Running Head: Single-Case Methods

Neuropsychologia, in press.

Reviews and Perspectives

On comparing a single case with a control sample: An alternative perspective

John R. Crawford[1], Paul H. Garthwaite[2], and David C. Howell[3]

[1]School of Psychology, University of Aberdeen, UK, [2]Department of Statistics, The Open University, UK, and [3]Department of Psychology, University of Vermont, Burlington, USA.

Address correspondence to: Professor John R. Crawford, School of Psychology, King's College, University of Aberdeen, Aberdeen AB24 3HN, UK. Tel: +44 (0)1224 272231, e-mail: j.crawford@abdn.ac.uk

Abstract

Corballis (2009) offers an interesting position paper on statistical inference in single-case studies. The following points arise: (1) Testing whether we can reject the null hypothesis that a patient's score is an observation from the population of control scores can be a legitimate aim for single-case researchers, not just clinicians. (2) Counter to the claim made by Corballis (2009), Crawford and Howell's (1998) method *does* test whether we can reject the above null hypothesis. (3) In all but the most unusual of circumstances Crawford and Howell's method can also safely be used to test whether the mean of a notional patient population is lower than that of a control population, should neuropsychologists wish to construe the test in this way. (4) In contrast, the method proposed by Corballis is not legitimate for either of these purposes because it fails to allow for uncertainty over the control mean (as a result Type I errors will not be under control). (5) The use of a mixed ANOVA design to compare a case to controls (with or without the adjustment proposed by Corballis) is beset with problems but these can be overcome using alternative methods.

Introduction

Caramazza (1988) noted that, if single-case studies are to provide sustainable advances in cognitive theory, they "… must be based on unimpeachable methodological foundations" (p. 619). Although the logical basis of the single-case approach has been the subject of detailed scrutiny (e.g., Coltheart, 2001; Shallice, 1988), questions surrounding the statistical analysis of single-case data have received much less attention. Therefore, the position paper offered by Corballis (2009) is to be welcomed and will hopefully generate a debate that goes beyond this current response. It is also the case that Neuropsychologia is an appropriate forum for such a debate as, notwithstanding the existence of journals specialising in single-case research (e.g., Cognitive Neuropsychology, Neurocase), it probably publishes more single-case studies than any other neuroscience journal. However, although we agree that the issues raised in the paper are of fundamental importance for single-case research, we disagree with most of the conclusions drawn and the solutions proposed.

Two forms of null hypothesis when comparing a single case to controls

Corballis (2009) states that, "Crawford and Howell implicitly assume that the investigator wishes to make inferences about the population to which the case belongs, whereas the aim may be simply to determine whether the case itself belongs in the control population" (p. 3). He also proposed a modification to the Crawford and Howell method as a means of achieving this second aim and suggested that this modified method would primarily be of relevance to clinical practice, rather than single-case research.

We agree with Corballis that there are two possible forms of null hypothesis when comparing a case to controls. It is also true that the method proposed by

Crawford and Howell has been construed by some authors as a test for a difference in population means, both in a broader context (Sokal & Rohlf, 1995) and specifically when comparing a patient's score to a control sample (Mycroft, Mitchell, & Kay, 2002). However, this is decidedly not the position taken by Crawford and colleagues (Crawford, Garthwaite, Howell, & Gray, 2004). For example, Crawford and Garthwaite (2006) state, "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" (p. 891).

Crawford and colleagues were primarily motivated to adopt this position by statistical considerations. However, a number of influential theorists have come to exactly the same conclusion based on neuropsychological and theoretical considerations (e.g., Caramazza, 1986; Caramazza & McCloskey, 1988; Coltheart, 2001). They argue 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).

The above discussion demonstrates two points: First, it makes it clear that the Crawford and Howell method need not be seen as a test for a difference in population means, and second, it highlights that testing whether a patient's scores are observations from a population of control scores can be a legitimate aim for the cognitive neuroscientist, not just the clinician.

If, as is common among advocates of the single-case approach, we are willing to make the assumption of a universal cognitive architecture (Caramazza, 1986; Caramazza & McCloskey, 1988), then the performance of a single-case can constitute a Popperian "Black Swan". For example, a patient may exhibit a dissociation between two tasks and this single piece of evidence is sufficient to overturn any amount of prior evidence that the tasks measure a unitary (i.e., non-dissociable) function. That is, we do not need to be able to generalize such a result to a notional population of patients for it to have implications for cognitive theory. Of course, we must be confident that the black swan is genuine, rather than a white swan that has had a chance encounter with an oil slick. In other words, the logic may be sound but we must be sure that our inferential statistical methods are rigorous (this issue is discussed in a later section).

