The Clinical Neuropsychologist
1998, Vol. 12, No. 4, pp. 482-486

1385-4046/98/1204-482$12.00

Comparing an Individual's Test Score Against Norms Derived from Small Samples*

J.R. Crawford[1] and David C. Howell[2]
[1]University of Aberdeen, UK, and [2]University of Vermont, Burlington

ABSTRACT

The standard method for comparing an individual's test score with a normative sample involves converting the score to a *z* score and evaluating it using a table of the area under the normal curve. When the normative sample is small, a more appropriate method is to treat the individual as a sample of $N = 1$ and use a modified *t* test described by Sokal and Rohlf (1995). The use of this *t* test is illustrated with examples and its results compared to those from the standard procedure. It is suggested that the *t* test be used when the *N* of the normative sample is less than 50. Finally, a computer program that implements the modified *t*-test procedure is described. This program can be downloaded from the first author's website.

Comparison of an individual's test score against a normative sample is a fundamental feature of the assessment process in clinical neuropsychology. The procedure for statistical inference in this situation is well known; when it is reasonable to assume that scores are normally distributed, 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). Thus, if the clinician has formed a directional hypothesis concerning the individual's score prior to testing (e.g., that the score will be below the mean), then a *z* score which fell below -1.64 would be considered statistically significant (using the conventional .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 procedure provides the clinician with information on the rarity or abnormality of the individual's score.

In the standard procedure described above the normative sample is treated as if it were a population; that is, 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, 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 which have poor normative data. It should also be borne in mind that 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. For example, Geffen, Moar, O'Hanlon, Clark, and Geffen (1990) presented norms for the Auditory Verbal Learning Test (AVLT) based on a sample of 153 participants. As gender and age were both found to exert an influence on AVLT scores, the sample was broken down by these variables. The modal *n* in the resultant normative tables was 10.

Second, many clinicians have gathered local norms for neuropsychological instruments. This

* Address correspondence to: John R. Crawford, Department of Psychology, King's College, University of Aberdeen, Aberdeen AB24 3HN, UK. E-mail: crawford@abdn.ac.uk.
Accepted for publication: July 12, 1998.

may have stemmed from a concern that the demographics of the sample used to derive normative data for particular tests were too dissimilar from the population from which their patients are drawn; for example, the normative sample may consist predominantly of urban dwellers yet their patients come from isolated rural locations. Although the collection of local norms is to be actively encouraged, the time and expense involved is such that the sample sizes are often modest.

Third, in recent years there has been an enormous resurgence of interest within academic neuropsychology in single case studies (Ellis & Young, 1996; McCarthy & Warrington, 1990; Shallice, 1988). 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 less than 10 and often less than 5. It may be noted that the control group in such studies need not be healthy participants; for example, the hypothesis tested may be that a particular individual case differs in one important respect from a sample of patients having other clinical features in common.

When population parameters are known, or when we have normative samples of sufficient size to make highly reliable estimates of these parameters, probabilities for normally distributed scores are easily estimated by using *z*:

$$z = (X - \mu) / \sigma$$

One problem with using the standard deviation of a small normative sample as if it were the parameter (σ) is that, although the sample variance is a maximum likelihood estimator of σ^2, the sampling distribution of the variance is positively skewed. This means that we will be more likely to underestimate σ than to overestimate it. Thus we are more likely to overestimate *z* and the rarity of the observation.

When we are not in a position to assume a highly stable estimate of the necessary parameters, one solution is to use the standard deviation of the (small) normative sample as an estimate of the population standard deviation, but to calculate probabilities using the *t* distribution in place of *z*. The formula would otherwise be the same, but the distribution would have wider tails.

Sokal and Rohlf (1995), writing for biometricians, describe 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. Sokal and Rohlf's (1995) formula, with minor changes in notation, is as follows:

$$t = \frac{X_1 - \overline{X_2}}{s_2 \sqrt{\frac{N_2 + 1}{N_2}}},$$

where, for our purposes, X_1= the individual's score, $\overline{X_2}$ = the mean of the normative sample, S_2 = the standard deviation of the normative sample, and N_2 = the sample size. To avoid any uncertainty the standard deviation referred to here is the estimated population standard deviation, that is, it should be calculated with *N*-1 in the denominator not *N*. With large samples the two *SD*s will be very similar but this is not the case with small *N*s. When working with summary data (i.e., the output of a statistical package or published data) it can be assumed that the *SD* reported will be the estimated population *SD*.The degrees of freedom for *t* are $N_2 + N_1 - 2$ which reduces to $N_2 - 1$.

