ELSEVIER

Neuropsychologia 44 (2006) 666–677

NEUROPSYCHOLOGIA

www.elsevier.com/locate/neuropsychologia

Testing for a deficit in single-case studies: Effects of departures from normality

John R. Crawford [a,*], Paul H. Garthwaite [b], Adelchi Azzalini [c], David C. Howell [d], Keith R. Laws [e]

[a] *School of Psychology, King's College, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 2UB, UK*
[b] *Department of Statistics, The Open University, Milton Keynes, UK*
[c] *Department of Statistical Sciences, University of Padua, Padua, Italy*
[d] *Department of Psychology, University of Vermont, Burlington, USA*
[e] *School of Psychology, University of Hertfordshire, Hatfield, UK*

Received 2 March 2005; received in revised form 1 June 2005; accepted 16 June 2005
Available online 25 July 2005

Abstract

In neuropsychological single-case research inferences concerning a patient's cognitive status are often based on referring the patient's test score to those obtained from a modestly sized control sample. Two methods of testing for a deficit (z and a method proposed by Crawford and Howell [Crawford, J. R. & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. The Clinical Neuropsychologist, 12, 482–486]) both assume the control distribution is normal but this assumption will often be violated in practice. Monte Carlo simulation was employed to study the effects of leptokurtosis and the combination of skew and leptokurtosis on the Type I error rates for these two methods. For Crawford and Howell's method, leptokurtosis produced only a modest inflation of the Type I error rate when the control sample N was small-to-modest in size and error rates were lower than the specified rates at larger N. In contrast, the combination of leptokurtosis and skew produced marked inflation of error rates for small Ns. With a specified error rate of 5%, actual error rates as high as 14.31% and 9.96% were observed for z and Crawford and Howell's method respectively. Potential solutions to the problem of non-normal data are evaluated.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Neuropsychology; Single-case methods; Statistical methods; Non-normality; Robustness; Monte Carlo simulation

1. Introduction

In neuropsychological single-case research, inferences concerning a patient's cognitive status are commonly based on referring the patient's test scores to a control sample. The most common method of forming inferences about the presence of a deficit in such scenarios is to convert the patient's score on a given task to a z-score based on the mean and S.D. of the control sample and then refer this score to a table of the areas under the normal curve (Howell, 2002). Thus, if a neuropsychologist has formed a directional hypothesis for the patient's score prior to testing (i.e., that the patient's score will be below the control sample mean), then a score that fell below −1.645 would be considered statistically significant ($p < 0.05$) and would be taken as an indication that the patient had a deficit on the task in question.

One potential problem with this approach is that it treats the control sample as if it was a population; i.e., the mean and standard deviation are used as if they were *parameters* rather than *sample statistics*. This is not a problem if the control sample is large as then the sample statistics should provide sufficiently accurate estimates of the parameters. However, the control samples in single-case studies in neuropsychology typically have modest Ns; $N < 10$ is not unusual and Ns < 20 are very common (Crawford & Howell, 1998). With samples

* Corresponding author. Tel.: +44 1224 272231; fax: +44 1224 273426.
E-mail address: j.crawford@abdn.ac.uk (J.R. Crawford).

0028-3932/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2005.06.001

of this size it is not appropriate to treat the mean and S.D. as though they were parameters. The effect of using z with small control samples is to exaggerate the rarity/abnormality of a patient's score and to inflate the Type I error rate (Crawford & Garthwaite, 2005a); in this context a Type I error occurs when a case that is drawn from the control population is incorrectly classified as not being a member of this population; i.e., they are incorrectly classified as exhibiting a deficit.

A solution to the problem outlined above is to use a method proposed by Crawford and Howell (1998) that treats the control sample statistics as sample statistics (see also Crawford & Garthwaite, 2002). This method, based on Sokal and Rohlf (1995), uses the t-distribution (with $n-1$ degrees of freedom), rather than the standard normal distribution, to estimate the abnormality of the patient's scores and to test whether it is significantly lower than the scores of the control sample. The formula for this test is

$$t = \frac{x^* - \bar{x}}{s\sqrt{(n+1)/n}}, \quad (1)$$

where x^* is the patient's score, $\bar{x}$ and s the mean and standard deviation of scores in the control sample, and n is the size of the control sample.

The p value obtained when this test is applied is used to test significance, but it also simultaneously provides an unbiased point estimate of the *abnormality* of the patient's score; that is, it is an estimate of the proportion of the control population that would obtain a lower score. A formal proof of this statement is provided in Appendix 1 of Crawford and Garthwaite (2005b). As a concrete example, if the one-tailed p is 0.013 then we know that the patient's score is significantly ($p < .05$) below the control mean and that it is estimated that 0.013 (i.e., 1.3%) of the control population would obtain a score lower than the patient's.

An assumption underlying the use of both z and Crawford and Howell's method is that the control samples against which a case is compared have been drawn from a normal distribution. However, it is not at all uncommon for the scores of controls on neuropsychological tests to depart from normality (Capitani & Laiacona, 2000; Crawford & Garthwaite, 2005a). It is therefore important to examine the effects of departures from normality on these methods to examine whether they are robust. That is, there is the danger that the Type I error rate will be inflated when the assumption of normality is violated. (As noted, in this context, a Type I error would occur if we wrongly classified a case as not having been drawn from the control population; i.e., we incorrectly conclude they have a deficit.) Data on the effects of departures from normality would either provide reassurance for researchers, should the effects be mild, or serve as a warning, should the effects be substantial (in the latter scenario strategies for dealing with the problem should also be addressed).

Unfortunately, to date, little empirical work has addressed these important issues. An exception is the recent study by Crawford and Garthwaite (2005a) in which Monte Carlo simulation was used to examine the effects of skew on the Type I error rate for z and for Crawford and Howell's method. For both tests, skew produced an inflation of the Type I error rate but the effects were surprisingly modest; for Crawford and Howell's method the use of a more conservative critical value kept the error rate at or below 5% in all of the scenarios examined. In contrast, the Type I error rate for z was unacceptably high when the control sample N was small to modest in size. However, this was largely attributable to the inappropriate treatment of the control sample statistics as parameters; the presence of skew further inflated the error rate but this effect was relatively modest.

Another potential problem that will arise in the conduct of single-case research is that the distribution of control data will be overly peaked and have heavier tails than a normal distribution. That is, in practice, the distribution of the control data may be leptokurtic. An illustration of a leptokurtic distribution, superimposed on a normal distribution, is provided in Fig. 1. It can be seen from Fig. 1a that leptokurtic distributions (shaded) are more peaked than a normal distribution (unshaded) and have thinner "shoulders". Fig. 1b shows the same leptokurtic distribution with the right-hand tail area magnified to show the heavy tails.

