Psychology Press
Taylor & Francis Group

Methods of testing for a deficit in single-case studies: Evaluation of statistical power by Monte Carlo simulation

John R. Crawford
University of Aberdeen, Aberdeen, UK

Paul H. Garthwaite
Department of Statistics, The Open University, Milton Keynes, UK

Testing for the presence of a deficit by comparing a case to controls is a fundamental feature of many neuropsychological single-case studies. Monte Carlo simulation was employed to study the statistical power of two competing approaches to this task. The power to detect a large deficit was low to moderate for a method proposed by Crawford and Howell (1998; ranging from 44% to 63%) but was extremely low for a method proposed by Mycroft, Mitchell, and Kay (2002; ranging from 1% to 13%). The effects of departures from normality were examined, as was the effect of varying degrees of measurement error in the scores of controls and the single case. Measurement error produced a moderate reduction in power when present in both controls and the case; the effect of differentially greater measurement error for the single case depended on the initial level of power. When Mycroft et al.'s method was used to test for the presence of a classical dissociation, it produced very high Type I error rates (ranging from 20.7% to 49.3%); in contrast, the rates for criteria proposed by Crawford and Garthwaite (2005b) were low (ranging from 1.3% to 6.7%). The broader implications of these results for single-case research are discussed.

INTRODUCTION

In neuropsychological single-case research inferences concerning a patient's cognitive status are often based on comparing the patient's test scores to those of a control sample. In the present study the performance of two inferential methods of testing for acquired deficits is evaluated: a modified *t* test proposed by Crawford and Howell (1998; see also Crawford & Garthwaite, 2002) and a modified analysis of variance (ANOVA) proposed by Mycroft, Mitchell, and Kay (2002).

Crawford and Howell's (1998) method has been widely used to test for acquired deficits in single-case research (e.g., see Bird, Castelli, Malik, Frith, & Husain, 2004; Howard & Nickels, 2005; Papps, Calder, Young, & O'Carroll, 2003; Robinson, Shallice, & Cipolotti, 2005; Rosenbaum, Fuqiang, Richards, Black, &

Correspondence 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, UK (Email: j.crawford@abdn.ac.uk).

© 2006 Psychology Press Ltd

http://www.psypress.com/cogneuropsychology
DOI:10.1080/02643290500538372

Moscovitch, 2005; Rusconi, Priftis, Rusconi, & Umiltà, 2006; Schindler et al., 2004). Mycroft et al.'s (2002) method was developed more recently but it too has been used in a number of single-case studies for the same purpose (see Bobes et al., 2004; Farrer, Franck, Paillard, & Jeannerod, 2003; Forti & Humphreys, 2004; Miller & Swick, 2003). There are fundamental differences in the rationale behind these two methods, and they can produce radically different results when applied to the same data set. This raises the alarming possibility that results obtained in single-cases studies (and their apparent implications for theory) could be more a reflection of the inferential method employed than genuine characteristics of the constructs under investigation.

Crawford and Howell's (1998) proposed method of testing for a deficit in single-case studies is based on a procedure described by Sokal and Rohlf (1995) and takes the form of a modified t test. The formula for this test is

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

where x^* is the patient's score, $\bar{x}$ and s are the mean and standard deviation of scores in the control sample, and n is the size of the control sample. If the t value obtained from this test exceeds the one-tailed 5% critical value for t on $(n - 1)$ df then it can be concluded that the patient's score is sufficiently low to reject the null hypothesis that it is an observation from the scores of the control population, and the patient is considered to exhibit an impairment on the task in question. A one-tailed test is employed because the hypothesis tested (that the patient has a deficit) is directional.

The p value obtained from this test also provides a point estimate of the abnormality of the patient's score. (In previous work on this topic we have simply stated this without proof; in the present paper we provide a brief mathematical proof in Appendix A.) For example, if the one-tailed p is .013 then we know that the patient's score is sufficiently low to render it unlikely ($p <$ .05) that it has come from the control population and that only $(p \times 100) = 1.3\%$ of the control population would be expected to obtain a score lower than the patient's. This point estimate can be supplemented with confidence limits on the abnormality of the patient's score using a method developed by Crawford and Garthwaite (2002).

The starting point for Mycroft et al.'s (2002) method is a modified ANOVA that is the direct equivalent of the modified t test outlined above (i.e., the test would yield identical p values because F on $[1, n - 1]$ $df = t^2$ on $(n - 1)$ df. However, Mycroft et al. make a further modification to this ANOVA by replacing the standard critical values for F with values that are larger and therefore more conservative.

The rationale for this modification is as follows. First, Mycroft et al. (2002) argue that when an individual patient is compared to a control sample, this should be conceptualized as a test for a difference in population means (i.e., a notional population of patients should be invoked). Second, they argue that, relative to the control population, the notional patient population will have markedly increased variance. Thus Mycroft et al. test whether the case comes from a notional population whose mean score differs from the mean score of the control population and explicitly assume that the variance that should be attached to the score of the case is bigger than the variance for controls.

Mycroft et al. (2002) also argue that if the increase in variance is ignored, there will be an inflation of the Type I error rate. (In this context a Type I error occurs when it is incorrectly concluded that a patient has a deficit, or, in Mycroft et al.'s terms, that the population means differ.)

Mycroft et al.'s (2002) use of modified F values deals with this perceived problem. They suggest that, in contrast, Crawford and Howell's (1998) method "fails to note the consequences of unequal variance ... and cannot be considered reliable when there are differences in variability between patients and controls" (p. 294). That is, they argue that Crawford and Howell's method will not control the Type I error rate.

Crawford, Garthwaite, Howell, and Gray (2004) have argued that Mycroft et al.'s (2002) concern over Type I errors is misplaced (this issue is returned to in a later section). They also suggested that (a) in general, the statistical power to detect an acquired deficit in single-case studies will be low (because an individual patient, rather than a sample, is compared to a control sample that will often itself have a modest *n*), and (b) Mycroft et al.'s method will produce a further (unnecessary) diminution of power and should thus be avoided.

Both these statements, however, were made without supporting empirical evidence. Therefore, in the present study, Monte Carlo simulation is used to quantify the extent to which Mycroft et al.'s (2002) suggested modification to Crawford and Howell's (1998) approach reduces the statistical power to detect a deficit. In the course of conducting this simulation it will also be possible to more broadly examine the issue of the statistical power to detect a deficit in single-case studies. Given that a prima facie case can be made that power will be low, it is surprising that little attention has been given to such a fundamental issue. One recent study (Crawford & Garthwaite, 2005a) looked at power in single-case studies but was concerned solely with the power to detect dissociations.

Before presenting the Monte Carlo simulation studies it is worth noting that the statistically sophisticated reader may wonder why simulation methods were used in preference to a direct analytical treatment of these issues. We had two reasons for this. First, although it would have been relatively straightforward to tackle some of the simpler scenarios under study using an analytic approach, this is not the case for some of the more complex scenarios (e.g., studying the effects of non-normal data in Study 2). Second, we took the view that empirical demonstrations of the level of statistical power in various scenarios would be more accessible to the nonstatistician. Finally, note that in all the simulations that follow, alpha for the tests of Crawford and Howell (1998) and Mycroft et al. (2002) was set at .05.

STUDY 1: STATISTICAL POWER TO DETECT A DEFICIT IN SINGLE-CASE STUDIES

As sample size is an important determinant of statistical power, power will almost inevitably be modest in single-case studies (Crawford, 2004; Crawford, Garthwaite, & Gray, 2003). As noted above, an individual patient (rather than a sample of patients) is compared to a control sample, and, furthermore, this sample will commonly have a modest *n*. An additional factor that serves to reduce power is the wide variability in cognitive abilities in the general population. A neurological patient's performance on a given cognitive task will reflect not only the effects of any insult but will also be strongly influenced by her/his premorbid competency (Crawford, 1992, 2004; Deary, 1995; Lezak, 1995). Take the example of a patient whose premorbid ability on a task was high relative to demographically matched controls (say 1 *SD* above the control mean). Any acquired deficit would need to be large before we have any realistic possibility of detecting it; for example, a 1-*SD* deficit would simply place the patient's postmorbid score exactly at the mean.

To an outside observer these points may lead to a pessimistic view of the prospects for the single-case enterprise. In reality, of course, deficits are routinely detected because the effect sizes in this area of enquiry can be very large; that is, neurological damage can have catastrophic effects on cognitive functioning.

In the first study we run a Monte Carlo simulation to examine the power of Crawford and Howell's (1998) method and that of Mycroft et al. (2002) to detect an acquired deficit. The bar is set low in this simulation and in those that follow on from it. That is, we take it as a given that the power to detect small-to-moderate effects will be very low and thus limit attention to the ability of the methods to detect deficits that are large (or very large) in magnitude but, nevertheless, not uncommon following neurological damage. This first simulation in part also serves as an introduction for the reader to the basic rationale and methods used in the simulations ahead of modelling more complex scenarios in later studies.

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 0 and 1), and these 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 values of n (the sample size of the control sample): 5, 10, 20, 50, and 100. For each of these values of n, 1,000,000 samples of $n + 1$ observations were drawn randomly from a standard normal distribution. On each Monte Carlo trial, the first n observations were used to represent the scores of the control sample, and the $(n + 1)$th observation was used to represent the single case. The single case was then "lesioned" to impose a (large) 2-*SD* deficit (as the observations were sampled from a standard normal distribution, this simply required subtracting 2 from the score).

In order to avoid any potential confusion it should be made explicit that this simulation procedure is designed to model patients with acquired deficits: It does not produce patients with scores that are simply 2 *SD*s below the mean of controls. Rather, the method recognizes that (a) patients are initially members of the healthy control population until the onset of their lesion, and (b) there will be premorbid differences in competency on the task.

Having sampled observations to represent controls and a single case, Crawford and Howell's (1998) test was then applied to these data; t values that were negative (i.e., where the score of the single case was below the control sample mean) and exceeded the one-tailed critical value for t ($\alpha = .05$) on the appropriate degrees of freedom $(n - 1)$ were recorded. That is, we recorded whether Crawford and Howell's method had successfully detected that the case had a deficit.

Two versions of Mycroft et al.'s (2002) method (hereafter labelled as intermediate and extreme) were also applied to the same data. Mycroft et al. tabulated a range of modified critical values for F; in practice the value employed would depend on a user's estimate of the extent to which the standard deviation of the notional population of patients would be larger than that of the controls. Critical values were provided to cover estimates ranging from 1.25 times larger through to 5 times larger. We applied an intermediate critical value (notional patient population $SD = 2.5$ times that of controls) and an extreme critical value ($SD = 5$ times that of controls). These critical values are for a two-tailed test (in keeping with Mycroft et al.'s advocacy of a two-tailed test, modified critical values for a one-tailed test were not tabulated in their paper).

Finally, we then repeated the procedure described but imposed a (very large) 3-*SD* deficit on the case. Thus, in total, 10 million Monte Carlo trials were run; 1 million trials for each of the five sample sizes combined with the imposition of either a 2- or a 3-*SD* deficit on the single case.

