CORTEX FORUM

WANTED: FULLY OPERATIONAL DEFINITIONS OF DISSOCIATIONS IN SINGLE-CASE STUDIES

John R. Crawford[1], **Paul H. Garthwaite**[2] **and Colin D. Gray**[1]

([1]Department of Psychology, University of Aberdeen, Aberdeen, UK; [2]Department of Statistics, The Open University, Milton Keynes, UK)

ABSTRACT

In contrast to the careful consideration given to the issue of what we can infer from dissociations in single-case studies, the more basic question of how we decide whether a dissociation is present has been relatively neglected. Proposals are made for fully operational definitions of a deficit, classical and strong dissociations, and double dissociations. In developing these definitions it was assumed that they should be based on the use of inferential rather than descriptive statistical methods. The scope of these definitions is limited to typical single-case studies in which patients are compared to control samples of a modest size. The operational definition of a classical dissociation incorporates a requirement that a patient's performance on Task *X* is significantly different from Task *Y*, in addition to the "standard" requirement that the patient has a deficit on Task *X* and is within normal limits on Task *Y*. We ran a simulation to estimate the Type I error rates when the criteria for dissociations are applied and found these to be low (Type I errors were defined as identifying an individual from the control population as having a dissociation). The inferential methods for testing whether the various criteria are met make use of *t*-distributions. These methods are contrasted with the widespread use of *z* to test for a deficit or a difference between tasks. In the latter approach the statistics of the control sample are treated as parameters; this is not appropriate when, as is normally the case, the control sample size is modest in size.

Key words: double dissociation, single case studies, inferential statistics

INTRODUCTION

In the recent Cortex forum on double dissociations Dunn and Kirsner (2003) and other contributors provide a detailed examination of the logical foundations of the single-case approach in cognitive neuropsychology and its potential pitfalls. These contributions build on a number of earlier thoughtful reviews on this topic (e.g., Caramazza and McCloskey, 1988; Shallice, 1988; Ellis and Young, 1996; Coltheart, 2001; Sartori, 1988). However, in contrast to the careful and detailed scrutiny of assumptions in Dunn and Kirsner (2003) and in preceding reviews, our view is that the two key constructs in this field of enquiry, the dissociation and the double dissociation, continue to be defined very loosely. Thus, while there has been much concern with what we can and cannot infer from a dissociation, much less attention has been paid to the more basic, methodological question of how we decide whether we have one in the first place.

Typically, a dissociation is defined as occurring when 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). A double dissociation is said to occur when another patient is found to show the opposite pattern. For example, Coltheart (2001) states, "With double dissociations we need two patients: patient A who is impaired on task X but normal on task Y, and patient B who is normal on task X but is impaired on task Y" (p. 12).

Such definitions do not contain an explicit (i.e. fully operational) definition of a "deficit" or "impairment", nor of what should be considered "unimpaired", "normal" or "within normal limits". Indeed, in most reviews, it is unclear whether the criteria for deficits and dissociations should be based on descriptive or inferential statistics. Some reviews refer to the need for a patient's performance on Task *X* to be "significantly different" from controls, but the inferential statistical method that should be used to test this is not specified.

The present treatment of these issues is predicated on the assumption that we should have fully explicit criteria for a dissociation and that these criteria should be based on the application of *inferential* statistical methods. Within the neurosciences, any field of enquiry that limited itself to the use of descriptive statistics would be an outlier. Furthermore, Newcombe and Marshall (1988) noted that, in the 1920s, "single-case studies, no matter how well-conducted, … began to be described as 'merely anecdotal'" (p. 549). A reliance solely on descriptive statistics would carry the risk that a similarly harsh judgement could be passed upon the current generation of single-case studies.

In the present treatment, some proposals for operational definitions of dissociations will be developed. We shall limit our attention to the typical single-case study, in which a patient's performance is compared with that of a modestly-sized control sample.

Operational Definition of a Deficit

Before we can arrive at fully operational definitions of dissociations and double dissociations, we must take a step back and tackle the more fundamental issue of defining a deficit. By far the most common method of inferring a deficit is to use the normal deviate *z*. A patient's performance is converted to a *z*- score based on the mean and *SD* of the control sample and this *z* is then referred to a table of areas under the normal curve. If the probability of a *z*-score at least as large as the one obtained is less than 0.05 (or some other selected alpha value, such as the more stringent .01 level) then the patient's score is considered to be significantly poorer than the controls and it is inferred that the patient has a deficit.