To illustrate the application of this procedure in neuropsychological assessment, take the example of a patient who obtains a score of 33 on a test of spatial memory. The normative sample for someone of her / his age has an *N* of 15 and a mean and standard deviation of 50 and 10 respectively. Entering this data into the formula yields the following:

$$t = \frac{X_1 - \overline{X_2}}{s_2 \sqrt{\frac{N_2 + 1}{N_2}}} = \frac{33 - 50}{10 \sqrt{\frac{15 + 1}{15}}} = \frac{17}{10 \sqrt{1.066}} =$$

$$\frac{17}{10.32} = 1.647$$

$df = N_2 - 1 = 15 - 1 = 14$

If we assume that the clinician hypothesised the presence of a deficit a priori and therefore employ a one-tailed test, the critical value for *t* at the .05 level with 14 degrees of freedom is 1.76; thus the individual's score is not significant at the .05 level. However, it does achieve significance at the more liberal .10 level (critical value = 1.35). The exact *p* in this example is .062. Therefore, the expectation is that only 6% of individuals in the population from which the normative sample was drawn would obtain a score as low as that observed for the patient.

Table 1 is designed to provide further insight into the modified *t*-test procedure. It records the test score below which an individual's score must fall to be considered significantly different at the .05 level, one-tailed. The critical values are those required when the scores of a normative sample are expressed on one of three generic measurement scales; these scales were chosen so as to be readily familiar to clinicians. Thus, the scores are expressed as *z* scores ($\bar{X}$ = 0, *SD* = 1), *T* scores ($\bar{X}$ = 50, *SD* = 10) and IQ or Memory Quotients (MQ) ($\bar{X}$= 100, *SD* = 15). These three scales are simple linear transformations of one another, reflecting the differing choices we make in how we present scores to a wider audience.

To illustrate the data in Table 1, if scores were expressed as *T* scores, an individual's score would have to fall below 31 to have a probability of less than .05 when *N* = 10 in the normative sample. The results from using the modified *t*-test procedure can be compared against the results obtained by ignoring the size of the normative sample and treating the sample statistics as parameters (i.e., evaluating a score using tables of the area under the normal curve). The results of applying the latter procedure are recorded in bold in the last row of Table 1.

This table illustrates a number of points other than the obvious one that the smaller the normative sample the larger the difference required for significance. We have noted that for clinicians the emphasis will primarily be on obtaining an estimate of the rarity or abnormality of an individual's test score rather than on whether it is significant at a given significance level. It is informative to contrast the estimates of the rarity of an individual's score produced by the two procedures. Take the case where *N* = 10 for the normative or control sample, scores are expressed as *z* scores, and the individual's score is -1.92. Referring this *z* score to a table of the nor-

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

N_2	*Z* score ($\bar{X}$ = 0)	*T* Score ($\bar{X}$ = 50)	IQ / MQ ($\bar{X}$ = 100)
5	–2.33	27	65
7	–2.08	29	69
10	–1.92	31	71
15	–1.82	32	73
20	–1.77	32	73
25	–1.74	33	74
30	–1.73	33	74
50	–1.69	33	75
70	–1.68	33	75
120	–1.66	33	75
	–1.64	**34**	**75**

[a] *Note*. Table shows the value below which an individual's score must fall to be significantly different at the .05 level (one-tailed) when compared against normative samples of varying *N*s; the individual's score is expressed as a *z* score, T score and IQ / MQ. For comparison purposes the last row records the values that would be required if the sample statistics were treated as population parameters. *T* scores and IQs / MQs are rounded up to the nearest integer.

mal curve would imply that only 2.7% of the population would exhibit a score this low. In contrast the *t*-test procedure estimates the percentage to be 5%; -1.92 is the critical value for an individual's score (at the .05 level, one-tailed) when the score is expressed as a *z* score, see Table 1.

A further appreciation can be gained by approaching this issue from the opposite direction; that is, by examining what percentage of the population is estimated to exhibit a score as low as an individual's score when the standard procedure estimates it at 5%. Again using the case of $N = 10$, the estimate provided by the *t* test is 7.6%. The message here is that, when *N* is small, use of the standard procedure will overestimate the rarity of an individual's score.

