Journal of Clinical and Experimental Neuropsychology
1998, Vol. 20, No. 6, pp. 898-905

1380-3395/98/2006-898$12.00

METHODOLOGICAL COMMENTARY

Payne and Jones Revisited: Estimating the Abnormality of Test Score Differences Using a Modified Paired Samples *t* Test*

J. R. Crawford[1], David C. Howell[2], and Paul H. Garthwaite[3]
[1]University of Aberdeen, UK, [2]University of Vermont, Burlington, and [3]Department of Mathematical Sciences, University of Aberdeen, UK

ABSTRACT

Payne and Jones (1957) presented a useful formula for estimating the abnormality of differences between an individual's scores on two tests. Extending earlier work by Sokal and Rohlf (1995) and Crawford and Howell (in press), we developed a modified paired samples *t* test as an alternative to this formula. Unlike the Payne and Jones formula, the new method treats data from a normative or control sample as sample statistics rather than as population parameters. Technically, the new method is more appropriate for any comparison of an individual's difference score against normative data. However, it is most useful when the normative data is derived from samples with modest *N*s; in these circumstances the Payne and Jones method overestimates the abnormality of differences. We suggest that the modified *t* test can be a useful tool in clinical practice and in single-case research. A computer program is made available that automates the calculations involved and can be used to store relevant data for future use.

In the assessment of acquired cognitive deficits, normative comparison standards have limitations because of the large individual differences in premorbid competencies. For example, an average test score can represent a marked decline from the premorbid level for a gifted individual whereas a low score can represent an entirely normal level of functioning for someone with modest premorbid abilities. Thus, considerable emphasis is placed on *intra*-individual comparisons when attempting to detect and quantify acquired deficits (Crawford, 1992; Lezak, 1995; Walsh, 1991). In the simplest case the clinician may wish to compare an individual's scores on two tests; a fundamental consideration in assessing the clinical significance of any discrepancy between scores on the tests is the extent to which it is rare or abnormal.

* Address correspondence to: John R. Crawford, Department of Psychology, King's College, University of Aberdeen, Aberdeen AB24 2UB, UK. E-mail: crawford@abdn.ac.uk
Accepted for publication: July 30, 1998.

USES OF THE PAYNE AND JONES (1957) FORMULA FOR THE ABNORMALITY OF DIFFERENCES

Payne and Jones (1957) presented a formula for the abnormality of a difference between scores on two tests; the formula estimates the percentage of the population that will equal or exceed a given discrepancy. A number of authors have noted the usefulness of this formula for assessment in clinical psychology and clinical neuropsychology (Crawford, 1996; Jones, 1970; Ley, 1972; Miller, 1993; Neufeld, 1977; Silverstein, 1981). Applications of the formula include estimation of the abnormality of discrepancies between Verbal and Performance IQs (Grossman, Herman, & Matarazzo, 1985), and factor scores (Atkinson, 1991) on the Wechsler Adult Intelligence Scale Revised (WAIS-R; Wechsler, 1981), and estimation of the abnormality of discrepancies between memory indices (Mittenberg, Thompson, & Schwartz, 1991) on the Wechsler Memory Scale- Revised (Wechsler, 1987).

Although the formula has mainly been used as an aid in clinical practice it is also a useful research tool. In recent years there has been a resurgence of interest within academic neuropsychology in single-case studies (Ellis & Young, 1996; McCarthy & Warrington, 1990; Shallice, 1988). A principal aim of this work is to fractionate the cognitive system into its constituent parts by attempting to establish the presence of double dissociations of function. Miller (1993) draws a distinction between studies in which patients are selected a priori on the basis of an independent criterion (normally anatomical localisation of lesion), and studies in which patients showing a contrasting pattern of preserved and impaired performance are identified serendipitously from among large samples of patients with impairment in the general area of interest. He argues that, in the latter case, it is incumbent upon investigators to demonstrate that the dissociations observed were unlikely to have arisen from chance variations in performance. Miller (1993) suggests that the Payne and Jones formula can be used to estimate the rarity of an observed discrepancy and provides examples that suggest the *normal* range of variability may often be larger than investigators commonly suppose.