Results and discussion

The full results of the Monte Carlo simulation are presented in Table 1; the basic pattern of these results can readily be assimilated by referring to Figure 1, which plots power to detect a 2-*SD* deficit as a function of control sample n and the method employed. It can be seen that, for Crawford and Howell's (1998) test, the statistical power to detect a large (2-*SD*) deficit ranges from low (when the control n is small; minimum power = 44.83%) to moderate (for larger n; maximum power = 63.0%).

In contrast, power is very low for the intermediate version of Mycroft et al.'s (2002) method (ranging from 11.14% to 13.33%) and is extremely low for the extreme version of their method (ranging from 1.07% to 4.16%). Two factors contribute to the very low power of Mycroft et al.'s method relative to that observed for Crawford

Table 1. *Simulation results: Power[a] to detect a deficit as a function of control sample size, size of deficit, and inferential method*

		Crawford and Howell	*Mycroft et al. (2002)*	
Deficit[b]	*n*	*(1998)*	*Intermediate*	*Extreme*
2	5	44.83	11.14	4.16
	10	54.63	12.51	2.56
	20	59.46	13.09	1.74
	50	62.06	13.33	1.15
	100	63.00	13.33	1.07
3	5	72.80	25.90	10.77
	10	83.87	34.86	10.12
	20	87.88	40.12	9.36
	50	90.11	43.19	8.77
	100	90.73	44.81	8.92

[a] In percentages. [b] Number of standard deviations.

and Howell's (1998) method. First, because of Mycroft et al.'s concern over inflation of the Type I error rate, they employ conservative critical values. Second, their test is two-tailed, whereas Crawford and Howell employ a one-tailed test.

Although it was appropriate to incorporate both these features in the simulation (i.e., the procedure recommended by Mycroft et al., 2002, should be faithfully implemented), it would have been useful to study their effects in isolation. To achieve this, the simulation was rerun, substituting two-tailed for one-tailed critical values for Crawford and Howell's (1998) test. In this scenario, the only difference between the two methods lies in Mycroft et al.'s use of modified (i.e., conservative) critical values. For the two-tailed version of Crawford and Howell's method power ranged from 29.04% for a *n* of 5 to 50.43% for a *n* of 100. Although it can be seen that power has been lowered by the use of a two-tailed test, these percentages are still very much higher than the equivalent percentages for the intermediate and extreme versions of Mycroft et al.'s method. Hence it is clear that the modified critical values must carry most of the blame for the low levels of power of Mycroft et al.'s test.

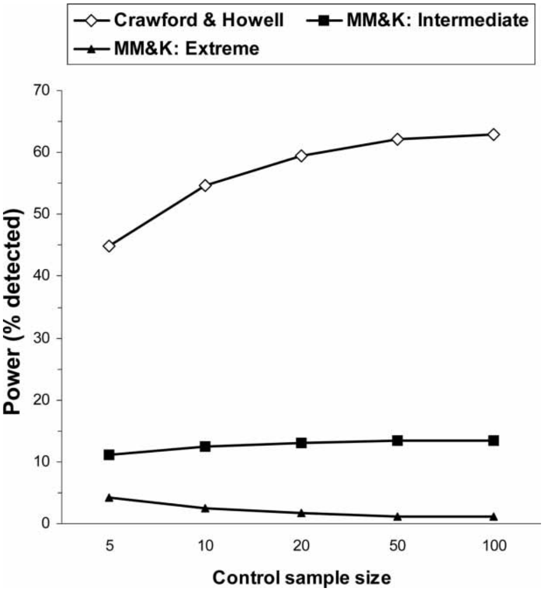

Figure 1. *Power to detect a large (2-SD) deficit as a function of control sample size and method employed.*

In the present simulation, the power of Crawford and Howell's (1998) and Mycroft et al.'s (2002) methods to detect a very large (3-*SD*) deficit was also examined. It can be seen from Table 1 that for Crawford and Howell's method power is high (above 80%, with the exception of a control sample *n* of 5) and is over 90% for control sample *n*s of 50 and 100. However, it must be stressed that control samples of this size are very rare in single-case studies, and, moreover, a 3-*SD* deficit represents a very severe, catastrophic, impairment.

In contrast, for Mycroft et al.'s (2002) method, although power is obviously higher for a 3- rather than a 2-*SD* deficit, it remains the case that in absolute terms power is still low. For the intermediate version, maximum power observed was 44.81% (for a control sample size of 100); for the extreme version, maximum power was 10.77% (for a sample size of 5).

Finally, the present results have some additional implications for the conduct of single-case research. For Crawford and Howell's (1998) method it can be seen that, for both 2- and 3-*SD* deficits, power increases appreciably with increasing *n* up to an *n* of 20. Thereafter, although (as expected) power continues to increase, it is subject to diminishing returns; for example, for a

2-*SD* deficit, power is 59.46% for an *n* of 20 but only rises to 63% when *n* is increased by a factor of 5 to 100. This suggests that, in the interests of achieving reasonable power to detect large effects, control sample *n*s should be larger than those employed in most existing single-case studies (where *n*s in the range of 3 to 15 are most typical) but that recruiting *n*s > 30 is not liable to be worth the additional expenditure of effort.

STUDY 2: THE EFFECTS OF DEPARTURES FROM NORMALITY ON POWER TO DETECT A DEFICIT

An assumption underlying the use of both Crawford and Howell's (1998) method and that of Mycroft et al. (2002) 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, 2005b).

Ideally, researchers would carefully select the measures they employ in single-case studies so as to avoid potential problems arising from non-normal control data. However, for many published single-case studies, it is clear from even a cursory inspection of the control sample means and standard deviations that the control data are negatively skewed. That is, the standard deviations tell us that, were the data normally distributed, a substantial percentage of scores would lie above the maximum obtainable score on a particular task, yet we know that this is impossible; hence the data must be heavily skewed (Crawford & Garthwaite, 2005a).

Skew will be almost inevitable when the tasks employed measure abilities that are largely within the competence of most healthy individuals. In this situation, negative skew will occur when the measure of interest is based on the number of items passed (i.e., there will be ceiling effects) and positive skew when the measure is an error rate (i.e., there will be floor effects). Evidence of severely skewed control data can be found in the literature on recognition of facial expression of emotion (Milders, Crawford, Lamb, & Simpson, 2003) and in the extensive single-case literature on category-specific object naming. For example, in a recent review of single-case studies of the living versus nonliving distinction, it was reported that the accuracy of naming in controls was in excess of 95% in the vast majority of these studies (Laws, Gale, Leeson, & Crawford, 2005).

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 would a normal distribution; that is, the control data will be leptokurtic (it follows from the fact that leptokurtic distributions are more peaked and have heavier tails that they also have thinner "shoulders" than a normal distribution). Leptokurtic distributions are pervasive in many areas of scientific enquiry including psychology, economics, and biology (DeCarlo, 1997; Lange, Little, & Taylor, 1989). For example, IQ tests are 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, measured IQ commonly exhibits highly significant leptokurtosis (Burt, 1963).

Single-case researchers also have to face the possibility that their control data may have both these aforementioned features simultaneously. That is, the control data may be skewed and leptokurtic. Indeed, it is likely that control data more commonly possess both these characteristics rather than either alone. As noted, 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; this will produce negative skew and leptokurtosis (i.e., the distribution will be overly peaked as scores will accumulate at, or near, the maximum possible score).

The effects of non-normal control data on inferential methods for detecting a deficit in single-case studies have recently been examined using Monte Carlo methods. These studies, however, were solely concerned with examining the effects on Type I error rates; that is, they estimated the percentage of the (cognitively intact) control population that would incorrectly be identified as exhibiting a deficit. The results were, in general, fairly reassuring for single-case researchers.

For Crawford and Howell's (1998) method, the presence of skew raised the Type I error rates above the specified error rate (of 5%) but the effects were by no means catastrophic, even when skew was very extreme. For example, when skew was severe ($\gamma_1 = -0.7$) the error rate for a control sample size of 20 was 6.95%. Similarly, when the distribution of scores in the control population was leptokurtic, error rates were not seriously affected. The combination of skew and leptokurtosis produced the most serious inflation of the error rate but, even so, Crawford and Howell's method was more robust than might have been anticipated; for example, when severe skew was combined with severe leptokurtosis, the error rate was 7.45% for a control sample size of 20.

As noted, these studies were solely concerned with examining control of the Type I error rate. To our knowledge, there have been no previous attempts to quantify the effects of departures from normality on the power to detect deficits in single-case studies (i.e., the extent to which inferential methods avoid committing Type II errors has not been examined). Therefore, in Study 2 we quantify the effects of skewed and/or leptokurtic control data on the power of Crawford and Howell's (1998) and Mycroft et al.'s (2002) methods to detect a deficit.

Method

Simulations were run using a similar approach to that employed in Study 1—that is, 1,000,000 samples of $n + 1$ observations were drawn for five different sample sizes, and a deficit was applied to the single case. However, instead of sampling observations from a normal distribution, observations were sampled from distributions that were skew, leptokurtic, or both.

Sampling from leptokurtic distributions

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 (1998) 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 (and Mycroft et al.'s, 2002, test) 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) *df*. Kurtosis (β_1) is 5 for a *t* distribution on 7 *df* compared to a value of 3 for a normal distribution; the kurtosis for a *t* distribution on 4 *df* is even more extreme but is undefined (because the denominator in the formula for kurtosis requires subtracting 4 from the *df* and is hence zero).

To sample from these distributions, observations representing the controls and the single case were sampled initially from a normal distribution. Each observation was then divided by $\surd(\chi^2/7)$ or $\surd(\chi^2/4)$ where χ^2 is a random draw from a chi-square distribution on 7 or 4 *df*, respectively. The resultant quantities are observations from *t* distributions on 7 or 4 *df*; that is, they are observations that are drawn from moderately or severely leptokurtic distributions.

Sampling from skew distributions

Three negatively skewed distributions were specified, ranging from a distribution with moderate skew ($\gamma_1 = -0.31$), through one with severe skew (-0.70), to one with extreme skew (-0.99). The method used to sample observations representing controls and the single case from these distributions (termed skew-normal distributions) was that of Azzalini and colleagues (Azzalini & Capitanio, 1999; Azzalini & Dalla Valle, 1996); the technical details are presented in Appendix B.

Sampling from distributions that are both skew and leptokurtic

To sample from distributions that possessed both skew and leptokurtosis we followed the procedure outlined in the section on skew distributions but, after obtaining skew-normal observations, these observations were then divided by $\sqrt{(\chi^2/7)}$ (moderate leptokurtosis) or $\sqrt{(\chi^2/4)}$ (severe leptokurtosis) where χ^2 was a random draw from a chi-square distribution on 7 or 4 *df*, respectively. The resultant distributions are skew *t* distributions; they depart from a normal distribution in that they are both leptokurtic and (negatively) skewed (Azzalini & Capitanio, 2003).