One problem with this approach is that it treats the control sample as if it were a population: the mean and standard deviation are used as if they were *parameters* rather than *statistics*. In other areas of psychology, this is often not

a problem in practice, since the normative or control sample is large and therefore affords sufficiently accurate estimates of the parameters. In single-case, neuropsychological studies, however, the control samples typically have modest *n*s: $n < 5$ is not unusual and $n < 15$ is very common (Crawford and Garthwaite, 2002). With samples of this size, it is not appropriate to treat the mean and *SD* as though they were parameters.

A solution to this problem is to use a method proposed by Crawford and Howell (1998), in which the score of the patient is compared with the controls by means of a modified *t*-test described by Sokal and Rohlf (1995). Thus Crawford and Howell's approach uses the *t* distribution (with $N - 1$ degrees of freedom), rather than the standard normal distribution, to estimate the abnormality of the patient's scores and to test whether it is significantly lower than the scores of the control sample.

The practical effect of using *z* with small control sample is to exaggerate the rarity/abnormality of a patient's score and to inflate the Type I error rate (a Type I error occurs when we falsely conclude that the patient differs from controls when they do not). This occurs because the normal distribution has "thinner tails" than *t*-distributions. Intuitively, the less that is known, the less extreme should be statements about abnormality/rarity. The *z* score method treats the variance of controls as being known, when it is not, and consequently makes statements that are too extreme (Crawford and Howell, 1998).

To illustrate, suppose that an individual obtains a score of 23 on a test and that the mean and *SD* for this test in a control sample are 40 and 10 respectively. If the *n* of the control sample were 6 (which is not unusual in a single-case study), the modified *t*-test procedure shows that approximately 8.8% of the population would obtain a score at least as low as the patient's score, and the conclusion is that the patient's score is not significantly different from the controls ($p > .05$, one-tailed). According to the *z*-test, however, only 4.5% of the population would obtain a score at least as low as this, and the conclusion is that the patient is significantly different from controls ($p < .05$, one-tailed).

It is proposed that Crawford and Howell's (1998) method be used to provide an operational definition of a deficit (see Table I). If the modified *t*-test reveals that the patient is significantly lower than controls ($p < 0.05$, one-tailed), we can conclude that the patient is "impaired" or "has a deficit". If the test is not significant then, for present purposes, we consider that the patient is "not impaired" or "within normal limits".

A one-tailed test is advocated because, in practice, researchers always have a directional hypothesis when testing patients: they want to test the hypothesis that their patient is significantly *lower* than controls against the null hypothesis that they are not lower, that is, unimpaired. Essentially, matched controls are used to represent a patient's likely premorbid performance. Situations in which *improvement* in performance over premorbid levels occurs as the result of brain damage are highly specific and very rare. Finally, although in the proposed definition, alpha is set at the conventional value of 0.05, some researchers may prefer the more stringent 0.01 level.

TABLE I
Proposed Operational Definitions of Deficits and Dissociations

Term/Construct	Operational definition
Deficit	1) Patient's score is significantly lower than controls using Crawford and Howell's (1998) method and a one-tailed test; $p < 0.05$ or 0.01 depending on level of stringency considered appropriate.
Within normal limits	1) Patient's score not significantly lower ($p > 0.05$, one-tailed) than controls using Crawford and Howell's (1998) method.
Classical Dissociation	1) Patient's score on Task X significantly lower than controls (i.e., score meets the above criterion for a deficit 2) Patient's score on Task Y not significantly lower than controls (i.e., score fails to meet criterion for a deficit and is therefore considered to be within normal limits). 3) Patient's score on Task X significantly lower ($p < .05$; two-tailed) than patient's score on Task Y using Crawford, Howell and Garthwaite's (1998) test. The test is two-tailed to allow for the fact that the data are examined before deciding which task is X and which is Y)
Strong Dissociation (i.e., differential deficit)	1) Patient's score on Task X significantly lower than controls (i.e., score meets above criterion for a deficit). 2) Patient's score on Task Y is also significantly lower than controls (i.e., score meets above criterion for a deficit). 3) Patient's score on Task X significantly lower ($p < .05$, two-tailed) than patient's score on Task Y using Crawford, Howell and Garthwaite's (1998) test.
Classical Double dissociation	1) Patient 1 meets the criterion for a deficit on Task X, and meets the criteria for a classical dissociation between this task and Task Y. 2) Patient 2 meets the criterion for a deficit on Task Y and meets the criteria for a classical dissociation between this task and Task X.
Strong Double Dissociation	1) Patient 1 meets the criterion for a deficit on Task X, and meets the criteria for a classical or strong dissociation between this task and Task Y. 2) Patient 2 meets the criterion for a deficit on Task Y and meets the criteria for a classical or strong dissociation between this task and Task X. 3) Both dissociations are not classical (otherwise we have a classical double dissociation).