Leptokurtic distributions are pervasive in many areas of scientific enquiry including psychology, economics, and biology (Lange, Little, & Taylor, 1989). For example, IQ tests would be regarded as prototypical examples of normally distributed psychological data; moreover, transformations are routinely applied to these tests to force them to conform to a normal distribution. Despite this, IQ tests have been shown to exhibit highly significant leptokurtosis (Burt, 1963).

Leptokurtosis is of concern because of the danger of inflating the Type I error rate when parametric tests are applied to data that possess this characteristic. Leptokurtic distributions are associated with the presence of outliers and one potential method of reducing leptokurtosis is to simply remove cases that meet some criterion for an outlier. An argument against such a course of action is that, if outliers have not arisen from coding errors, then they are a genuine characteristic of the phenomenon under investigation and should be incorporated in any analysis.

In the present context, that of a neuropsychologist attempting to make inferences concerning an individual patient, there are even more compelling reasons not to remove outliers. For example, suppose a single-case researcher sets out to obtain support for their hypothesis that a patient is impaired on measure X, and obtains data from a sample of healthy controls. Further, suppose that the researcher finds that some controls perform unexpectedly poorly on this measure such that their scores approach that of the patient and would be classed as outliers if referred to a normal distribution. Removing these cases would be dubious from both a scientific and ethical point of view unless additional, independently obtained, evidence of previously unsuspected pathology is uncovered. Thus, if leptokurtosis is liable to be a common feature of neuropsychological data and there are reasons not to deal with it

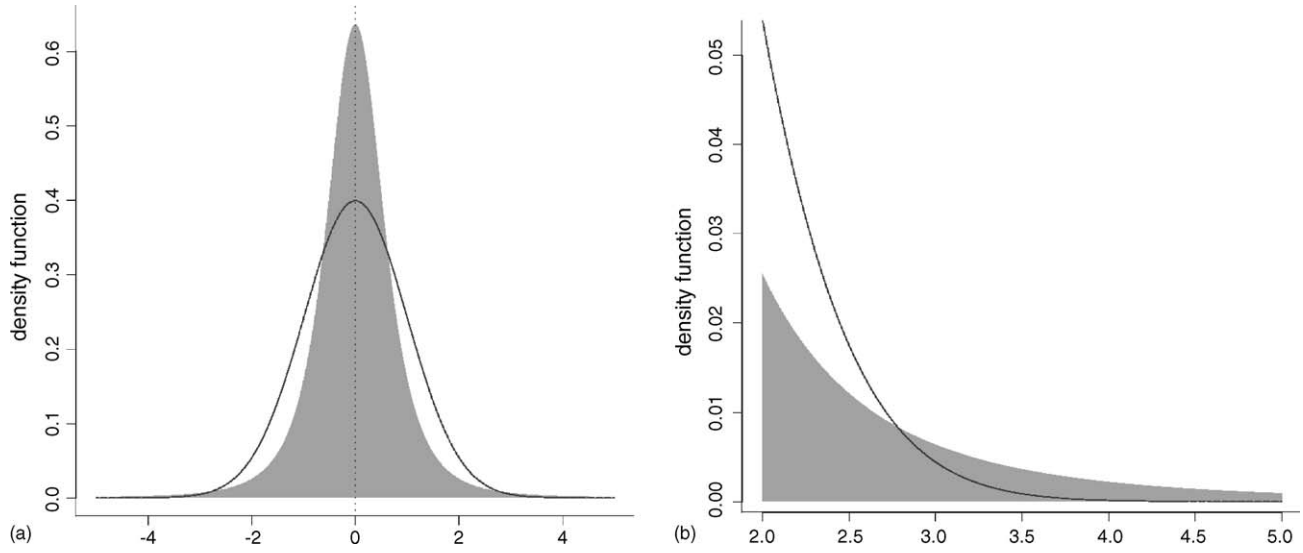

Fig. 1. Graphical illustration of a leptokurtic distribution (shaded area) superimposed upon a normal distribution (in (b) the right-hand tail area has been magnified to illustrate the heavy tails).

by removing outliers, we should at least study its likely effects on inferential methods for testing for a deficit and, should it prove to be problematic, implement alternative strategies for dealing with it.

2. Study 1

2.1. The effects of leptokurtosis in the control population on Type I error rates

As noted, a large literature indicates that leptokurtosis is a pervasive feature of data in many areas of scientific enquiry. However, its effects on inferential methods used in single-case studies have not been examined. In Study 1, Monte Carlo simulation is employed to estimate Type I error rates for both methods of testing for a deficit (z and Crawford and Howell's method) when the control data are leptokurtic.

2.1.1. Method

The most common approach to modelling the effects of leptokurtic distributions on test statistics is to sample from t-distributions (Lange et al., 1989). This is potentially confusing as Crawford and Howell's method uses the t-distribution to test for a significant difference between the case and controls. However, as noted, the assumption in applying this test statistic is that the controls were drawn from a normal distribution; in the present study we examine the effects of violating this assumption by drawing controls from leptokurtic distributions and it so happens that t-distributions have this required characteristic.

In the present study we sampled from t-distributions on 7 (moderate leptokurtosis) and 4 (severe leptokurtosis) degrees-of-freedom. Kurtosis (β_2) is 5 for a t-distribution on 7 d.f. compared to a value of 3 for a normal distribution; the kurtosis for a t-distribution on 4 d.f. is even more extreme but is undefined (because the denominator in the formula for kurtosis requires subtracting 4 from the d.f. and is hence zero).

Monte Carlo simulations were run on a PC and implemented in Borland Delphi (Version 4). The simulations were run with five different values of N (the size of the control sample): 5, 10, 20, 50 and 100. For each of these values of N, 1,000,000 samples of observations were drawn from a standard normal distribution. Each observation was divided by $\sqrt{\chi^2/7}$ or $\sqrt{\chi^2/4}$ where χ^2 is a random draw from a chi-square distribution on 7 or 4 degrees-of-freedom respectively. The resultant quantities are observations from t-distributions on 7 or 4 d.f.; that is they are observations that are drawn from moderately or severely leptokurtic distributions. The first N observations in each sample were taken as the control sample and the observation was taken as the individual control case.

On each Monte Carlo trial, Crawford and Howell (1998) test was applied to compare the control case with the control sample and t-values that were negative (i.e., where the control case was below the control sample) and exceeded the one-tailed critical value for t on the appropriate degrees of freedom ($n-1$) were recorded as Type I errors; z was also computed and the result recorded as a Type I error if it exceeded the one-tailed critical value of -1.645. One-tailed tests were employed because, in the vast majority of cases, the (directional) hypothesis tested by neuropsychologists is that their patient's score is below that of controls.