Testing for a difference in population means

As noted above, we view Crawford and Howell's method as testing whether we can reject the null hypothesis that a patient's score is an observation from the population of control scores. However, it is worth considering whether this method could also be seen as a legitimate test of whether the mean of a notional patient population is lower than that of the control population (i.e., can we use it to test the null hypothesis of equivalence of means against the directional hypothesis that the patient population mean is lower?)

As Corballis (2009) points out, the problem is that, although we have an estimate of the variance of the control population, we only have a single patient in hand and so we do not have an estimate of the variance of the notional patient population. He also suggests that researchers must be willing to make the assumption

that this latter variance is equal to the variance of the control population if the Crawford and Howell method is to be used to test for a difference in means. Moreover, Mycroft, Mitchell and Kay (2002) have suggested that this assumption will not hold: rather we should expect the variance of a notional patient population to be (considerably) larger than the control population. They argue that the Crawford and Howell method will therefore fail to control Type I errors in this scenario. That is, the method will incorrectly conclude that the patient mean is lower than the control mean at a rate greater than the nominal rate set by the investigator.

In view of the foregoing, the prospects for using the Crawford and Howell method to test for a difference in population means do not look good. However, these apparent problems are paper tigers. The first point to appreciate is that we need only be concerned with increased variance in the notional patient population when this occurs in the absence of a difference in the population means. If the patient population mean is lower than the control mean then (regardless of how minimal this difference may be) the issue of a Type I error does not arise: There *is* a difference in the means and the question becomes one of the power to detect it. Therefore, the only scenario in which we need be concerned with inflated Type I errors is when the variance of the patient population is increased *without* a corresponding lowering of the patient population mean relative to the control mean.

Fortunately, this scenario is very unrealistic: neurological damage would have to have the effect of increasing the variability of patients' scores relative to matched controls without producing even a minimal lowering of the patient mean. Thus, every patient who suffers impairment on a given cognitive task would need to be balanced by another patient for whom neurological damage enhanced their performance to a comparable degree (if the impairments and enhancements of performance are not in

balance then, as noted, the means will differ and the concern over Type I errors evaporates). It can be concluded therefore that, although we do not construe the Crawford and Howell method as testing whether a patient population mean is lower than that of controls, it can be used for this purpose without being concerned with Mycroft et al's. misplaced warnings over inflation of the Type I error rate.

The Crawford & Howell method versus the Corballis method: statistical theory

It is still quite common in single-case studies for inferences to be based on converting a patient's raw score to a *z*-score and obtaining a probability for this *z* from a table of areas under the normal curve (or algorithmic equivalent). The problem with this method is that the mean and standard deviation of the control sample are both inappropriately treated as population parameters. In contrast, the Crawford and Howell method treats both the mean and standard deviation as what they are: *sample* means and standard deviations. The formula for this test is

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

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 is negative and its magnitude exceeds the one-tailed 5% critical value for t on n - 1 degrees of freedom, then it can be concluded that the patient's score is sufficiently low to enable rejection of the null hypothesis that it is an observation from the scores of the control population (the patient is therefore considered to exhibit an impairment on the task in question). The method proposed by Corballis (2009) differs only in that it omits the right hand term in the numerator, i.e.,

$$\sqrt{\frac{n+1}{n}}. \qquad (2)$$

With the omission of this term the formula reduces to that for a *z*-score although, in common with Crawford and Howell's method, it is proposed that the result is evaluated against a *t* distribution on $n-1$ df, rather than against a standard normal distribution.

What then is the effect of dropping the term in (2)? For controls, the population mean and standard deviation are μ and σ, while the sample mean and sample standard deviation are $\bar{x}$ and *s*. The observed value for the case is x^* and we suppose the population distribution is normal. Suppose *X* is the value for a randomly selected control. One of the following three quantities should be calculated to decide if x^* is too far from the population/sample mean to be the value of *X*.