Graphical illustration of some of the distributions used are presented as Figure 2; the shaded areas show the densities for distributions possessing both skew and leptokurtosis (i.e., skew-*t* distributions), the unshaded lines show the densities for the equivalent distributions possessing skew alone (i.e., skew-normal distributions). In all cases the distributions have been rescaled to have a variance of 1 (so that visual comparison of their shapes is meaningful).

Imposing a deficit on the single case

In Study 1, a 2- or 3-*SD* deficit could be applied to the score of the single case by simply subtracting 2 or 3 from their initial score, because observations were drawn from a standard normal distribution. Things are a little more complicated in the present study because, as alluded to in the last section, the standard deviations of the non-normal distributions are not 1; the quantity subtracted therefore differed according to the distribution used. For example, the standard deviation of a *t* distribution on 4 *df* (used to represent severe leptokurtosis) is 1.4142. Therefore, in this case, to impose a 2-*SD* deficit we subtracted 2.8284 from the single-cases' initial scores. Similarly, the standard deviation of a skew-normal distribution with $\gamma_1 = -0.99$ (used to represent extreme negative skew) is 0.6035, and, therefore, 1.207 was subtracted from the single-cases' initial scores.

Because of the larger number of simulations involved in the present study we limited attention to a 2-*SD* deficit. In total 60 million Monte Carlo trials were run: 1 million trials for each combination of five sample sizes, four levels of skew (absent to extreme), and three levels of leptokurtosis (absent to severe).

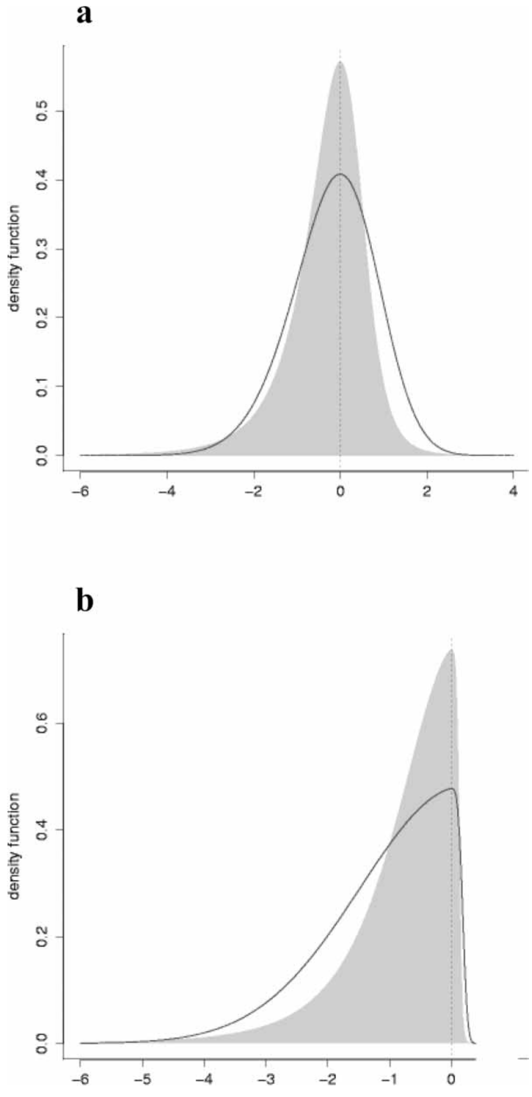

Figure 2. *Graphical illustration of some of the distributions employed in Study 2; the shaded area shows the density for distributions possessing both skew and leptokurtosis (skew-t), and the unshaded line shows the density for the equivalent distributions possessing skew alone (skew-normal). (a) =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

Results and discussion

The results of the simulations are presented in Table 2. This table presents power to detect a large (2-*SD*) deficit as a function of the inferential method applied, the control sample size, degree of skew, and degree of leptokurtosis. The first block of this table reproduces the results from Study 1—that is, it presents power when the control population distribution was neither skew nor leptokurtic; these data provide a comparison standard against which to assess the effects of departures from normality.

Attending first to the results for Crawford and Howell's (1998) method when the control distribution is leptokurtic, it can be seen that power is higher than when normality holds. However, the effects are relatively modest; for example, for a control sample of size 20, the power is 59.46% for a normal distribution but is 64.11% when the control distribution is severely leptokurtic. This pattern of results arises because of the heavy tails; more controls are already in the lower tail of this distribution than would be in the tail of a normal distribution, and so when a deficit is imposed (i.e., they move from being a previously cognitively intact, healthy control to being a patient with a deficit) they are more likely to be detected.

In contrast to the results for leptokurtosis, negative skew lowers the power of Crawford and Howell's (1998) method to detect a deficit, although again the results are not dramatic. For example, when skew is extreme, power is 53.88% for a control sample size of 20 compared to 59.46% for a normal distribution. This pattern of results can be attributed to the fact that when negative skew is present, more controls are found in the upper region of the distribution, so that when a deficit is imposed, their resultant score (i.e., their premorbid score minus the 2-*SD* deficit) is not sufficiently low to be detected.

When the control distribution is both skew and leptokurtic, as may be common in single-case research, the effects observed when distributions feature either of these characteristics alone tend to cancel each other out. The net result is that, as can be seen from Table 2, power to detect a deficit when distributions are both skew and leptokurtic does not differ to any great extent from power when the control distribution is normal. For example, power in the face of extreme skew and severe leptokurtosis is 63.51% for a control sample size of 20, compared to 59.46% when normality holds. Note, however, that there is an interaction between the presence of skew and leptokurtosis and sample size: With large sample sizes power is marginally lower than that for a normal distribution, whereas with very small sample sizes, power is appreciably higher (e.g., power is 58.32% for extreme skew and severe leptokurtosis for a control *n* of 5 compared to 44.83% for a normal distribution).

Turning to the results for Mycroft et al.'s (2002) method, it can be seen from Table 2 that, in both its intermediate and extreme versions, leptokurtosis raises the power to detect a deficit, but the effects are relatively modest, and power remains very low in absolute terms (e.g., power is 16.18% when leptokurtosis is severe for a control sample size of 20 compared to 13.09% for a normal distribution). For skew, it can be seen that, in contrast to the results for Crawford and Howell's (1998) method, power is also higher than that observed for a normal distribution. The effects are quite large, particularly for the extreme version of their method, although, again, power is still very low in absolute terms.

When the combination of skew and leptokurtosis is examined it can be seen that the effects are additive; in general power is markedly higher than when the control distribution is normal. For example, power is 21.93% for a control sample *n* of 20 for the intermediate version, compared to 13.09% for a normal distribution. Again, however, power remains low in absolute terms even in these circumstances and is well below that observed for Crawford and Howell's (1998) method. For Mycroft et al.'s (2002) method it can also be seen that the increase in power is related to sample size; the effects are marked for small sample sizes but become marginal for large *n*s.

Table 2. *The effects of departures from normality*[a] *on power to detect a 2-SD deficit*

		No skew			*Moderate skew*			*Severe skew*			*Extreme skew*		
	n	*C&H*	*M-I*	*M-E*	*C&H*	*M-I*	*M-E*	*C&H*	*M-I*	*M-E*	*C&H*	*M-I*	*M-E*
No leptokurtosis	5	44.83	11.14	4.16	44.90	12.35	4.87	45.17	14.18	6.18	45.37	15.82	7.48
	10	54.63	12.51	2.56	53.60	13.58	3.28	52.22	14.96	4.42	51.34	16.01	5.34
	20	59.46	13.09	1.74	57.86	13.81	2.40	55.67	14.79	3.34	53.88	15.38	4.04
	50	62.06	13.33	1.15	60.46	13.73	1.80	57.70	14.32	2.54	55.22	14.88	3.09
	100	63.00	13.33	1.07	61.45	13.57	1.65	58.44	14.15	2.40	55.51	14.54	2.92
Moderate leptokurtosis	5	47.77	13.32	5.36	48.47	15.82	6.95	49.73	18.23	8.82	50.67	20.38	10.63
	10	56.67	14.46	3.64	55.32	16.41	5.18	55.06	18.25	6.66	55.17	19.76	7.92
	20	60.89	13.94	2.56	58.40	15.47	3.92	57.24	16.59	5.00	56.53	17.46	5.79
	50	63.70	12.90	1.79	60.24	13.87	2.91	58.10	14.51	3.70	56.35	14.97	4.17
	100	64.67	12.32	1.59	60.94	12.99	2.63	58.26	13.44	3.29	55.98	13.81	3.64
Severe leptokurtosis	5	52.61	16.75	7.27	54.66	20.71	10.08	56.64	24.07	12.68	58.32	26.86	15.12
	10	60.51	17.67	5.23	60.24	21.09	7.81	61.37	23.69	9.88	62.59	25.90	11.62
	20	64.11	16.18	3.65	62.37	18.69	5.73	62.68	20.49	7.17	63.51	21.93	8.28
	50	66.23	13.79	2.40	63.00	15.27	3.97	62.29	16.26	4.84	62.53	17.07	5.39
	100	67.10	12.24	2.03	62.94	13.37	3.39	61.42	14.02	3.98	60.93	14.55	4.41

Note: Results are presented for Crawford and Howell's (1998) method (C&H), and the intermediate (M-I) and extreme (M-E) versions of Mycroft et al.'s (2002) method.

[a] Negative skew alone, leptokurtosis alone, or their combination.

In summary, for Crawford and Howell's (1998) method, skew and leptokurtosis, either alone or in combination, do not exert a marked effect on power to detect a deficit (the exception being for very small sample sizes when the degree of departure is very marked). Given that such departures from normality are liable to be common features of single-case studies, the present results, when taken with the corresponding results from studies examining Type I errors (Crawford & Garthwaite, 2005b; Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006), provide reassurance for researchers. It would appear that for Crawford and Howell's method, the effects of violating the assumption of normality are, in general, fairly modest; that is, the method is surprisingly robust. For Mycroft et al.'s (2002) method, departures from normality exert a greater influence. With small to moderately sized control samples power is increased markedly. However, given that power remains universally low in absolute terms for this method, these latter effects are of limited practical importance.

STUDY 3: THE EFFECTS OF MEASUREMENT ERROR ON POWER TO DETECT A DEFICIT

In introducing Study 1 it was noted that a number of factors conspire to lead to low power to detect a deficit in single-case studies. A further factor that will serve to reduce power is the presence of measurement error. From the perspective of classical test theory, an individual's observed score is an amalgam of their true score (the average score obtained if the individual was administered an infinite number of parallel versions of the task) and random measurement error. If a neurological insult reduces the true score by a given amount this true score deficit will not be faithfully reflected in the observed score. In the present study we conduct a Monte Carlo simulation to examine power to detect a large (2-*SD*) or very large (3-*SD*) deficit in the presence of measurement error.

We also quantify the size of deficit required to achieve 80% power to detect a deficit as a function of control sample size, degree of measurement error, and inferential method. A criterion of 80% power is widely taken as indicating a high, or at least acceptable, level of statistical power for group studies (e.g., in clinical trials, etc.). It is therefore of interest to apply this yardstick to single-case research and to examine the relative contributions of control sample size, measurement error, and method to determining the size of deficit required. Such an analysis is in keeping with our general aim of subjecting single-case methods to a degree of scrutiny similar to that applied to group-based research. It was impractical to study this latter issue using simulation, and therefore the problem is tackled directly using an analytic approach.