For comparison purposes we also repeated the above procedure but sampled from a *normal* distribution; the results from this latter simulation provide reference values for interpreting the effects of leptokurtosis. Thus, in total, 15 million Monte Carlo trials were run; i.e., 1 million trials for each combination of five sample sizes and three types of distribution.

Table 1
Simulation results: percentage of Type I errors (i.e., percentage of control cases classified as exhibiting a deficit) using z and Crawford and Howell's method for a specified error rate of 5% when sampling from leptokurtic distributions

Control N	Normal distribution		Moderate leptokurtosis		Severe leptokurtosis	
	z	Crawford and Howell	z	Crawford and Howell	z	Crawford and Howell
5	10.36	5.02	10.42	5.42	10.34	5.64
10	7.53	4.98	7.47	5.23	7.42	5.39
20	6.26	5.01	6.10	5.06	5.88	4.97
50	5.48	4.99	5.20	4.80	4.91	4.59
100	5.23	4.98	4.93	4.73	4.53	4.37

2.1.2. Results and discussion

The results of the Monte Carlo simulation are presented in Table 1. The first two columns provide the results when sampling from a normal distribution. It can be seen that, as is predicted by theory and by results from previous simulations (Crawford & Garthwaite, 2005a), the Type I error rate is controlled when Crawford and Howell's test is applied. That is, the specified error rate was set at 5% and the observed error rates cleave closely to this value for all values of N (the small deviations from 5% are of the order expected solely from Monte Carlo variation). In contrast, it can be seen that, when the control sample is small, control of the Type I error rate is poor when z is used to test for a significant difference between a case and controls. For example, the error rate is 10.36% for a N of 5, more than double the specified rate of 5% (with large Ns, z-values more closely approximate t-values so that the error rate is under satisfactory control; however control sample Ns of this magnitude are rare in single-case studies in neuropsychology).

These values for sampling from a normal distribution provide reference values for studying the effects of leptokurtosis on the Type I error rate. Looking first at Crawford and Howell's method, it can be seen that, when the control sample is small (i.e., 10), the presence of leptokurtosis inflates the Type I error rate but only marginally, even when leptokurtosis is severe. At larger Ns the observed Type I error rate falls *below* the specified rate. It can be concluded that, at least when unaccompanied by other departures from normality (i.e., skew), leptokurtosis is not a serious cause for concern when employing Crawford and Howell's method of testing for a deficit. That is, the method is robust even when leptokurtosis is severe.

Turning to z, it can be seen that with one exception (moderate leptokurtosis coupled with a N of 5), the observed Type I error rates are *all lower* than those obtained when sampling from a normal distribution. This is a rather bizarre outcome. In the present scenario there are two problems with the use of z to test for a deficit: (1) z inappropriately treats the sample statistics as parameters, and (2) the use of z for inferential purposes makes the assumption of normality and this assumption is violated. However, it transpires that the presence of leptokurtosis serves to ameliorate the inflation of the Type I error rate caused by the former problem. It should be noted nevertheless that the observed error rate in the presence of leptokurtosis is still above the specified rate of 5% for $N \leq 20$.

As noted, a probability distribution has greater leptokurtosis if it has thicker tails and is more peaked. It follows that, between the tails and the peak, the distribution must be lower (the total area under the distribution must equal one); that is, a leptokurtic distribution has thinner shoulders than a normal distribution (see Fig. 1b). In some of the present scenarios where we (a) compare an individual with a sample rather than compare two groups, (b) employ a one-tailed test and (c) set alpha at 0.05 (rather than at a more conservative value), the effects of the thin shoulders are in evidence; the statistical tests are applied in a region of the distribution that is not sufficiently far out in the tails to produce inflated Type I errors.

In the case of Crawford and Howell's method, the extent to which Type I errors undershoot the specified rate reduces as sample size decreases until, in the case of severe leptokurtosis, the error rates are higher than the specified rate for a N of 10. Thus, it is near to this point that the density of the normal distribution falls below that of the leptokurtic distribution; i.e., we see the effects of the heavy tails reflected in the Type I error rates. It follows from this that, if a more conservative alpha were specified for the Crawford and Howell method, inflation of the Type I error rate would become more pronounced. That is, the effects of leptokurtosis at small Ns would be increased and the crossing point referred to above would occur at a larger N.

This was confirmed by re-running the simulation and testing for significance at the 1% level rather than 5%. For example, for a N of 10, the error rate when sampling from the severely leptokurtic distribution was 1.88%, i.e., 88% higher than the specified rate of 1%, whereas the corresponding figure (5.39%) from the original simulation was only 7.8% higher than the specified rate of 5%. In addition, the error rate for the leptokurtic distribution did not fall below the specified rate even for a N of 100.

3. Study 2

3.1. The effects of combinations of leptokurtosis and skew in the control population on Type I error rates

Given that many neuropsychological instruments do not yield normally distributed data (Capitani & Laiacona, 2000), the results to date from simulation studies are reassuring. That

is, from Crawford and Garthwaite (2005a) examination of the effects of skew and the present examination of the effects of leptokurtosis, it can be concluded that, on their own, neither form of departure from normality seriously compromises inferential methods for testing for a deficit.

However, single-case researchers have to face the possibility that their control data may have *both* these features simultaneously. Indeed, it is likely that control data more commonly possesses both these characteristics rather than either alone. Many neuropsychological tasks (particularly those developed for use in single-case studies) measure abilities that are largely within the competence of many healthy individuals and thus yield ceiling or near-ceiling levels of performance in control samples.

For example, in a review of single-case studies of the living versus non-living distinction in object naming, it was reported that the mean accuracy of naming in the control samples was 95% or greater in the vast majority of these studies (Laws, Gale, Leeson, & Crawford, 2005). Similar ceiling or near-ceiling performance in controls is also a characteristic of most stimulus sets used to test for deficits in the recognition of specific emotions from facial expressions or prosody (Milders, Crawford, Lamb, & Simpson, 2003).

The effects of ceiling, or near-ceiling, performance is that the distribution of control data will be both negatively skewed *and* leptokurtic. In Study 2, we quantify the control over the Type I error rate for z and Crawford and Howell's test when the control data possess both these characteristics. We limit ourselves to examination of negative skew as this will be the more common problem in practice. (Positive skew can often be a feature of control data when performance on a task is expressed as errors rather than number correct. However, the results of the present study will be equally applicable in such a scenario because, if the data were reflected, they would possess an equivalent degree of negative skew.) Although it is obvious that tasks subject to ceiling effects will be negatively skewed, they will also tend to be leptokurtic: if the task is well within the competence of a large proportion of the controls then scores will accumulate at the maximum obtainable score and hence the distribution will be more peaked than a normal distribution.