Criteria for Establishing a Dissociation

Armed with these proposed operational definitions of a "deficit" and "within normal limits" is the typical definition of a dissociation now acceptable? Clearly not, for two related reasons. First, one half of the typical definition of a dissociation essentially requires us to prove the null hypothesis (we must demonstrate that a patient is not different from the controls), whereas, as is well known, we can only fail to reject it. This is particularly germane to single-case studies, where the power to reject the null hypothesis is inevitably low: an individual patient (rather than a group) is compared with a control group, which itself is usually of very modest size.

The second problem with the typical definition is that a patient's score on the "impaired" task could lie just below the critical value for defining impairment and the performance on the other test lie just above it. That is, the difference between the patient's relative standing on the two tasks of interest could be very

trivial; in this situation we would not want to infer the presence of a dissociation, despite the results of the formal statistical tests. This remains a problem, irrespective of whether the critical value is derived from Crawford and Howell's (1998) method or by the *z* method. Therefore, we need to impose an *additional* criterion for a dissociation that focuses on examining the *difference* between the patient's relative standing on the two tasks. It is proposed that this criterion should take the form of a statistically significant difference between the patient's scores on the two tasks (i.e., it should involve an inferential statistical method) and that this test should be conducted by comparing the difference between the patient's *z* scores on Tasks *X* and *Y* with the distribution of differences in the control sample.

This criterion not only deals with the problem of trivial differences, but it also provides us with a *positive* test for a dissociation; otherwise we must rely heavily on what boils down to an attempt to prove the null hypothesis of no deficit on Task *Y*. One possible method for testing the difference between a patient's *z* scores is to divide it by the standard deviation of the difference in the controls (SD_{X-Y}) to obtain a *z* score for the difference (z_D). The z_D can then be referred to a table of the areas under the normal curve to test whether it exceeds the required critical value (i.e., 1.64 for one-tailed significance at the 0.05 level or 1.96 for a two-tailed test). This method, originally proposed by Payne and Jones (1957), has been widely used in clinical research and practice. The standard deviation of the difference is

$$SD_{X-Y} = \sqrt{2 - 2r_{XY}}, \quad (1)$$

where r_{XY} is the correlation between the two tests in the control sample used to obtain the patient's *z*-scores, and the first value under the square root sign (2) is the sum of the *SD*s for the two tests in the control sample (*z* scores have *SD*s of 1).

There is a better alternative to this method that treats the control sample statistics *as* statistics rather than parameters (see below) but, before discussing it, attention will first be given to what is essentially a variant upon the Payne and Jones approach proposed by Shallice (1979). Shallice's (1979) criteria for a dissociation are that a patient should be impaired on Task *X*, within normal limits on Task *Y* (neither of these first two criteria were operationally defined), and that the difference between the patient's *z* scores should be at least 2.0. It can be seen that, unlike the typical definition of a dissociation, this definition does recognise the importance of examining the magnitude of the difference between Tasks *X* and *Y*. However, while acknowledging that "theoretically" (p. 190) it is the *z*-score for the difference (z_D), rather than the simple difference between *z* scores, that should be compared with a critical value, Shallice suggests that the latter difference be used as a proxy for this quantity.

A z_D of 2.0 would often allow us to conclude that the patient's difference is significantly greater ($p < .05$) than the differences between the tasks in controls (subject to the important caveat that the control sample has a large *n* so that the statistics provide good estimates of the parameters, see below). On the other hand, Shallice's (1979) suggestion that we use the simple difference between *z*

scores as a proxy for z_D is difficult to justify. The effort required to calculate z_D is minimal, particularly when set against the effort expended in running and writing up a single-case study. Furthermore, the suggestion is liable to receive little sympathy from group researchers, who typically have to run rigorous and often elaborate statistical analyses in their studies.