Method

Simulation study

A modified version of the simulation procedure described in Study 1 was used to quantify the effect of measurement error on the power to detect a deficit. That is, five sample sizes were used (5, 10, 20, 50, and 100), and on each trial an additional observation was drawn to represent the single case; a 2- or 3-*SD* deficit was then applied to the single case. As in Study 1, sampling of control cases and the single case was from a standard normal distribution; in the present simulation these observations are used to represent the true scores of controls and the case. For each control and the single case, a further random draw was then made from a normal distribution having a mean of zero and a variance corresponding to the degree of measurement error required (these observations represented the error scores for controls and the case).

For example, to model the effects of a task reliability of .6, the variance of this latter distribution was set at 0.66667. These scores were then added to the true scores of controls and the single case to obtain observed scores. (Given that the true scores had a variance of 1, it can be seen that, in this example, the population variance of observed scores is 1.66667, and the required 60:40 ratio of true score variance to error variance is thereby achieved.) It should be noted that the 2- or 3-*SD* deficit was imposed on the true score rather than the observed score.[1]

As in previous simulations, on each Monte Carlo trial the three methods of testing for a deficit were applied (i.e., Crawford & Howell's, 1998, method and the intermediate and extreme versions of Mycroft et al.'s, 2002, method), and the percentage of cases correctly identified was recorded.

Analytic approach to quantifying the size of deficit required for 80% power

Looking first at Crawford and Howell's (1998) method, a deficit is detected in the case if, using the notation of Equation 1,

$$\frac{\bar{x} - x^*}{s\sqrt{(n+1)/n}} > t_{n-1;\alpha} \qquad (2)$$

where $t_{n-1;\alpha}$ is the critical value of a t distribution on $(n - 1)$ *df* for a one-sided hypothesis test with significance level α. Suppose the case has a deficit of η, and the reliability is r_{xx}. Also, let $t_{n-1}(\delta)$ denote a variate that has a noncentral t distribution on $(n - 1)$ *df* with noncentrality parameter δ. Let δ^* be the value of δ for which

$$\Pr(t_{n-1}(\delta^*) > t_{n-1;\alpha}) = 0.8 \qquad (3)$$

In Appendix C we show that $\eta^* = \delta^*\sqrt{[(n + 1)/(r_{xx}n)]}$ is the minimal deficit that Crawford and Howell's (1998) method detects with a power of 0.8.

The test of Mycroft et al. (2002) assumes that the case comes from a population whose variance differs from the variance of the control population. Define k as the ratio, $k =$ (variance of case)/(variance of controls). In Appendix C we show that if δ^* is the value of δ for which

$$\Pr\left(t_{n-1}(\delta^{\#}) > t_{n-1,\alpha/2}\sqrt{\frac{kn+1}{n+1}}\right) + \Pr\left(t_{n-1}(\delta^{\#}) < -t_{n-1,\alpha/2}\sqrt{\frac{kn+1}{n+1}}\right) = 0.8, \qquad (4)$$

then $\eta^{\#} = \delta^{\#}\sqrt{[(n + 1)/(r_{xx}n)]}$ is the minimal deficit that the test of Mycroft et al. detects with a power of 0.8. The test of Mycroft et al. is two-tailed so Equation 4 contains two probabilities on its left-hand side and uses critical values for a significance level of $\alpha/2$ rather than α. While deriving these formulae, other theoretical results were developed that relate to the test of Mycroft et al. In particular, the Appendix gives a formula for the exact critical values of the test (Mycroft et al. estimated critical values by simulation and only tabulated critical values for limited values of k and n).

Results and discussion

The simulation results obtained when measurement error was present are presented in Table 3; the results obtained in the absence of measurement error (i.e., when $r_{xx} = 1$) are also incorporated. The basic pattern of results can readily be appreciated by referring to Figure 3. This figure presents power for the methods as a function of the reliability of scores; the values plotted are limited to those obtained for a control sample n of 20.

It can be seen that for Crawford and Howell's (1998) method, measurement error exerts an appreciable effect on the power to detect a deficit; for example, power was 59.46% in the absence of measurement error for a n of 20 but falls to 53.64% when task reliability was .85 and falls to

[1] Some readers will realise that it was unnecessary to sample separately from true score and error distributions. For instance, in the example just given, the simulation could have been run simply by sampling from a single normal distribution with a variance of 1.66667 (*SD* = 1.291). However, in the former approach it is made explicit that we are modelling the effects of varying degrees of measurement error and that the deficit is imposed on the true score (i.e., subtracting 2.0 from the case's score imposes a 2-*SD* deficit on the true score). In other words, the approach was used for didactic purposes.

Table 3. *Simulation results: Power to detect a deficit as a function of control sample size, reliability of scores, size of deficit, and inferential method*

		r_{xx}											
		Crawford and Howell (1998)				*Mycroft et al. (2002): Intermediate*				*Mycroft et al. (2002): Extreme*			
Deficit[a]	*n*	*1.0*	*.85*	*.7*	*.6*	*1.0*	*.85*	*.7*	*.6*	*1.0*	*.85*	*.7*	*.6*
2	5	44.83	40.25	35.45	32.02	11.14	9.51	7.93	6.88	4.16	3.48	2.88	2.47
	10	54.63	49.23	43.26	38.99	12.51	10.25	8.13	6.72	2.56	1.98	1.50	1.20
	20	59.46	53.64	47.16	42.62	13.09	10.43	7.95	6.47	1.74	1.25	0.86	0.66
	50	62.06	56.16	49.55	44.73	13.33	10.45	7.82	6.22	1.15	0.80	0.52	0.39
	100	63.00	57.11	50.29	45.42	13.33	10.35	7.64	6.09	1.07	0.71	0.45	0.32
3	5	72.80	66.69	59.69	54.35	25.90	21.73	17.73	15.08	10.77	8.82	7.04	5.85
	10	83.87	78.50	71.31	65.57	34.86	28.44	22.29	18.24	10.12	7.57	5.43	4.15
	20	87.88	83.09	76.26	70.59	40.12	32.33	24.78	20.00	9.36	6.66	4.37	3.18
	50	90.11	85.46	78.99	73.41	43.19	34.92	26.46	21.03	8.77	5.77	3.57	2.41
	100	90.73	86.20	79.84	74.22	44.81	35.68	26.86	21.28	8.92	5.74	3.43	2.31

Note: r_{xx} = reliability of scores.
[a]Number of standard deviations.

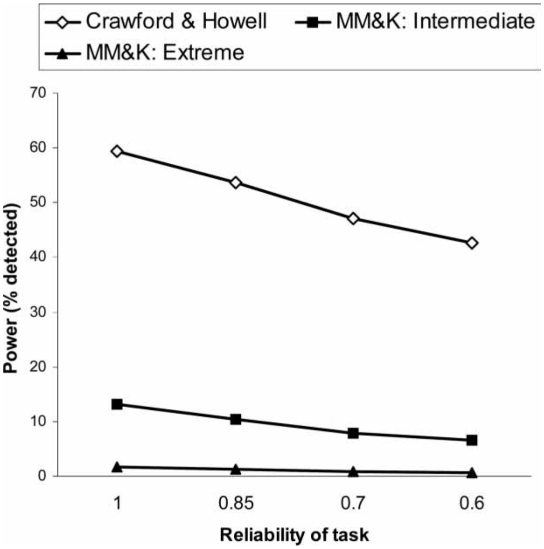

Figure 3. *Power to detect a large (2-SD) deficit as a function of task reliability and method employed (n for these data = 20).*

42.62% when reliability was .6. However, it can also be said that the effect of measurement error is by no means catastrophic; a moderate level of power is retained if the control sample is of a reasonable size. Furthermore, it is to be hoped that single-case researchers will take sufficient care in selecting or developing their tasks, such that reliabilities below 0.7 will be rare in practice.

It can be seen that power is also reduced when measurement error was present for Mycroft et al.'s (2002) method although, particularly for the extreme version, the effects are attenuated because of the low baseline rate. Nevertheless, it can also be seen that even if measurement error is relatively modest, power can fall below 1% (for the extreme version, power was below 1% in 8 of the 20 scenarios examined). These levels of power are extraordinarily low; for example, with a control sample *n* of 100 and a task reliability of .7, it can be estimated that 99.55% of patients with large (2-*SD*) acquired deficits would be missed.

Analytic approach to quantifying the size of deficit required for 80% power

The results of the power analysis are presented in Table 4. This table lists the size of deficit, in standard deviation units, required for 80% power as a function of sample size, degree of measurement error, and inferential method. To illustrate, for Crawford and Howell's (1998) method, with a control sample size of 20, and a task reliability of .8, a score that was 2.87 *SD*s below individuals' premorbid scores would be required to achieve 80% power to detect an acquired deficit.

These results complement those presented in Table 3 and demonstrate that, regardless of the method employed, the probability of detecting a deficit will only be high if the deficit is very large. Deficits are only routinely detected because neurological illness or disease can have catastrophic (i.e., very large) effects on cognition. In passing, note that using the analytic approach for Crawford and Howell's (1998) method with a task reliability of .7 and control sample size of 100, a deficit of approximately 3 *SD*s (3.01) is required for 80% power to detect a deficit. This accords very closely with the results of the Monte Carlo simulation in which 79.8% of cases were detected when a 3-*SD* deficit was imposed.

Although, as noted, the deficits required for high power can all be classified as very large, it can also be seen from Table 4 that there are nevertheless marked differences in the size of deficit required for 80% power as a function of the method employed. For Crawford and Howell's (1998) method, the size of deficit required ranged from 2.52 *SD*s (for a control sample size of 100 combined with no measurement error) to 4.30 *SD*s (for a control sample size of 5 and task reliability of .6). For the extreme version of Mycroft et al.'s (2002) method, the equivalent deficits were 5.32 and 10.28 *SD*s, respectively.

Finally, we should stress that the concern in this study (and in those that preceded it) was with the size of an acquired deficit or impairment, not with a case's obtained score (a case's obtained score is a function of the imposed deficit and the case's premorbid ability). Thus, for example, for Crawford and Howell's (1998) method with a control sample size of 20, any cases with an obtained score of 1.772 *SD*s or more below the mean of the control sample will be recorded as exhibiting a deficit (t will be $\geq$ 1.729, and hence the

Table 4. *Size of deficit required*[a] *to achieve 80% power to detect a deficit as a function of control sample size, reliability of scores, and inferential method*

	r_{xx}											
	Crawford and Howell (1998)				*Mycroft et al. (2002): Intermediate*				*Mycroft et al. (2002): Extreme*			
n	*1*	*.85*	*.7*	*.6*	*1*	*.85*	*.7*	*.6*	*1*	*.85*	*.7*	*.6*
5	3.33	3.61	3.98	4.30	5.87	6.37	7.02	7.58	7.97	8.64	9.52	10.28
10	2.83	3.07	3.38	3.65	4.68	5.07	5.59	6.04	6.29	6.83	7.52	8.13
20	2.65	2.87	3.16	3.41	4.27	4.63	5.11	5.51	5.71	6.20	6.83	7.37
50	2.55	2.76	3.04	3.29	4.07	4.55	4.86	5.25	5.41	5.87	6.47	6.99
100	2.52	2.73	3.01	3.25	4.00	4.34	4.78	5.17	5.32	5.77	6.35	6.86

Note: r_{xx} = reliability of scores.
[a]Number of standard deviations.

one-tailed p is $< .05$); that is, the power to detect such cases as exhibiting a deficit is 100%. However, for many cases—that is, those of high premorbid ability on the task in question—even a substantial (e.g., 2-*SD*) deficit will not be sufficiently large to produce an obtained score that is 1.772 *SD*s or more below the mean of the control sample.

