Neuropsychology
2005, Vol. 19, No. 3, 318–331

0894-4105/05/$12.00 DOI: 10.1037/0894-4105.19.3.318

Testing for Suspected Impairments and Dissociations in Single-Case Studies in Neuropsychology: Evaluation of Alternatives Using Monte Carlo Simulations and Revised Tests for Dissociations

John R. Crawford
University of Aberdeen

Paul H. Garthwaite
The Open University

In neuropsychological single-case studies, a patient is compared with a small control sample. Methods of testing for a deficit on Task X, or a significant difference between Tasks X and Y, either treat the control sample statistics as parameters (using z and z_D) or use modified t tests. Monte Carlo simulations demonstrated that if z is used to test for a deficit, the Type I error rate is high for small control samples, whereas control of the error rate is essentially perfect for a modified t test. Simulations on tests for differences revealed that error rates were very high for z_D. A new method of testing for a difference (the revised standardized difference test) achieved good control of the error rate, even with very small sample sizes. A computer program that implements this new test (and applies criteria to test for classical and strong dissociations) is made available.

Keywords: single-case studies, statistical methods, dissociations

In many single-case studies in neuropsychology, the performance of a patient on a series of tasks is compared with that of a control sample. By far the most common method of forming inferences about the presence of a possible impairment in such studies is to convert the patient's score on a given task to a z score based on the mean and standard deviation of the controls and then refer this score to a table of the areas under the normal curve. 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 ($p < .05$) and would be taken as an indication that the patient had an impairment on the task in question.

One problem with this approach is that it treats the control sample as if it were a population; that is, the mean and standard deviation are used as if they were parameters rather than sample statistics. In other areas of psychology, this is often not a problem in practice because the normative or control sample is large and, therefore, should provide sufficiently accurate estimates of the parameters. However, the control samples in single-case studies in cognitive neuropsychology typically are modest: $N < 10$ is not unusual, and Ns < 20 are very common (Crawford & Howell, 1998). With samples of this size, it is not appropriate to treat the mean and standard deviation as though they were parameters.

A solution to this problem is to use a method described by Crawford and Howell (1998) that treats the control sample statistics as sample statistics. Their approach uses a formula for a modified t test given by Sokal and Rohlf (1995). This method uses the t distribution (with $n - 1$ df), rather than the standard normal distribution, to estimate the abnormality of the patient's score and to test whether it is significantly lower than the scores of the control sample. The practical effect of using z with a small control sample is to exaggerate the rarity and/or abnormality of a patient's score and to inflate the Type I error rate (in this context, a Type I error occurs when an individual who is drawn from the control population is incorrectly classified as not being a member of this population; i.e., he or she is incorrectly classified as exhibiting an impairment). This occurs because the normal distribution has thinner tails than do t distributions. Intuitively, the less that is known, the less extreme should be statements about abnormality and/or rarity. The z-score method treats the variance of controls as being known, when it is not, and consequently makes statements that are too extreme (Crawford & Howell, 1998). The formula for Crawford and Howell's test is

$$t = \frac{X^* - \overline{X}}{S\sqrt{\frac{n+1}{n}}}, \tag{1}$$

where X^* is the patient's score, $\overline{X}$ and S are the mean and standard deviation, respectively, 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 provides a point estimate of the abnormality of the patient's score; for example, if the one-tailed p is .013, then one knows that the patient's score is significantly ($p < .05$) below the control mean and that it is estimated that 1.3% of the control population would obtain a score lower than the patient's. As Crawford and Howell (1998) noted, this point estimate of abnormality is a useful complement to the significance test given that the use of an alpha of .05 is essentially an arbitrary convention (albeit one that has, in general, served science well).

John R. Crawford, School of Psychology, University of Aberdeen, Aberdeen, Scotland; Paul H. Garthwaite, Department of Statistics, The Open University, Milton Keynes, England.

Correspondence concerning this article should be addressed to John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 2UB, Scotland. E-mail: j.crawford@abdn.ac.uk

Study 1: Tests Aimed at Detecting an Impairment When a Case Is Compared With a Control Sample

In the first study, we ran a Monte Carlo simulation to quantify and compare control of the Type I error rate when the two alternative methods of detecting an impairment are used to compare individual control cases against control samples. The statistically sophisticated reader may consider that running this simulation is unnecessary because theory would predict that the use of z will fail to control Type I errors, whereas the modified t test will achieve adequate control. However, we had two reasons for conducting it. First, the use of z to detect an impairment in a patient is very widespread (Crawford & Garthwaite, 2002; Crawford, Garthwaite, & Gray, 2003), so clearly, many researchers are either unaware or have chosen to ignore the issue of inflated Type I errors. Quantifying the magnitude of this inflation may help to raise awareness of the problem, and doing so using an empirical method may be more convincing than appeal to theory alone.

Second, all readers will be familiar with the use of independent-sample t tests to test for a difference in population means in which two samples are compared. In this standard situation, variance estimates are obtained from two samples that are then pooled (or alternatively, separate variance estimates are used when the variances differ). However, many readers will not be familiar with the modified t test in which the concern need only (and can only) be with the variance estimate of the control population. Under the null hypothesis, the patient is an observation from a distribution with the same mean and variance as the controls. Because, unlike a standard t test, the patient does not contribute to a pooled variance estimate (nor contribute a separate variance estimate), readers may appreciate reassurance that control of Type I errors is adequate in this nonstandard use of a t test.

Method

The Monte Carlo simulation was run on a PC and implemented in Borland Delphi (Version 4). The algorithm ran3.pas (Press, Flannery, Teukolsky, & Vetterling, 1989) was used to generate uniform random numbers (between zero and one), which were transformed by the polar variant of the Box–Muller method (Box & Muller, 1958) to sample from a normal distribution. The simulation was run with five different control sample sizes: For each of these values, 1,000,000 samples of $N + 1$ observations were drawn from a normal distribution. The first N observations in each sample were taken as the control sample, and the $N + 1$th item was taken as the individual control case. Crawford and Howell's (1998) test was then applied to these data, and t values that were negative (i.e., when 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 used because, in the vast majority of cases, the (directional) hypothesis tested by neuropsychologists is that their patient's score is below that of controls.

Results and Discussion

The results of the Monte Carlo simulation are presented in Table 1. It can be seen from Table 1 that, when the size of 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 10.37% with a sample size of 5, more than double the specified rate of 5%. Therefore, if z is used in a single-case study with a control sample of 5, it is to be expected that more than 10% of individuals from the control population would be incorrectly identified as not having come from this population (i.e., they would be considered to exhibit an impairment). With large sample sizes, z values more closely approximate t values so that the error rate is under satisfactory control. However, it will be appreciated that control sample sizes of this magnitude are rare in single-case studies in neuropsychology.

Table 1
Results From a Monte Carlo Simulation Study of the Percentage of Control Cases Classified as Exhibiting a Deficit (i.e., Percentage of Type I Errors) Using z and a Modified t Test When the Specified Error Rate Is 5%

	Percentage of Type I errors		
Control sample N	z	t	z required[a]
5	10.37	5.01	−2.335
10	7.57	5.00	−1.923
20	6.25	5.00	−1.772
50	5.53	5.03	−1.693
100	5.28	4.98	−1.669

[a] Records the value of z required to maintain the Type I error rate at the specified (5%) level.

In contrast to the inflated error rates when z is used, it can be seen that there is immaculate control of the Type I error rate when the modified t test is used; the error rates for all of the sample sizes examined are all at, or very close to, the specified rate of 5% (the magnitude of the differences from 5% is of the order expected solely from Monte Carlo variation). Having verified empirically that the Type I error rate is controlled when the modified t test is used, we can use the fact that the z score satisfies

$$z = t\sqrt{\frac{n+1}{n}}, \tag{2}$$

to record the actual value of z that would be required to maintain the Type I error rate at 5%. These values of z are presented in the final column of Table 1. It can be seen that the values of z required to maintain the Type I error rate at the specified level are markedly greater than the nominal critical value of -1.645; for example, with a control sample size of 5, a z of -2.335 would be required. This example also highlights the extent to which z will tend to provide an exaggerated estimate of the rarity of a patient's score. Suppose a patient obtained a z score of -2.335; using a table of the areas under the normal curve, we would estimate that 0.98% of the control population would obtain a lower score (i.e., the patient's score is estimated to be very rare), yet the unbiased estimate provided by t is that 5% of the population would be expected to obtain a lower score.