3.1.1. Method

Simulations were run using a similar approach to that employed in Study 1, i.e., 1,000,000 samples of $N+1$ observations were drawn for five different sample sizes. However, instead of the initial sampling of observations being from a standard (i.e., symmetrical) t-distribution, sampling was from skew-t distributions with varying degrees of negative skew. The method used to sample from these distributions was that of Azzalini and Capitanio (2003); the technical details are presented in Appendix 1.

Four skew distributions were specified ranging from a distribution with moderate skew ($\gamma = -0.31$) through severe (-0.70), very severe (-0.93) to extreme skew (-0.99). Thereafter, the procedure was the same as that followed in Study 1; i.e., each observation was divided by $\sqrt{\chi^2/7}$ (moderate leptokurtosis) or $\sqrt{\chi^2/4}$ (severe leptokurtosis). The resultant distributions are *skew-t* distributions; they depart from a normal distribution in that they are both leptokurtic *and* (negatively) skewed.

Finally, in order to provide a reference for the combined effects of leptokurtosis and positive skew, the above procedure was followed but the observations were sampled from skew-*normal* distributions (Azzalini & Dalla Valle, 1996); i.e., these distributions had the same degree of skewness as the former distributions but were *not* leptokurtic (note that reference results for the opposite scenario, i.e., leptokurtosis in the absence of skew, are already provided in Table 1). This latter procedure also constitutes an attempt to replicate Crawford and Garthwaite (2005a) findings for the effects of skew using a different method of forming the skew distributions (these authors used two-piece normal distributions to model the effects of skew; see Kimber, 1985).

Illustrative, graphical representations of the distributions employed in the present study are presented as in Fig. 2. Note that this figure presents the skew-t distributions together with the equivalent skew-normal distributions; for each pairing, the variances were rescaled so that they had a common variance of 1.

As in Study 1, on each Monte Carlo trial of each scenario, Crawford and Howell (1998) test and z were applied to the score of the $N+1$th control case and the result recorded as a Type I error if it fell in the respective one-tailed critical region. A total of 60 million Monte Carlo trials were performed; i.e., 1 million trials for each combination of five sample sizes, four levels of skew, and three levels of leptokurtosis (absent, moderate, and severe).

3.1.2. Results and discussion

The complete simulation results for z and for Crawford and Howell's test are presented in Table 2. For reference purposes, the first block of rows present the results obtained when the control data were skew but were not also leptokurtic. These former results are for all practical purposes indistinguishable from those obtained by Crawford and Garthwaite (2005a); thus their results on the effects of skew on Type I error rates (obtained using two-piece normal distributions) are replicated using Azzalini and Dalla Valle (1996) form of skew-normal distributions.

The general pattern of results for Crawford and Howell's method is most readily appreciated by referring to Fig. 3. This figure presents the results for moderate or extreme skew coupled with either no leptokurtosis or severe leptokurtosis; results obtained when sampling from a distribution with severe leptokurtosis alone and when sampling from a normal distribution are also presented (using the data obtained in Study 1). Fig. 4 presents the corresponding results for z; it should be noted that a different scale is employed for the ordinate because of the higher error rates for z.

The results for Crawford and Howell's method can be summarized as follows: it can be seen that, in the absence

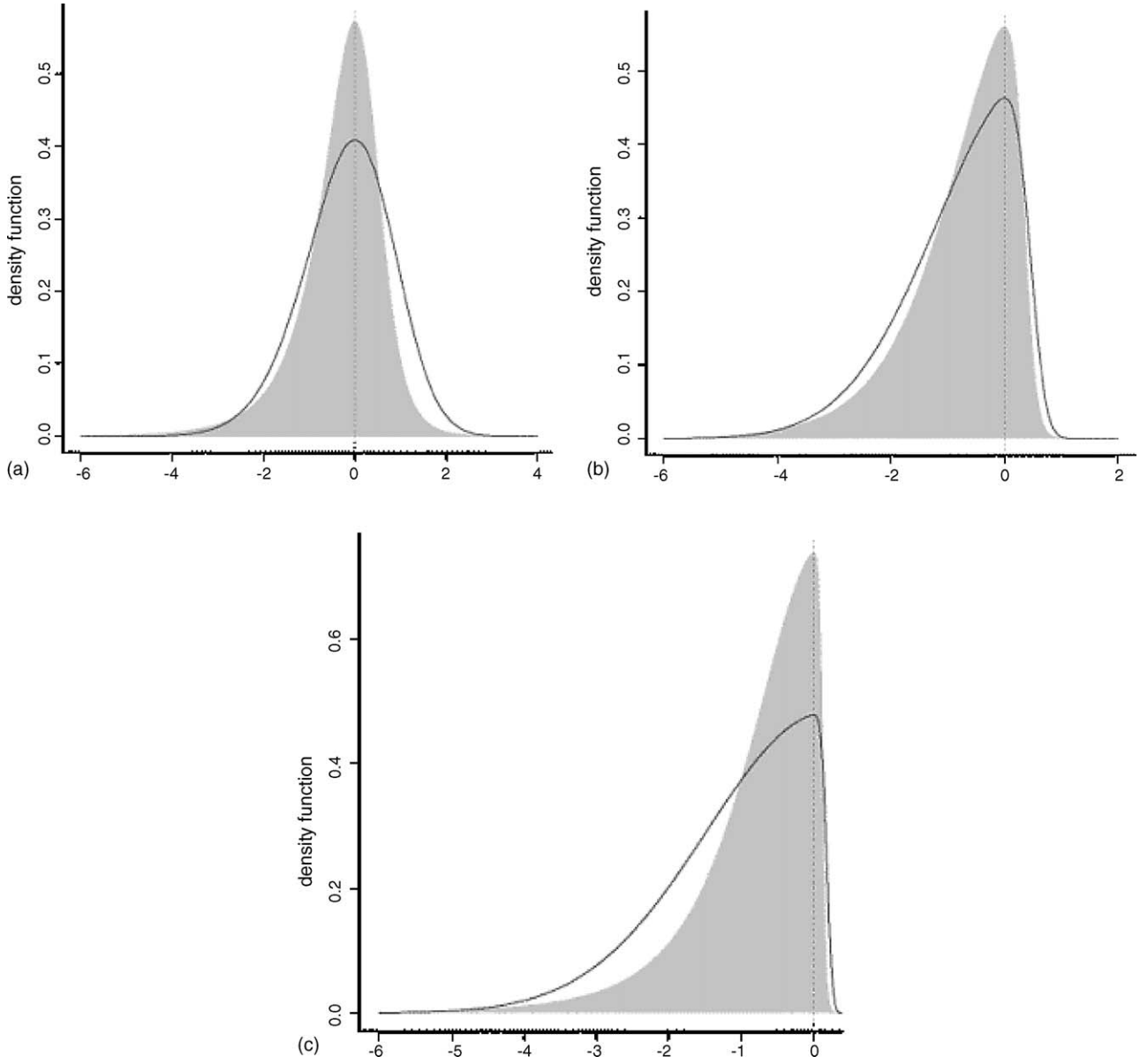

Fig. 2. Graphical illustration of some of the distributions employed in Study 2; the shaded area shows the density for distributions possessing both skew and leptokurtosis (skew-*t*), the unshaded line shows the density for the equivalent distributions with skew alone (skew-normal). (a) Moderate skew/severe leptokurtosis; (b) very severe skew/moderate leptokurtosis; (c) extreme skew/severe leptokurtosis. Note that the skew-*t* and skew-normal distributions have been scaled to have a common variance of 1.