STUDY 4: THE EFFECTS OF INCREASED MEASUREMENT ERROR FOR THE SINGLE CASE ON POWER TO DETECT A DEFICIT

Crawford and Howell's (1998) method poses the following question: Is the patient's score sufficiently below those of the controls to allow us to reject the null hypothesis that the patient is an observation from the control population? Therefore, for this method, it is neither necessary nor appropriate to be concerned with a notional patient population.

Crawford and colleagues were primarily motivated to adopt this perspective by statistical considerations. However, a number of influential theorists have come to exactly the same conclusion based on neuropsychological considerations (e.g., Caramazza, 1986; Caramazza & McCloskey, 1988; Coltheart, 2001). It is argued that because (a) the functional architecture of cognition is enormously complex, and (b) there is substantial variability in the site and extent of naturally occurring lesions, each single case should be considered to be unique (Vallar, 2000). McCloskey (1993) provides an unequivocal expression of this position when he states, "In the single-patient approach, patients are not identified as members of patient populations" (p. 729).

In contrast to Crawford and colleagues approach, Mycroft et al. (2002) require a notional patient population so that the variance that should be associated with the score of the case is defined. They suggest that this population should consist of patients who are "equivalent" (p. 295) to the patient of interest (Mycroft et al., 2002, p. 295), but there has been subsequent debate on what the term "equivalent" should mean in this context (Crawford et al., 2004; Mitchell, Mycroft, & Kay, 2004).

Further discussion may not resolve this issue so it is worth searching for an alternative conception that would be less contentious. One possible way forward is to move from considering the variance of a hypothetical population of patients to considering the variance of an individual case's scores. This is the position that Mitchell et al. (2004) appear to take at some points in their original paper; that is, they shift emphasis from variability between patients to intraindividual variability in performance.

Therefore we should consider the case in which the task performance of a patient with neurological damage will be less reliable than that of matched controls; that is, the patient's scores within a test session (or across sessions) will contain more measurement error than would those of control participants. This can be reasonable, provided that the loss of reliability occurs in the context of impaired performance. That is, allowing different reliabilities for the case and controls is inappropriate when considering Type 1 errors, but can be reasonable when considering Type 2 errors or evaluating power; see later section. In Study 4 we model the effects of a patient's score containing more measurement error than controls on statistical power in single-case research. As was the case for the previous studies, we believe that the study of this issue has implications for single-case research that extend well beyond the specific comparison of the two methods.

Method

Simulations were run using a similar approach to that employed in Study 1—that is, 1,000,000 samples of $n + 1$ observations were drawn for five different sample sizes. However, in this simulation the reliability was set at .85 for the scores of controls on all Monte Carlo trials but was varied for the scores of the single-cases: Three levels of reliability were incorporated (.7, .6, and .5). Thus, for example, when task reliability was .7 for a single case, the error attached to a case's score was sampled from a normal distribution with a mean of zero and variance of 0.4286; this error was then added to the true score to obtain the observed score. As in Study 3, the deficit was applied to the true scores of the single cases rather than their observed scores (i.e., it represents the true level of deficit). In this simulation we limit attention to a 2-*SD* deficit.

To summarize, in this simulation the cases have a deficit (and the concern is to study how successfully such cases can be identified; i.e., it is a power study) but, unlike Study 3, the cases also have more variable performance than controls.

Results and discussion

The full results of the simulation are presented in Table 5. The main focus of interest is the effect on power of imposing differential amounts of measurement error on the single cases. Therefore, it is necessary to compare the present results against those obtained when error was present in controls and cases in equal measure. This is provided by the results obtained from Study 3 in which the reliabilities of the scores of controls and the single cases were both set at 0.85. To avoid the reader having to switch between two tables, these results are reproduced as the first column of results for each method in Table 5.

It can be seen from Table 5 that for Crawford and Howell's (1998) method, when the scores of the single cases are more variable (i.e., unreliable)

Table 5. *Simulation results: Power to detect a 2-SD deficit as a function of control sample size, reliability of patient's scores (r_{xx}) and inferential method*

	r_{xx}											
	Crawford and Howell (1998)				*Mycroft et al. (2002): Intermediate*				*Mycroft et al. (2002): Extreme*			
n	*.85*	*.7*	*.6*	*.5*	*.85*	*.7*	*.6*	*.5*	*.85*	*.7*	*.6*	*.5*
5	40.25	40.73	41.06	41.54	9.51	10.28	10.87	11.70	3.48	3.87	4.22	4.63
10	49.23	49.21	49.20	49.32	10.25	11.59	12.64	14.18	1.98	2.50	2.96	3.64
20	53.64	53.09	53.09	52.73	10.43	12.11	13.64	15.36	1.25	1.75	2.32	3.13
50	56.16	55.62	55.29	54.84	10.45	12.48	14.11	16.17	0.80	1.33	1.88	2.76
100	57.11	56.36	55.91	55.45	10.35	12.45	14.21	16.38	0.71	1.25	1.85	2.73

Note: r_{xx} = reliability of scores. The reliability of scores in the control sample was set at .85 for all these simulations.

the effect on power to detect a deficit is marginal; this holds even when the differential is large (i.e., a reliability of .85 for controls vs .5 for the single cases).

In contrast, it can be seen that for Mycroft et al.'s (2002) method, in both its extreme and intermediate forms, power increases consistently as measurement error for the case increases (although power is still very low in absolute terms in all scenarios). This latter result runs counter to the broad principle that random measurement error will lower power (Schmidt & Hunter, 1996). These results illustrate that the effects of lower reliability of the scores of single cases (i.e., more variable performance) on the power to detect a deficit is conditional upon the baseline level of power obtainable when the scores of the single case possess the same level of reliability as controls.

With both the test of Mycroft et al. (2002) and the test of Crawford and Howell (1998), the test statistic does not involve the variance of the case's score. Consequently, an increase in the random variability of a case's score will change results in both directions. That is, some occasions when a Type II error would have occurred will now result in a deficit being detected, and vice versa. Moreover, which effect is the greater is largely determined by whether there are originally more Type II errors or more occasions when a deficit would have been detected. When the baseline power is below 50%, as is the case for Mycroft et al.'s method (and would be for Crawford and Howell's method if the effect size to be detected, i.e. the deficit, was smaller in magnitude), power will be increased. In contrast, when power is above 50% (e.g., if the effect size was very large) the expected pattern will then be observed—that is, power will be reduced. From Table 5 we can see this beginning to kick in for Crawford and Howell's method (e.g., with a control *n* of 50 power was 56.16% when control and patient reliabilities were equivalent but falls to 54.84% when the differential is at its largest). When baseline power is at or around 50%, as is the case for Crawford and Howell's method in many of the scenarios examined, the effects are marginal.

As far as we are aware this paradoxical effect of measurement error (i.e., that power can be raised when measurement error is increased) has not been discussed in the existing literature on power. This is presumably because the general topic of differential reliability across individuals or samples (as opposed to as across tasks) has itself received little attention and because such effects will only be obvious when power is low in absolute terms. In any event, these results serve to illustrate a broader point: General principles that can normally be relied on to provide researchers with sound guidance can fail when applied to single-case research.

Type I errors when testing for a deficit

The focus on these simulations has been on the power to detect a deficit. In contrast, the issue of control over the Type I error rate has not been addressed (in the present context Type I errors occur when cases that do not have a deficit are classified as exhibiting a deficit). The reasons for this is that theory tells us that Crawford and Howell's (1998) method should control Type I errors in the absence of differentially greater measurement error for cases; this was confirmed by a Monte Carlo simulation conducted by Crawford and Garthwaite (2005b). We should note, though, that Mycroft et al. (2002) examined Type I errors for the scenario in which cases are more variable than controls. Their results from Monte Carlo simulation suggested that large percentages of single cases would be incorrectly identified as exhibiting a deficit. However, it can be logically argued that Mycroft et al.'s concern over inflation of the Type I error rate is groundless. Specifically, the results of their simulation are not informative because of the unrealistic assumptions made.

As noted, Mycroft et al. (2002) consider that when a single case is compared to a control sample we should invoke a notional population of patients and treat the situation as one in which we are testing for a difference in population means. Although we see things differently, we can explore the implications of such a view. The best way to illustrate Mycroft et al.'s position

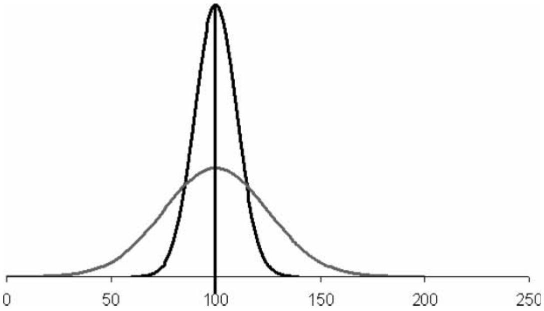

Figure 4. *Control and patient distributions in which there is a common mean but different standard deviations (note that a common mean is essential for Type I errors to be an issue).*

is with the help of Figure 4. This figure plots the distribution of a control population in which the mean is 100, and the standard deviation is 10; this distribution can be used to represent scores on any task of cognitive ability.

Superimposed on this distribution is a notional patient population that has the same mean but a larger standard deviation. (One can think of this situation as one in which the patient population and the control population were one and the same population until the former suffered neurological damage.) Mycroft et al.'s (2002) intermediate example is used in this figure so the standard deviation for the patient population is 25 (i.e., 2.5 times that of the control population). This represents the situation that Mycroft et al. consider a cause for concern; that is, they argue that because Crawford and Howell's method does not factor in the increased variability in the patient population, it will not control Type I errors.

This scenario, however, is not credible: Neurological damage has had absolutely no effect on the mean score of the patients (i.e., it has not produced deficits) but has markedly increased the variability of the patients' scores. Note that it is absolutely central to Mycroft et al.'s (2002) argument that neurological damage has not lowered the mean of the patient population (i.e., produced deficits). If there is any lowering of the patient population mean, no matter how small, then the issue of a Type I error does not arise: The population means differ, and the only remaining question is whether this effect can be detected (i.e., the question becomes one of the power to detect deficits).