More importantly, using the simple difference in *z* scores as a proxy for z_D can lead researchers to make both Type I and Type II errors: that is the practice can lead them not only to conclude that a difference is present when it is not, but also to conclude that a difference is absent when it is present. This is because the simple difference between *z* scores and z_D only have the same value when the correlation between tests is exactly 0.5. From formula (1) it can be seen that, when $r_{XY} = 0.5$, $SD_{X-Y} = 1.0$ and hence a simple difference of 2.0 between *z* scores corresponds exactly to z_D (2.0/1.0 = 2.0). If the correlation is less than 0.5, then using the simple difference will lead to an increase in Type I errors. For example, if the correlation between tests is modest, say 0.2, then $SD_{X-Y} = 1.26$, and dividing the simple difference of 2 by this quantity yields a z_D of 1.58. This is not significant (two-tailed p = 0.11); yet, if the simple difference was used as a proxy, the conclusion would be that the difference was significant.

When the correlation between tests is above 0.5, the opposite problem occurs; there will be an increase in Type II errors. For example, if the two tasks are very highly correlated ($r = 0.85$), $SD_{X-Y} = 0.55$ and, when the raw difference between z scores is 2.0, z_D is 3.65. This difference is highly significant: the two-tailed *p* value is 0.00026. Therefore we can be very confident (much more than 95% confident!) that the patient's difference has not come from the distribution of differences in the healthy population. The crucial point is that, in this scenario, lesser differences would also be highly significant and yet using the simple difference as a proxy would lead us to conclude that we cannot reject the null hypothesis. For example, if the simple difference between *z* scores was 1.90 (thereby failing Shallice's criterion), then $z_D = 3.46$ and the two-tailed p value is 0.00054. It was noted earlier that statistical power is inevitably low in single-case studies. In the scenario just described, a further, severe, unnecessary, and arbitrary reduction in power has been imposed.

Proposed Operational Definitions for Dissociations

To summarise the foregoing discussion: it has been argued that the typical definition of a dissociation needs to be tightened to include fully operational definitions of a "deficit" and "within normal limits". It has also been argued that we need an additional criterion for testing whether the *difference* between the patient's performance on the two tasks is statistically significant. A candidate method for establishing whether this additional criterion is met is that proposed by Payne and Jones (1957). Although the method is clearly preferable to Shallice's (1979) suggestion that we use the simple difference between *z* scores, it too has a problem. This problem is the same as that encountered when examining the use of *z* to test whether a patient's performance on a single task

is significantly different from the controls, namely that the control sample is treated as if it were a population and the statistics of the sample as if they were parameters. While the Payne and Jones method can be used with confidence when an individual is compared to a very large control or normative sample, it would not be appropriate to use it in the vast majority of single-case studies where the samples are generally very small.

To overcome this problem, Crawford et al. (1998) devised a method that does not treat the statistics as parameters. As in the Payne and Jones (1957) method, the patient's scores on the two tasks are converted to z scores based on the control sample's means and SDs. However, the difference between the patient's z scores is divided by the standard *error* of the difference (rather than SD_{X-Y}), yielding a quantity that is distributed as t with $n - 1$ degrees of freedom.

It is proposed that a significant result ($p < 0.05$) on Crawford et al.'s (1998) test should be used to provide the test on the difference between a patient's scores on Task *X* and *Y*. We now have the necessary criteria in place for a fully operational definition of what Shallice (1988) has termed a "classical dissociation" (p. 227). These criteria are (1) that the patient's score on Task *X* meets the criteria for a deficit; (2) the score on Task *Y* fails to meet this criterion (and is therefore considered to be within normal limits); (3) performance on Task *X* is significantly poorer than on Task *Y* (i.e., the difference between the patient's *z* scores is significant). Shallice (1988) identifies another form of dissociation, which he terms a "strong dissociation"[1]. In a strong dissociation, the patient exhibits a deficit on both tasks, but the deficit on Task *X* is greater than that on Task *Y* (i.e., the patient exhibits a differential deficit). A fully operational definition of a strong dissociation can readily be incorporated into the proposed scheme. We need only require that the patient should meet the criteria for a deficit on *both* tasks, and that the performance on Task *X* is significantly poorer than that on Task *Y*.

It is worth stressing that in our proposed criteria the requirement for a significant difference between tasks applies to both strong *and* classical dissociations. In contrast, both Coltheart (2001, p. 12) and Ellis and Young (1996, p. 5) appear to suggest that this requirement need only be invoked when the patient is classified as having a deficit on both tasks (i.e., when the criteria for a classical dissociation have not been met and the issue is with whether the patient nevertheless exhibits a strong dissociation).