of leptokurtosis, Type I error rates are related to skew by a monotonic increasing function, regardless of the size of the control sample (see Table 2). However, when leptokurtosis is also present, the effects of skew are exaggerated for small-to-moderate *N*s (i.e., $N \leq 20$) but are attenuated for larger *N*. This is most easily seen in Fig. 3 by comparing the results for the combination of extreme skew and severe leptokurtosis with those for extreme skew alone.

Unfortunately, it is the results for small-to-moderate *N*s that are of most relevance to the single-case researcher (i.e., control sample *N*s > 20 are rare in the single-case study literature). Within this range of *N*, it can be seen that, in extreme circumstances, error rates can be close to double the specified rates (the maximum error rate was 9.96%).

Turning now to the results for *z*: it can be seen from Table 2 and Fig. 4 that the effects of combined skew and leptokurtosis are broadly similar to that seen for Crawford and Howell's method; the difference being that, for *z*, these effects are superimposed on the inflating effect of treating the control sample statistics as parameters. As was the case for Crawford and Howell's method, Type I error rates are related to skew by a monotonic increasing function but, when leptokurtosis is also present, error rates are further inflated for small-to-moderate *N*s but the effects are attenuated at larger *N*s.

Table 2
Simulation results: percentage of Type I errors using *z* and Crawford and Howell's method for a specified error rate of 5%; effects of combinations of skewness and kurtosis

Control *N*	Moderate skew		Severe skew		Very severe skew		Extreme skew	
	z	Crawford and Howell	*z*	Crawford and Howell	*z*	Crawford and Howell	*z*	Crawford and Howell
No leptokurtosis								
5	11.28	5.93	12.51	7.27	13.24	8.07	13.37	8.27
10	8.50	5.99	9.59	7.17	10.22	7.80	10.34	7.93
20	7.10	5.88	8.12	6.95	8.74	7.56	8.85	7.70
50	6.31	5.84	7.29	6.84	7.79	7.33	7.97	7.51
100	6.08	5.85	7.01	6.79	7.56	7.33	7.64	7.42
Moderate leptokurtosis								
5	12.06	7.12	13.11	8.31	13.78	9.09	13.97	9.31
10	8.99	7.42	9.98	7.85	10.47	8.37	10.65	8.56
20	7.56	6.98	8.34	7.35	8.70	7.71	8.85	7.86
50	6.54	6.16	7.24	6.87	7.54	7.17	7.64	7.26
100	6.19	6.01	6.86	6.67	7.12	6.95	7.16	6.98
Severe leptokurtosis								
5	12.56	7.93	13.59	9.10	14.18	9.81	14.31	9.96
10	9.37	7.39	10.24	8.35	10.60	8.71	10.74	8.85
20	7.62	6.74	8.32	7.45	8.62	7.78	8.69	7.84
50	6.43	6.11	6.95	6.64	7.22	6.91	7.20	6.89
100	5.92	5.78	6.38	6.24	6.61	6.46	6.68	6.48

It can be seen from Table 2 that, for *z*, the Type I error rate is seriously inflated in many of the scenarios examined and rises as high as 14.31%. That is, on average, up to 14.31% of healthy controls would be incorrectly classified as having a deficit. In these circumstances we have three factors contributing to inflation of the error rate: treatment of the control sample statistics as parameters, skewness, and leptokurtosis. Although this combination of factors is extreme, it is not, however, necessarily *unusual*. That is, ceiling effects in controls are common in single-case studies (Crawford, Garthwaite, & Gray, 2003). Moreover, although Crawford and Howell's method is increasingly used in single-case research (e.g., Bird, Castelli, Malik, Frith, & Husain, 2004; Di Pietro, Laganaro, Leemann, & Schnider,

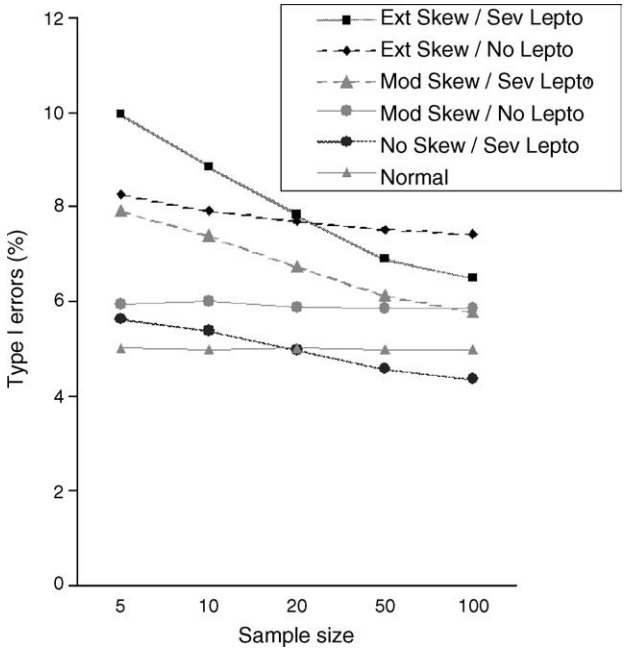

Fig. 3. Effect of departures from normality on Type I errors for Crawford and Howell's method.

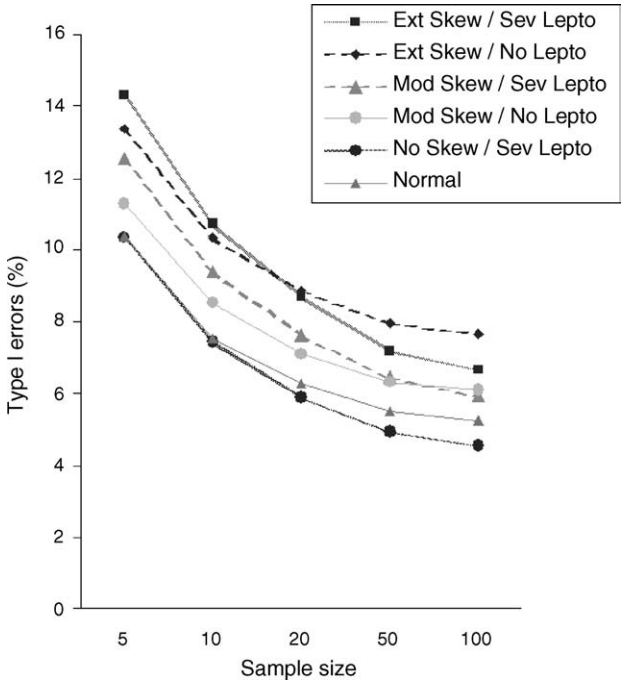

Fig. 4. Effect of departures from normality on Type I errors for *z*.

