because of the large individual differences in premorbid competencies. For example, an average score on a test of mental arithmetic would represent a marked decline from the premorbid level in a patient who was a qualified accountant. Conversely, a score that fell well below the normative mean does not necessarily represent an acquired deficit in an individual who had modest premorbid abilities [5,15,28].

Because of the foregoing, considerable emphasis is placed on *intra*-individual comparison standards when attempting to detect and quantify the severity of acquired deficits [6,24,39]. In the simplest case, the neuropsychologist may wish to compare an individual's score on two tests; a fundamental consideration in assessing the importance of any discrepancy between scores on the two tests is the extent to which it is rare or abnormal. Payne and Jones [29] developed a formula for this purpose. The method requires the mean and S.D. of the two tests in a normative sample and the correlation between them. The two tests must be on the same metric, or they must be converted to a standard metric (z scores are normally used). The formula provides an estimate of the percentage of the population that would exhibit a discrepancy that equals or exceeds the discrepancy observed for a patient.

A number of authors have commented on the usefulness of this formula in neuropsychology [7,23,26,32,34], and it has been applied to the analysis of differences on a variety of tests [1,19,27]. However, just as was the case for the standard method of comparing a single score with a normative sample, the Payne and Jones [29] formula treats the statistics of the normative or control sample as if they were population parameters. This limits the valid use of the method to comparisons of an individual with a large normative sample.

Crawford et al. [11] developed a method that treats the normative statistics as statistics. Like the Payne and Jones [29] method, it requires that the normative or control sample mean are converted to a common metric (z scores). The patient's difference is divided by the standard error of the difference, yielding a quantity that is distributed as t with $N-1$ d.f. (where N is the sample size, i.e. it does not include the individual). Essentially then this is a modified paired samples t-test. The formula for this test is presented in Appendix A.2.

Technically, this method is more appropriate than the Payne and Jones [29] method for comparison of an individual's test score difference with differences from *any* size of normative or control sample (i.e. our test norms are always obtained from a sample rather than a population). However, its usefulness lies in its ability to deal with comparisons involving normative or control samples that are modest in size; the Payne and Jones [29] method systematically overestimates the abnormality of an individual's test score difference in such comparisons.

Crawford et al. [11] suggest that their method is particularly useful in single case studies where, as noted, the control samples against which a patient is compared usually has a small N. A common aim in neuropsychological case studies is to fractionate the cognitive system into its constituent parts and it proceeds by attempting to establish the presence of dissociations of function. Typically, if a patient obtains a score in the impaired range on a test of a particular function

and is within the normal range on a test of another function, this is regarded as evidence of a dissociation. However, a more stringent test for the presence of a dissociation is to also compare the *difference* between tests observed for the patient with the distribution of differences in the control sample. For example, a patient's score on the "impaired" task could lie just below the cut point for defining impairment and the performance on the other test lie just above it.

Crawford et al. [11] method can be used in such studies to test if the difference observed in the patient is significantly different from the differences in the controls. Their method is also useful in the converse situation where a patient's scores are within the impaired range on both tasks. When this pattern is observed, the researcher can still test whether the magnitude of the difference between the two tasks is abnormal; i.e. evidence can be sought for the presence of a *differential* deficit on the test of one of the functions.

The above methods are designed to yield *point* estimates of the rarity or abnormality of either an individual's single test score, or the difference between an individual's scores on two tests. In the present paper, we extend this work by providing methods for obtaining *confidence limits* on the abnormality of test scores and test score differences. This is in keeping with the contemporary emphasis in statistics, psychometrics, and biometrics on the use of confidence limits [14,18,44]. Gardner and Altman [18] for example, in discussing the general issue of the error associated with sample estimates note that, "these quantities will be imprecise estimates of the values in the overall population, but fortunately the imprecision itself can be estimated and incorporated into the findings" (p. 3).

Neuropsychologists are aware that estimates of the rarity/abnormality of a test score or score difference are subject to sampling error and will have an intuitive appreciation that less confidence should be placed in them when N for the normative sample is small. However, the advantage of the procedures to be outlined is that they *quantify* the degree of confidence that should be placed in these pestimates.

In the following sections, we present the methods for obtaining confidence limits on the abnormality of a single test score and the difference between a pair of test scores. These methods and their applications are illustrated with examples relevant to academics who pursue single case research and to clinical neuropsychologists. We also include a method for obtaining confidence limits on the abnormality of the difference between an individual's mean score on k-tests and a test score entering into that mean. The existing method of obtaining a *point* estimate of the abnormality of such a difference [35,36] treats the normative sample against which the individual is compared as if it were a population. Therefore, we also develop a method for obtaining a point estimate of the abnormality of the difference that treats the normative sample statistics as statistics rather than as parameters. This is achieved by a straightforward extension of Crawford et al. [11] method for obtaining a point estimate of the abnormality of a pair of test scores.

The methods to be described for obtaining confidence limits require non-central t-distributions. As readers may not be familiar with such distributions a brief description is provided before formally presenting the methods. Both the t and non-central t-distributions are derived from a ratio of the distribution of sample means and that of sample variances drawn from a normal population. The sampling distribution of the mean is normal (and symmetrical), while the sampling distribution of the variance is skewed (and follows a χ^2 distribution). When the sampling distribution of the mean has a mean of 0 (i.e. when the population distribution has a mean of 0) sample variances are combined equally often with positive and negative sample means. Effectively the asymmetry of the sampling distributions of the variance occurs equally often facing in positive and negative directions and so the resulting central t-distribution is symmetrical.

When the sampling distribution of the mean has a non-zero mean (i.e. when the population distribution itself has a non-zero mean) then the asymmetry of the sampling distribution of the variance is not balanced equally between positive and negative sample means and so the resulting non-central t-distribution is asymmetrical. The extent of its skew depends upon the mean and variance of the population distribution. The upshot for calculating confidence intervals is that one cannot simply shift a t-distribution along an axis in order to find a confidence interval around a mean, one has to find and use non-central t-distributions with specified properties.