Study 2: The Effects of Skew in the Control Population on Type I Error Rates

An assumption underlying the use of z or Crawford and Howell's (1998) test is that the controls have been drawn from a normal

distribution. However, it is not uncommon for the scores of controls on neuropsychological tests to depart from normality. In Study 2 we examined the effects on the Type I error rate of violating the assumption of normality. For a number of reasons, the focus of this study was on the effects of negative skew. One reason for concentrating on skewness is that, a priori, skew is liable to have a greater effect on Type I errors than other forms of departure from a normal distribution. This is because low-order moments are important (the first-order moment is the mean, and the second-order moment is the variance), and skewness is the lowest order moment that does not correspond to a parameter of a normal distribution. Also, empirical studies of error rates for independent-sample *t* tests confirm that skew is the most important parameter (e.g., Boneau, 1960; see Howell, 2002, for a brief review). In addition, it has been shown that the effect of skew is particularly pronounced when combined with large imbalances in sample sizes; as the present methods involve comparing an individual with a sample, this underlines the need to study this issue.

We focused on negative skew because it is clear that, in practice, the scores of controls are often negatively skewed. As Crawford, Garthwaite, and Gray (2003) noted, "*z* has been used for inferential purposes in numerous single-case studies when it is obvious from the means and *SD*s of their control samples that the data are highly negatively skewed (i.e., the *SD* tells us that, were the data normally distributed, a substantial percentage of scores should lie above the maximum obtainable score yet we know that none did)" (p. 367).

Negative skew is common in control data because the tasks used often measure abilities that are largely within the competence of most healthy individuals and thus yield ceiling or near-ceiling levels of performance. For example, in a review of single-case studies of the living versus nonliving distinction in object naming, it was reported that the accuracy of naming in controls was 95% or greater in the vast majority of these studies (Laws, Gale, Leeson, & Crawford, in press).

It will be appreciated that when (as is less common) performance is expressed as the number of errors on a task, then the opposite situation to that described above will often occur; that is, the distribution of scores for controls will be positively skewed. However, by reflecting scores, a positively skewed distribution can be converted into an equivalently negatively skewed distribution, so results obtained in the present study are equally applicable to scenarios in which the control scores are positively skewed.

Method

We ran simulations using an approach identical to that of Study 1 except that instead of sampling from a normal distribution, we sampled observations from distributions with varying degrees of negative skew. We achieved this by using two-piece normal distributions (Gibbons & Mylroie, 1973; Kimber, 1985); these distributions have also been termed *joined half-Gaussian* or *binormal distributions*. In comparison to alternative methods of modeling the effects of skewness, two-piece normal distributions have been shown to possess a number of desirable properties, including their suitability when there is a requirement (as in the present study) to draw small samples (Garvin & McClean, 1997).

Skewness is most commonly quantified using the statistic γ_1; this statistic is obtained by dividing the third central moment of a distribution by the cube of its standard deviation. We formed four distributions that varied in their degree of skew, ranging from moderate ($\gamma_1 = -0.31$) to severe ($\gamma_1 = -0.70$), very severe ($\gamma_1 = -0.93$), and extreme ($\gamma_1 = -0.99$). These distributions were obtained by setting the standard deviation of the normal distribution used to form the right-hand side of the two-piece distribution to 1.0 and the standard deviation of the normal distribution used for the left-hand side to values of 1.5, 3.0, 10.0, and 100.0, respectively. The resultant two-piece distributions are illustrated in Figure 1. It can be seen that the degrees of skew for the latter two of these distributions are exceptionally large, to the extent that the gross appearance of the distribution with extreme skew is that of a normal distribution in which all of the right-hand side is absent.

One million samples of $N + 1$ pairs of observations were drawn from each of the four skew distributions. As in Study 1, this was done for five sample sizes: 5, 10, 20, 50, and 100. Also as in Study 1, Crawford and Howell's (1998) test and *z* were applied to the score of the $N + 1$th control case, and the result was recorded as a Type I error if it exceeded the respective one-tailed critical value.

Results and Discussion

The simulation results for *z* and for Crawford and Howell's (1998) test are presented in Table 2. It can be seen that when *z* is used to test for an impairment, the control of the Type I error rate is poor; the error rates range from a low of 6.20% ($N = 100$ combined with moderate skew) to a high of 13.39% ($N = 5$ combined with extreme skew). However, at the small control sample sizes commonly used in single-case studies, the poor control of the error rate mainly stems from the treatment of the control sample statistics as population parameters. That is, although the presence of skewness has further inflated the error rate over that observed in Study 1, the increment is relatively modest. For example, when sampling from a normal distribution with a sample size of 5, the error rate was over twice the nominal rate of 5% (10.37%), and even the presence of extreme skew raised this error rate to only 13.40%.

As noted, the effect of skew on *t* tests has been examined in previous simulation studies. However, in all of these prior studies, the focus was on the use of *t* to test for a difference in population means. In contrast, Crawford and Howell's (1998) procedure tests the hypothesis that an individual patient did not come from a population of controls; under the null hypothesis, the individual is an observation from a distribution with the same mean and variance as the controls (Crawford, Garthwaite, Howell, & Gray, 2004). The effects of skew have therefore not been previously investigated in this nonstandard application of a *t* test.

It can be seen from Table 2 that there is modest inflation of the Type I error rate using Crawford and Howell's (1998) test when skew is moderate or not very severe (e.g., for a control sample size of 10, the error rates are 6.04% for moderate and 7.14% for severe skew). The Type I error rate rises as high as 8.27% when skew is extreme. However, it is clear that the rate of increase in Type I error rate becomes attenuated as the distributions become more severely skewed; the increase in the degree of skew from very severe to extreme is substantial, and yet the concomitant increase in the Type I error rate is modest (e.g., with a control sample size of 10, the error rate is 7.80% for very severe skew and 7.94% for extreme skew). We reiterate that the degree of skew in these latter two distributions is exceptionally large.

Although the inflation of the Type I error rate is not very acute, even with very severe skew (i.e., the test is more robust than might have been predicted), it remains the case that the error rate is not under control. Therefore, some guidance should be offered for researchers when it is suspected that control data are skew.

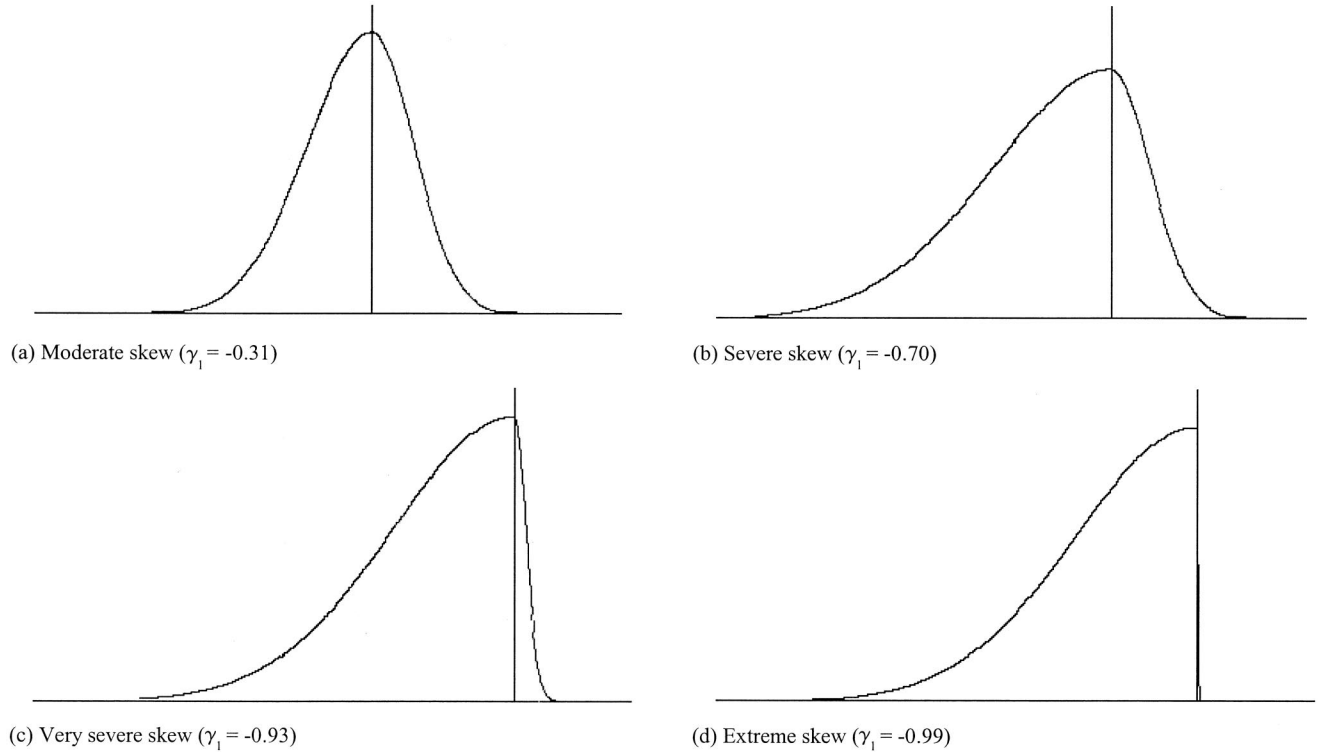

Figure 1. Graphical illustration of the four negatively skewed distributions used in Study 2.