The scenario can be seen to be even more unlikely when we consider that if there is no difference in the means (as there cannot be if the concern is with Type I errors), and variability is much higher in the patient population, then any observation from the patient distribution that lies below that of the control population must be balanced exactly by observations that lie above it. In other words, and as can be seen from Figure 4, in Mycroft et al.'s (2002) scenario, patients will frequently obtain cognitive test scores that are vastly higher than those of controls (or, equivalently, scores that are vastly higher than their premorbid scores).

To conclude this discussion of Type I errors: In essence Mitchell et al. (2004) argue (p. 758) that if a reliable result is obtained using Crawford and Howell's (1998) method, then this may be because (a) the patient does not have a deficit, and her/his performance is simply more variable, (b) the patient does have a deficit, and her/his performance is also more variable, or (c) the patient has a deficit but is no more variable than controls. In view of the points made above we can safely discount possibility (a), and in both (b) and (c) a Type I error cannot be made as the patient has a deficit.

Of course, even after ruling out the possibility of inflated Type I errors, it could still be argued that a patient might have a very minor deficit on a task but be much more variable in their performance (i.e., a Type I error has not occurred but the result would potentially mislead investigators as to the fundamental nature of the patient's problem). One response to this is simply to reiterate that the patient still has a deficit. However, researchers may be uneasy were this scenario to occur in the real world.

Fortunately such a scenario is little more credible than the previous scenario. For example, picture a slight leftward movement of the patient distribution in Figure 4 (i.e., the patient population mean is lower than the control but only minimally); the patient population would still contain very many scores that exceed those of the control population. Thus, it is hard to envisage a situation in which neurological damage has caused a patient's performance to become markedly more variable than controls without also markedly lowering the overall level of performance. This is mainly because it cannot be expected that following neurological damage, patients will routinely obtain scores on a cognitive task that exceed their premorbid scores. As a result, any increased variability will occur below the ceiling imposed by the limits of their premorbid ability, and it is then necessarily the case that their average postmorbid level of performance will be well below their premorbid level.

Unreliability of the performance of single cases and increased variability of a patient population

In the present power study the effect of increased unreliability in the performance of single cases has been examined. It should be stressed that this study can equally readily be conceived as a study of the effects on power of increased variability in a notional patient population. That is, Mycroft et al. (2002) were concerned with Type I errors when the variability of a patient population was larger than a control population, and the present study can be seen as extending this to study power under these circumstances.

One difference, however, is that we have not modelled situations in which the increase in variability for patients is as extreme as the scenarios examined by Mycroft et al. (2002) in their study of Type I errors. As noted, Mycroft et al. examined scenarios in which the standard deviation of patients was up to five times that of controls (and provided modified critical value of F for use by researchers to cover this scenario). However, if we translate this increased variability into reliabilities it stretches credibility. For example, if the reliability of scores for controls was .85, then the reliability of scores for patients would have to be vanishingly low (.034) in order to produce a situation in which the standard deviation of observed scores was 5 times that of the observed scores for controls. Even if we leave aside this issue, the scenario can be seen to be unrealistic on grounds similar to those advanced above when considering Type I errors. That is, even when a case has a large deficit, if their scores were more variable than controls by a factor of five, their obtained scores would frequently exceed their premorbid scores or those of matched controls by a very large amount.

An ultraconservative model of single-case research?

To summarize the empirical findings thus far: Although power to detect a large deficit for Crawford and Howell's (1998) method is at best moderate, this method has vastly greater power to detect a deficit than Mycroft et al.'s (2002) method in all scenarios examined: that is, when the control population is normal, when it departs from normality, in the absence of random measurement error, when error is present in equal measure for controls and the single case, and when error is greater for the single case (as noted, this latter scenario can also be conceived of in Mycroft et al.'s terms as an increase in the variability of the patient population).

Although single-case researchers will be rightly concerned that the use of Mycroft et al.'s (2002) method will result in a failure to detect many deficits that are of importance for cognitive theory,[2] observers of the single-case enterprise may be more sanguine. That is, they may take the position that very rigorous standards of proof should be applied in single-case research, and if that means that many interesting deficits will be missed, then so be it. In other words, it might be argued that avoidance of Type I errors should be paramount; single-case researchers should err strongly on the side of being conservative when interpreting their data.

It is certainly true that, if Mycroft et al.'s (2002) method yielded a significant result, we could be very confident that the patient truly has a deficit

[2] Note that potentially important deficits will also be missed when Crawford and Howell's (1998) method is used but power is particularly low for Mycroft et al.'s (2002) method.

(i.e., the critical values employed make their method very conservative). However, there are at least two reasons why an appeal to conservatism is unsatisfactory.

First, if there is concern over standards of proof, then a decision to toughen standards should be implemented through more conventional means. That is, with Crawford and Howell's (1998) method it is very easy to reduce the Type I error rate simply by adopting a more conservative value of alpha (e.g., one could require that the difference between a patient and controls was significant at the .01 level rather than the .05 level). This approach has the major advantage that researchers would know that having opted for a Type I error rate they regard as acceptable, this is the error rate that will apply when the test is used in practice (subject to the proviso that the assumption of normality is not violated; as noted, both methods make this assumption). That is, we know from theory and from Crawford and Garthwaite's (2005b) simulation study that the observed Type I error rates for Crawford and Howell's method will match the specified error rates.

Moreover, such a strategy does not require that researchers attempt to estimate the variance of a notional patient population in order to select a critical value (there is also the prospect that one researcher's guess at this variance will differ radically from that of another such that if they applied Mycroft et al.'s, 2002, method to the same dataset, they would arrive at radically different conclusions).

It should be stressed that we are not advocating that researchers should adopt a more stringent value of alpha. The present results demonstrate that power will not be high, even for Crawford and Howell's (1998) method, unless the deficit to be detected is very large. Therefore, use of the conventional .05 level (one-tailed) will generally strike a reasonable balance between controlling Type I and Type II errors.

There is a second reason why an appeal to conservatism should be regarded with scepticism. Because of the nature of many of the questions posed in single-case studies, methods that are apparently conservative may, paradoxically, lead researchers to claim erroneous support for their hypotheses. This possibility, which also has broader implications for single-case research, is explored in Study 5.

STUDY 5: TYPE I ERRORS FOR CLASSICAL DISSOCIATIONS IN SINGLE-CASE STUDIES

Although identifying deficits is a fundamental feature of single-case studies, such deficits are normally of limited theoretical interest unless they are accompanied by performance in the normal range on other tasks. That is, much of the focus in single-case studies is on establishing dissociations of function (Caramazza & McCloskey, 1988; Coltheart, 2001; Crawford et al., 2003; Ellis & Young, 1996; Shallice, 1988).

A classical dissociation (Shallice, 1988, p. 227) is conventionally defined as occurring when, with reference to the performance of matched healthy controls (or a healthy normative sample), 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) state, "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). Similarly, Coltheart (2001) states that a classical dissociation is established when a patient "is impaired on task X but normal on task Y" (p. 12).

In practice, when the research design is one in which a patient is compared to matched controls, the patient is considered to have met these conventional criteria for a classical dissociation if her/his performance is significantly different from that of controls on task *X* but is not significantly different on task *Y*. The danger here is that if a method with low power is used to test whether these criteria are met, then a patient with a genuine deficit on task *Y* will not differ significantly from controls. Thus a spurious classical dissociation will be recorded. Ironically then, low power to detect a deficit (i.e., a high Type II error rate) can lead to a high Type I error rate (i.e., falsely concluding that a case has a classical dissociation).

The final simulation study is designed to subject the above argument to empirical scrutiny. That is, although such paradoxical effects are clearly possible in theory, it is important to examine the extent to which they are liable to pose a threat in practice. To study this we adopt a procedure developed by Crawford and Garthwaite (2005a) to model cases that have a strictly equivalent level of acquired impairment on both tasks of interest (*X* and *Y*) and to record the number misclassified as exhibiting a classical dissociation.

We use Mycroft et al.'s (2002) method to test whether the conventional criteria for a classical dissociation are met (i.e., a significant difference between the case and controls on either task *X* or task *Y* but not both), and we compare the results with those obtained when Crawford and Howell's (1998) method is used for the same purpose. In addition, we compare both sets of results to those obtained when an alternative set of criteria (Crawford & Garthwaite, 2005b) for a classical dissociation is applied. These criteria, which stem from a critique of the conventional criteria made by Crawford et al. (2003), incorporate a test on the standardized difference between the patient's *X* and *Y* scores. That is, the conventional criteria are supplemented with a requirement that the difference between the patient's scores significantly exceed the differences observed for controls. (This additional criterion specifies that it is the standardized differences that should be compared because, typically, tasks *X* and *Y* will differ in their means and standard deviations.) The test on the standardized difference is achieved using the Revised Standardized Difference Test (RSDT; Crawford & Garthwaite, 2005b; Garthwaite & Crawford, 2004).[3]

Method

As in previous studies, the simulation was implemented in Delphi. However, for each control and the single case there is now a pair of scores, and so sampling is from a bivariate normal distribution. Also, an additional factor needs to be added to the design in order to study performance of the inferential methods at different magnitudes of the population correlation between *X* and *Y*.

A total of 1,000,000 samples of $n + 1$ pairs of observations were drawn from each of four bivariate standard normal distributions in which the population correlation (ρ) was set at .0, .2, .5, and .8. As in the previous simulations, this was done for five values of n: 5, 10, 20, 50, and 100.

The first n pairs of observations were taken as the control sample's scores on *X* and *Y* and the ($n + 1$)th pair taken as the scores of the single case. The single case was then "lesioned" by imposing an acquired impairment of 3 *SD*s on both *X* and *Y*. As the observations are sampled from a standard normal bivariate distribution, the standard deviation is 1.0 for both *X* and *Y*, and therefore this required simply that 3.0 was subtracted from the case's *X* and *Y* scores. These cases are used to represent patients who have suffered very large, but strictly equivalent, deficits on *X* and *Y*; that is, they do not exhibit a classical dissociation. Note that 3-*SD* deficits were applied in this simulation rather than the 2-*SD* deficits applied in Crawford and Garthwaite's (2005a) previous work on dissociations. This was done because it is clear from the foregoing power studies that Mycroft et al.'s (2002) method will only be capable of identifying deficits if they are very large; for the same reason only the intermediate version of Mycroft et al.'s method is examined.

The simulation procedure is designed to model patients with equivalent acquired deficits: It does not produce cases with scores that are simply 3 *SD*s below the control mean on *X* and *Y*, nor does it produce cases with equivalent scores on *X* and *Y*. Rather, the method

[3] Obtaining a sound inferential method of examining the difference between an individual's standardized scores has proved to be much more difficult than might be anticipated as the problem is one of testing for a difference between two *t* variates. The RSDT was developed using asymptotic expansion methods and, unlike previously available methods, achieves control of the Type I error rate across all values of the control sample n and the correlation between tasks.

recognizes that (a) patients are initially members of the healthy control population until the onset of their lesion, (b) there will be premorbid differences in competencies on X and Y, and (c) the magnitude of premorbid differences between X and Y will be a function of the population correlation between the two tasks (i.e., the magnitude of such differences will, on average, be smaller when the population correlation is high than when it is low).