2. Obtaining confidence limits for the abnormality of a test score

Letting P_1 denote the percentage of the population that will fall below a given individual's score (X_0), we suppose we require a $100(1-\alpha)\%$ confidence interval for P_1. Let $(X_0 - \bar{X})$ represent the difference between the individual's score and the mean score of the normative or control sample, let S be the standard deviation in the normative sample, and let N be the size of the normative sample. We assume scores for the control population are normally distributed. If we put

$$c_1 = \frac{X_0 - \bar{X}}{S}, \tag{1}$$

then c_1 is an observation from a non-central t-distribution on $N-1$ d.f. Non-central t-distributions have a non-centrality parameter that affects their shape and skewness. We find a value of this parameter, δ_U say, such that the resulting non-central t-distribution has $c_1\sqrt{N}$ as its $100\alpha/2$ percentile. Then we find the value δ_L such that the resulting distribution has $c_1\sqrt{N}$ as its $100(1-\alpha/2)$ percentile. From tables for a standard normal distribution we obtain $\Pr(Z < \delta_L/\sqrt{N})$ and $\Pr(Z < \delta_U/\sqrt{N})$. These probabilities are multiplied by 100 to express them as percentages. They depend upon α, c_1 and N and we denote the percentages by $h(\alpha/2;\ c_1;\ N)$

and $h(1-\alpha/2;\ c_1;\ N)$, respectively. Then a $100(1-\alpha)\%$ confidence interval for the percentage P_1 may be written as

$$(h(\alpha/2;\ c_1;\ N),\quad h(1-\alpha/2;\ c_1;\ N)). \tag{2}$$

Details of the derivation of h are given in Appendix B.

As a worked example, suppose that a neuropsychologist has administered a new measure of spatial short-term memory to a patient and the patient obtains a score of 30. The control or normative sample for this test has an N of 15 and a mean and standard deviation of 50 and 10, respectively. Using Crawford and Howell's [10] method, the *point* estimate of the percentage of the population that would be expected to obtain a score lower than the patient (P_1) is 3.7%. To obtain 95% confidence limits on P_1 we proceed as follows:

$$c_1 = \frac{X_0 - \bar{X}}{S} = \frac{30.0 - 50.0}{10} = -2.0$$

$$c_1\sqrt{N} = -2\sqrt{15} = -7.746$$

We want a non-central t-distribution on $N-1=14$ d.f. that has -7.746 as its 0.025 quantile. This determines the non-centrality parameter to be -4.252 so we put $\delta_U = -4.252$. We also want a non-central t-distribution on 14 d.f. that has -7.746 as its 0.975 quantile. This gives $\delta_L = -11.151$. A graphical representation of the two non-central t-distributions and the process of obtaining the non-centrality parameters is provided in Fig. 1. Then

$$\Pr\left(Z < \frac{-11.151}{\sqrt{15}}\right) \times 100 = 0.2$$

and

$$\Pr\left(Z < \frac{-4.252}{\sqrt{15}}\right) \times 100 = 13.6.$$

Hence, the 95% lower confidence limit for the percentage P_1 is 0.2% and the upper limit is 13.6%.

As a further illustration, Table 1 records confidence limits on the percentage of the population falling below a given observed score as a function of both N in the normative or control sample, and the extremity of the observed score. The scores (c) for which limits are tabulated are all below the sample mean. Confidence limits on the equivalent positive values of c can be obtained by subtracting the percentages from 100 in which case the upper and lower limits are reversed. Table 1 also presents *point* estimates of the percentage falling below the score; these estimates were obtained using Crawford and Howell's [10] method. The 95% confidence interval of (0.02%, 13.6%) for the last example may be read from the middle of the third row of the table.

As noted, the confidence limits quantify the degree of uncertainty surrounding the estimate of the abnormality (i.e. rarity) of a given test score. It can be seen from Table 1 that the limits are wide with small sample sizes. However, even with more moderate Ns the limits are not insubstantial and thus serve as a useful reminder of the fallibility of normative data. The table also illustrates that the limits are not symmetrical around the point estimate; the point estimate is nearer the lower limit when c is negative. This happens because, as noted, a non-central t-distribution is skew. It can also be seen that the limits become more asymmetric about

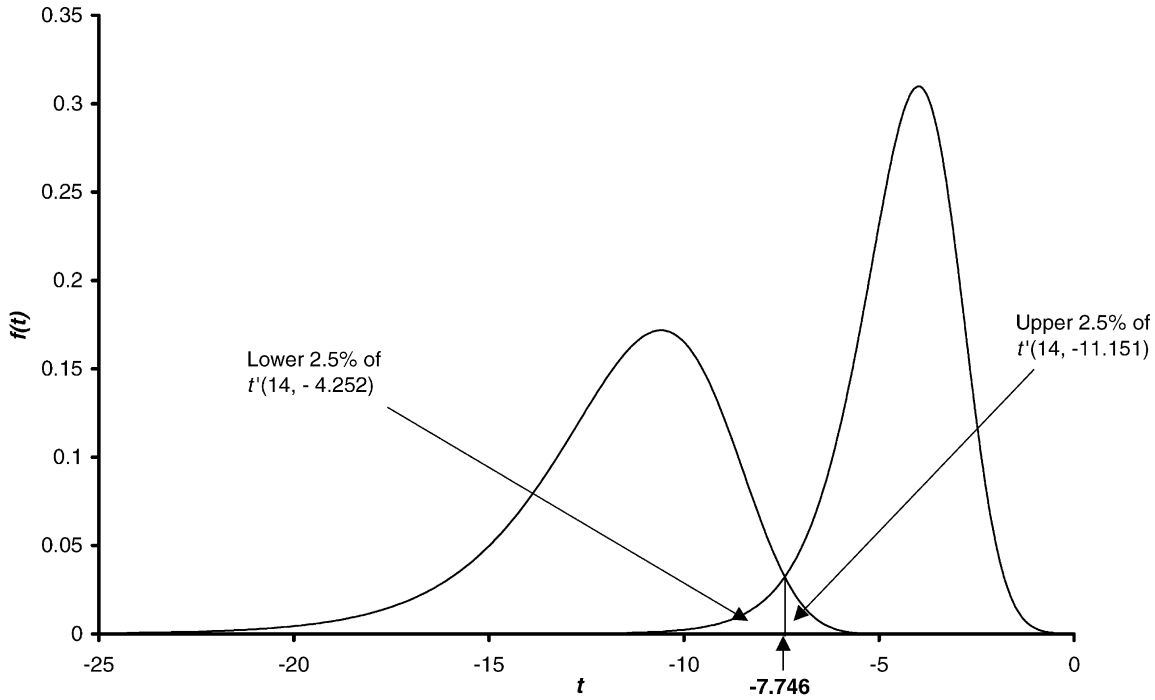

Fig. 1. Graphical illustration of the non-central t-distributions used in the worked example.