One potential alternative to Crawford and Howell's (1998) parametric test would be to use nonparametric tests (e.g., randomization tests). However, there are two limitations to this potential solution. First, these methods are, by necessity, completely insensitive to the degree to which a patient's score is extreme and therefore will have low power (e.g., a patient whose score on a task was eight standard deviations below the control mean would be treated identically to a patient whose score was two standard deviations below the mean, provided that their rank order relative to controls was the same). Power is inevitably low in single-case studies because an individual rather than a sample is compared with 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, & Gray, 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 typical in single-case studies. A minimum of 20 controls would be required to reject the null hypothesis even when the alternative hypothesis is directional ($p < .05$, one-tailed), and such an outcome would occur only if the score of every control was higher than the patient's.

When the control data are skew, one possibility would be to transform the scores of controls and the patient in an attempt to normalize the control score distribution. For example, in the case of moderate negative skew, the scores could be reflected and a

Table 2
Simulation Results: Percentage of Type I Errors (i.e., Percentage of Control Cases Classified as Exhibiting a Deficit) Using z and a Modified t Test for a Specified Error Rate of 5% When Sampling From (Negatively) Skewed Distributions

	Skew							
	Moderate ($\gamma_1 = -0.31$)		Severe ($\gamma_1 = -0.70$)		Very severe ($\gamma_1 = -0.93$)		Extreme ($\gamma_1 = -0.99$)	
N	*z*	*t*	*z*	*t*	*z*	*t*	*z*	*t*
5	11.48	6.06	12.50	7.23	13.23	8.04	13.39	8.27
10	8.59	6.04	9.64	7.14	10.23	7.80	10.23	7.94
20	7.23	5.97	8.20	6.97	8.72	7.53	8.72	7.66
50	6.50	6.00	7.37	6.90	7.85	7.37	7.85	7.47
100	6.20	5.97	7.11	6.87	7.56	7.32	7.56	7.32

logarithmic transformation applied (for further guidance, see Howell, 2002). Alternatively, however, even if Crawford and Howell's (1998) test is applied to untransformed scores, the researcher can still have a high degree of confidence that the patient's score did not come from the control distribution if the result is highly significant. That is, even with very severe skew, the observed error rate for a specified rate of 5% never rose above 8.27%; thus, t values that are markedly larger than the critical value would be sufficient to warrant rejection of the null hypothesis. To study this suggestion more formally, we reran the simulation (for Crawford & Howell's, 1998, test alone given its demonstrated superiority over the use of z in both this study and Study 1) but substituted the critical value of t required for significance at the 2.5% level (one-tailed) rather than at 5%. The Type I error rate was below 5% for all sample sizes at all levels of γ_1, with the exception of sample sizes of 5 and 10 when skew was extreme (i.e., $\gamma_1 = -0.99$). Even in these two latter cases, the error rates (5.21% and 5.18%, respectively) were only marginally above 5%. For the other sample sizes and γ_1 examined, the error rate ranged from 3.23% to 4.99%. These results suggest that if the p value obtained from Crawford and Howell's test is below .025, then a researcher could be 95% confident that the patient's score did not come from the control population even in the face of extreme skewness.

Study 3: Tests on the Difference Between a Patient's Performance on Two Tasks

Although the detection of suspected impairments is a fundamental feature of single-case studies, evidence of an impairment on a given task usually becomes of theoretical interest only if it is observed in the context of less impaired or normal performance on other tasks. That is, much of the focus in single-case studies is on establishing dissociations of function (Caramazza & McCloskey, 1988; Crawford, Garthwaite, & Gray, 2003; Ellis & Young, 1996; Shallice, 1988).

In the typical definition of a dissociation, the requirement is that a patient is "impaired" or shows a "deficit" on Task X, but is "not impaired," "normal," or "within normal limits" on Task Y. For example, Ellis and Young (1996) stated, "If patient X is impaired on Task 1 but performs normally on Task 2, then we may claim to have a dissociation between tasks" (p. 5). Shallice (1988) has termed this form of dissociation a *classical* dissociation.

It has been argued that the typical definition of a classical dissociation is insufficiently rigorous (Crawford, 2004; Crawford, Garthwaite, & Gray, 2003) for two related reasons. First, one half of the typical definition essentially involves an attempt to prove the null hypothesis (we must demonstrate that a patient is not different from the controls), whereas, as is well known, we can only fail to reject it. This is particularly germane to single-case studies, in which, as noted, the power to reject the null hypothesis is inevitably low: An individual patient (rather than a group) is compared with a control group, which itself is usually of very modest size (Crawford, 2004; Crawford, Garthwaite, & Gray, 2003).

The second problem is that a patient's score on the impaired task could lie just below the critical value for defining impairment, and the performance on the other test could lie just above it. That is, the difference between the patient's relative standing on the two tasks of interest could be very trivial; in this situation, we would not want to infer the presence of a dissociation (Crawford & Garthwaite, 2002; Crawford, Garthwaite, & Gray, 2003).

Crawford, Garthwaite, and Gray (2003) have developed formal criteria for a classical dissociation that, in addition to the standard requirement of a deficit on Task X and normal performance on Task Y, incorporated a requirement that the patient's performance on Task X should be significantly poorer than performance on Task Y. This criterion not only deals with the problem of trivial differences, but also provides a positive test for a dissociation (thereby lessening reliance on what boils down to an attempt to prove the null hypothesis of no deficit or impairment on Task Y).

Criteria for what Shallice (1988) termed a "strong" dissociation were also developed. A strong dissociation refers to the case in which a patient is impaired on both tasks but is more severely impaired on one (i.e., he or she exhibits a differential deficit). Crawford, Garthwaite, and Gray's (2003) criteria for a strong dissociation require that the patient has a significant deficit on Task X and on Task Y and a significant difference between Tasks X and Y. Note that previous definitions of a strong dissociation (e.g., Coltheart, 2001; Ellis & Young, 1996) also require a significant difference between Tasks X and Y (although the method to be used to test for this is rarely specified). Crawford, Garthwaite, and Gray's (2003) criteria differ from previous definitions in that these also require such a difference for a classical dissociation (and fully specify the methods used to determine whether all of the criteria for either type of dissociation are met).

Given the importance of testing for a significant difference between a patient's performance on two (or more) tasks, there is the need to select an appropriate inferential method for conducting such a test. One candidate is the long-established Payne and Jones (1957) method, in which the following formula is used:

$$z_D = \frac{Z_X - Z_Y}{\sqrt{2 - 2r_{xy}}}, \tag{3}$$

where Z_X and Z_Y are the patient's scores on the two tasks expressed as z scores (based on the means and standard deviations of the controls), and r_{xy} is the correlation between the tasks in the control sample (the denominator represents the standard deviation of the difference in controls when scores are expressed as z scores). The value of z_D is referred to a table of the areas under the normal curve to determine whether there is a significant difference between the patient's performance on the two tasks; for example, if a two-tailed test is required, then the difference would be significant ($p < .05$) if z_D exceeded 1.96.

A problem with the Payne and Jones (1957) formula is that, just as was the case when z is used to infer the presence of a deficit on a single task, it treats the control sample as if it were a population. In an attempt to overcome this problem, Crawford, Howell, and Garthwaite (1998) proposed a modified paired-sample t test to replace the Payne and Jones formula in single-case studies. Their formula is

$$t_D = \frac{Z_X - Z_Y}{\sqrt{(2 - 2r_{xy})\left(\frac{n+1}{n}\right)}}, \tag{4}$$

where all terms have been previously defined.

The modified paired-sample *t* test differs from a conventional paired-sample *t* test in three respects. First, a conventional paired *t* test is used to test for a difference in means obtained from the same sample—for example, to compare before and after scores on a task or to compare scores on a task under two different experimental conditions. In contrast, the modified *t* test is used to test whether the difference between scores on two tasks for an individual is sufficiently large such that it is unlikely to have come ($p < .05$) from the distribution of differences in the population of controls.