1. If μ and σ are both known:

$$\Pr(X-\mu > |x^*-\mu|) = \Pr\left(z > \frac{|x^*-\mu|}{\sigma}\right).$$

2. If μ is known but σ is unknown:

$$\Pr(X-\mu > |x^*-\mu|) = \Pr\left(z > \frac{|x^*-\mu|}{s}\right).$$

3. μ and σ are both unknown:

$$\Pr(X-\bar{x} > |x^*-\bar{x}|) = \Pr\left(t > \frac{|x^*-\bar{x}|}{s\sqrt{\frac{n+1}{n}}}\right).$$

Thus the test statistics for these three tests are:

$$z = \frac{x^*-\mu}{\sigma}, \quad t = \frac{x^*-\mu}{s} \text{ and } t = \frac{x^*-\bar{x}}{s\sqrt{\frac{n+1}{n}}}.$$

The first of these is the *z*-score given by Corballis in his equation (1) and the third is the test statistic given by the Crawford and Howell method, i.e., equation (1) in this commentary and equation (3) in Corballis. The second test statistic, $t = (x^* - \mu) / s$, appears in neither paper, but replacing μ by $\bar{x}$ in this formula gives $t = (x^* - \bar{x}) / s$, which is the test statistic recommended by Corballis. That is, the effect of dropping the term in (2) is to treat the sample mean, $\bar{x}$, as though it were the population mean, μ. The other obvious consequence from dropping the term in (2) is to increase the value of the test statistic, by the factor $\sqrt{(n+1)/n}$. This will make the case's value seem more unusual than is warranted.

It can be seen then that by omitting the term in (2) the Corballis method ignores the uncertainty over the population mean when comparing the case to controls. Given the control sample sizes that typify single-case research, this uncertainty can be considerable (i.e., the sample mean will often provide a poor estimate of the population mean). The upshot is that the Type I error rate for the Corballis method will be inflated. That is, the method will incorrectly reject the null hypothesis that the case's score is drawn from the population of scores of controls at a rate greater than the rate specified by the investigator (for example, if an investigator sets alpha at the conventional level of 5%, the actual rate will be above 5%).

The difference in the results provided by the two methods will be relatively modest but there is no reason why a sound test should be replaced with an approximate test, particularly when the existing method does not require complicated calculations. (Moreover, for Crawford and Howell's test, a computer program relieves the neuropsychologist of the need to perform any of the calculations involved; see Crawford and Garthwaite, 2002 for details).

In summary, when *z* is used to draw inferences concerning the score of a single-case there is a failure to allow for the uncertainty over the control population mean and standard deviation. Although the Corballis (2009) method allows for the uncertainty over the population standard deviation, it fails to allow for the uncertainty over the population mean. In contrast, Crawford and Howell's method allows for both uncertainties. In passing, note that a further method for drawing inferences concerning an individual case has been proposed by Bridges and Holler (2007) that "completes the set". That is, the method attempts to allow for the uncertainty over the control mean whilst ignoring the uncertainty over the control standard deviation; see Crawford and Garthwaite (2008) for a mathematical and empirical critique of this latter method.

Crawford & Howell method versus the Corballis method: empirical evaluation using Monte Carlo simulation

The foregoing treatment indicates that the method proposed by Corballis (2009) is flawed in that it will not maintain the Type I error rate at the specified level. However, for neuropsychologists with a limited knowledge or interest in statistical matters, an empirical comparison of the performance of the Crawford and Howell and Corballis methods may be more convincing than an appeal to statistical theory alone. Therefore, in this section, a Monte Carlo simulation is performed to evaluate control of the Type I error rate for both methods.

The simulation was conducted using control sample sizes of 3, 5, 10, 20, and 50. For each of these five sample sizes, four million trials were performed in which the requisite number of controls were drawn from a normal distribution which (with no loss of generality) had a mean of 50 and standard deviation of 10. On each trial a

further observation (representing a case) was drawn from the same distribution. Crawford and Howell's test and the Corballis test were then applied: the number of trials in which the case was significantly ($p < 0.05$, one-tailed) below the controls was recorded separately for each test. Note that, in this simulation, both the control sample and the case were drawn from the same population. Therefore, any statistically significant results constitute a Type I error (the null hypothesis, that the case's score is an observation from the population of control scores, has incorrectly been rejected). It can also be seen that this simulation directly tests the assertion of Corballis that his method tests "whether the case itself belongs in the control population" (p. 3).