Table 1
Point estimates and 95% confidence limits on the percentage of the population falling below a given observed score as a function of N in the normative sample and the extremity of the score

N	$c = -2.5$			$c = -2.0$			$c = -1.0$		
	Point	Lower	Upper	Point	Lower	Upper	Point	Lower	Upper
5	4.2	0.00	27.2	7.1	0.02	35.2	20.7	1.94	55.5
10	2.0	0.01	11.6	4.5	0.10	18.8	18.3	4.00	41.5
15	1.4	0.02	7.4	3.7	0.20	13.6	17.5	5.33	35.8
20	1.2	0.03	5.6	3.3	0.29	11.1	17.1	6.28	32.6
25	1.1	0.05	4.6	3.1	0.37	9.6	16.8	7.00	30.5
30	1.0	0.06	3.9	2.9	0.44	8.5	16.7	7.57	30.0
50	0.8	0.11	2.7	2.7	0.66	6.5	16.3	9.06	25.6
70	0.8	0.15	2.2	2.6	0.81	5.6	16.2	9.94	23.9
125	0.7	0.22	1.6	2.4	1.07	4.5	16.1	11.3	21.7
250	0.7	0.30	1.2	2.4	1.34	3.7	16.0	12.5	19.8
500	0.6	0.37	1.0	2.3	1.57	3.2	15.9	13.4	18.6

the point estimate as the extremity of the scores (measured by c) increases in magnitude.

Although not the primary focus of the present paper, it is apparent from Table 1 that the point estimates of the rarity of the score are markedly different as a function of N. For example, with an N of 5 the estimated percentage of the population that would fall below a value of -2 for c is approximately 7%; this is more than twice the point estimate (2.3%) obtained with a normative sample of 500. If, as is commonly done in single case studies, the patient's score was evaluated using a table of the normal curve (i.e. the statistics from the normative samples were treated as parameters), the point estimate of the abnormality of the score would be 2.28%. This is essentially the same as the estimate obtained with a sample size of 500 (because a t-distribution on large d.f. is virtually indistinguishable from a normal distribution). The important point to reiterate here is that, in the majority of cases, the commonly used (z score) method will *exaggerate* the rarity of an observed score; when the normative or control sample is small this effect can be substantial.

3. Confidence limits for the abnormality of a difference between pairs of tests

The methods of the previous section may be used, with slight modification, to obtain lower and upper limits for the percentage of the population that will fall below a given difference score between two tests. In many situations, the means and S.D. of the two tests in the normative or control sample will differ, and the scores need to be converted to a common metric. We will use z scores and we assume that differences in scores are normally distributed in the normative population. We let X_0 and Y_0 denote the original scores of an individual on the two tests and we let X_{z_0} and Y_{z_0} denote their values in z score form. A $100(1-\alpha)$% confidence interval is required for P_2, the percentage of the population whose difference in scores will fall below $-|X_{z_0} - Y_{z_0}|$. We suppose a sample of size N has been taken and summary statistics from these data are available in one of the following forms.

(a) A difference $D_z = X_z - Y_z$ has been calculated for each individual in the normative or control sample and the standard deviation of this difference (S_{D_z}) has been obtained.
(b) The summary statistics are the sample mean scores for each test, $\bar{X}$ and $\bar{Y}$, the sample standard deviations S_X and S_Y, and the sample correlation r_{XY}. It will be appreciated that, by providing a formula for dealing with summary data in this form, it will be possible to use the method with data reported by other researchers (i.e. the raw data are not required).

For (a) put

$$c_2 = \frac{-|X_{z_0} - Y_{z_0}|}{S_{D_z}}, \tag{3}$$

and for (b) put

$$c_2 = \frac{-|((X_0 - \bar{X})/S_X) - ((Y_0 - \bar{Y})/S_Y)|}{\sqrt{2 - 2r_{XY}}}. \tag{4}$$

These two formulae are equivalent: in (4) the observed scores of the patient are converted to z scores (whereas in (3) they are already in this form), and the difference between these z scores divided by the standard deviation of the difference (the correlation between two measures provides sufficient information to calculate the standard deviation of the difference when scores have been expressed in z score form). It is also worth making explicit that, in the numerator of both (3) and (4), we are actually subtracting the mean difference in the controls from the patient's difference. However, as the mean difference between the scores in the control sample will be zero (we are using z scores) there is no need to include this term.

It will also be noted that in both (3) and (4) the absolute value of the difference between z scores is taken and then

set to be negative. This is because which test should be designated as *X*, and which as *Y*, is arbitrary. By setting the difference to be negative (or, more precisely, non-positive) the confidence limits will be on the percentage of the population that will obtain a difference more extreme than the patient's and in the same direction.

In Appendix B.2, it is shown that the $100(1-\alpha)\%$ confidence interval for P_2 is

$$(h(\alpha/2;\ c_2;\ N),\quad h(1-\alpha/2;\ c_2;\ N)), \tag{5}$$

where h is the same function used earlier. Hence, we find two non-central t-distributions on $N-1$ d.f., such that $c_2\sqrt{N}$ is the $100(1-\alpha/2)$ percentile for one of the distributions and the $100(\alpha/2)$ percentile for the other. We denote their non-centrality parameters by δ_{L} and δ_{U}, respectively, and then

$$h(\alpha/2;\ c_2;\ N) = \Pr(Z < \delta_{\mathrm{L}}/\sqrt{N}) \times 100 \tag{6}$$

and

$$h(1-\alpha/2;\ c_2;\ N) = \Pr(Z < \delta_{\mathrm{U}}/\sqrt{N}) \times 100 \tag{7}$$

where Z has a standard normal distribution.

To illustrate the calculation of these confidence limits, suppose that a patient has been administered novel tests of spatial and verbal short-term memory. In the interest of generality we will work with data that are in form (b); i.e. we will work with summary data rather than assuming that we have access to the scores for individual's in the normative or control sample. Suppose that these tests have been administered to a control sample of $N = 20$, that the mean and S.D. of the verbal memory task are 50 and 10, respectively (we will designate this task as test *X*), and that the mean and S.D. of the spatial task (test *Y*) are 40 and 7, respectively. Further suppose that the correlation between these tests in the control sample is 0.7 and that the patient obtained a score of 40 on test *X* and 44 on test *Y*. Using Crawford et al.'s method [11], the point estimate of the abnormality of this difference (i.e. the point estimate of the percentage of the population that would exhibit a difference more extreme than the patient's difference (P_2) is 3.1%. We now obtain confidence limits on the abnormality of the difference (P_2):

$$\begin{aligned} c_2 &= \frac{-|((40-50)/10) - ((44-40)/7)|}{\sqrt{2-1.4}} \\ &= \frac{-|(-1.0) - (+0.5714)|}{0.7746} \\ &= \frac{-|-1.5714|}{0.7746} = -2.029. \end{aligned}$$

$$c_2\sqrt{N} = -2.029\sqrt{20} = -9.074$$

We want a non-central t-distribution on $N-1 = 19$ d.f. that has -9.074 as its 0.025 quantile. This determines the non-centrality parameter to be -5.565 so we put $\delta_{\mathrm{U}} = -5.565$. We also want a non-central t-distribution on 19 d.f. that has –9.074 as its 0.975 quantile. This gives $\delta_{\mathrm{L}} = -12.504$ Then,

$$\Pr\left(Z < \frac{-12.504}{\sqrt{20}}\right) \times 100 = 0.26$$

and

$$\Pr\left(Z < \frac{-5.565}{\sqrt{20}}\right) \times 100 = 10.7.$$

Hence, the 95% lower confidence limit for P_2 is 0.26% and the upper limit is 10.7%.