Second, in the modified *t* test, the scores on the two tasks are standardized; the individual's performance on Tasks X and Y is expressed as *z* scores based on the mean and standard deviations of the control sample. This is obviously never done when applying a conventional paired *t* test; the difference in means would necessarily be zero. Expressing the patient's score as a standard score is normally required in neuropsychological single-case studies because researchers attempt to establish the presence of a dissociation between two tasks of different cognitive functions, and the tasks normally have different means and standard deviations (indeed, the means and standard deviations are essentially arbitrary in most cases). Third, the probability value for *t* provides a point estimate of the abnormality of the patient's difference score (i.e., it quantifies the proportion of the control population that would exhibit a difference more extreme than the patient's).

In Study 3, we run a Monte Carlo simulation to test and compare control of Type I errors when the Payne and Jones (1957) test and Crawford et al.'s (1998) modified paired *t* test are used to test for a difference between an individual's performance on two tasks.

Method

Simulations were run using the same uniform random number generator as in Study 1. The Box–Muller transformation generates pairs of normally distributed observations, and by further transforming the second of these pairs, it is possible to generate observations from a bivariate normal distribution with a specified correlation (e.g., see Kennedy & Gentle, 1980). One million samples of $N + 1$ pairs of observations were drawn from each of four bivariate normal distributions in which the correlations (ρ) were set at .0, .2, .5, and .8. As in Study 1, this was done for five sample sizes: 5, 10, 20, 50, and 100.

The first *N* pairs of observations were taken as the control sample's scores on Tasks X and Y, and the $N + 1$th pair was taken as the scores of the individual control case. Crawford et al.'s (1998) test was then applied to these data, and *t* values that exceeded the two-tailed critical value for *t* on the appropriate degrees of freedom ($n - 1$) was recorded as a Type I error. The Payne and Jones test (1957) was also applied to these same data, and the result was recorded as a Type I error if z_D exceeded the two-tailed critical value of -1.96.

Results and Discussion

The simulation results obtained when the Payne and Jones (1957) formula was applied are presented in the first four columns of Table 3. It can be seen that the error rates are very high when the size of the control sample is modest (and are much higher than the rates obtained when *z* is used to compare a patient's score with that of controls on a single task). For example, when the control sample size is 5 and $\rho = .8$, the error rate is very inflated; that is, more than 25% of the control cases were identified as exhibiting a significant difference between Tasks X and Y. Indeed, it can be seen that, with a sample size of 5, the error rate does not fall below 21% for any value of ρ. It can also be seen that, even with larger sample sizes, the error rate is inflated; that is, when $N = 20$ and $\rho = .8$, the error rate is still 8.59%.

The results of applying Crawford et al.'s (1998) test are presented in the next four columns of Table 3. It can be seen that the control of the Type I error rate is substantially better than the rates obtained using the Payne and Jones (1957) formula (z_D) and that, with samples of 20 and above, the error rate is only marginally above the specified rate. However, it can be seen that control of the error rate is unsatisfactory with small sample sizes. Indeed, when $N = 5$, the error rate averaged over values of ρ is around twice the specified rate, with a high of 12.3%. It can also be seen that, as is the case for the Payne and Jones test, the Crawford et al. (1998) test has the undesirable characteristic that error rates vary as a function of the correlation between the tasks; error rates rise as the correlation rises. This feature of both tests is unfortunate because, as Shallice (1979) pointed out, much of the search for dissociations is focused on tasks that are at least moderately and even highly correlated in the general population (i.e., tasks for which there is a prima facie case that they tap a unitary function and therefore may not be dissociable).

In summary, it is clear that Crawford et al.'s (1998) test represents a considerable improvement over the Payne and Jones (1957) formula; the error rates for the latter test were alarmingly high. However, it is also apparent that the test statistic in Crawford et al.'s test does not follow a *t* distribution when the sample size is small and that the result is an inflation of the Type I error rate. It also follows that the point estimate of the abnormality of a pa-

Table 3
Simulation Results: Percentage of Control Cases Exhibiting Significant Differences Between Tasks X and Y (i.e., Percentage of Type I Errors) When Using Three Inferential Tests Under Different Values of N of the Control Sample and Correlations Between Tasks

	Payne & Jones (1957) test (z_D)				Crawford et al.'s (1998) test (t_D)				Unstandardized difference test (t_{UD})			
N	.0	.2	.5	.8	.0	.2	.5	.8	.0	.2	.5	.8
5	21.02	22.05	23.78	25.70	9.18	9.91	10.97	12.31	5.00	5.01	5.01	5.04
10	11.57	11.94	12.58	13.18	6.55	6.82	7.31	7.75	5.01	4.98	5.04	5.03
20	7.91	8.05	8.30	8.59	5.66	5.77	6.00	6.23	4.99	4.99	5.01	5.03
50	6.09	6.11	6.21	6.28	5.27	5.26	5.36	5.43	5.03	4.96	5.01	5.00
100	5.53	5.55	5.59	5.60	5.12	5.15	5.18	5.19	5.00	5.02	5.00	4.97

tient's difference score is biased with small control sample sizes; that is, the rarity of the patient's difference is exaggerated.

Study 4: Revised Tests for Differences

The limitations of Crawford et al.'s (1998) test stem from the fact that two "hidden" quantities in Equation 4 are still treated as parameters rather than as sample statistics: The standard deviations of the raw scores for controls on Tasks X and Y are used to convert the patient's raw scores on Tasks X and Y to z scores. There are two potential solutions to this problem.

As noted, in most situations in which neuropsychologists wish to test the difference between a patient's performance on two tasks, it is necessary to standardize the patient's scores. However, there are some scenarios in which this standardization is unnecessary, such as when a patient's performance on the same or parallel version of a task is compared with that of controls under two different experimental conditions. For example, a neuropsychologist might want to compare performance on the same task (or parallel version thereof) under monocular versus binocular viewing or when stimuli are viewed in the left versus right visual field. Similarly, the aim may be to compare a patient's performance when the same task is performed with the dominant versus nondominant hand or under single- versus dual-task conditions. In these situations, it is possible to apply the modified t test, but with the standardized scores replaced by unstandardized scores. The resultant test statistic takes the following form:

$$t_{UD_{n-1}} = \frac{(X^* - \overline{X}) - (Y^* - \overline{Y})}{\sqrt{(s_X^2 + s_Y^2 - 2s_X s_Y r_{xy})\left(\frac{n+1}{n}\right)}} \quad (5)$$

where X^* and Y^* are the scores of the patient on Tasks X and Y, respectively, and $\overline{X}$ and $\overline{Y}$ are the corresponding control means. The first bracketed term under the radical sign is the variance of the difference for controls, and it is obtained from the variance of Tasks X and Y in the controls (s_X^2 and s_Y^2) and the covariance of X and Y ($s_X s_Y r_{xy}$) in controls; as was the case for the original modified t test, the patient's score does not contribute to the variance estimate. The test statistic in Equation 5 should follow a t distribution on $n - 1$ *df.*

The potential solution outlined above is limited in its applicability because, as noted, it is more common for neuropsychologists to attempt to demonstrate dissociations between tasks of different cognitive functions in which the two tasks also have radically different means and standard deviations. In this more common situation, it is necessary to standardize the patient's score against the control's performance in order to conduct a meaningful test on the difference between a patient's performance on the two tasks. Therefore, it would be very useful if a method could be found that permits standardization of the patient's scores while also maintaining control of the Type I error rate. That is, we would like a test statistic that will closely approximate a t distribution when the patient's score has been standardized. To achieve this, we need a method in which, unlike Crawford et al.'s (1998) test, none of the control sample statistics are treated as parameters.

Starting with results for bivariate t distributions given by Siddiqui (1967), Garthwaite and Crawford (2004) used a computer algebra package (Maple) to perform asymptotic expansions. They obtained the statistic

$$\psi = \frac{\frac{(X^* - \overline{X})}{s_X} - \frac{(Y^* - \overline{Y})}{s_Y}}{\sqrt{\left(\frac{n+1}{n}\right)\left\{(2 - 2r) + \frac{2(1-r^2)}{n-1} + \frac{(5+y^2)(1-r^2)}{2(n-1)^2} + \frac{r(1+y^2)(1-r^2)}{2(n-1)^2}\right\}}}, \quad (6)$$

where all terms are as defined earlier except y, which is the critical two-tailed value for t on $n - 1$ *df.* They showed that the probability $\text{Prob}(\psi > y)$ is approximately equal to $\text{Prob}(t > y)$, where t has a standard t distribution on $n - 1$ *df.* This result can be used to test whether the difference between the patient's scores on Tasks X and Y is sufficiently large such that the patient differs significantly from controls. That is, if ψ exceeds the selected two-tailed critical value for t on $n - 1$ *df,* then the patient is significantly different from controls. Hereafter, this test is referred to as the revised standardized difference test (RSDT).