In the foregoing section on statistical theory it was suggested that the Corballis method will fail to control the Type I error rate because it fails to allow for the uncertainty in estimating the population mean for controls from a sample mean. To demonstrate that this is indeed the cause of its problems a further simulation was run in which, on each trial, the population mean of 50 was substituted for the sample mean for that trial. That is, the *population* mean was subtracted from the case's score and the result divided by the *sample* standard deviation; the figure thus obtained was evaluated against a *t*-distribution on $n-1$ df. If the problem with the Corballis method has been correctly identified then the Type 1 error rate should be 5% for this modified version at all sample sizes.

The results of the main simulation are plotted in Figure 1 and are also presented in Table 1. This table records the percentage of Type I errors that occurred at each control sample size for both methods; the results of the second simulation are also presented (these will be discussed later). If the tests are sound then the error rates should closely match the specified Type I error rate of 5% for all sample sizes. It can

be seen that the results for the Crawford and Howell test equal the specified rate to two decimal places (a minor exception occurs for a control sample size of 20, but discrepancies of this magnitude are attributable to Monte Carlo variation). For the Corballis method the Type I error rate is inflated above the specified rate at all sample sizes. It can be seen, however, that the effects are not severe and that the error rates converge on the specified rate as sample size increases. This convergence occurs because, when *n* grows large, the term in (2) approaches 1 and therefore the effects of omitting it are attenuated.

When the Corballis method was modified by substituting the control population mean for the sample mean it can be seen (from the last row of Table 1) that, for all sample sizes, the Type I error rates are at, or very close to, the specified rate. Thus, this second simulation provides an empirical demonstration that the inflation of the error rate for the standard Corballis method occurs because the method fails to account for the uncertainty over the population mean. Note that the modification made to the Corballis formula in this second simulation was introduced purely for didactic reasons and should not be seen as "fixing" the method. In practice, a researcher does not know the value of the population mean.

Estimating the proportion of the population with lower scores than a case

A very useful feature of Crawford and Howell's method is that the *p* value obtained from the test is also the optimal point estimate of the proportion of the control population that will obtain a lower score (it is probably more convenient to multiply this proportion by 100 to express it as a percentage). That is, the method provides a point estimate of the level of abnormality of a patient's score (Crawford et al., 2004). Thus, for example, if the *p* value is 0.023, then 2.3% of the control population are

expected to obtain lower scores. For a mathematical proof that the *p* value serves this dual function see Crawford and Garthwaite (2006).

It has been shown that the Corballis method does not yield the same results as the Crawford and Howell method and so it follows that it cannot provide this optimal point estimate. Were the *p* value from the Corballis test to be used for such a purpose it would exaggerate the abnormality of a case's score. As was the case for the significance tests, this effect will be relatively modest, even for modest control sample sizes. However, again it makes no sense to replace a method that performs correctly with an incorrect method.

Before leaving this topic note that the point estimate of the abnormality of the patient's score provided by Crawford and Howell's (1998) method can be supplemented with an interval estimate using methods developed by Crawford and Garthwaite (2002). The provision of these interval estimates is in keeping with the contemporary emphasis in psychology on the use of confidence limits in research (APA, 2001). Moreover, the methods developed in Crawford and Howell (1998) and Crawford and Garthwaite (2002) are classical statistical methods but recent work has shown that a Bayesian analysis of this problem yields the same results (Crawford & Garthwaite, 2007). This convergence is reassuring regardless of whether a researcher is classical, Bayesian, or eclectic in their orientation. In contrast, the Corballis method does not provide interval estimates, nor will it exhibit convergence with the results from a Bayesian analysis.

Extension of the Corballis method to ANOVA

Corballis (2009) suggests extending the principal of adjusting the Crawford and Howell formula to more complex ANOVA designs (e.g., mixed factorial

designs). As the term in (2) is applied automatically to the Mean Squares when running an ANOVA, multiplying the Mean Squares by this term would cancel it out. The *F* values required to test for the significance of effects would then be calculated using these adjusted Mean Squares. The suggestion of Corballis is that, as was the case for the proposed adjustment to the Crawford and Howell method, this will allow researchers to test whether the patient's score, or combination of scores, were drawn from the control population, rather than test whether there are differences between control and patient populations.