4. Point estimates of the abnormality of a difference between an individual's mean score on *k*-tests and score on a test entering into that mean

Up to this point we have been concerned with point estimates and confidence limits on the abnormality of a single test score, or difference between scores on two tests. However, in neuropsychology there is an emphasis on examining an individual's relative strengths and weaknesses across a wide range of cognitive domains [13,24]. This necessitates using a large number of neuropsychological tests; as a result, there is a problem of how to reduce the number of potential comparisons between scores to a manageable proportion. For example, if a neuropsychologist administers a total of 12 tests to a patient, then there are 66 potential pairwise comparisons between tests. It is clearly difficult to assimilate such a large amount of information when attempting to arrive at a formulation of the patient's difficulties. Against this must be set the need to retain theoretically (and/or clinically) significant attributes of the patient's profile [9].

A method that strikes an appropriate balance between these competing demands was developed by Silverstein [35,36]. This approach provides an estimate of the abnormality of the difference between an individual's score on a test and the mean of the individual's score on a series of *k*-tests (including the test of interest). Thus, if 12 tests are administered, there are 12 comparisons rather than the 66 involved in a full pairwise comparison. Silverstein [36] presented the following formula to estimate the abnormality of the difference between an individual's mean test score and one of the tests entering into that mean (changes have been made to the notation to render it consistent with the rest of this paper):

$$z_{\mathrm{Da}} = \frac{X_{\mathrm{a}} - \bar{X}_k}{S_{\mathrm{Da}}}, \tag{8}$$

where X_{a} is the individual's score on test a, $\bar{X}_k$ the individual's mean score on the *k*-tests (including test a), and S_{Da} the standard deviation of the difference between individuals' scores on test a and individuals' mean scores on the *k*-tests; S_{Da} is obtained from the formula below:

$$S_{\mathrm{Da}} = S\sqrt{1 + \bar{G} - 2\bar{T}_{\mathrm{a}}}, \tag{9}$$

where S is the (common) standard deviation of the tests; $\bar{G}$ the mean of all the elements in the test correlation matrix (including unities in the diagonal) and $\bar{T}_a$ is the mean of the elements in row a or column a of the matrix (again including the diagonal). It can be seen from this formula that, if scores on the tests involved are expressed on different metrics, then they must be transformed to a common one (i.e. test scores or z scores) before being entered into formula (8). It will also be noted that if scores are converted to z scores then the S term falls out of formula (9) as z scores have a S.D. of 1.

To estimate the percentage of the population that would obtain a score which is more extreme and in the same direction as that observed for an individual, the probability of exceeding $|z_{Da}|$ is obtained from a table of the area under the normal curve and multiplied by 100 [35]; see Silverstein [36] for a derivation of the formula.

The Silverstein method then provides the clinician or single case researcher with a point estimate of the abnormality of the difference between an individual's mean score on a series of tests and his/her score on a single test (in which the individual test under consideration enters into the mean). The method has been widely endorsed [8,22,31]; its principal application has been to the analysis of subtest profiles on the WAIS-R [40] although it has also been used for the analysis of test profiles on other instruments such as the Test of Everyday Attention [12,30].

In Silverstein's formula the data from the control or normative sample are treated as population parameters rather than as sample statistics. This is not problematic if the normative sample is very large, as in the case of the Wechsler scales. However, in the present paper the intention is to develop methods that are applicable to comparisons of individual's with normative or control samples of any size and, in particular, to comparisons involving small samples.

Therefore, before dealing with confidence limits on the abnormality of such differences, we modify Silverstein's approach to develop a method for obtaining a *point* estimate of the abnormality of the difference that treats sample statistics as sample statistics rather than as population parameters. The formula for this method is a straightforward extension of Crawford et al. [11] formula (see Appendix A) and is presented below:

$$t = \frac{X_a - \bar{X}_k}{S_{Da}\sqrt{(N+1)/N}}, \tag{10}$$

where N is the size of the normative or control sample and all other terms are as defined in formula (8). In Appendix B.3, we show that the quantity obtained in formula (10) has a t-distribution with $N-1$ d.f. (rather than the standard normal distribution). If $|t|$ obtained from formula (10) exceeds the critical value for a specified significance level (e.g. 0.05) then the patient's discrepancy is significantly different from the mean discrepancy in the normative or control sample. In addition, if the precise probability of t is obtained, then multiplying this probability by 100 yields the percentage of the population that would be expected to obtain a discrepancy larger and in the same direction as that observed for the patient.

Silverstein's original formula, and the modified version presented here, use summary statistics from the normative sample (i.e. the common S.D. of the k-tests, and the correlations between the k-tests). This approach was adopted to allow maximum flexibility; i.e. the formulae can be used when only the summary statistics are available. However, if the researcher or clinician has the raw data for the normative or control sample in a spreadsheet or statistics package, then it would be easier to calculate the standard deviation of the difference between the k-tests and each of the individual tests directly from the raw data. That is, assuming that the scores have already been converted to a common metric, the normative sample's scores on an individual test should be subtracted from the normative sample means on the k-tests and the standard deviation of this new variable (the standard deviation of the difference) entered into formula (10).

A worked example for this method will illustrate the steps involved. In the interest of generality the example is based on the use of summary statistics from the normative or control sample. Suppose that a research group has developed tests of the ability to recognise emotions from facial expressions. For each of five basic emotions (anger, happiness, sadness, fear, and disgust) they have prepared tests consisting of 20 photographs of faces displaying the relevant emotion. Suppose also that, within each test, the difficulty level is sufficient to avoid ceiling effects in healthy participants (e.g. the sets contain mild expressions of the relevant emotion or ambiguous poses). This scenario is based on one of a series of related studies in which the aim has been to identify selective or differential deficits in the processing of emotion from facial expression [2]. In Calder et al.'s [2] study of single cases who had suffered amygdala lesions, 10 healthy participants were recruited as a control group. We will take this as the sample size for the current example.

The first step is to convert the scores on each test to a common metric; for simplicity suppose that z scores are used. Secondly, the matrix of correlations between the five tests is used to obtain the standard deviation of the difference between a test and the mean of the five tests; we will illustrate the calculations for test a (disgust). The correlation matrix is presented as Fig. 2.