Although Miller emphasised that investigators may overestimate the rarity of a given discrepancy, there is also the opposite danger that clinicians may dismiss a discrepancy that appears modest but is, in fact, highly abnormal (this latter situation is liable to arise when the two tests are highly correlated in the healthy population).

PSYCHOMETRIC ASPECTS OF THE PAYNE AND JONES FORMULA

There are a number of variants on the generic Payne and Jones (1957) formula but the simplest version, both conceptually and computationally, requires that an individual's raw scores on the two tests of interest are first converted to *z* scores (Z_X and Z_Y) based on a normative sample. The difference between the two component scores is in turn expressed as a *z* score (Z_D); this *z* score can then be referred to a table of the area under the normal curve to estimate the percentage of the population which would exhibit a difference as large as that observed. The formula is presented below:

$$Z_{\mathrm{D}} = \frac{Z_X - Z_Y}{\sqrt{2-2r_{xy}}} \qquad (1)$$

where r_{xy} represents the correlation between the two tests and all other terms are as defined above. Like any *z* score, Z_D is obtained by dividing a deviation score (which in the present instance is the difference score between two component deviation scores) by the standard deviation of the deviation score (the denominator in Equation (1)). A formal derivation for this formula can be found in Ley (1972).

To illustrate its use take the example of an individual who has been administered tests of verbal and spatial short-term memory, and suppose that the correlation between these tests is 0.75. For simplicity assume that both tests are expressed as *T* scores (i.e., they have a mean of

50 and a *SD* of 10). If the scores obtained on the verbal and spatial tests were 55 and 40, respectively, these would correspond to *z* scores of 0.5 and –1.0. Entering the relevant data into the formula yields the following:

$$Z_D = \frac{0.5-(-1.0)}{\sqrt{2-2(0.75)}} = \frac{1.5}{\sqrt{0.5}} = 2.12$$

Referring the Z_D of 2.12 to a table of the area under the normal curve yields a *p* of .017; thus it is estimated that 3.4% of the population would exhibit a discrepancy, in either direction, as large as that observed; 1.7% would be expected to exhibit a discrepancy in the *direction* observed (i.e., a discrepancy in favour of verbal memory).

The Payne and Jones formula assumes that scores on the two tests are normally distributed. This should be borne in mind when using the formula although, for many tests used in clinical practice (particularly fully standardized instruments), the assumption will not be an unreasonable one. However, another feature of the formula is that the data used in the computations are treated as if they were *population parameters* rather than as *sample statistics*. The effect of this is that the formula will systematically overestimate the abnormality of the difference between an individual's test scores. When the sample used to generate the statistics has a large *N*, this effect will be minimal.

However, situations arise where the only available data come from a small sample but a clinician nevertheless wishes to estimate the abnormality of a discrepancy between test scores. For example, most clinicians use tests derived from a variety of sources so that, although the individual standardization samples for any particular pair of tests may both have large *N*s, the tests may only have been administered *together* in a much smaller sample. Even when the tests have been standardised together it may still be necessary to rely on a smaller sample for the necessary data because the test manual may not report the correlation between them.

The sample size issue is particularly relevant to single-case studies where, as noted, the principal aim is to establish dissociations of functions. 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 very modest (*N* is often < 10 and sometimes < 5).

The remainder of this paper is concerned with developing a method of estimating the abnormality of test score differences which treats the sample statistics as statistics rather than population parameters.

DEVELOPMENT OF A MODIFIED PAIRED SAMPLES *T* TEST FOR THE ABNORMALITY OF TEST SCORE DIFFERENCES

Sokal and Rohlf (1995), writing for biometricians, described a modification to the independent samples *t* test which can be used to compare a single specimen with a sample. In this modification the individual specimen (or person!) is treated as a sample of *N* =1 and, therefore, does not contribute to the estimate of the within-group variance. Crawford and Howell (in press) have recently illustrated the use of this procedure in clinical practice to compare an individual with norms derived from samples with modest *N*s. They contrast this approach with the standard procedure in which the sample statistics are treated as parameters; that is, the individual's score is converted to a *z* score and evaluated using tables of the area under the normal curve (Howell, 1997; Ley, 1972). Using Crawford and Howell's (in press) notation, Sokal and Rohlf's (1995) formula is as follows:

$$\frac{X_1-\overline{X_2}}{S_2\sqrt{\frac{N_2+1}{N_2}}}, \qquad (2)$$

where, for our purposes, X_1 = the individual's score, $\overline{X}$ = the mean of the normative sample, S_2 = the standard deviation of the normative sample, and N_2 = the sample size. The degrees of freedom for *t* are $N_2 = N_1 - 2$ which reduces to