2004; Rosenbaum, McKinnon, Levine, & Moscovitch, 2004; Temple & Sanfilippo, 2003; Westmacott, Black, Freedman, & Moscovitch, 2004), it remains the case that *z* is still widely used for inferential purposes.

One obvious implication of the results obtained to date is that Crawford and Howell's method should be used in preference to *z*. Not only is the Type I error rate under control when the assumption of normality is met but also the error rates are below those observed for *z* in all scenarios where this assumption was violated. Moreover, the differences are most marked for control sample *N*s that are typical of those used in single-case studies.

The emphasis in the present studies has been on the use of Crawford and Howell's method as a significance test. However, as noted in the Introduction, the *p* value from this test simultaneously provides a point estimate of the abnormality of the patient's score. It follows that, if the distribution in the control population is leptokurtic and/or negatively skewed, there will be a corresponding exaggeration of the point estimate of abnormality.

For example, when sampling from a distribution with severe skew and severe leptokurtosis, the Type I error rate for Crawford and Howell's method was 8.35% for a control sample *N* of 10 (see Table 2). Thus, if application of Crawford and Howell's test to a patient's score yielded a *p* value of exactly 0.05, the corresponding point estimate of the abnormality of the patient's score (i.e., that 5% of the control population would obtain a more extreme score) will exaggerate the rarity of the patient's score; if the control distribution has the specified degree of skew and leptokurtosis, it is to be expected that 8.35% of controls would obtain lower scores. Like Crawford and Howell's method, the complementary method for setting confidence limits on the abnormality of the patient's score developed in Crawford and Garthwaite (2002) also assumes normality. Thus, if the control distribution has marked skew and/or leptokurtosis, the upper confidence limit will be too conservative and the lower confidence limit too liberal. However, the most striking feature of the present results is that the basic method is surprisingly robust; i.e., the Type I error rates and estimates of abnormality are relatively insensitive to departures from normality even in the most extreme of the scenarios modelled in the present study.

It should also be noted that this exaggeration of the abnormality of the patient's score is less than occurs when z is used to provide the point estimate. For example, a *z* of −1.645 provides the point estimate that 5% of controls would obtain lower scores but, in the same circumstances as those discussed above (i.e., severe skew and leptokurtosis and a control *N* of 10), Table 2 shows that 10.34% of controls would be expected to obtain lower scores.

3.2. Strategies for dealing with non-normal control data

Although, as noted above, even in the most extreme of the scenarios studied, the effects of skew and leptokurtosis on Crawford and Howell's method are by no means catastrophic (i.e., the test is more robust than might have been predicted), it remains the case that Type I error rate is not under control in these circumstances. Therefore, it is appropriate to consider strategies to deal with this problem.

One potential solution would be to abandon Crawford and Howell's parametric test in favour of non-parametric alternatives; e.g., conventional randomisation tests or computer intensive resampling methods (see Howell, 2002 for an introduction to these latter methods). However, as noted by Crawford and Garthwaite (2005a) there are two serious limitations to this strategy.

First, non-parametric methods are necessarily completely insensitive to the degree to which a patient's score is extreme; therefore they will tend to have low power (e.g., a patient whose score on a task was 5S.D.s below the control mean would be treated identically to a patient whose score was 2S.D.s below the mean if their rank order relative to controls was the same). Power will typically be low in single-case studies because an individual rather than a sample is compared to a control sample that is itself typically modest in size; therefore any treatment that imposes a further reduction in power should be avoided if at all possible (Crawford & Garthwaite, in press; Crawford et al., 2003).

Second, the size of sample required before a researcher has any possibility of rejecting the null hypothesis of no difference between patient and controls is larger than is typically available in single-case studies. A minimum of 20 controls would be required to be able to reject the null hypothesis even when the alternative hypothesis is directional ($p < 0.05$, one-tailed) and such an outcome would only occur if the score of every control was higher than the patient's.

An anonymous reviewer took issue with this conclusion and suggested that five controls would be sufficient to reject the null hypothesis (one-tailed) that the patient's score came from the same distribution as those of the controls. The argument is that the probability is 0.5 that a randomly drawn control would obtain a higher score than the patient and thus, using the multiplicative probability rule for independent events, the probability is $0.5^5 = 0.03125$ (i.e., $p < 0.05$) that all five controls would obtain higher scores. Unfortunately, this gives the wrong answer as the independence assumption does not hold; whether the patient's score is less than the score of one control is *not independent* of whether the patient's score is less than the scores of other controls. To obtain the correct probability, note than under the null hypothesis the patient's score comes from the same distribution as the scores of the controls. If there are six scores each randomly drawn from the same distribution, any one of which may be the patient's, then 1 in 6 (0.1666) is the probability that the lowest score is the patient's (i.e., the Type I error rate would be 16.66%). Note also that this result will hold regardless of the shape of the control distribution.

As the reviewer's argument has been offered by other researchers in informal discussions, a small, additional, simulation was run to provide an empirical demonstration of the

Table 3
Simulation results: Percentage of Type I errors for a non-parametric (multiplicative probability) method of testing for a deficit

Distribution	Percentage of Type I errors
Normal	16.64
Extreme skew	16.68
Severe leptokurtosis	16.71
Extreme skew and severe leptokurtosis	16.65

high Type I error rate. Five controls plus an additional control case were drawn from a normal distribution, a distribution with extreme skew, a distribution with severe leptokurtosis, and a distribution that possessed both extreme skewness *and* severe leptokurtosis (these labels correspond to the distributions used in the earlier simulations). One million trials were run for each distribution. The number of control cases whose scores fell below all five members of the control sample was recorded. The results of this simulation are presented in Table 3, from which it can be seen that the empirical error rates match the theoretical error rates, regardless of the shape of the control distribution (i.e., all error rates are very close to 16.66%).

Before leaving the topic of non-parametric methods, it is worth noting that, although computer intensive resampling approaches are useful when comparing two modestly sized samples, they are no more helpful than traditional non-parametric approaches in the present context. These methods work by repeatedly shuffling cases between the two samples and quantifying how unusual the difference observed for the original samples is relative to the differences obtained between the randomly created alternative samples. Even with small numbers in each sample, enough unique samples can be formed to make this viable. However, in the present context, where a single case is compared to a sample, the maximum number of unique alternative samples is equal to the *N* of the control sample.