The standard deviation of the difference is calculated from formula (9). The mean of the elements in the correlation matrix ($\bar{G}$), including the unities in the diagonal, is 0.616. The mean of the elements in the row or column of the matrix that records the correlation between the test and all other tests ($\bar{T}_a$) is 0.640. The common standard deviation of the scales is 1 as z scores are used. Entering these values into formula (9) gives the following result:

$$S_{Da} = 1\sqrt{1 + 0.616 - (2 \times 0.640)} = \sqrt{0.336} = 0.580.$$

This procedure is repeated for each of the four other tests to obtain their standard deviation of the difference. These

	a	*b*	*c*	*d*	*e*
a	**1.00**	0.50	0.60	0.40	0.70
b	0.50	**1.00**	0.30	0.50	0.60
c	0.60	0.30	**1.00**	0.40	0.50
d	0.40	0.50	0.40	**1.00**	0.70
e	0.70	0.60	0.50	0.70	**1.00**

Fig. 2. Correlation matrix for the five tests of facial emotion recognition.

standard deviations are recorded in the first row of Table 2. Up to this point the procedure is identical to that involved in using Silverstein's original formula. Now however, we obtain the standard error of the difference between an individual and the normative sample by multiplying the standard deviation of the difference (S_D) for each test by $\sqrt{(N+1)/N}$, as in formula (10). In the present case, in which N is 10, this value is 1.0488. Therefore, for test a the standard error is 0.608; standard errors for the other four tests are presented in the second row of Table 2. We now have the required constants in formula (10) and can compare a patient's performance on any of the individual tests with his/her mean score on the five tests.

Let us suppose that a neuropsychologist wishes to assess the ability of a patient with Huntington's disease (HD) to process emotion from facial expressions. In particular the neuropsychologist wishes to discover if there is evidence for a differential deficit in the processing of disgust; such a deficit has been reported in HD cases [38]. The third row of Table 2 records the patient's performance in *z* score form for each of the five tests; the mean of these scores is −1.20. The discrepancies between the patient's score on each test and the patient's mean are recorded in the fourth row of Table 2. The fifth row records the *t*-value obtained when these discrepancies are divided by their respective standard errors. Multiplying the one-tailed probabilities for each *t* recorded in row five by 100 provides the point estimate of the percentage of the population that would obtain a difference score lower than the patient's (i.e. it estimates the abnormality of the patient's difference score). These percentages are recorded in row six of Table 2.

From the third row of Table 2 it can be seen that the patient was below the control mean on recognition of all five emotions and that performance on the disgust task is particularly poor. Using Crawford and Howell's [10] method for comparing a single score against a control sample mean, the deficit for disgust would be statistically significant. However, as noted, the patient was below the mean on *all* facial expressions and, therefore, the question of whether the deficit is a *differential* deficit remains; i.e. is performance on the disgust task significantly poorer than the patient's averaged performance. From row six of Table 2 it can be seen that only 2.1% of the population would be expected to exhibit a discrepancy on disgust that was lower than the patient's. In addition, the one- and two-tailed 5% critical values for *t* with 9 d.f. are 1.83 and 2.26, respectively. It can be seen that the *t*-value for disgust (2.37) exceeds these values; thus, the deficit on disgust is significantly greater than the patient's averaged deficits on the other emotions.

In this worked example, the tests were measures of ability within the same cognitive domain. However, it will be appreciated that the method is just as applicable to analysing strengths and weaknesses on measures of diverse functions. Indeed, the original Silverstein formula has typically been used for this latter purpose, e.g. to compare performance across the 11 subtests of the WAIS-R [36].

Finally, in this particular example the patient was below the control sample means on all tasks. However, in a patient of high premorbid ability, it will often be the case that scores on a number of the *k*-tests will be above the normative mean (particularly if the tests measure diverse functions and any neurological damage is relatively focal). In these circumstances, the use of the present method can provide evidence of acquired deficits that would not be apparent using normative standards. That is, a patient's score on a particular test may not be significantly lower than the normative

Table 2
Figures for the worked example of assessing the abnormality of the difference between a patient's mean score on *k*-tests and a test score entering into that average

	Tests				
	a, disgust	b, anger	c, happiness	d, sadness	e, fear
S.D. of the difference	0.580	0.675	0.704	0.645	0.465
S.E.	0.608	0.708	0.738	0.676	0.488
Patient's *z* scores	−2.73	−0.76	−0.24	−1.14	−1.23
Discrepancies from patient's mean *z* score	−1.51	0.46	0.98	0.08	−0.01
The *t*-values for discrepancies	−2.48	0.65	1.33	0.12	−0.02
Estimated %age of population falling below patient's discrepancies	1.74	73.4	89.1	54.6	49.2
Lower 95% confidence limits	0.004	48.9	68.7	30.1	26.1
Upper 95% confidence limits	10.5	91.3	98.8	77.1	72.5

sample mean for that test, but the discrepancy between the patient's score and her/his mean score on the *k*-tests would be statistically significant.

The calculations involved in using this method may appear somewhat laborious. However, most of the calculations are those required to obtain the standard error of the difference. These calculations need only be carried out once, so that the labour involved to use the method subsequently is minimal. Furthermore, a computer program that automates the calculations is available (details are provided in a later section).

5. Confidence limits on the abnormality of a difference between an individual's mean score on *k*-tests and score on a test entering into that mean

Having presented the method for obtaining a point estimate of the abnormality of the difference between an individual's mean test score and the score on a test entering into that mean, attention can now be turned to obtaining confidence limits on this estimate of abnormality. Let P_3 denote the percentage of the population that will fall below the difference score observed for the individual. The formula for the confidence limits for P_3 are easily obtained from the results that gave formula (1). We put

$$c_3 = \frac{X_a - \bar{X}_k}{S_{Da}}. \tag{11}$$

Then a $100(1-\alpha)$% confidence interval for P_3 is

$$(h(\alpha/2;\ c_3;\ N), \quad h(1-\alpha/2;\ c_3;\ N)). \tag{12}$$

A worked example is not provided because of the similarity between formulae (11) and (1). Note though, that in (11), $\bar{X}_k$ is the mean of the individual's scores on the *k*-tests, while in (1), $\bar{X}$ is the mean score of the normative sample on a single test. A total of 95% confidence limits obtained using formula (12) are presented in rows seven and eight of Table 2.

6. Use of the confidence limits in single case studies and clinical practice

We believe the confidence limits presented in the present paper will be of benefit to both single case researchers and clinicians. Firstly, they serve the useful purpose of reminding us of the fallibility of our normative or control data. As such they are in keeping with the contemporary emphasis on using confidence limits in many areas of statistics and psychometrics.