However, this proposal suffers from exactly the same problem as those outlined earlier. This is most easily appreciated by taking the simplest case of a one-way ANOVA comparing the score of a patient to controls. The *p* value obtained from the standard ANOVA will be identical to the *p* value obtained from application of Crawford and Howell's method, because *F* on $[1, n-1]$ df equals t^2 on $n-1$ df. Similarly, the *p* value obtained from the ANOVA with the Corballis adjustment will be identical to the *p* value obtained from the Crawford and Howell method with the Corballis adjustment. Therefore, the Type I error rate for this adjusted ANOVA will not be maintained at the specified rate and, as Corballis proposes making exactly the same adjustment to the Mean Squares for factorial ANOVAs, the same holds for more complicated designs.

Of cows and canaries: The use of factorial ANOVA (without the Corballis adjustment) to test for dissociations in the single-case

Having seen that the ANOVA adjustment Corballis (2009) suggests is not sound, attention is now given to his broader suggestion that the interaction term from a mixed ANOVA (without the Corballis adjustment) provides a means of drawing inferences

concerning a single-case. Consider the simplest example of such a design: A two-factor mixed ANOVA with a between-subjects factor (group, i.e., case vs. controls) and a within-subjects factor (Task *X* vs. Task *Y*). A significant interaction between group and task could be interpreted as demonstrating that the case performed significantly more poorly on (say) Task *Y* than on Task *X*. That is, it could be claimed that the case exhibits a dissociation.

This is indeed a tempting prospect for the single-case researcher given that (a) dissociations have a central role in neuropsychology (Crawford, Garthwaite, & Gray, 2003), and (b) as Corballis points out, ANOVAs can easily be set up and run in all standard statistical packages. However, as detailed below, the use of a mixed ANOVA for this purpose is beset with problems.

In single-case research it will typically be the case that the control sample standard deviations for the two tasks will differ, and often these differences will be substantial. For example, a neuropsychologist may wish to test for a dissociation between performance on say a measure of executive functioning (suppose this task has a mean of 100 and a standard deviation of 15) and a Theory of Mind task (suppose this latter task has a mean of 20 and standard deviation of 5.5). Although in these circumstances it is obvious that the main effect of task is meaningless, it might be thought that the interaction term can still safely be interpreted because it is somehow immune to the effects of the differences between the task's standard deviations. This is decidedly not the case and is an example of a more general problem that Capitani, Laiacona, Barbarotto and Cossa (1999) label the "Cows and Canaries Problem". When two tasks vary in their standard deviations, the task with the larger standard deviation will exert a larger effect on the distribution of difference scores derived from them. Indeed, when the difference in standard deviations is

extreme, the difference scores will be almost perfectly correlated with scores on the task having the larger standard deviation (it is the difference in the standard deviations that is the root of the problem, not the difference in means).

The problems this causes for attempts to detect dissociations using a mixed ANOVA are best illustrated with concrete examples. Table 2 presents examples of data from a control sample and a single-case for two tasks (Tasks *X* and *Y*). Without loss of generality the control sample size is 20 and the correlation between tasks in the controls is 0.5 in all these examples. In scenario A the means (100) and standard deviations (10) of the controls are the same for both tasks and the patient is one standard deviation below the mean on Task *X* and three standard deviations below the mean on Task *Y*. Analysis of these data using a mixed ANOVA yields a non-significant *F* value for the group by task interaction ($p > 0.05$). Scenario B is identical in every respect except that, on Task *Y*, the control mean is 1000 and the control standard deviation is 100. The interaction is highly significant ($p < 0.01$). Note that, crucially, the standing of the patient relative to controls on both tasks is identical in these two examples (see the columns presenting the patient's *z*-scores on the tasks), and yet radically different results are obtained. Note also that Scenarios A and B could be obtained from exactly the same experiment, Scenario B would be obtained if a neuropsychologist chose to award a credit of 10 points for each item passed on Task *Y*, Scenario A would be obtained if only one point was awarded per item.

Scenario C is designed to provide a concrete illustration that the Cows and Canaries Problem stems from differences in the control sample standard deviations, and not from differences in means. In this scenario the control means are identical on the two tasks, only the standard deviations differ. It can be seen that the *F* value and

its associated probability value is identical to Scenario B (in which both the standard deviations and means differed).