Proposed Operational Definitions for Double Dissociations

Having obtained an operational definition for a classical dissociation, the operational definition for a classical double dissociation follows naturally: two patients meet the criteria for a classical dissociation; but the tasks on which the

[1] This choice of term is potentially confusing because, in Shallice's scheme, a strong dissociation constitutes weaker evidence of functionally distinct processes than a classical dissociation; Shallice used it to distinguish the deficit pattern described from an even weaker form of evidence that he termed a "trend dissociation".

deficit is observed are reversed (see Table I). When one or both patients fail to meet the criteria for a classical dissociation but nevertheless meet the criteria for a strong dissociation we have what could be termed a "strong double dissociation" (see Table I).

The Probability of an Individual from the Control Population Fulfilling the Proposed Criteria for Dissociations

The criteria for classical and single dissociations are based on the pattern of results obtained from the application of three inferential tests (two to test for the presence of deficits on Tasks *X* and *Y*, and one on the difference between *X* and *Y*). There is a separate Type I error rate for each of these individual tests but there will also be an *overall* Type I error rate for the criteria we propose for dissociations.

If we define a Type I error as occurring when a member of the control population is classified as exhibiting a dissociation (other definitions of a Type I error are possible) we can estimate this error rate. At first glance it might be thought that the Type I error rate will be high (and will exceed the error rate of the individual tests) because multiple tests have been performed. However, as the series of tests function as a set of hurdles, and you must pass a complete set to enter a particular dissociation classification, this assumption is wrong. It also follows that Type I errors, as defined above, will be very small for *strong* dissociations as the score on Y must be low enough to be rejected at the 0.05 level *and X* must be substantially lower again, in order for the score on task *X* to be significantly lower than *Y*.

Under the assumption that the scores for controls are drawn from a bivariate normal distribution, the overall Type I error rate in identifying a dissociation will be a function of the size of the control sample and the magnitude of the correlation between *X* and *Y*. To estimate the error rates under different values of these variables, we ran a Monte Carlo simulation in which 100,000 samples of N +1 were drawn from each of four bivariate normal distributions. These distributions were specified by setting the correlation between *X* and *Y* at 0.0, 0.2, 0.5, and 0.8 (for each distribution the means of *X* and *Y* were set equal to 0 and their variances to 1). We did this for five values of N; N =5, 10, 20, 50 and 100. The first N items in each sample were taken as the control sample and the sample means and variances of *X* and *Y* were calculated, together with their correlation.

These summary statistics were used to test whether the $N + 1th$ item, which was taken as the control case, exhibited either form of dissociation. To do this, the same sequence of tests were applied to the simulated sample as would be used with real data. The number of occurrences of a control (i.e., the N + 1th case) meeting the criteria for a classical dissociation (< .05 one-tailed on either *X* or *Y* but not both; < .05 two-tailed on difference between *X* and *Y*) was recorded and expressed as a percentage of the total. The same was done for the criteria for a strong dissociation (< .05 one-tailed on both *X* and *Y*; < .05 two-tailed on difference between *X* and *Y*).

TABLE II

Results from a Monte Carlo Simulation Study: Percentage of Control Cases Classified as Exhibiting Dissociations under Different Values of N of the Control Sample and Correlations between Tasks

	Strong Dissociation				Classical Dissociation			
N	$r = 0.0$	0.2	0.5	0.8	$r = 0.0$	0.2	0.5	0.8
5	0.02	0.04	0.11	0.37	3.41	3.16	2.71	2.04
10	0.00	0.01	0.04	0.15	3.00	2.66	2.08	1.34
20	0.00	0.00	0.02	0.08	2.71	2.34	1.79	1.10
50	0.00	0.00	0.01	0.05	2.61	2.14	1.64	0.93
100	0.00	0.00	0.00	0.03	2.45	2.12	1.48	0.90

The results of the simulation are presented in Table II. As a check on the robustness of these results, a second simulation was performed using a different programming language, algorithm for generating random numbers, and method of sampling from the bivariate normal distributions; the results closely matched those reported. It can be seen from Table II that for a strong dissociation, for all the values of the correlation and sample size that were examined, only a tiny number of cases were classified as having a Type 1 error (less than 0.4% in each case). In addition, it can be seen that the proportion showing a classical dissociation was always comfortably less than the nominal value of 5%. This illustrates the conservatism inherent in the sequence of tests for a dissociation.