They will also directly assist neuropsychologists in their attempts to achieve a valid assessment of a patient's relative strengths and weaknesses. Consider the situation in which an individual patient's test score is estimated to be rare (e.g. the point estimate is that only 2.5% of the population would be expected to obtain a score lower than that observed). If the upper limit on this percentage is still extreme, then the neuropsychologist can be confident that the score lies beyond the normal range and is likely to represent an acquired deficit. In contrast, if the upper limit indicates that the observed score may not be uncommon in the healthy population, then the neuropsychologist would require more in the way of convergent evidence from other sources before inferring impairment.

Although the methods will be of greatest use when used with normative or control data from small samples, they are applicable to *any* normed test, irrespective of sample size. It was shown in Table 1 that the width of these limits can be substantial even with moderate sized normative samples. Furthermore, neuropsychologists typically use tests drawn from a variety of sources, and hence use norms obtained from a variety of samples. Therefore, it is as important to know that the limits on the abnormality of scores obtained on particular tests in the neuropsychologist's battery are narrow (i.e. the point estimates of the abnormality/rarity are liable to be accurate), as it is to know that, for others, the limits will be wide. For example, when tests yield conflicting findings, the width of the confidence limits on the point estimates provide one source of information when weighting this evidence.

The confidence limits presented in the present paper are confidence limits on the estimated *rarity* or *abnormality* of a given score or difference between scores. As noted, they allow the user to quantify the effects of error arising from using a sample in place of the population; i.e. they quantify the fallibility of normative or control sample data. The methods of obtaining these confidence limits do not factor in measurement error in the instruments. Indeed, it is important that these confidence limits are not confused with confidence limits that *do* quantify the effect of measurement error in a test instrument (or instruments) on an individual's score (or score differences). The latter confidence limits are obtained by multiplying the standard error of measurement of a test (or standard error of measurement of the difference in the case of test score differences) by a value of *z* corresponding to the desired limits (i.e. 1.96 for 95% limits). The distinction between the reliability and the abnormality of test score differences is a particularly important one in clinical neuropsychology [7,12,13,34,43].

A reliable difference between an individual's test scores is one unlikely to have arisen from measurement error in the tests. However, many healthy individuals will have reliable differences among their abilities in different cognitive domains and, therefore, a reliable difference cannot be taken as indicating acquired impairment (e.g. see [8]). Therefore, for many purposes, particularly assessments conducted for medico-legal purposes, the abnormality of differences are more directly relevant to detecting and quantifying impairment. Furthermore, when dealing with control or normative samples that have small *N*s it is not practical to examine the reliability of differences. The methods for such comparisons treat the reliability coefficients of the tests as parameters

whereas such coefficients are, of course, subject to sampling error [17]; the reliability estimates obtained from samples that have *N*s of the magnitude with which we are concerned would be very unstable.

Finally, it is important to be aware of caveats attached to the use of the point estimates and confidence limits presented in the present paper. The methods for obtaining point estimates of the abnormality of a test score [10,11], including the new method presented here for the difference between a score and the individual's mean score, involve assumptions about the underlying distributions from which the normative data were sampled. [1] In the case of the comparison of a score with a normative mean, the assumption is that the control data were sampled from a normal distribution. For the difference between (a) a pair of tests and (b) a score and the individual's mean, it is assumed that the differences follow a normal distribution. The assumption holds for (a) if scores on the two tests follow a bivariate normal distribution and it holds for (b) if scores follow a multivariate normal distribution, but these conditions are not essential.

These same assumptions apply to the corresponding methods for obtaining confidence limits on the abnormality of scores or score differences. Therefore, these methods should not be employed when it is known or suspected that the normative or control data are markedly skewed or platykurtic/leptokurtic. When the control or normative samples are small, the neuropsychologist should also be particularly alert to the presence of outliers. For example, in elderly control or normative samples it is not uncommon to observe occasional cases who perform very poorly despite the absence of any other evidence that suggests the presence of a brain pathology (e.g. early stage dementia).

7. Computer programs for confidence limits on the abnormality of test scores and test score differences

Although all of the calculations described in the present paper could be carried out by hand or calculator it would clearly be more convenient if the methods were automated. In addition, tables for the non-central *t*-distribution (or a computer package that contains an algorithm for non-central *t*-distributions) would be needed for the calculations and these may not be readily accessible. Because of these considerations the methods have been implemented in computer programs for PCs. Aside from their convenience, the use of these programs reduces the chance of clerical and arithmetic errors.

The program SINGLIMS.EXE provides confidence limits on the abnormality of a single test score. The user enters the mean and S.D. for the test in the normative or control sample along with the sample size. The individual patient's score on the test is then entered. The output consists of the point estimate of the percentage of the population that will fall below the individual's score, obtained using the modified *t*-test procedure outlined by Crawford and Howell [10], and the accompanying 95% confidence limits for this percentage using the method outlined in the present paper (formula 2).

The program DIFFLIMS.EXE provides confidence limits on the abnormality of a difference between a pair of tests. The user enters the means and S.D. for the tests in the normative or control sample, along with the sample size and the correlation between the two tests. The individual patient's scores on the two tests are then entered and these are converted by the program to *z* scores. The output consists of the point estimate of the percentage of the population that will fall below the individual's difference score, obtained using Crawford et al.'s [11] procedure, and the accompanying 95% confidence limits for this percentage using the method outlined in the present paper (formula 5).

The program PROFLIMS.EXE provides the point estimate and confidence limits on the abnormality of the difference between a test and an individual's mean score on *k*-tests (formulae 8 and 10). The user enters the number of tests involved (*k*) and the size of the normative sample, and then the sample means and S.D. for the tests. The program then prompts for entry of the correlations between the *k*-tests (into a lower diagonal matrix). The individual patient's scores on the tests are then entered.

For each of the *k*-tests the output consists of the raw score, the score expressed as a *z* score, the difference between the *z* score on the test and the individual's mean *z* score, and the point estimate of the percentage of the population that would obtain a difference that is lower than the individual's. Finally, 95% confidence limits on this percentage are provided.

A useful feature of these programs is that the statistics from the normative sample are saved to a file. Therefore, subsequently, the neuropsychologist can rapidly generate point estimates and confidence limits on the rarity/abnormality of test scores, or test score differences, for other patients. Compiled versions of these programs can be downloaded from the first author's web site at the following address: http://www.psyc.abdn.ac.uk/homedir/jcrawford/abnolims.htm.