On each Monte Carlo trial, Mycroft et al.'s (2002) method, in its intermediate form, was applied to test whether the conventional criteria for a classical dissociation were met—that is, a significant difference in favour of controls on either task X or task Y but not on both. This was then repeated substituting Crawford and Howell's (1998) method as the means of testing whether the criteria were met. Finally, Crawford and Garthwaite's (2005b) set of criteria was applied. This required a significant difference, using Crawford and Howell's method, between the single case and controls on either X or Y, but not both, and also required that the standardized difference between the case's X and Y scores was significantly larger than the standardized differences of the controls ($p < .05$, two-tailed).

Results and discussion

Full results from the simulation are presented in Table 6. However, the pattern of results is more readily appreciated by referring to Figures 5 and 6. Figure 5 plots Type I errors as a function of the control sample size and the criteria applied; the results are those for an intermediate population correlation of .5 between tasks X and Y.

It can be seen that when Mycroft et al.'s (2002) method is used to test whether the conventional criteria are met, large percentages of the cases with strictly equivalent deficits on X and Y are misclassified as exhibiting a classical dissociation. The minimum Type I error rate was 20.69% (for a control sample n of 100 and population correlation between X and Y of .8), and this rose to a maximum of 49.33% for a control sample size of 100 and correlation of 0; that is, up to half of all cases with strictly equivalent deficits are liable to be wrongly classified as exhibiting a classical dissociation. It can also be seen that the Type I error rate is relatively unaffected by the control sample size; that is, larger control samples do not protect against misclassifications.

It can be seen from Table 6 and Figure 5 that when Crawford and Howell's (1998) method is used to test whether the conventional criteria are met, the percentage of cases misclassified as exhibiting a dissociation are, in general, much lower than those observed for Mycroft et al.'s (2002) method. It can also be seen that, unlike Mycroft et al.'s method, larger control sample sizes protect against misclassifications; for example, in the case of a correlation between X and Y of .5, misclassification rate falls from 29.74% for a n of 5 to 12.79% for a n of 100.

Although the use of Crawford and Howell's (1998) method to test whether the conventional criteria are met leads to lower rates of misclassification, it can be seen that these rates are nevertheless still uncomfortably high. The final columns of Table 6 present the percentage of cases misclassified when Crawford and Garthwaite's (2005b) criteria are applied (see also Figure 5). In contrast to the foregoing results, it can be seen that misclassification rates are low in all of the scenarios examined; the rates range from a maximum of 6.66% for a control sample n of 5 and a correlation between tasks of zero to a minimum of 1.32% for a n of 100 and correlation of 0.8. (Note that these criteria are based on the application of three statistical tests and function as a set of hurdles; therefore, unlike the results for an individual test statistic, it should not be expected that the misclassification rates will be at or around 5%.)

It can also be seen from Table 6 and Figure 6 that the percentage of cases misclassified as exhibiting a classical dissociation declines for all three methods as the population correlation between X and Y increases. This is encouraging because, as Shallice (1979) notes, in practice much of the search for dissociations is focused

Table 6. *Type I errors for a classical dissociation as a function of control sample size, correlation between tasks (ρ_{xy}), and criteria employed*

	ρ_{xy}														
	Mycroft et al. (2005): Intermediate					*Conventional criteria using t*					*Crawford and Garthwaite (2005b) criteria*				
n	*0*	*.3*	*.5*	*.7*	*.8*	*0*	*.3*	*.5*	*.7*	*.8*	*0*	*.3*	*.5*	*.7*	*.8*
5	38.20	34.99	31.37	25.72	21.66	39.66	34.55	29.74	23.46	19.28	6.66	6.43	5.79	4.67	3.85
10	45.25	39.28	33.87	26.75	22.01	26.95	23.39	20.16	15.81	13.05	4.82	4.27	3.71	2.86	2.35
20	48.00	40.35	34.12	26.41	21.55	21.03	18.35	15.83	12.47	10.28	3.96	3.37	2.81	2.14	1.72
50	49.02	40.23	33.62	25.71	20.92	17.94	15.68	13.59	10.69	8.77	3.57	2.86	2.36	1.76	1.41
100	49.33	40.16	33.44	25.54	20.69	16.96	14.82	12.79	10.11	8.34	3.40	2.70	2.19	1.65	1.32

Note: ρ_{xy} = correlation between tasks.

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). However, the rates are still high in absolute terms, except with Crawford and Garthwaite's (2005b) criteria, and this is the case even when the correlation between tasks is substantial.

These results illustrate the virtues of incorporating a test on the difference between a patient's *X* and *Y* scores when testing for a classical dissociation. That is, we suggest that the conventional criteria for a classical dissociation are fundamentally unsatisfactory because one half of the criteria relies on failing to find evidence of a deficit. The present results for Mycroft et al.'s (2002) method provide a particularly vivid demonstration of this problem (because of the low power of the method to detect deficits). However, it

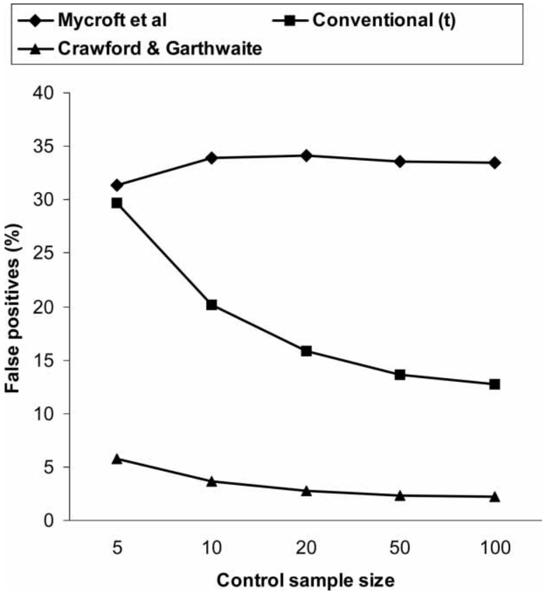

Figure 5. *Type I errors for a classical dissociation as a function of control sample size and criteria applied (for these data ρ_{xy} = .5).*

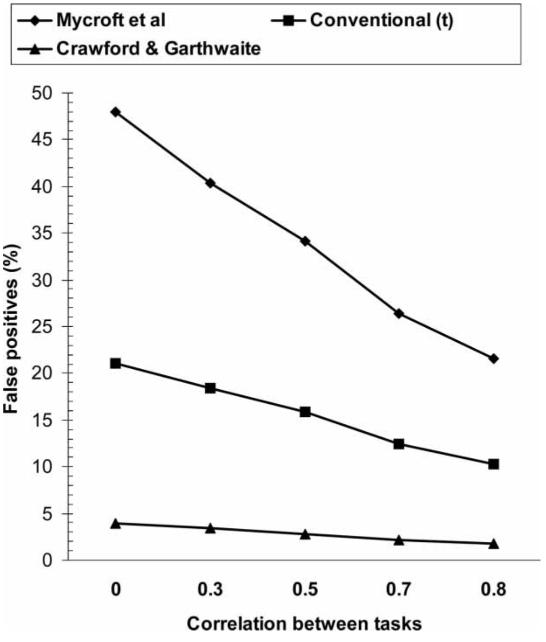

Figure 6. *Type I errors for a classical dissociation as a function of criteria applied and correlation between tasks (based on a n of 20 for the control sample).*

remains a problem regardless of the method used to test the conventional criteria because low or moderate power to detect deficits is an inherent feature of single-case studies.

As Crawford and Garthwaite (2005b) note, the additional requirement of a significant difference between the patient's X and Y scores provides a positive test for a classical dissociation rather than having to rely on failing to find evidence for a deficit on one of the tasks. When this additional criterion is imposed it is unlikely that a patient with equally severe acquired deficits will be misclassified unless she/he had an unusually large premorbid difference on the abilities measured by the two tasks.

GENERAL DISCUSSION

The present investigation was originally motivated by a simple question: Of two methods of testing for a deficit in single-case research, which should be preferred? The results obtained provide an unequivocal answer to this question but they also have broader implications for single-case research. To our knowledge, this is the first study to conduct an empirical examination of the power to detect a deficit in single-case studies. Although the results demonstrate that power will be particularly low for Mycroft et al.'s (2002) method, they also serve to demonstrate that low-to-moderate power will be an inherent feature unless the deficits to be detected are extremely large.

A commonly used alternative to the methods of Crawford and Howell (1998) and Mycroft et al. (2002) is to express the patient's score as a z score and refer it to a table of areas under the normal curve; for example, the patient's score is considered to be significantly lower than that of controls ($p < .05$, one-tailed) if z falls below -1.645. This method will identify more patients with true deficits than will either of the foregoing methods (particularly when the control sample size is small) but power will still be only moderate even for large deficits. Moreover, this increase in power occurs at the expense of inflation of the Type I error rate. For example, with a specified error rate of 5%, Crawford and Garthwaite (2005b) reported observed error rates for z as high as 10.37% with control sample sizes typical of those used in single-case studies (z was not included in the present study for this reason; i.e., the power of a method can only be meaningfully interpreted when the Type I error rate is close to the nominal level or more conservative).

The present study has also demonstrated that researchers cannot be sanguine about the fact that power will tend to be low to moderate in single-case studies. Low power not only increases the likelihood that researchers will fail to gain support for their hypotheses (or fail to detect unexpected but interesting deficits), but may also produce spurious support for hypotheses that specify that a deficit occurs in the context of unimpaired performance on other tasks. As noted, a positive test for a classical dissociation (i.e., a test on the difference between tasks) is required to avoid reliance on a null result. Although power to detect a classical dissociation will be at best moderate when this additional criterion is applied (Crawford & Garthwaite, 2005a), it has the virtue that it will markedly reduce the number of false positives.

The effects of measurement error on power to detect deficits also yielded results that are of broad interest. Measurement error, when present in equal measure for controls and the single case, will reduce the power to detect a deficit. However, if task reliabilities are moderate to high (i.e., $\geq .70$) the effects are relatively modest. More importantly perhaps, when a patient score is subject to greater error than that of controls, power will be broadly comparable to that achievable in the absence of such a differential, unless power deviates markedly from 50% in the latter case. For Crawford and Howell's (1998) method (and for z, although we do not recommend this test for the reasons outlined above) power to detect a large deficit will typically be in the range of 40 to 60%. Thus it appears that differential measurement error does not pose a serious threat to the validity of inferences drawn in single-case studies given that we can also discount the possibility that its presence will inflate the Type I error rate.

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.

Manuscript received 13 April 2004
Revised manuscript received 12 October 2005
Revised manuscript accepted 7 December 2005
First published online 6 April 2006

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.

Bobes, M. A., Lopera, F., Comas, L. D., Galan, L., Carbonell, F., Bringas, M. L., et al. (2004). Brain potentials reflect residual face processing in a case of prosopagnosia. *Cognitive Neuropsychology, 21*, 691–718.