Confidence Intervals on the Abnormality of a Patient's Score and Score Differences

The foregoing methods are aimed at providing an inferential method for testing for deficits and dissociations. However, the methods also simultaneously provide estimates of the abnormality of the patient's score or score difference. That is, in the case of Crawford and Howell's proposed method for comparing a patient's score against controls, the one-tailed p value also provides an unbiased point estimate of the proportion (or percentage if multiplied by 100) of the control population that would obtain a score lower than the patient's. Crawford and Garthwaite (2002) have recently provided methods for obtaining confidence limits on this percentage.

The confidence limits quantify the uncertainty arising from using sample statistics to estimate population parameters and, in combination with the methods discussed above, provide researchers with information of the form "the estimated percentage of the healthy population that would obtain a score (or score difference) lower than the patient is 2.1% and the 95% CI on the percentage is from 0.2% to 6.7%". The provision of these confidence limits is in keeping with the contemporary emphasis in statistics, psychometrics, and biometrics on the use of confidence limits in research (American Psychological Association, 2001; Daly et al., 1995; Gardner and Altman, 1989; Zar, 1996). For example, the American Psychological Association (2001) states that, "The use of confidence intervals … is strongly recommended" (p. 22). The specific

motivation behind the development of these confidence limits was to provide tools for single-case research that would parallel those taken for granted, or even viewed as mandatory, in group-based research.

A potential area of confusion over the use of these confidence intervals centres round their relationship with the inferential tests for deficits and dissociation. For example, it would be wrong to expect that, if the *p* value for the inferential test on a deficit was less than 0.05 (and consequently the estimate of the percentage of the population obtaining a lower score than the patient was less than 5%), then the upper limit of the CI should also be less than 5%. If, for instance, it was estimated that 4.99% of the population would obtain a lower score than the patient, then *p* equals 0.0499 but even trivial uncertainty about the figure of 4.99% would result in 5% being included in the confidence interval for the percentage of the population with a lower score than the patient.

Implementation of the Inferential Methods

With the exception of the methods for finding confidence limits on the abnormality of a patient's score, the statistical methods outlined above do not require complicated computations. However, as we realised that a few hours work on our part would save many potential hours of work on the part of researchers, we made available computer programs (for PCs) that implement all the methods discussed in this paper. Apart from saving researcher's time, these programs are specifically tailored to these methods, thereby reducing the chance of a user making procedural or computational errors. The programs can be downloaded from the first author's web site (www.psyc.abdn.ac.uk/homedir/jcrawford/abnolims.htm). They have also been placed on the Cortex web site (www.cortex-online.org.).

Some Caveats and One Red Herring

As was noted at the outset, our aim was to develop operational definitions for dissociations in which patients are compared with control samples. Thus the double dissociations with which we are concerned are based on what Shallice (1988) has termed "complementary" dissociations. The vast majority of double dissociations in the cognitive neuropsychology literature are of this form. However, Shallice (1988) argues that complementary dissociations do not necessarily allow us to conclude that we have uncovered separate cognitive subsystems (because of the possibility of resource artefacts). He argues that a firm conclusion is only possible when two patients are *directly* compared with each other, and the comparison reveals that, "…on task I, patient A performs significantly better than patient B, but on task II, the position is reversed" (p. 235).

Shallice's (1988) argument raises many important questions that lie beyond the scope of the present work. However, one point, which can be anticipated from the foregoing discussion, is worth noting. It is necessary to specify how we should test whether an individual patient's test score is significantly different

from that of another individual patient. The present authors are unaware of any inferential method that would allow us to conduct such a test.

It was argued that operational definitions for deficits and dissociations should incorporate inferential statistical methods to test whether the particular criteria are met for an individual patient. The inferential methods used in our proposals are both modified *t*-tests (Crawford and Howell, 1998; Crawford et al., 1998). These tests assume that the control sample data are normally distributed. Therefore, these methods should not be used when the control sample distributions depart substantially from normality, i.e., when the data are highly skewed. Importantly, the more commonly used alternative methods (i.e., the use of *z* to test for a deficit, or z_D to test for a difference between tasks) make exactly the same assumption and will be equally compromised when this assumption is violated.

Skewed control data are not uncommon in single-case studies when the tasks that measure abilities are largely within the competence of most healthy individuals. In this situation, negative skew will occur when the measure of interest is number correct (i.e., there will be ceiling effects) and positive skew when the measure is an error rate (i.e., there will be floor effects). This issue has not received sufficient attention in the single-case literature. For example, *z* has been used for inferential purposes in numerous single-case studies when it is obvious from the means and *SD*s of their control samples that the data are highly negatively skewed (i.e., the *SD* tells us that, were the data normally distributed, a substantial percentage of scores should lie above the maximum obtainable score yet we know that none did). In these studies *two* factors combine to inflate the Type I error rate and exaggerate the patient's deficit: the treatment of control statistics as parameters and the presence of skew. The problems arising from non-normal control data in single-case studies and potential solutions to these problems are currently being addressed in detail elsewhere (Crawford et al., in preparation).