It is also desirable to obtain a precise probability for this test. Moreover, this would also allow users to obtain a point estimate of the abnormality of the difference observed for a patient. To obtain a p value, we solve $\psi = y$, which is a quadratic equation in y^2. Choosing the positive root gives

$$y = \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)^{1/2}, \quad (7)$$

where

$$a = (1 + r)(1 - r^2),\ b = (1 - r)\{4(n-1)^2 + 4(1+r)(n-1) + (1+r)(5+r)\}, \text{ and} \quad (8)$$

$$c = -2\left[\frac{X^* - \overline{X}}{s_X} - \frac{Y^* - \overline{Y}}{s_Y}\right]^2\left(\frac{n(n-1)^2}{n+1}\right). \quad (9)$$

Then the p value equals $\text{Prob}(t > y)$, where t has a standard t distribution on $n - 1$ *df.* The derivation of Equation 6 and Equation 7 is long and technical. In addition, the formulas can potentially be applied to test hypotheses other than those that are the focus of this article; that is, they can be used in any situation in which there is a need to test the difference between two variables that are themselves distributed as t. Because of these considerations, the derivation of the formulas are the subject of a separate article (Garthwaite & Crawford, 2004).

In the present article, the aim is to (a) examine the control of the Type I error rate when these revised tests are used for the specific purpose of comparing an individual patient's difference to the differences in controls, (b) examine the effect of using results from these tests in criteria for dissociations (see Study 5), and finally, (c) provide worked examples of their use in single-case studies.

Method

To examine the Type I error rate for the unstandardized difference test (Equation 5), we repeated the simulation procedure used for the Payne and Jones (1957) test and Crawford et al.'s (1998) original modified paired t test (see Study 3) using the same sample size and ρ but substituting the unstandardized difference test for these latter tests. A similar procedure

was followed for the RSDT (Equation 6) in that the same sample size and ρ were used. However, in addition, control of the error rates was examined for a larger range of specified error rates in order to examine in more breadth the accuracy of the approximation given by Equation 6.

Results and Discussion

The simulation results for the unstandardized difference test (Equation 5) are presented in the final four columns of Table 3. It can be seen from Table 3 that control of the error rate is impeccable at all sample sizes, including the small sample sizes that produced marked inflation of the error rate with the Payne and Jones (1957) test and Crawford et al.'s (1998) test. For example, the error rate is 5.04% for the unstandardized difference test when $N = 5$ and $\rho = .8$, compared with 25.7% and 12.31%, respectively, for the latter tests.

The simulation results for the RSDT (Equation 6) are presented in Table 4. Turning first to the results when the specified error rate was 5% (i.e., the error rate used in all of the other simulations), it can be seen from Table 4 that control of the error rate is good at all sample sizes, including the small samples that produced marked inflation of the error rate with the Payne and Jones (1957) test and Crawford et al.'s (1998) test. The error rates ranged from 5.17% to 5.59% for the RSDT when $N = 5$, compared with a range of 21.02% to 25.7% for the Payne and Jones test and a range of 9.18% to 12.31% for Crawford et al.'s. (1998) test. It can also be seen that the error rate is under control at all values of ρ in the table, unlike the latter tests for which the error rates became more inflated as larger values of ρ were specified.

The differences in the pattern of results for the Payne and Jones (1957) test, Crawford et al.'s (1998) test, and the RSDT can clearly be appreciated by examining Figure 2. This figure displays the Type I error rates for the three tests as a function of the control sample size. For clarity, the results are limited to those in which the population correlation between tasks (ρ_{XY}) was .5; the differences in control of the error rates would be even more extreme for $\rho_{XY} > .5$ and less extreme for $\rho_{XY} < .5$.

As Table 4 shows, in the case of the RSDT, control of the error rates was examined for a range of specified error, and it can be seen that the observed error rates all cleave closely to the specified error rates. For example, for $N = 5$, the observed rates for a specified rate of 1% ranged from 0.97% to 1.12%. A similarly close correspondence was obtained at the other extreme; that is, the range of observed rates for a specified rate of 40% was from 40.10% to 40.22% for $N = 5$. With larger sample sizes, the correspondence is even closer. The accuracy of the approximation will not hold as well for correlations that closely approach unity (unlikely in practice) and error rates well below 1% (although additional analysis showed that control was very good even at a specified rate of 0.05%). Otherwise, it is the case that error rates will be approximated very satisfactorily.

In conclusion, the present results, when taken together with those from Study 3, indicate that the RSDT should replace previously available alternative methods of testing for a difference between a patient's performance on two tasks.

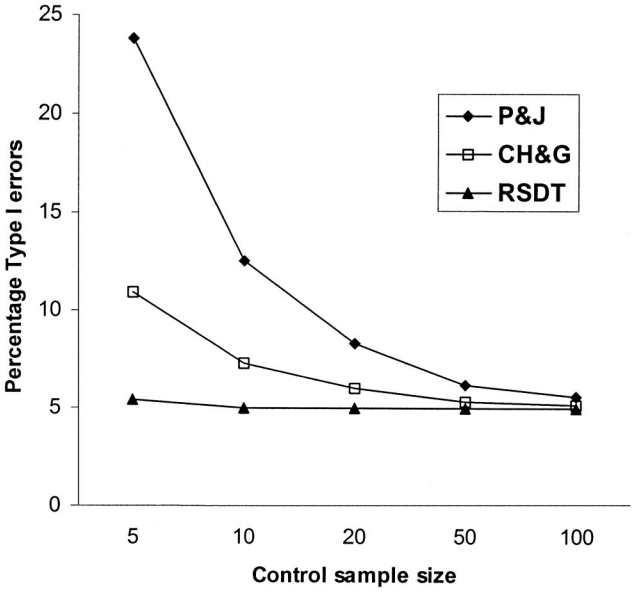

Figure 2. Monte Carlo simulation results: Type I errors for three tests on the difference between a patient's scores on two tasks (results are for a ρ_{XY} of .5). P&J = Payne & Jones test (Payne & Jones, 1957); CH&G = Crawford, Howell, & Garthwaite test (Crawford, Howell, & Garthwaite, 1998); RSDT = revised standardized difference test.

Table 4
Simulation Results: Percentage of Control Cases Exhibiting Significant Differences Between Tasks X and Y (i.e., Percentage of Type I Errors) When Using the RSDT (t_{RD}) Under Different Values of N of the Control Sample, Correlations Between Tasks, and Specified Error Rates

	Specified error rate																			
	1%				5%				10%				20%				40%			
N	.0	.2	.5	.8	.0	.2	.5	.8	.0	.2	.5	.8	.0	.2	.5	.8	.0	.2	.5	.8
5	0.97	1.00	1.05	1.12	5.17	5.30	5.42	5.59	10.25	10.42	10.55	10.78	20.28	20.41	20.53	20.68	40.10	40.18	40.19	40.22
10	0.99	0.99	0.99	1.02	5.00	5.02	5.04	5.06	9.99	10.04	10.06	10.08	19.97	20.06	20.02	20.08	40.00	39.99	40.02	40.06
20	1.00	1.01	1.01	1.00	4.99	4.99	5.01	5.03	9.99	9.96	10.01	10.04	20.01	19.95	19.99	20.01	40.03	39.99	40.02	40.02
50	1.00	0.99	1.00	1.02	5.01	5.00	4.98	5.01	10.03	10.00	9.97	10.03	20.04	20.00	19.93	20.03	40.02	39.99	39.98	40.00
100	0.99	0.99	1.01	0.99	4.97	5.00	5.00	5.00	10.00	10.00	10.01	10.00	19.97	20.00	20.05	19.96	40.02	40.00	40.01	39.96

Note. RSDT = revised standardized difference test.

Study 5: Use of the RSDT in Setting Criteria for Dissociations

As noted, Crawford, Garthwaite, and Gray (2003) proposed formal criteria for dissociations for use in single-case studies. Their criteria for classical and strong dissociations were based on the pattern of results obtained from the application of three inferential tests: two to test for the presence of deficits on Tasks X and Y using Crawford and Howell's (1998) test, and one on the difference between Tasks X and Y using Crawford et al.'s (1998) test.