Another potential solution would be to transform the scores of controls and the patient in an attempt to normalise the control score distribution (for a brief introduction to transformations see Howell, 2002). For example, to deal with moderate negative skew the scores could be reflected and a logarithmic transformation applied. A more flexible alternative would be to seek the optimal Box–Cox (Box & Cox, 1964) normalising power transformation. A problem here is that, with the small samples typically employed in single-case studies, there are little data with which to assess the true form of the underlying control distribution and thereby select the appropriate normalising transformation.[1] Furthermore, the granularity of scores encountered in single-case studies poses a further problem. That is, there may be a limited number of possible scores; no transformation will ever adequately normalise such data.

In summary, we recommend that researchers attempt to find a normalising transformation but recognise that this may not be possible in practice. Note that, although the search for an appropriate normalising transformation should be conducted using the control data alone, if a reasonable transformation is found, this should then be applied to the data of the controls *and* the patient before running Crawford and Howell's test.

In view of the actual or potential difficulties with alternative approaches, it may be often more practical to rely on Crawford and Howell's method even when there are concerns over departures from normality: a researcher can still have a high degree of confidence that the patient's score did not come from the control distribution if the test result for Crawford and Howell's method is highly significant. That is, even with the combination of extreme skew and severe leptokurtosis, the observed error rate for a specified rate of 5% never rose above 10%; thus *t*-values that are markedly larger than the critical value would be sufficient to warrant rejection of the null hypothesis that the patient's score is an observation from the control population.

To study this suggestion formally, we partially re-ran the simulation but substituted the critical value of *t* required for significance at the 2.0% level (one-tailed) rather than 5%. This was only done for Crawford and Howell's method because of its demonstrated superiority over the use of *z* in both this study and Study 1. As can be seen from Table 4, the observed Type I error rate was below 5% for the vast majority of combinations of *N*, degree of skew and degree of leptokurtosis. These results indicate that, if the *p* value obtained from

Table 4
Simulation results: percentage of Type I errors using Crawford and Howell's method with a more stringent specified error rate of 2%

N	Moderate skew	Severe skew	Very severe skew	Extreme skew
No leptokurtosis				
5	2.61	3.61	4.32	4.52
10	2.73	3.76	4.34	4.48
20	2.80	3.74	4.24	4.34
50	2.79	3.72	4.13	4.27
100	2.76	3.64	4.09	4.20
Moderate leptokurtosis				
5	3.67	4.70	5.42	5.60
10	3.85	4.79	5.29	5.46
20	3.78	4.55	4.95	5.05
50	3.59	4.28	4.54	4.63
100	3.47	4.12	4.37	4.42
Severe leptokurtosis				
5	4.55	5.59	6.25	6.40
10	4.59	5.52	5.88	6.02
20	4.31	5.00	5.31	5.39
50	3.91	4.45	4.70	4.70
100	3.69	4.14	4.36	4.37

[1] The public domain software package R (www.cran.r-project.org) provides a very useful routine that finds the optimal Box–Cox normalising transformation by the method of maximum likelihood. As noted, however, although it will find the best available transformation, this by no means ensures that the data will be normalised successfully.

Crawford and Howell's test is below 0.02, then a researcher could be 95% confident that the patient's score did not come from the control population except in the most extreme of circumstances (i.e., very small *N*, coupled with very marked skew and leptokurtosis). Even in these latter circumstances the error rate is only marginally above 5% (i.e., the maximum error rate is 6.4%).

Crawford and Garthwaite (2002) developed a computer program (singlims.exe) to implement Crawford and Howell's method and made it freely available over the internet.[2] This program provides a *precise p* value (one-tailed) for the difference between the case and control sample. It is therefore straightforward for researchers to incorporate the suggestions made above into their practice. That is, if there is concern over the control distribution, then the results can be examined to determine not only whether it exceeds the conventional 0.05 level but also whether it exceeds the 0.02 level. If the latter holds then the user can have a high degree of confidence that the result is not an artefact of non-normality. A result that does not meet this second criterion should not be dismissed but should be treated with more caution, particularly if the control distribution appears to be markedly non-normal.

Finally, it should also be noted that the strategies of applying a transformation to the data and using a more conservative *p* value are not mutually exclusive. That is, a transformation may fall short of normalising the data but nevertheless be an improvement on the distribution of raw scores. If when Crawford and Howell's test is applied to the transformed data and *p* is <0.02 the researcher can be particularly confident that the result is not an artefact.

4. General discussion

Very useful and elegant methods have been devised for drawing inferences concerning an individual patient's performance on fully standardized neuropsychological tests; i.e., on tests that have been normed on very large, representative samples of the population (e.g., Capitani & Laiacona, 2000; De Renzi, Faglioni, Grossi, & Nicheli, 1997; Willmes, 1985). When these methods are used in single-case research, the patient is compared against normative values rather than against controls. In such approaches, error arising from sampling from the control population is ignored; this is entirely justified because the samples are large enough for such error to be minimal. The methods developed by Capitani and colleagues have the added advantage that they are non-parametric and are therefore immune to the problems that are the focus of the present study.

Although these latter approaches have much to commend them, there are many circumstances in which they cannot be used because (a) the questions posed in many single-case studies cannot be fully addressed using existing standardized neuropsychological tests, (b) new constructs are constantly emerging in neuropsychology, and (c) the collection of large-scale normative data is a time-consuming and arduous process (Crawford, 2004). Therefore, there is a continued need for methods that can be used when a patient is compared to a modestly sized control sample rather than a large normative sample.

The single-case approach in neuropsychology has made a significant contribution to our understanding of the functional architecture of human cognition. However, as Caramazza and McCloskey (1988) note, if advances in theory are to be sustainable they "... must be based on unimpeachable methodological foundations" (p. 619). The statistical treatment of single-case study data is one area of methodology that has been relatively neglected.

Given that most single-case studies employ control samples with modest *N*s, and that skew and leptokurtosis are likely to be common features of the control data in such studies, the present study has gone some way to tackling what are fundamental issues in the statistical treatment of single-case studies. From one perspective the results are reassuring: the effects of departures from normality, even when severe, did not produce a drastic inflation of the Type I error rate (i.e., Crawford and Howell's method is more robust than was expected).

Nevertheless, the Type I error rates were raised above the specified rates, to the extent that they were close to double the specified rate for Crawford and Howell's method in the most extreme scenarios examined (even higher rates were observed for *z* but, as noted, most of this effect can be attributed to the treatment of control sample statistics as parameters). Of the potential approaches to dealing with this problem we suggest that our proposed solution is both practical and rigorous; i.e., it provides single-case researchers (and their peers) with sufficient reassurance that any deficits they record are not artefacts of non-normal data.