The advantage of the proposed inferential tests for deficits and dissociations over the use of *z* and z_D for the same purpose, is that they account for sampling error in the controls. However, neither the proposed methods, nor the more common use of *z*, allow for the effects of measurement error in the instruments. Very useful and elegant methods have been devised for drawing inferences concerning an individual patient's performance on fully standardized neuropsychological tests; i.e., on tests that have been normed on a large representative sample of the population (e.g., Capitani, 1997; De Renzi et al., 1997; Willmes, 1985). Some of these methods *do* factor in the effects of measurement error but treat the reliability of a test as a parameter. In reality, of course, reliability coefficients are themselves subject to error (Feldt and Brennan, 1983). This is not a problem with very large normative samples but would be a problem were the reliability estimates obtained from samples that have *N*s of the magnitude with which we are concerned (i.e. the reliability estimates would be very unstable).

It is also important that the current methods for testing for dissociations should not be confused with methods that are *solely* concerned with testing for *reliable* differences between an individual's scores on two tasks (such methods

involve variants upon the basic procedure of computing the standard error of measurement of the difference between tasks). Many healthy individuals will have reliable differences among their abilities in different cognitive domains. Indeed, if the tests concerned have high reliability, reliable differences will be *very* common and therefore cannot be taken as indicating an acquired deficit (Crawford, 2003). Therefore, methods that quantify the probability that a patient's score difference was drawn from the distribution of score differences in the healthy population (and thereby focus on the *abnormality* of the patient's score) are more germane to single-case researchers (Crawford and Garthwaite, 2002).

It should be noted that if the proposed criteria for dissociations (or their equivalents that use z or z_D) were applied repeatedly in a study to test many patients there would be inflation of the Type I error rate. In this situation consideration should be given to applying a Bonferroni correction to maintain the experimentwise error rate at the desired level. Because, as noted above, the criteria for dissociations are based on a set of tests that must each be passed, it is not necessary to apply the Bonferroni correction to all of them. Our favoured choice would be to apply the correction to the test on the difference between *X* and *Y* (any of the tests could be chosen, as long as it was selected *a priori*, except the test on *Y* in the criteria for a classical dissociation). Then the Type I error for incorrectly classifying one of the patients as having a dissociation would be no greater than the specified experimentwise level.

Finally, Mycroft et al. (2002) have recently argued that our methods "fail to note the consequences of unequal variance" (p. 294) between controls and a notional population of patients, and have proposed an alternative method. We have responded to this in detail elsewhere (Crawford et al., submitted) but a few observations can be made here. The perceived need for an alternative method is based on Mycroft et al.'s. view that (1) the patient in a single-case study should be considered to have been drawn from a notional population of patients, and (2) that such a notional population is liable to have markedly greater between-subject variability than the population of controls. However, our methods are designed to test the hypothesis that an *individual* patient did not come from a population of controls; therefore the variance of a hypothetical population of patients is not an issue. (Under the null hypothesis, the individual is an observation from a distribution with the same mean and variance as the controls). In contrast, the hypothesis that Mycroft et al. propose should be tested is that the mean of a notional population of patients (from which they have a sample of one) is different from the mean of a population of controls.

Mycroft et al.'s proposed alternative statistical method presents the researcher with the impossible task of arriving at an estimate of the variance of a hypothetical population (whereas for our methods no such estimate is required or appropriate) and is also liable to seriously lower statistical power (i.e., increase Type II errors). Furthermore, even if the modified *t*-test were conceptualised as a test for a difference between population means, Mycroft et al.'s position can be regarded as untenable. They argue that a significant test result could arise in the absence of a difference in population means because of putative differences in variances (the patient population having the larger variance). However, if the

population means do not differ, then the variance in the patient population could only be larger if patients with scores lower than controls were *balanced exactly* by patients with higher scores than controls. Thus we are asked to accept that neurological damage impairs performance in some patients but enhances performance to an equivalent degree in others (Crawford et al., submitted).