Crawford, Garthwaite, and Gray (2003) ran a Monte Carlo simulation to estimate the percentage of control cases that would be incorrectly classified as exhibiting a dissociation when their criteria were applied. The results were encouraging in that the percentage of control cases classified as exhibiting a classical dissociation was low (below 5%) and was even lower for strong dissociations. However, in the present study we have shown that the RSDT is superior to Crawford et al.'s (1998) original difference test in controlling Type I errors. Therefore, this suggests that Crawford, Garthwaite, and Gray's criteria should be modified so that the test on the difference between a patient's scores on Tasks X and Y is provided by the RSDT rather than Crawford et al.'s original test. The purpose of Study 5 was to rerun Crawford, Garthwaite, and Gray's simulation to estimate the percentage of control cases that will be misclassified when the revised criteria are applied. The revised criteria are set out formally in Table 5. Although the focus of the present study is on evaluating the criteria for single dissociations, the criteria for double dissociations (i.e., dissociations involving 2 patients with opposite patterns of performance) stem naturally from these former criteria. Therefore, for completeness, Table 5 also includes the revised criteria for double dissociations.

Table 5
Revised Criteria for Classical and Strong Dissociations Obtained by Modifying Crawford, Garthwaite, and Gray's (2003) Original Criteria

Dissociation	Criteria
Classical	1. Patient's score on Task X significantly lower than that of controls ($p < .05$, one-tailed) on Crawford & Howell's (1998) test; that is, score meets the criterion for an impairment. 2. Patient's score on Task Y not significantly lower than that of controls ($p > .05$, one-tailed) on Crawford & Howell's test; that is, score fails to meet criterion for an impairment and is therefore considered to be within normal limits. 3. Patient's score on Task X significantly lower ($p < .05$, two-tailed) than patient's score on Task Y with the use of the RSDT. The test is two-tailed to allow for the fact that the data are examined before deciding which task is X and which is Y.
Strong (i.e., differential deficit)	1. Patient's score on Task X significantly lower than that of controls ($p < .05$, one-tailed) on Crawford & Howell's test; that is, score meets the criterion for an impairment. 2. Patient's score on Task Y is also significantly lower than that of controls ($p < .05$, one-tailed) on Crawford & Howell's test; that is, score meets the criterion for an impairment. 3. Patient's score on Task X significantly lower ($p < .05$, two-tailed) than patient's score on Task Y with the use of the RSDT.
Classical double	1. Patient 1 meets the criterion for a deficit on Task X and meets the criteria for a classical dissociation between this task and Task Y. 2. Patient 2 meets the criterion for a deficit on Task Y and meets the criteria for a classical dissociation between this task and Task X.
Strong double	1. Patient 1 meets the criterion for a deficit on Task X and meets the criteria for a classical or strong dissociation between this task and Task Y. 2. Patient 2 meets the criterion for a deficit on Task Y and meets the criteria for a classical or strong dissociation between this task and Task X. 3. Only one of the above dissociations is classical (otherwise we have a classical double dissociation).

Note. RSDT = reversal standardized difference test.

Method

The simulation procedure was similar to that used in Study 3. That is, 1,000,000 samples of $N + 1$ pairs of observations were drawn from each of four bivariate normal distributions in which the correlations were specified as .0, .2, .5, and .8. This was done for the five sample sizes used in Study 3. The first N pairs of observations were taken as the control sample's scores on Tasks X and Y, and the $N + 1$th pair was taken as the scores of the individual control case. Crawford and Howell's (1998) test was applied to the scores of the control case on Tasks X and Y (using a one-tailed test), and the RSDT based on the statistic in Equation 6 was applied to the standardized difference score of the control case (using a two-tailed test).

The percentage of control cases that met the criteria for a classical dissociation was recorded (i.e., a significant result on either Task X or Task Y, but not both, and a significant result on the revised difference test). The percentage of control cases that met the criteria for a strong dissociation was also recorded (i.e., a significant result on Tasks X and Y and a significant result on the revised difference test). Note that these classifications are mutually exclusive. The procedure followed in the present study was the same as that in Crawford, Garthwaite, and Gray's (2003) study except that 1,000,000 samples were drawn for each sample size and ρ (rather than 100,000), and crucially, the RSDT was substituted for Crawford et al.'s (1998) original modified paired t test.

Results and Discussion

The results of the simulation are presented in Table 6. It can be seen from Table 6 that, in the case of a strong dissociation, for all of the values of the correlation and sample size that were examined, a very small number of control cases were incorrectly classified as having a strong dissociation (less than 0.22% for all sample sizes and ρ and much smaller than this quantity in the majority of cases). In addition, it can be seen that the percentage showing a classical dissociation was comfortably below 5% (maximum = 2.51%) and much smaller than this in the majority of cases. Therefore, the results indicate that when these criteria are applied in single-case research, it would be unlikely that a member of the control (i.e., healthy) population would be misclassified as exhibiting either form of dissociation.

The percentage of controls exhibiting a classical dissociation, although small, is necessarily higher than that for a strong disso-

Table 6
Results From a Monte Carlo Simulation Study: Percentage of Control Cases Incorrectly Classified as Exhibiting Strong and Classical Dissociations When t_{RD} Is Used to Test for Differences Between Tasks X and Y Under Different Values of N of the Control Sample and Correlations Between Tasks

	Strong dissociation				Classical dissociation			
N	.0	.2	.5	.8	.0	.2	.5	.8
5	0.01	0.02	0.07	0.22	2.32	2.03	1.64	1.12
10	0.00	0.01	0.03	0.12	2.41	2.06	1.56	0.98
20	0.00	0.00	0.01	0.07	2.48	2.06	1.50	0.90
50	0.00	0.00	0.01	0.04	2.51	2.06	1.49	0.84
100	0.00	0.00	0.00	0.04	2.49	2.10	1.48	0.84

ciation. This is because, for a strong dissociation, scores must be extreme on both Tasks X and Y, and therefore, the score on one of these two tasks must be very extreme to meet the criteria.

Comparison of the results for Crawford, Garthwaite, and Gray's (2003) original criteria and the revised criteria demonstrates that the latter have reduced the probability of misclassifying a member of the control population. The percentages for the revised criteria are lower at all values of ρ and sample size but, as is to be expected given the results of Studies 2 and 3, are particularly marked with small sample sizes. The percentage of controls misclassified as exhibiting a strong dissociation ranged from a low of 0.02% (when $\rho = .0$) to a high of 0.37% (when $\rho = .8$) in Crawford, Garthwaite, and Gray's simulation for a sample size of 5. The corresponding figures in the present study using the RSDT were 0.01% and 0.22%.

The reduction in misclassification of controls was also evident for a classical dissociation. In Crawford, Garthwaite, and Gray's (2003) simulation, the percentage of controls misclassified as exhibiting a classical dissociation for a sample size of 5 ranged from a low of 2.04% (when $\rho = .8$) to a high of 3.41% (when $\rho = .0$). The corresponding figures in the present study using the RSDT were 1.12% and 2.32%.

In summary, the present results clearly illustrate the conservatism inherent in the sequence of tests for dissociations; that is, application of these criteria will rarely misclassify individuals drawn from the control population. Furthermore, the superior results obtained in Study 4 for the RSDT over its alternatives (Study 3) have carried over to the present study in which it was incorporated into a revised set of criteria for dissociations. This reinforces our recommendation that the RSDT should replace previously available alternatives; that is, Crawford et al.'s (1998) test and the Payne and Jones (1957) test in single-case research.

General Discussion

Worked Examples for the Revised Difference Tests and Revised Criteria for Dissociations

Worked examples of both of the revised tests for differences are provided next, although researchers need never perform the calculations as a computer program is available to accompany this article (see next section). To illustrate the use of the unstandardized difference test (Equation 5), suppose that a neuropsychologist examines the performance of a patient on a distance estimation task under monocular versus binocular viewing. The patient's score in the monocular condition was 40 and 76 in the binocular condition (high scores equal good performance). Suppose also that 12 matched controls had been recruited and administered the same task under the same two conditions; the mean score in controls was 80 ($SD = 14.0$) under binocular conditions and 78 ($SD = 15.0$) under monocular conditions, and the correlation between performance on the tasks in controls was .7.

$$t_{UD_{n-1}} = \frac{(76 - 80) - (40 - 78)}{\sqrt{(196 + 225 - 2 \times 14 \times 15 \times 0.70)\left(\frac{12 + 1}{12}\right)}} = \frac{34}{\sqrt{(127)(1.0833)}} = \frac{34}{11.7296} = 2.899. \quad (10)$$

The two-tailed probability for this t value on 11 df is .014. Therefore, there is a significant difference ($p < .05$) between the patient and controls; that is, it is highly unlikely that the difference between performance under binocular versus monocular viewing observed for the patient was drawn from the distribution of differences in the control population. The one-tailed p value (.007) also provides researchers with a point estimate of the abnormality of the patient's difference; in this example, it is estimated that only 0.7% of the control population would exhibit a difference of this magnitude in favor of binocular viewing.