The emphasis throughout this paper has been on single-case research in which a patient is compared to a control sample. However, it is worth noting that the sample need not be healthy controls. For example, it may be that a researcher or clinician wants to compare a patient against a sample of other patients with whom the patient shares some characteristics. Similarly, and as noted in Section 1, the methods are useful in comparing a patient to small-scale normative data (e.g., "provisional" norms collected for a new instrument or locally collected norms for existing instruments). In both these other applications there would also be reason to be concerned with departures from normality (for example, with provisional data, the user cannot assume that there has been the careful item selection and/or application of normalising transformations that one would expect with fully standardized tests). Thus, the present results also provide reassurance in these circumstances (i.e., the methods are relatively robust) and, just as was the case for the typical single-case study, they also provide a remedy if there is concern over departures from

[2] This program can be downloaded from the following website address: http://www.abdn.ac.uk/~psy086/dept/abnolims.htm.

normality (note that, with normative data from a third party, the alternative strategy of attempting to normalise the data is not usually an option).

Finally, Monte Carlo methods offer a means of examining methodological issues in single-case research that would be difficult or impossible to address by other means; it is to be hoped that the present study will encourage further use of such methods.

Appendix A. Sampling from skew-normal and skew-*t* distributions

The method used to sample from skew-normal distributions is based on work by Azzalini and colleagues (Azzalini & Capitanio, 1999; Azzalini & Dalla Valle, 1996). The starting point for this method is the generation of two independent standard normal variates u_0 and u_1 (u_1 is used to form the X observations and u_0 is used to control the degree of skew in X). Then u_2 is determined from the formula:

$$u_2 = \rho_{u_0u_1} + \sqrt{1 - \rho^2_{u_0u_1}}u_1. \tag{A.1}$$

The value of $\rho_{u_0u_1}$ required to introduce the desired degree of skew (γ_1) can be obtained by algebraic manipulation of Azzalini and Dalla Valle (1996) formulae for γ_1 to solve for $\rho_{u_0u_1}$. That is, put

$$a = \left(\frac{2\gamma_1}{4-\pi}\right)^{1/3} \tag{A.2}$$

and

$$\rho_{u_0u_1} = a\left(\frac{\pi}{2+2a^2}\right)^{1/2}. \tag{A.3}$$

Then

$$x = \begin{cases} u_2 & \text{if } u_0 \geq 0 \\ -u_2 & \text{otherwise} \end{cases} \tag{A.4}$$

is an observation from the skew-normal distribution with skewness γ_1. To sample from the equivalent skew-*t* distribution the above steps are followed by dividing x by $\sqrt{\chi^2/\nu}$, where χ^2 is a random draw from a Chi-square distribution on ν degrees-of-freedom (e.g., $\nu = 4$ if severe leptokurtosis is required).

References

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society Series B*, *61*, 579–602.

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-*t* distribution. *Journal of the Royal Statistical Society Series B*, *65*, 367–389.

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*, 715–726.

Bird, C. M., Castelli, F., Malik, O., Frith, U., & Husain, M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. *Brain*, *127*, 914–928.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B*, *26*, 211–246.

Burt, C. (1963). Is intelligence distributed normally? *British Journal of Statistical Psychology*, *16*, 175–190.

Capitani, E. (1997). Normative data and neuropsychological assessment. Common problems in clinical practice and research. *Neuropsychological Rehabilitation*, *7*, 295–309.

Capitani, E., & Laiacona, M. (2000). Classification and modelling in neuropsychology: From groups to single cases. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology: vol. 1* (2nd ed., pp. 53–76). Amsterdam: Elsevier.

Caramazza, A., & McCloskey, M. (1988). The case for single-patient studies. *Cognitive Neuropsychology*, *5*, 517–528.

Crawford, J. R. (2004). Psychometric foundations of neuropsychological assessment. In L. H. Goldstein & J. E. McNeil (Eds.), *Clinical neuropsychology: A practical guide to assessment and management for clinicians* (pp. 121–140). Chichester: Wiley.

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

Crawford, J. R. & Garthwaite, P. H. (in press). Evaluation of criteria for classical dissociations in single-case studies by Monte Carlo simulation. *Neuropsychology*.

Crawford, J. R., & Garthwaite, P. H. (2005a). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology*, *19*, 318–331.

Crawford, J. R. & Garthwaite, P. H. (2005b). Methods of testing for a deficit in single case studies: Evaluation of statistical power by Monte Carlo simulation, submitted for publication.

Crawford, J. R., Garthwaite, P. H., & Gray, C. D. (2003). Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex*, *39*, 357–370.

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

De Renzi, E., Faglioni, P., Grossi, D., & Nicheli, P. (1997). Apperceptive and associative forms of prosopagnosia. *Cortex*, *27*, 213–221.

Di Pietro, M., Laganaro, M., Leemann, B., & Schnider, A. (2004). Receptive amusia: temporal auditory processing deficit in a professional musician following a left temporo-parietal lesion. *Neuropsychologia*, *42*, 868–877.

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

Kimber, A. C. (1985). Methods for the two-piece normal distribution. *Communications in Statistics: Theory and Methods*, *14*, 235–245.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modelling using the *t*-distribution. *Journal of the American Statistical Association*, *84*, 881–896.

Laws, K. R., Gale, T. M., Leeson, V. C., & Crawford, J. R. (2005). When is category specific in dementia of Alzheimer's type? *Cortex*, *41*, 452–463.

Milders, M., Crawford, J. R., Lamb, A., & Simpson, S. A. (2003). Differential deficits in expression recognition in gene-carriers and patients with Huntington's disease. *Neuropsychologia*, *41*, 1484–1492.

Rosenbaum, R. S., McKinnon, M. C., Levine, B., & Moscovitch, M. (2004). Visual imagery deficits, impaired strategic retrieval, or memory loss: disentangling the nature of an amnesic person's autobiographical memory deficit. *Neuropsychologia*, *42*, 1619–1635.

Schindler, I., Rice, N. J., McIntosh, R. D., Rossetti, Y., Vighetto, A., & Milner, A. D. (2004). Automatic avoidance of obstacles is a dorsal stream function: Evidence from optic ataxia. *Nature Neuroscience*, *7*, 779–784.

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

Temple, C. A., & Sanfilippo, P. M. (2003). Executive skills in Klinefelter's syndrome. *Neuropsychologia*, *41*, 1547–1559.

Westmacott, R., Black, S. E., Freedman, M., & Moscovitch, M. (2004). The contribution of autobiographical significance to semantic memory: evidence from Alzheimer's disease, semantic dementia, and amnesia. *Neuropsychologia*, *42*, 25–48.

Willmes, K. (1985). 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*, *7*, 331–352.