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

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

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. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition, 5*, 41–66.

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. (1992). Current and premorbid intelligence measures in neuropsychological assessment. In J. R. Crawford, D. M. Parker, & W. W. McKinlay (Eds.), *A handbook of neuropsychological assessment* (pp. 21–49). Hove, UK: Lawrence Erlbaum Associates Ltd.

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, UK: 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. (2005a). Evaluation of criteria for classical dissociations in single-case studies by Monte Carlo simulation. *Neuropsychology, 19*, 664–678.

Crawford, J. R., & Garthwaite, P. H. (2005b). 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., Azzalini, A., Howell, D. C., & Laws, K. R. (2006). Testing for a deficit in single case studies: Effects of departures from normality. *Neuropsychologia, 44*, 666–676.

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

Deary, I. J. (1995). Age-associated memory impairment: A suitable case for treatment. *Ageing and Society, 15*, 393–406.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292–307.

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

Farrer, C., Franck, N., Paillard, J., & Jeannerod, M. (2003). The role of proprioception in action recognition. *Consciousness and Cognition, 12*, 609–619.

Forti, S., & Humphreys, G. W. (2004). Visuomotor cuing through tool use in unilateral visual neglect. *Journal of General Psychology, 131*, 379–410.

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

Howard, D., & Nickels, L. (2005). Separating input and output phonology: Semantic, phonological, and orthographic effects in short-term memory impairment. *Cognitive Neuropsychology, 22*, 42–77.

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 Alzheimer's disease? *Cortex, 41*, 452–463.

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

McCloskey, M. (1993). Theory and evidence in cognitive neuropsychology: A "radical" response to Robertson, Rafal, and Shimamura (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 718–734.

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.

Miller, K. M., & Swick, D. (2003). Orthography influences the perception of speech in alexic patients. *Journal of Cognitive Neuroscience, 15*, 981–990.

Mitchell, D. C., Mycroft, R. H., & Kay, J. (2004). Comparing a single case to a control sample: Differences in distribution versus difference in means. *Cognitive Neuropsychology, 21*, 756–760.

Mycroft, R. H., Mitchell, D. C., & Kay, J. (2002). An evaluation of statistical procedures for comparing an individual's performance with that of a group of controls. *Cognitive Neuropsychology, 19*, 291–299.

Owen, D. B. (1968). A survey of properties and applications of the noncentral *t*-distribution. *Technometrics, 10*, 445–478.

Papps, B. P., Calder, A. J., Young, A. W., & O'Carroll, R. E. (2003). Dissociation of affective modulation of recollective and perceptual experience following amygdala damage. *Journal of Neurology, Neurosurgery, and Psychiatry, 74*, 253–254.

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

Robinson, G., Shallice, T., & Cipolotti, L. (2005). A failure of high level verbal response selection in progressive dynamic aphasia. *Cognitive Neuropsychology, 22*, 661–694.

Rosenbaum, R. S., Fuqiang, G., Richards, B., Black, S. E., & Moscovitch, M. (2005). "Where to?" remote memory for spatial relations and landmark identity in former taxi drivers with Alzheimer's disease and encephalitis. *Journal of Cognitive Neuroscience, 17*, 446–462.

Rusconi, E., Priftis, K., Rusconi, M. L., & Umiltà, C. (2006). Arithmetic priming from neglected numbers. *Cognitive Neuropsychology, 23*, 227–239.

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.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

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, UK: Cambridge University Press.

Sokal, R. R., & Rohlf, F. J. (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. 53–76). Amsterdam: Elsevier.

APPENDIX A

A proof that, for Crawford and Howell's (1998) method, the estimated proportion of controls who have a lower score than a case equals the significance level of the one-tailed test

The proportion of controls who have a lower score than x^*, the score for a case, is

$$\Pr(x < x^*) \tag{A.1}$$

Adding $\bar{x}$ to both sides and dividing them by the same thing,

$$\Pr(x < x^*) = \Pr\left(\frac{x - \bar{x}}{\sqrt{s_x^2((n+1)/n)}} < \frac{x^* - \bar{x}}{\sqrt{s_x^2((n+1)/n)}}\right). \tag{A.2}$$

Now

$$\frac{x - \bar{x}}{\sqrt{s_x^2((n+1)/n)}}$$

has a t distribution on $(n - 1)$ *df*, so that

$$\Pr(x < x^*) = \Pr\left(t_{n-1} < \frac{x^* - \bar{x}}{\sqrt{s_x^2((n+1)/n)}}\right). \tag{A.3}$$

Also, the test statistic for testing whether x^* is from the same normal distribution as the control x's, is

$$\frac{x^* - \bar{x}}{\sqrt{s_x^2((n+1)/n)}} \tag{A.4}$$

and this is compared with a t distribution on $(n - 1)$ *df.* Comparison of Equations A.3 and A.4 shows that $\Pr(x < x^*)$ is equal to the significance level for the one-tailed test.

APPENDIX B

Sampling from skew-normal and skew-*t* distributions

The methods used to sample from skew-normal and skew-t distributions were based on work by Azzalini and colleagues (Azzalini & Capitanio, 1999; Azzalini & Dalla Valle, 1996). The starting point for sampling from skew-normal distributions 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

$$u_2 = \rho^2_{u_0 u_1} + \sqrt{1 - \rho^2_{u_0 u_1}}\, u_1 \tag{A.5}$$

The value of $\rho_{u_0 u_1}$ required to introduce the desired degree of skew (γ_1) can be obtained by algebraic manipulation of Azzalini and Dalla Valle's (1996) formulae for γ_1 to solve for $\rho_{u_0 u_1}$. That is, put

$$a = \left(\frac{2\gamma_1}{4 - \pi}\right)^{1/3} \tag{A.6}$$

and

$$\rho_{u_0 u_1} = a\left(\frac{\pi}{2 + 2a^2}\right)^{1/2}. \tag{A.7}$$

Then

$$x = \begin{cases} u_2 & \text{if}\, u_0 \geq 0 \\ -u_2 & \text{otherwise} \end{cases} \tag{A.8}$$

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 *df* (e.g., $\nu = 4$ if severe leptokurtosis is required).

APPENDIX C

Formulae for power calculations for Crawford and Howell's (1998) test and the test of Mycroft et al. (2002)

Crawford and Howell's (1998) test

When the population variance is 1, and the reliability is r_{xx}, then the variance of an observed score is $1/r_{xx}$. If the case has a deficit of η, then the power of the test of Crawford and Howell is

$$\Pr(\bar{x} - x^*/(s\sqrt{1 + 1/n}) > t_{n-1;\alpha}), \tag{A.10}$$

where

$$\bar{x} - x^* \sim N\left(\eta, \frac{n+1}{nr_{xx}}\right).$$

Now $r_{xx}(n - 1)s^2$ has a chi-squared distribution on $(n - 1)$ *df*, where s^2 is the sample variance of the controls. Hence, $t_{n-1}(\delta) = (\bar{x} - x^*)/\{s\sqrt{[(n + 1)/n]}\}$ has a noncentral t distribution on $(n - 1)$ *df* with noncentrality parameter δ, where $\delta = \eta\sqrt{[r_{xx}n/(n + 1)]}$. (See, for example, Owen, 1968.) To obtain η^*, the minimal value of η for which the power is 0.8, we first find δ^* such that $\Pr(t_{n-1}(\delta^*) > t_{n-1;\alpha}) = 0.8$. Then the probability in Equation A.10 also equals 0.8, so we put $\eta^* = \delta^* \surd[(n + 1)/r_{xx}n]$.

The test of Mycroft et al. (2002)

The critical values for the test of Mycroft et al. (2002) were determined by simulation in their original paper. To gain understanding, we first derive critical values using distribution theory. The premise for the test is that the variance for a control is some unknown value, ϕ say, while the variance for the case is $k\phi$, where k is known. We assume, for the moment, that there is no measurement error. Then, under the null hypothesis that the case comes from some population whose mean equals the mean of the controls' population,

$$\bar{x} - x^* \sim N(0,(k + 1/n)).$$

Also, $(n - 1)s^2/\phi$ has a chi-squared distribution on $(n - 1)$ *df*, so $(\bar{x} - x^*)/\{s\surd[k + (1/n)]\}$ follows a t distribution on $(n - 1)$ *df*. Hence, for the two-tailed test of Mycroft et al., the null hypothesis is rejected if

$$\left|\frac{\bar{x} - x^*}{s\sqrt{k + 1/n}}\right| > t_{n-1;\alpha/2}.$$

Equivalently, we could use a different test statistic and reject the null hypothesis if

$$\left|\frac{\bar{x} - x^*}{s\sqrt{(n + 1)/n}}\right| > t_{n-1;\alpha/2}\sqrt{\frac{kn + 1}{n + 1}}.$$

A further equivalent alternative is to reject the null hypothesis if

$$\left(\frac{\bar{x} - x^*}{s\sqrt{(n + 1)/n}}\right)^2 > \left(\frac{kn + 1}{n + 1}\right)F_{1,n-1;\alpha/2},$$

where $F_{1,n-1;\alpha/2}$ is the $(1 - \alpha/2)$ quantile of an F distribution on 1 and $(n - 1)$ *df*. This last alternative is the test described in Mycroft et al. Their test statistic, obtained through an ANOVA, equates to $[(\bar{x} - x^*)/\{s\surd[(n + 1)/n]\}]^2$, and the critical values given in Table 2 of that paper are equal to $\{(kn + 1)/(n + 1)\}F_{1,n - 1;\alpha/2}$. (Slight differences arise, however, because the figures given in Mycroft et al. are affected by Monte Carlo variation and hence are approximate. For that paper's notation, k must be replaced by σ^2.)

For power calculations, it is simplest to choose $(\bar{x} - x^*)/\{s\surd[(n + 1)/n]\}$ as the test statistic, since $t_{n-1}(\delta) = (\bar{x} - x^*)/\{s\surd[(n + 1)/n]\}$ has a noncentral t distribution. Its noncentrality parameter is $\delta = \eta\surd[r_{xx}n/(n + 1)]$, the distribution has $(n - 1)$ *df*, and, when the reliability is r_{xx}, its noncentrality parameter is $\delta = \eta\surd[r_{xx}n/(n + 1)]$. The test is two-tailed and the test statistic is $t_{n-1;\alpha/2}\surd[(\mathrm{k}n + 1)/(n + 1)]$, so we find $\delta^{\#}$ for which

$$\Pr\left(t_{n-1}(\delta^{\#}) > \frac{\sqrt{kn + 1}}{\sqrt{n + 1}}t_{n-1,\alpha/2}\right) + \Pr\left(t_{n-1}(\delta^{\#}) < -\frac{\sqrt{kn + 1}}{\sqrt{n + 1}}t_{n-1,\alpha/2}\right) = 0.8.$$

Then, $\eta^{\#} = \delta^{\#}\surd[(n + 1)/(r_{xx}n)]$ is the minimal deficit that the test of Mycroft et al. detects with a power of 0.8.