In deciding whether it is appropriate to compare a patient with controls using the unstandardized difference test, researchers need only pose the following question: "Would it be legitimate to use a paired t test to compare the performance of controls under the two different conditions?" If the answer is yes, then it is equally legitimate to use t_{UD} to test if the difference between performance under condition Task X versus Task Y observed for a patient is significantly different from the distribution of differences in controls.

The RSDT provides a much more general method of testing for differences between Tasks X and Y; that is, it can be used to compare a patient's performance on diverse tasks. To illustrate its use, suppose that a researcher administered a theory of mind (ToM; Baron-Cohen, Leslie, & Frith, 1985) task and a task of executive ability (e.g., set-shifting) to a patient and wished to determine whether his or her performance on these two tasks was significantly different (i.e., the researcher wished to determine whether the null hypothesis that the difference observed for the patient was drawn from the population of control differences could be rejected). Suppose also that 20 healthy controls matched to the patient on basic demographic variables had been recruited. The mean score for controls on the ToM task was 60 ($SD = 7$), and the mean score on the executive task was 24 ($SD = 4.8$); the correlation between the two tasks in controls was .68. Suppose the patient's raw scores on these two tasks were 33 and 15, respectively.

In applying either of the formulas for RSDT (i.e., Equation 6 or Equation 7), the first step is to convert the patient's scores to z scores; in this example, the patient's z score on the ToM task is -3.857, and the z score on the executive task is -1.875. When using the Equation 6 formula, we have to enter the two-tailed critical value for t on $n - 1$ df for our selected value of alpha. If

we set alpha at the conventional 5% level, then the critical value is 2.093. Entering these data into Equation 6, we have

$$\psi = \frac{(-3.857) - (-1.875)}{\sqrt{\left(\frac{20+1}{20}\right)\left\{(2 - 2 \times 0.68) + \frac{2(1-.68^2)}{20-1} + \frac{(5+2.093^2)(1-.68^2)}{2(20-1)^2} + \frac{.68(1+2.093^2)(1-.68^2)}{2(20-1)^2}\right\}}}$$

$$= \frac{-1.982}{\sqrt{(1.05)\left\{(0.64) + \frac{1.0752}{19} + \frac{(9.3806)(0.5376)}{722} + \frac{.68(5.3806)(0.5376)}{722}\right\}}}$$

$$= \frac{-1.982}{\sqrt{1.05\{0.64 + 0.056589 + 0.006985 + 0.002724\}}}$$

$$= \frac{-1.982}{\sqrt{0.7416}} = -2.302. \quad (11)$$

As $\psi = -2.302$ exceeds the critical value of ± 2.093, we conclude that the patient's performance on the ToM task is significantly more impaired than performance on the executive task (at the 5% level).

As noted, Equation 7 permits researchers to obtain a precise probability for the difference between patient and controls and thereby also provides a point estimate of the abnormality of the patient's difference. Entering the data from the current example into Equation 7,

$$a = (1 + 0.68)(1 - 0.68^2) = 0.90317, \quad (12)$$

$$b = (1 - 0.68)\{4(20-1)^2 + 4(1+0.68)(20-1) + (1+0.68)(5+0.68)\} = 505.99, \text{ and} \quad (13)$$

$$c = -2[(-3.857) - (-1.875)]^2\left(\frac{20(20-1)^2}{20+1}\right) = -2701.19. \quad (14)$$

and therefore,

$$y = \left(\frac{-505.99 + \sqrt{505.99^2 - 4(0.90317)(-2701.19)}}{2(0.90317)}\right)^{1/2} = 2.300. \quad (15)$$

The two-tailed p value for a t of 2.30 on 19 df is .033; therefore, we come to the same conclusion as that reached when we use Equation 6: The patient's ToM performance is significantly more impaired ($p < .05$) than her or his performance on the executive task. To obtain a point estimate of the abnormality of the patient's difference, we use the one-tailed p value for t. The p value is .0165, and therefore, we estimate that only 1.65% of the control population would exhibit a discrepancy in favor of the executive task of this magnitude and direction.

We can also use this example to illustrate the application of the revised criteria for dissociations (see Table 5). Application of Crawford and Howell's (1998) test (Equation 1) reveals that the patient is significantly different (one-tailed) from controls on the ToM task, $t(19) = 3.76$, $p = .001$, and on the executive task, $t(19) = 1.83$, $p = .042$. The patient is therefore considered to have an impairment on both tasks and does not meet the criteria for a classical dissociation. However, the patient does meet the criteria for a strong dissociation; performance on both tasks is impaired, but the ToM deficit is significantly greater (i.e., the ToM deficit is a differential deficit).

Finally, the RSDT provides a very flexible method of testing for dissociations as its use need not be limited to cases in which performance is quantified by simple test scores (such as number of items correct). For example, a patient's memory for temporal order is typically assessed by computing the rank-order correlation between the order reported by a patient and the actual order of presentation. Similarly, in estimation tasks, such as distance, weight, or time estimation, performance is commonly assessed by the slope of the regression line relating an individual's estimates to the actual distances, weights, or elapsed times. Crawford, Garthwaite, Howell, and Venneri (2003) and Crawford and Garthwaite (2004) have recently developed methods that allow single-case researchers to test whether a patient is significantly different from a control sample when performance is quantified by a parametric or nonparametric correlation coefficient or slope.

These authors noted that Crawford et al.'s (1998) test could be used to test whether there was a dissociation between constructs measured by slopes or correlations or dissociations between such constructs and constructs measured by conventional means. The present results suggest that the RSDT should be used for this purpose instead. For example, it could be used to test if a patient exhibits a dissociation between temporal order memory for verbal material and free recall of such material; details of the treatment of patient and control data that are in the form of slopes or correlations can be found in the aforementioned articles (Crawford & Garthwaite, 2004; Crawford, Garthwaite, Howell, & Venneri, 2003).

Computer Program to Implement the Revised Difference Tests and Revised Criteria for Dissociations

The calculations involved in applying the unstandardized difference test (Equation 5) or RSDT (Equation 7), and thereby also obtaining a point estimate of the abnormality of the patient's difference, could be performed by hand or calculator. However, the calculations for the RSDT are tedious and prone to human error. For the foregoing reasons, we have implemented the statistical methods in a computer program (dissocs.exe) for PCs.

The program prompts the user to select either the unstandardized difference test or the RSDT. The data inputs required are the means and standard deviations for Tasks X and Y and the correlation between Tasks X and Y in controls, the size of the control sample, and the patient's scores on Tasks X and Y. The results of applying the selected difference test are reported: namely, the t value and its associated two-tailed probability and the point estimate of the abnormality of the patient's difference.

The program also applies the revised criteria for dissociations presented in the present article. That is, it applies Crawford and

Howell's (1998) test to test for deficits on Tasks X and Y (point estimates and confidence limits for the abnormality of the patient's scores are also reported), and uses these results together with the results of the RSDT (or unstandardized difference test if the latter has been selected) to establish whether the patient's results fulfill the criteria for a classical or strong dissociation. The results of these analyses can be viewed on screen, printed, or saved to a file. The program can be downloaded from the following web page address: http://www.abdn.ac.uk/~psy086/dept/dissociations.htm.

Some Comments and Caveats on the Use of Single-Case Methods

The revised inferential methods for differences presented in this article are both modified *t* tests. As is the case for Crawford and Howell's (1998) test, they assume that the control sample data are normally distributed. Examining the robustness of these tests in the face of skew is more complicated than was the case for the former test as it is necessary to sample from skewed bivariate distributions and a larger variety of scenarios needs to be covered (e.g., investigating robustness when Tasks X and Y are both skew, or only one of X and Y, and studying effects of skew in opposite directions for X and Y). However, we have conducted some provisional analysis of this issue for the RSDT and obtained results that are as encouraging as those reported in Study 2 for Crawford and Howell's test (Garthwaite & Crawford, 2004). Nevertheless, the results from applying these tests should be treated cautiously when the data exhibit severe skew unless the resultant *p* value is well beyond .05 (i.e., $<.025$). It is important to note that the more commonly used alternative methods, for example, the use of z_D or Crawford et al.'s (1998) method to test for a difference between tasks, make exactly the same assumption and will be equally compromised when this assumption is violated.