$N_2 - 1$. Although this method can be used to determine if an individual's score is significantly different from that of the normative or control sample (and, thus, may be useful in single-case studies), Crawford and Howell (in press) emphasised that its value was primarily in providing an unbiased estimate of the *abnormality* of the individual's score (i.e., if the *p* value for *t* was calculated to be 0.03 then it can be estimated that only 3% of the healthy population would exhibit a score as extreme as that observed).

The above formula is for comparison of an individual's score on a *single* test with the mean of a normative sample. However, we can extend the approach to a paired *t* test and use it to compare the *difference* between an individual's scores on *two* tests with the mean difference between these tests in a normative sample. To achieve this we follow Payne and Jones and transform the individual's scores, and the scores of the normative sample, to *z* scores based on the mean and standard deviation of the normative sample. In the following formula the subscripts 1 and 2 identify the individual and normative sample, respectively, and *X* and *Y* identify the two tests (derivation of the formula is given in Appendix 1):

$$t = \frac{(Z_{X_1} - Z_{Y_1}) - (\overline{Z}_{X_2} - \overline{Z}_{Y_2})}{\sqrt{2-2r_{xy}}\sqrt{\frac{N_2+1}{N_2}}} \qquad (3)$$

The left-hand term in the denominator represents the standard deviation of the differences in the normative or control sample and is multiplied by the right-hand term to obtain the standard error of the difference. The right-hand term in the numerator drops out because, from centring, $\overline{Z}_{X2}$ and $\overline{Z}_{Y2}$ are both zero, and we can gather together terms in the denominator to obtain the following:

$$t = \frac{Z_{X_1} - Z_{Y_1}}{\sqrt{(2-2r_{xy})\left(\frac{N_2+1}{N_2}\right)}} \qquad (4)$$

This Equation (4) resembles the original Payne and Jones Equation (1) but differs in that the statistics of the normative or control sample are now treated as *statistics* rather than as *population parameters*. In Appendix 1 we show that the quantity in Equation (4) has a *t* distribution with N_2-1 degrees of freedom (rather than the standard normal distribution). This should be used to evaluate the abnormality of the difference between the individual's scores. Returning to the worked example of verbal and spatial memory tests used earlier; suppose that the normative data had been obtained from a small healthy, control group in which *N* =10. Entry of these data into the formula for the modified *t* test (Equation (4)) yields the following:

$$t = \frac{1.0-(-0.5)}{\sqrt{(2-2(0.75))\left(\frac{10+1}{10}\right)}} = \frac{1.5}{\sqrt{(2-1.5)(1.1)}} =$$

$$\frac{1.5}{\sqrt{0.55}} = 2.023$$

The *p* value for a *t* of 2.023 with 9 *df* = 0.037. Therefore, it is estimated that, 7.4% of the population would exhibit a discrepancy, in either direction, as large as that observed; 3.7% would be expected to exhibit a discrepancy in the *direction* observed (i.e., a discrepancy in favor of verbal memory). This estimate of the abnormality of the patient's score would be preferred over the exaggerated estimate provided by the Payne and Jones formula (for which the corresponding figures were 3.4% and 1.7%), because the latter treats the control data as parameters.

Table 1 provides further data comparing the Payne and Jones formula with the modified *t*-test procedure. For consistency, the correlation between tests in this example was also taken as 0.75. Table 1 records the value that an individual's difference score must exceed to be significantly different from a normative or control sample at the 0.05 level, one-tailed; these difference scores correspond to the numerator in Equations (1) and (4). It can be seen that if *N* = 20 in the normative sample then an individual's difference score (Z_{X_1}–Z_{Y_1}) would have to exceed

1.25 to have a probability of less than 0.05; for example, if an individual's score on Test X was 0.1 (marginally above the normative sample mean) and was –1.1 on Test Y, the difference (–1.2) would not be large enough to be significantly different from the normative sample. The results from the *t*-test procedure can be compared with those from the Payne and Jones formula which, as noted, ignores the size of the normative sample and treats the sample statistics as parameters. The results from the Payne and Jones formula are presented in bold in the last row of Table 1.