8. Conclusion

The single case approach in neuropsychology has made a significant contribution to our understanding of the architecture of human cognition [3,4,16,21,25,33]. However, as Caramazza and McCloskey [3] notes, if advances in theory are to be sustainable they "... must be based on unimpeachable methodological foundations" (p. 619). The statistical analysis of single case data is an aspect of methodology that has been relatively neglected. This is to be regretted. Other methodological (and logical) considerations may have compelled many researchers to abandon group-based research,

[1] The equivalent methods that treat the sample statistics as population parameters [20,29,36] also make these assumptions.

but it is clear that the *statistical* problems associated with drawing inferences from single cases significantly exceed those of the former approach. Although there remains much to do, we believe that the methods presented here make a useful contribution to the process of developing valid, optimal, and practical statistical methods for single case research.

Acknowledgements

We are grateful to Dr. Sytse Knypstra of the Department of Econometrics, University of Groningen, The Netherlands, for providing an algorithm that finds the non-centrality parameter of a non-central *t*-distribution given a quantile, its associated probability, and the d.f. This algorithm is incorporated into the computer programs that implement the methods presented in this paper. We are also grateful to Professor David C. Howell (Department of Psychology, University of Vermont) for early discussions on the topics addressed in this paper. Finally, we would like to thank the anonymous referees whose suggestions have improved the paper. In particular, one referee provided a paragraph and a figure illustrating the properties and use of non-central *t*-distributions. We have incorporated this material with only small alterations.

Appendix A

A.1. Formula for modified t-test comparing an individual's score with the mean score for a normative or control sample (see [10,37])

$$t = \frac{X_1 - X_2}{S_2\sqrt{(N_2+1)/N_2}},$$

where X_1 is the individual's score, $\bar{X}_2$ the mean score in the normative sample, S_2 the standard deviation of scores in the normative sample, and N_2 the number of persons in the normative sample. The test statistic follows a *t*-distribution on $N_2 - 1$ d.f. Multiplying the one-tailed probability of t by 100 gives the point estimate of the abnormality of the individual's score (e.g. if P is 0.03 then the point estimate is that 3% of the population would obtain scores lower than that observed for the individual).

A.2. Formula for modified t-test comparing the difference between an individual's score on two tests with the mean difference for a normative or control sample (see [11])

$$t = \frac{Z_X - Z_Y}{\sqrt{(2-2r_{XY})((N_2+1)/N_2)}},$$

where Z_X and Z_Y are the scores of an individual on tests X and Y expressed as z scores formed using the means and S.D. of the normative sample, r_{XY} the correlation between tests X and Y in the normative sample and N_2 is as defined above. The test statistic follows a *t*-distribution on N_2-1 d.f. Multiplying the one-tailed probability of t by 100 gives the point estimate of the abnormality of the individual's score. A derivation for the formula can be found in Appendix A of [11].

Appendix B

B.1. Derivation of h

The confidence intervals given in this paper are derived from a non-central *t*-distribution. This distribution is defined by

$$T_\nu(\delta) = \frac{(Z+\delta)}{\sqrt{Y/\nu}},$$

where Z has a normal distribution with a mean of zero and variance 1, and Y is independent of Z with a Chi-square distribution on ν d.f. δ is referred to as the non-centrality parameter.

For a specified value X_0, let $P^* = \Pr(X < X_0) \times 100$ where $X \sim N(\mu, \sigma^2)$. We require a $100(1-\alpha)\%$ confidence interval for P^* based on sample data $\bar{X}$ and S^2, where $\bar{X} \sim N(\mu, \sigma^2/N)$ and $\nu S^2/\sigma^2 \sim \chi^2(\nu)$. (For the methods given in this paper, $\nu = N - 1$.) Put

$$c = \frac{(X_0 - \bar{X})}{S} \tag{B.1}$$

and let $c^* = (X_0 - \mu)/\sigma$. Then c is an estimate of c^*. Also, tables for the percentage points of a normal distribution determine P^* from c^*, so a $100(1-\alpha)\%$ confidence interval for c^* will yield the required confidence interval for P^*. Now,

$$c\sqrt{N} = \frac{(\mu - \bar{X})\sqrt{N}/\sigma + (X_0 - \mu)\sqrt{N}/\sigma}{\sqrt{S^2/\sigma^2}}$$

so $c\sqrt{N}$ has a non-central *t*-distribution with non-centrality parameter $\delta = c^*\sqrt{N}a$ and ν d.f. The $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ points of this distribution will depend on the value of δ. Let δ_L denote the value of δ for which the $100(1-\alpha/2)\%$ point is $c\sqrt{N}$. Similarly, let δ_U denote the value of δ for which the $100(\alpha/2)\%$ point is $c\sqrt{N}$. Then $(\delta_L/\sqrt{N}, \delta_U/\sqrt{N})$ is a $100(1-\alpha)\%$ confidence interval for c^*. Define $h(\alpha/2;\ c;\ \nu+1)$ and $h(1-\alpha/2;\ c;\ \nu+1)$ by

$$h(\alpha/2;\ c;\ \nu+1) = \Pr(Z < \delta_L/\sqrt{N}) \times 100$$

and

$$h(1-\alpha/2;\ c;\ \nu+1) = \Pr(Z < \delta_U/\sqrt{N}) \times 100$$

where Z is the standard normal variate (i.e. $Z \sim N(0,1)$). Then a $100(1-\alpha)\%$ confidence interval for P^* is $(h(\alpha/2;\ c;\ \nu+1),\ h(1-\alpha/2;\ c;\ \nu+1))$.

B.2. Derivations of formulae (2) and (5)

As $\nu + 1 = N$ formula (2) is obtained by applying the result in Appendix B.1. Formula (5) also follows immediately when S_{D} is given; we simply put $X_0 = -|X_{z0} - Y_{z0}|$, $\bar{X} = 0$ and $S = S_{D_z}$ in formula (B.1). When the summary statistics are $\bar{X}$, $\bar{Y}$, S_X^2, S_Y^2 and r_{XY}, we must show that

$$S_{\mathrm{D}}^2 = 2 - 2r_{XY}.$$

$\bar{X}_z = Y_z = 0$ and $S_{X_z} = S_{Y_z} = 1$ giving

$$S_{D_z}^2 = \frac{\Sigma(X_z - Y_z)^2}{(N-1)} = \frac{(\Sigma X_z^2 + \Sigma Y_z^2 - 2\Sigma X_z Y_z)}{(N-1)}$$
$$= S_{X_z}^2 + S_{Y_z}^2 - 2r_{XY}S_{X_z}S_{Y_z} = 2 - 2r_{XY}.$$

B.3. Distribution of t_{Da} in formula (10)

The difference observed for an individual is $X_{\mathrm{a}} - \bar{X}_k$. The average value of this difference over the normative sample is 0. Silverstein [36] shows that the standard deviation of the difference is S_{Da}, so the result follows from Eq. (2) in Crawford et al. [11].

References

[1] Atkinson L. Some tables for statistically based interpretation of WAIS-R factor scores. Psychological Assessment 1991;3:288–91.