CONCLUSION

The single case approach in neuropsychology has made a significant contribution to our understanding of the architecture of human cognition. However, as Caramazza and McCloskey (1988) note, if advances in theory are to be sustainable they "… must be based on unimpeachable methodological foundations" (p. 619). To establish those foundations, fully operational definitions of deficits and dissociations are required. While we realise that the reader may disagree with our specific proposals, we hope they will serve as a useful point of departure for a closer consideration of matters that have hitherto received less than their fair share of attention.

REFERENCES

AMERICAN PSYCHOLOGICAL ASSOCIATION. *Publication Manual of the American Psychological Association*. Washington DC: Author, 2001.

CAPITANI E. Normative data and neuropsychological assessment. Common problems in clinical practice and research. *Neuropsychological Rehabilitation, 7:* 000-000, 1997.

CARAMAZZA A and MCCLOSKEY M. The case for single-patient studies. *Cognitive Neuropsychology, 5:* 517-528, 1988.

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

CRAWFORD JR. *Psychometric foundations of neuropsychological assessment.* In LH Goldstein and J McNeil (Eds), *Clinical Neuropsychology: A Practical Guide to Assessment and Management for Clinicians*. Chichester: Wiley, in press, 2003.

CRAWFORD JR and GARTHWAITE PH. Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia, 40:* 1196-1208, 2002.

CRAWFORD JR, GARTHWAITE PH, HOWELL DC and GRAY CD. Inferential methods for comparing a single case with a control sample: Modified *t*-tests versus Mycroft et al.'s (2002) modified ANOVA. Manuscript submitted for publication.

CRAWFORD JR, GARTHWAITE PH, HOWELL DC and LAWS KR. Inferential statistical methods for neuropsychological single-case studies: The issue of non-normal control data. Manuscript in preparation.

CRAWFORD JR and HOWELL DC. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12:* 482-486, 1998.

CRAWFORD JR, HOWELL DC and GARTHWAITE PH. Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples *t*-test. *Journal of Clinical and Experimental Neuropsychology, 20:* 898-905, 1998.

DALY F, HAND DJ, JONES MC, LUNN AD and MCCONWAY KJ. *Elements of Statistics*. Wokingham: Addison-Wesley, 1995.

DE RENZI E, FAGLIONI P, GROSSI D and NICHELI P. Apperceptive and associative forms of prosopagnosia. *Cortex, 27:* 213-221, 1997.

DUNN JC and KIRSNER K. What can we infer from double dissociations? *Cortex, 39:* 1-7, 2003.

ELLIS AW and YOUNG AW. *Human Cognitive Neuropsychology: A Textbook With Readings*. Hove: Psychology Press, 1996.

FELDT LS and BRENNAN RL. Reliability. In RL Linn (Ed), *Educational Measurement*. New York: Macmillan, 1983.

GARDNER MJ and ALTMAN DG. *Statistics With Confidence-Confidence Intervals and Statistical Guidelines*. London: British Medical Journal, 1989.

MYCROFT RH, MITCHELL DC and KAY J. An evaluation of statistical procedures for comparing an individual's performance with that of a group of controls. *Cognitive Neuropsychology, 19:* 291-299, 2002.

NEWCOMBE F and MARSHALL JC. Idealisation meets psychometrics: The case for the right groups and the wrong individuals. *Cognitive Neuropsychology, 5:* 549-564, 1988.

PAYNE RW and JONES G. Statistics for the investigation of individual cases. *Journal of Clinical Psychology, 13:* 115-121, 1957.

SARTORI G. *From models to neuropsychological data and vice versa*. In G Denes, P Bisiacchi, C Semenza and E Andreewsky (Eds), *Perspectives in Cognitive Neurospychology*. Hove: Erlbaum, 1988, pp. 59-73.

SHALLICE T. Case study approach in neuropsychological research. *Journal of Clinical Neuropsychology, 3:* 183-211, 1979.

SHALLICE T. *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press, 1988.

SOKAL RR and ROHLF JF. *Biometry*. San Francisco, CA: W.H. Freeman, 1995.

WILLMES K. An approach to analyzing a single subject's scores obtained in a standardized test with application to the Aachen Aphasia Test (AAT). *Journal of Clinical and Experimental Neuropsychology, 7:* 331-352, 1985.

ZAR JH. *Biostatistical Analysis*. London: Prentice-Hall, 1996.

John R. Crawford, Department of Psychology, King's College, University of Aberdeen, Aberdeen AB24 2UB, United Kingdom. e-mail: j.crawford@abdn.ac.uk