The examples in Table 1 are intended to further illustrate that the Payne and Jones formula systematically overestimates the abnormality of an individual's difference score when used with data from samples with modest *N*s. To this end we took a significance testing approach and demonstrated that the critical values required for the Payne and Jones approach are smaller than those required with the modified *t* test. This inferential use of the modified *t* test may be appropriate for single-case studies in which there will be a legitimate concern with rejecting the null hypothesis. However, it should be stressed that we primarily see the value of the modified *t* test as simply providing the clinician with a less biased estimate of the abnormality of an individual's difference score.

Technically, the modified *t* test introduced in the present paper is more appropriate than the original Payne and Jones formula for *any* comparison of an individual's difference score against norms for difference scores. This is because the normative data with which we work are always derived from samples rather than populations. However, with large samples (e.g., > 250) the difference between the value of *t* and *z* becomes vanishingly small and ,thus, the estimates of the abnormality of a difference score also converge. Further, even with more moderate sample sizes (e.g., $N = 50$), the difference between the two are relatively trivial. Thus, we would suggest that the modified *t* test procedure be used with $N < 50$. With larger sample sizes the Payne and Jones formula approximates the technically more correct method, thus, either method could be used. However, by using the computer program accompanying this paper the *t*-test procedure can be run in under 30 seconds. Thus the bias inherent in the Payne and Jones method is not necessarily offset by any savings in time.

One of the assumptions underlying any form of *t* test is that the data is normally distributed. Monte Carlo simulations have revealed that *t* tests are surprisingly robust in the face of moderate violation of this assumption (Boneau, 1960; Howell, 1997). However, especially given the small *N*s with which we are concerned, the *t* test procedure and the Payne and Jones *z* procedure are best avoided when it is known or suspected that the normative data are markedly skewed.

When referring to discrepancies between an individual's scores on two tests, it may be best to reserve the term ‘‘significantly different’’ to describe the case in which the scores are *reliably* different. In that context a significant difference is taken to mean that the difference is unlikely (e.g., p <.05) to have occurred because of measurement error in the tests concerned. If the tests have very high reliabilities, it is quite possible that the majority of healthy individuals would exhibit reliable differences between their scores. Thus, the issue of the reliability of a difference should not be confused with the topic of the present paper which is primarily concerned with the *rarity* or *abnormality* of a patient's difference score; see Crawford, (1996), Crawford, Sommerville, and Robertson, (1997), Crawford, Venneri, and O'Carroll (1998), and Silverstein (1981) for related discussion of the distinction between reliable and abnormal differences.

COMPUTER PROGRAM FOR THE ABNORMALITY OF A DIFFERENCE BETWEEN A PAIR OF TEST SCORES

The calculations involved in the aforementioned procedure are relatively straightforward. However, a statistics package would be required to obtain the exact *p* corresponding to a given *t*. In view of this, and because of an awareness of the time pressures under which many of us operate,

Table 1. Cutoff Values to Attain Significance.[a]

N_2	$Z_{X1} - Z_{Y1}$
5	1.65
7	1.47
10	1.36
15	1.29
20	1.25
25	1.23
30	1.22
50	1.20
70	1.19
120	1.18
—	1.16

[a] Table shows the difference (Z_{X_1}–Z_{Y_1}) between an individual's scores on two tests required to be significantly different at the .05 level (one-tailed) when compared against normative samples of varying *N*s. For comparison purposes the last row records the values that would be required if the sample statistics were treated as population parameters, that is, if the Payne and Jones formula were applied. This specific example is based on a correlation of 0.75 between the two tests.

we have written a program for PCs that automates the process. Apart from saving time, use of this program reduces the chance of clerical and arithmetic errors. The user enters the mean and *SD* for the two tests of interest in the normative or control sample, followed by the tests' intercorrelation and the sample size. These data can be saved to a file (as can equivalent data for other pairs of tests) so that they need not be reentered each time the clinician wishes to use the procedure with future clients. Finally, the individual client's scores on the two tests are entered. The output consists of the original scores and the difference score in *z*-score form, and the estimate of the percentage of the population that will exceed the observed discrepancy (in either direction and in the observed direction). A compiled version of the program can be downloaded from the first author's website at the following address: http://www.psyc.abdn.ac.uk/homedir/jcrawford/pairabno.htm