The emphasis in the present article has been on evaluating the performance of the inferential tests for deficits and dissociations when single-case research is conducted with modestly sized control samples. To avoid any potential confusion, it should be noted that the methods can be used with control samples of any size and remain more valid than commonly used alternatives based on *z* when the sample size is large; in this situation, the researcher is still dealing with a sample not a population. Furthermore, although the methods achieve good control of Type I errors with small sample sizes, this does not mean that researchers should limit themselves to recruiting small control samples; the present article focuses on small samples simply because of the need to reflect the reality of current practice in many single-case studies. Indeed, as noted, statistical power is inevitably low in single-case studies (significant results are obtained because effects are often large enough to overcome this). Therefore, it makes sense to increase power by recruiting a large sample of controls when this is practical.

It should also be noted that very useful and elegant methods have been devised for drawing inferences concerning an individual patient's performance on fully standardized neuropsychological tests; that is, on tests that have been normed on very large, representative samples of the population (e.g., Capitani, 1997; 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, errors arising from sampling from the control population are ignored; this is justifiable because the samples are large enough for such errors to be minimal.

Although these latter approaches have much to commend them, they unfortunately can be used only in fairly circumscribed situations 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 with a modestly sized control sample.

At the other extreme, some single-case studies do not refer the patient's performance to either a control sample or a large normative sample. That is, conclusions on the presence of deficits and dissociations are based on intraindividual analysis. An example of this approach comes from the aforementioned literature on category specificity. It is quite common for conclusions of a dissociation between naming of living and nonliving things to be based on a significant result from a chi-square test; that is, a patient is administered an equal number of living and nonliving items and the number correctly named in each category is compared (Laws, in press).

However, aside from the fact that the independence assumption for a chi-square test is violated in these circumstances, there are further difficulties with this approach. For example, Laws et al. (in press) studied Alzheimer's disease patients who exhibited significant differences (on chi-square tests) between the number of living and nonliving items named and found that many of these raw differences were not unusual when standardized against control performance; that is, the intraindividual method yielded false-positive indications of a dissociation. The opposite pattern was also found; patients whose chi-square results were not significant showed strong evidence of a dissociation when their naming was referenced to control performance.

The focus of the present study has been on inferential methods for single tests (when attempting to detect deficits) or pairs of tests (when attempting to detect dissociations). However, it should be acknowledged that findings obtained from comparing the patient to a control sample are not interpreted in isolation. Rather, these findings are interpreted in the context of results from a prior assessment in which a broad characterization of the patient's strengths and weaknesses will have been achieved through the use of fully or partially standardized tests.

Furthermore, many single-case studies use multiple measures of the constructs under investigation (i.e., different but related Tasks X_1, X_2, etc., and Y_1, Y_2, etc., to measure constructs *X* and *Y*). That is, the patient is compared with controls over a series of tasks. This is in keeping with the fact that researchers are ultimately interested in dissociations between functions, not just in dissociations between specific pairs of indirect and imperfect measures of these functions (Crawford, Garthwaite, Howell, & Venneri, 2003; Vallar, 2000). Thus, researchers seek converging evidence of a deficit or dissociation (Vallar, 2000). The upshot of this is that the risk of drawing incorrect conclusions will typically be less than that associated with the results from a single inferential test (in the case of a deficit) or single application of a set of criteria (in the case of a dissociation).

However, the integration of these multiple sources of information is a complex and formidable task. It is fair to say that (a) currently there is little consistency across studies in how this task is approached and (b) existing attempts tend to be qualitative rather than quantitative. The development of a quantitative system, whereby the probabilities (e.g., of a dissociation) could be combined or updated as different stages of a study are completed, would make a very significant contribution to the discipline. The nature of this problem is such that an approach based on Bayesian rather than classical (i.e., frequentist) methods would be the obvious choice.

Finally, a central aim of the present study was to develop and evaluate more rigorous criteria for dissociations than those used previously. However, even if infallible criteria for identifying dissociations were available, there remains the wider and thornier issue of what dissociations allow researchers to conclude about the functional architecture of human cognition. Although this is a large topic, and one that lies beyond the scope of the present study, a few comments are in order.

It is generally acknowledged that a single dissociation implies that different cognitive functions underlie performance on the two tasks in question, but that such dissociations are prone to task difficulty artifacts. That is, a unitary cognitive function may contribute to performance on both Tasks X and Y, but only Task X is of sufficient difficulty to uncover an impairment of this function (Crawford, Garthwaite, & Gray 2003; Vallar, 2000). The identification of a double dissociation (i.e., patients who have opposite patterns of spared and impaired performance) is generally considered to largely rule out such artifacts. For this reason, the double dissociation is a central tool for the building and testing of theory in neuropsychology. As Vallar (2000) noted, the double dissociation provides "the most effective paradigm for investigating the modularity of the mental processes and their neural correlates" (p. 329). However, serious areas of debate remain (Dunn & Kirsner, 2003; Shallice, 1988). For example, Dunn and Kirsner (2003) argued that (a) researchers can only specify the characteristics of cognitive modules underlying a double dissociation if the cases involved are pure cases and the tasks are process pure and (b) there is no independent means of testing whether the former situation holds. Thus, their pessimistic conclusion is that "dissociations may tell us nothing more about mental functions other than that there are two of them" (Dunn & Kirsner, 2003, p. 5).

Conclusion

The single-case approach in neuropsychology has made a significant contribution to researchers' understanding of the functional architecture of human cognition. However, as Caramazza and McCloskey (1988) noted, if advances in theory are to be sustainable, they "must be based on unimpeachable methodological foundations" (p. 519). The statistical treatment of single-case study data is one area of methodology that has been relatively neglected. In the present article, the evaluation of inferential tests for comparing a patient to a control sample provides researchers with simulation results to guide their choice of methods and provides new methods that have significant advantages over the existing alternatives.

References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind. *Cognition, 21,* 37–46.

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

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 28,* 610–611.

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* (2nd ed., Vol. 1, pp. 53–76). Amsterdam: Elsevier.

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

Coltheart, M. (2001). Assumptions and methods in cognitive neuropsychology. In B. Rapp (Ed.), *The handbook of cognitive neuropsychology* (pp. 3–21). Philadelphia: Psychology Press.

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, England: 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. (2004). Statistical methods for single-case research: Comparing the slope of a patient's regression line with those of a control sample. *Cortex, 40,* 533–548.

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., Garthwaite, P. H., Howell, D. C., & Gray, C. D. (2004). Inferential methods for comparing a single case with a control sample: Modified *t*-tests versus Mycroft et al.'s (2002) modified ANOVA. *Cognitive Neuropsychology, 21,* 750–755.

Crawford, J. R., Garthwaite, P. H., Howell, D. C., & Venneri, A. (2003). Intra-individual measures of association in neuropsychology: Inferential methods for comparing a single case with a control or normative sample. *Journal of the International Neuropsychological Society, 9,* 989–1000.

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

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

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

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex, 39,* 1–7.

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

Garthwaite, P. H., & Crawford, J. R. (2004). The distribution of the difference between two *t*-variates. *Biometrika, 91,* 987–994.

Garvin, J. S., & McClean, S. I. (1997). Convolution and sampling theory of the binormal distribution as a prerequisite to its application in statistical process control. *The Statistician, 46,* 33–47.

Gibbons, J. F., & Mylroie, S. (1973). Estimation of impurity profiles in ion-implanted amorphous targets using half-Gaussian distributions. *Applied Physics Letters, 22,* 568–569.

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

Kennedy, W. J., & Gentle, J. E. (1980). *Statistical computing.* New York: Marcel Dekker.

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

Laws, K. R. (in press). Illusions of normality: A methodological critique of category-specific naming. *Cortex.*

Laws, K. R., Gale, T. M., Leeson, V. C., & Crawford, J. R. (in press). When is category *specific* in Alzheimer's disease? *Cortex.*

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

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal.* Cambridge, England: Cambridge University Press.

Shallice, T. (1979). Case study approach in neuropsychological research. *Journal of Clinical Neuropsychology, 3,* 183–211.

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

Siddiqui, M. M. (1967). A bivariate *t*-distribution. *Annals of Mathematical Statistics, 38,* 162–166.

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

Vallar, G. (2000). The methodological foundations of human neuropsychology: Studies in brain-damaged patients. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (2nd ed., Vol. 1, pp. 305–344). Amsterdam: Elsevier.

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.

Received April 23, 2003
Revision received April 29, 2004
Accepted May 20, 2004 ■