[2] Calder AJ, Young AW, Rowland D, Perret DI, Hodges JR. Facial emotion recognition after bilateral amygdala damage: differentially severe impairment of fear. Cognitive Neuropsyhology 1996;13:699–745.

[3] Caramazza A, McCloskey M. The case for single-patient studies. Cognitive Neuropsychology 1988;5:517–28.

[4] Code C, Wallesch C, Joanette Y, Lecours AR, editors. Classic cases in neuropsychology. Hove, UK: Psychology Press, 1996.

[5] Crawford JR. Estimation of premorbid intelligence: a review of recent developments. In: Crawford JR, Parker DM, editors. Developments in clinical and experimental neuropsychology. New York: Plenum Press, 1989. p. 55–74.

[6] Crawford JR. Current and premorbid intelligence measures in neuropsychological assessment. In: Crawford JR, Parker DM, McKinlay WW, editors. A handbook of neuropsychological assessment. London: Erlbaum, 1992. p. 21–49.

[7] Crawford JR. Assessment. In: Beaumont JG, Kenealy PM, Rogers MJ, editors. The Blackwell dictionary of neuropsychology. London: Blackwell, 1996. p. 108–16.

[8] Crawford JR, Allan KM. WAIS-R subtest scatter: base rate data from a healthy UK sample. British Journal of Clinical Psychology 1996;35:235–47.

[9] Crawford JR, Allan KM, McGeorge P, Kelly SM. Base rate data on the abnormality of subtest scatter for WAIS-R short-forms. British Journal of Clinical Psychology 1997;36:433–44.

[10] Crawford JR, Howell DC. Comparing an individual's test score against norms derived from small samples. The Clinical Neuropsychologist 1998;12:482–6.

[11] Crawford JR, Howell DC, Garthwaite PH. Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples *t*-test. Journal of Clinical and Experimental Neuropsychology 1998;20:898–905.

[12] Crawford JR, Sommerville J, Robertson IH. Assessing the reliability and abnormality of subtest differences on the Test of Everyday Attention. British Journal of Clinical Psychology 1997;36: 609–17.

[13] Crawford JR, Venneri A, O'Carroll RE. Neuropsychological assessment of the elderly. In: Bellack AS, Hersen M, editors. Comprehensive clinical psychology, vol. 7. Clinical geropsychology. Oxford, UK: Pergamon, 1998. p. 133–69.

[14] Daly F, Hand DJ, Jones MC, Lunn AD, McConway KJ. Elements of statistics. Wokingham, England: Addison-Wesley, 1995.

[15] Deary IJ. Age-associated memory impairment: a suitable case for treatment. Ageing and Society 1995;15:393–406.

[16] Ellis AW, Young AW. Human cognitive neuropsychology: a textbook with readings. Hove, UK: Psychology Press, 1996.

[17] Feldt LS, Brennan RL. Reliability. In: Linn RL, editor. Educational measurement. 3rd ed. New York: Macmillan, 1983.

[18] Gardner MJ, Altman DG. Statistics with confidence- confidence intervals and statistical guidelines. London: British Medical Journal, 1989.

[19] Grossman FM, Herman DO, Matarazzo JD. Statistically inferred vs. empirically observed VIQ-PIQ differences in the WAIS-R. Journal of Clinical Psychology 1985;41:268–72.

[20] Howell DC. Statistical methods for psychology. 4th ed. Belmont, CA: Duxbury Press, 1997.

[21] Humphreys GW, editor. Case studies in the neuropsychology of vision. Hove, UK: Psychology Press, 1999.

[22] Kaufman AS. Assessing adolescent and adult intelligence. Boston, MA: Allyn & Bacon, 1990.

[23] Ley P. Quantitative aspects of psychological assessment. London: Duckworth, 1972.

[24] Lezak MD. Neuropsychological assessment. 3rd ed. New York: Oxford University Press, 1995.

[25] McCarthy RA, Warrington EK. Cognitive neuropsychology: a clinical introduction. San Diego, CA: Academic Press, 1990.

[26] Miller E. Dissociating single cases in neuropsychology. British Journal of Clinical Psychology 1993;32:155–67.

[27] Mittenberg W, Thompson GB, Schwartz JA. Abnormal and reliable differences among Wechsler Memory Scale—revised subtests. Psychological Assessment 1991;3:492–5.

[28] O'Carroll R. The assessment of premorbid ability: a critical review. Neurocase 1995;1:83–9.

[29] Payne RW, Jones G. Statistics for the investigation of individual cases. Journal of Clinical Psychology 1957;13:115–21.

[30] Robertson IH, Ward T, Ridgeway V, Nimmo-Smith I. The Test of Everyday Attention. Bury St Edmunds: Thames Valley Test Company, 1994.

[31] Sattler J. Assessment of children. 3rd ed. San Diego: Sattler, 1988.

[32] Shallice T. Case study approach in neuropsychological research. Journal of Clinical Neuropsychology 1979;3:183–211.

[33] Shallice T. From neuropsychology to mental structure. Cambridge, UK: Cambridge University Press, 1988.

[34] Silverstein AB. Reliability and abnormality of test score differences. Journal of Clinical Psychology 1981;37:392–4.

[35] Silverstein AB. New formulas for evaluating the abnormality of test score differences. Journal of Psychoeducational Assessment 1984;2:79–82.

[36] Silverstein AB. Pattern analysis: the question of abnormality. Journal of Consulting and Clinical Psychology 1984;52:936–9.

[37] Sokal RR, Rohlf JF. Biometry. 3rd ed. San Francisco, CA: Freeman, 1995.

[38] Sprengelmeyer R, Young AW, Sprengelmeyer A, et al. Recognition of facial expressions: selective impairment of specific emotions in Huntington's disease. Cognitive Neuropsychology 1997;14:839–79.

[39] Vanderploeg RD, editor. Clinician's guide to neuropsychological assessment. Hillsdale, NJ: Erlbaum, 1994.

[40] Wechsler D. Manual for the Wechsler adult intelligence scale, revised. New York: Psychological Corporation, 1981.

[41] Wechsler D. Manual for the Wechsler adult intelligence scale, 3rd ed. San Antonio TX: The Psychological Corporation, 1997.
[42] Wechsler D, Wycherley RJ, Benjamin L, Crawford JR, Mockler D. Manual for the Wechsler adult intelligence scale, 3rd ed. UK, London: The Psychological Corporation, 1998.
[43] Willmes K. An approach to analyzing a single subject's scores obtained in a standardized test with application to the Aachen Aphasia Test (AAT). Journal of Clinical and Experimental Neuropsychology 1985;7:331–52.
[44] Zar JH. Biostatistical analysis. 3rd ed. London: Prentice-Hall, 1996.