CONCLUSION

There is much evidence that human judges are poor at estimating the rarity of differences involving correlated components (Hogarth, 1975; Slovic & Lichtenstein, 1971). Therefore, clinicians should, where possible, use a quantitative method to help them interpret an individuals' apparent neuropsychological strengths and weaknesses rather than rely solely on clinical intuition. The present method for estimating the abnormality of differences between an individual's test scores should assist in this process. Unlike the Payne and Jones method, it does not systematically overestimate the rarity of differences when the normative or control data has been derived from samples with modest *N*s.

REFERENCES

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

Boneau, C. A. (1960). The effect of violation of assumptions underlying the t-test. *Psychological Bulletin, 57*, 49-64.

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: Lawrence Erlbaum.

Crawford, J. R. (1996). Assessment. In J. G. Beaumont, P. M. Kenealy, & M. J. Rogers (Eds.), *The Blackwell dictionary of neuropsychology* (pp. 108-116). London: Blackwell.

Crawford, J. R., & Howell, D. C. (in press). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*.

Crawford, J. R., Sommerville, J., & Robertson, I. H. (1997). 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. (1998). 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.

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

Grossman, F. M., Herman, D. O., & Matarazzo, J. D. (1985). Statistically inferred vs. empirically observed VIQ-PIQ differences in the WAIS-R. *Journal of Clinical Psychology, 41*, 268-272.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association, 70*, 271-294.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.

Jones, H. G. (1970). Principles of psychological assessment. In P. J. Mittler (Ed.), *The psychological assessment of mental and physical handicaps* (pp. 1-25). London: Tavistock.

Ley, P. (1972). *Quantitative aspects of psychological assessment*. London: Duckworth.

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

McCarthy, R. A., & Warrington, E. K. (1990). *Cognitive neuropsychology: A clinical introduction*. San Diego, CA: Academic Press.

Miller, E. (1993). Dissociating single cases in neuropsychology. *British Journal of Clinical Psychology, 32*, 155-167.

Mittenberg, W., Thompson, G. B., & Schwartz, J. A. (1991). Abnormal and reliable differences among Wechsler Memory Scale – Revised subtests. *Psychological Assessment, 3*, 492-495.

Neufeld, R. W. J. (1977). *Clinical quantitative methods*. New York: Grune & Stratton.

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

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.

Silverstein, A. B. (1981). Reliability and abnormality of test score differences. *Journal of Clinical Psychology, 37*, 392-394.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organisational Behaviour and Human Performance, 6,* 649-744.

Sokal, R. R., & Rohlf, J. F. (1995). *Biometry* (3rd ed.). San Francisco, CA: W.H. Freeman.

Walsh, K. W. (1991). *Understanding brain damage* (2nd ed.). Melbourne, Australia: Churchill Livingstone.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation.

Wechsler, D. (1987). *Manual for the Wechsler Memory Scale-Revised*. San Antonio, TX: Psychological Corporation.

APPENDIX 1

Distribution of the test statistic in Equation (4)

Assume the centres and scales of *X* and *Y* have been so chosen that both their sample means are 0 in the normative sample and their sample variances are both 1. Let x_i and y_i denote the values of *X* and *Y* for the *i*th individual in the normative sample ($i = 1,..., N_2$). Then

$$\mathrm{r}_{xy} = \frac{1}{N_2-1}\sum_{i=1}^{N_2} x_i y_i$$

and

$$\frac{1}{N_2-1}\sum_{i=1}^{N_2} (x_i-y_i)^2 = 2-2r_{xy}$$

Define $d = X - Y$ and assume it is normally distributed. In the normative sample the mean and sample variance of *d* are 0 and $(2 - 2r_{xy})$, respectively. Hence analogous to Equation (2),

$$\frac{d-0}{\sqrt{(2-2r_{xy})\left(\frac{N_2+1}{N_2}\right)}}$$

follows a *t*-distribution on $N_2 - 1$ degrees of freedom. With the given choice of centres and scales, $Z_{X_1}-Z_{Y_1}$ = d, so the quantity in Equation (4) has a *t* distributiuon on $N_2 - 1$ degrees of freedom.