Given that the *t*-test procedure is rarely if ever used in clinical neuropsychology, it is ironic that technically it is more appropriate than the standard procedure for *any* comparison of an individual against a normative sample. This is because the norms we work with are always derived from samples rather than populations. However, with large samples (e.g., greater than 250) the difference between the value of *t* and *z* becomes vanishingly small. Further, even with more modest sample sizes, the difference between the two can be trivial. When an individual's score is expressed as a *z* score it can be seen that with a sample size of 50 there is still an appreciable difference between the *z* score needed to yield a significant *t* (1.69) and the *z* score needed for significance using the area under the normal curve (1.64). However, with the rounding involved in commonly used methods of expressing scores (such as *T* scores or MQs / IQs used in the present example) the two procedures have essentially converged. Thus, we would suggest that the modified *t* test be used with an *N* of less than 50; with larger sample sizes either method could be used but the standard method is more rapid.

One of the assumptions underlying any form of *t* test is that the data are normally distributed. Monte Carlo simulations have revealed that *t* tests are surprisingly robust in the face of moderate violation of this assumption (Boneau, 1960). However, especially given the small *N*s with which we were concerned, the *t*-test procedure is best avoided when it is known or suspected that the normative data are markedly skewed or platykurtic / leptokurtic. It should be noted that the standard procedure makes the same assumption of normality and is equally compromized by nonnormality.

It will be appreciated that the statistical power of any method of statistical inference will decline as sample size decreases. Thus with the small *N*s with which we are concerned it is inevitable that power will be low. The most obvious way of increasing power is to increase the size of the normative sample or control group against which the individual's score is to be compared. However, power can also be increased by adopting a more liberal significance level, for example, .15 or .20 rather than .05. Although this more liberal strategy will increase Type I errors (false positives) it will decrease Type II errors (false negatives). The decision to depart from the conventional .05 level should be based on the relative risks the clinician attaches to the occurrence of these two types of errors.

There is an important distiction between a reliable difference and a rare or abnormal one. There are many individuals whose mean body temperature is *reliably* above 36.9 ^{0}C and yet we would not class them as ill. However, if someone had a body temperature that was rare, for example, recorded a temperature that occurs in fewer than 5% of of the population, we would want to look closely at them. When working with the test score of an individual, the term ''significantly different from the mean'' is best used to describe the case in which the score is *reliably* different from the mean of a normative sample. Here a significant difference means that the difference is unlikely (e.g., $p < .05$) to have occurred because of measurement error in the instrument from which the score was derived. However, if a measurement instrument has high reliability, it would not be unusual for the majority of healthy individuals to be reliably different from the mean. 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 score. For further discussion of the dis-

tinction between reliable and abnormal differences, including the related issue of the reliability versus abnormality of discrepancies among an individual's scores on two or more tests, see Crawford (1996), Crawford, Sommerville, & Robertson (1997), Crawford, Venneri, & O'Carroll (1998), and Silverstein (1981).

The calculations involved in the modified *t*-test procedure are straightforward. However, as we are aware of the time pressures under which many clinicians operate, we have written a program for PCs which implements the formula. Apart from saving time, use of this program reduces the chance of clerical error. Furthermore, it provides an exact *p* for the test, whereas tabled values of *t* only record the *t* value which must be exceeded to achieve a given level of significance. An exact *p* is more useful for clinical purposes as the emphasis will be on the rarity or abnormality of the individual's score; for example, it is still of some value to have an estimate of the percentage of healthy individuals expected to exhibit a score as extreme as the patient's when the *t* value does not exceed the most liberal tabled value (e.g., .10 in many tables). It also does away with the need for interpolation when a value falls between two tabled critical values. The program will calculate the mean and standard deviation of the normative sample when these summary statistics are not already available to the user. A compiled version of this program can be downloaded from the first author's website at the address: http://www.psyc.abdn.ac.uk/homedir/jcrawford/singspec.htm.

To our knowledge the modified *t*-test procedure is not covered in any existing textbooks on statistics for psychology. Additionally, we are aware of only one paper in which the above method was employed (Venneri & Cubelli, 1998). In conclusion we believe that the modified *t*-test procedure can play a useful role in assisting the clinician or researcher with the quantitative analysis of test scores in the individual case.

REFERENCES

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

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., 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: Pergamon.

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

Geffen, G., Moar, K. J., O'Hanlon, A. P., Clark, C. R., & Geffen, L. B. (1990). Performance measures of 16- to 86-year-old males and females on the Auditory Verbal Learning Test. *The Clinical Neuropsychologist, 4*, 45-63.

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

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

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

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.

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

Venneri, A., & Cubelli, R. (1998). *The representation of geminate letters: Evidence from acquired dysgraphia*. (Manuscript submitted for publication).