By far the most striking illustration of the problem is provided by Scenario D. Here, unlike the three previous examples, the patient's standing on the two tasks is identical: The patient's scores are three standard deviations below the mean on both Task *X* and Task *Y*. Thus, there is absolutely no evidence of a dissociation in this case. However, the mixed ANOVA yields a significant interaction ($p = 0.012$). This result occurs because the interaction term is simply reflecting the fact that the patient has performed poorly on Task *Y*.

The scenarios set out in Table 2 were intentionally chosen to be extreme so that they provide a clear illustration of the problems associated with the use of mixed ANOVA to detect dissociations. However, these problems will occur, albeit in an attenuated form, whenever the standard deviations of the two tasks (or *k* tasks) differ. That is, a neuropsychologist may face a "Cows and Sheep" problem, or even just a "Cows and Smaller Breed of Cows problem", but it is still a problem.

A potential solution to this problem has probably already suggested itself to most readers: The scores of the controls and patient should be standardized (that is converted to z-scores), in both cases using the mean and standard deviation of the control sample to achieve this. The ANOVA would then be performed on these transformed scores. It can be seen that this "pre-standardized" ANOVA would prevent the entirely unsatisfactory outcome observed for Scenario D because the data would carry the fact that the patient's standing relative to controls is identical on both tasks (indeed the *F* value for the interaction would be 0).

Standardization improves matters but has its own problems. By standardizing the scores of the patient using the control mean and standard deviation we introduce

additional uncertainty. The patient's *z*-scores are, in reality, *t*-variates (because they have been obtained using sample means and standard deviations rather than population means and standard deviations). As a result the interaction for the "pre-standardized" mixed ANOVA will have an inflated Type I error rate.

Fortunately there are solutions to these problems, one based on classical statistics, the other based on a Bayesian approach. Considering the classical solution first: It was noted above that, following standardization, the patient's scores are *t*-variates. Therefore a method is required that tests for a difference between two *t*-variates. This was the approach adopted by Crawford and Garthwaite (2005) in developing the Revised Standardized Difference Test (RSDT). The test is based on asymptotic expansions performed to obtain the standard error for such differences (Garthwaite & Crawford, 2004). This is a useful test and Monte Carlo simulations demonstrate that it successfully controls the Type I error rate (Crawford & Garthwaite, 2005).

However, an even better solution is to use a Bayesian method developed by Crawford and Garthwaite (2007), the Bayesian Standardized Difference Test (BSDT). The RSDT is solely concerned with the standardized difference between the patient's scores and is blind as to how such a difference was obtained. For example, suppose the difference between a patient's standardized scores was 2.0; this difference could be obtained if the patient was one standard deviation above the control mean on Task *X* and one standard deviation below the mean on Task *Y*. However, it could also be obtained if the patient was two standard deviations below the mean on Task *X* and four standard deviations below on Task *Y*.

In the latter scenario there is more uncertainty over the patient's standardized scores. To appreciate this, consider the situation in which the control mean and

standard deviation for a task is 100 and 10 respectively and suppose that a patient obtains a score of either 98 or 80. These convert to *z*-scores of -0.2 and -2.0. The control standard deviation is a *sample* standard deviation and hence the real standing of the patient's score in the control population may have been underestimated (if the population standard deviation is larger than the sample standard deviation) or overestimated (if the population standard deviation is smaller than the sample standard deviation). Error in estimating the population standard deviation will have a larger effect on the more extreme *z*-score. For example, if the population standard deviation is actually 11 (in the interest of simplicity we suppose that by chance the sample mean equalled the population mean) then the *z*-score of -0.2 should actually be -0.18 (a modest difference), whereas the z-score of -2 becomes -1.8 (a much larger difference).

Fortunately using Bayesian Monte Carlo methods (as implemented in the BSDT) it is relatively straightforward to allow for the fact that the uncertainty over the z-scores varies as a function of their extremity (in contrast, classical statistical methods cannot cope with this variation in the degree of uncertainty).

Thus, in concluding this section, we suggest that the BSDT should be the test of choice when testing for a dissociation between a patient's performance on two tasks. (Simple to use software for performing a BSDT is freely available at www.abdn.ac.uk/~psy086/dept/BayesSingleCase.htm.) In contrast to the BSDT, the "unstandardized" mixed ANOVA suggested by Corballis (2009) will suffer from the Cows and Canaries Problem, the "pre-standardized" variant will fail to control Type I error rates, and the RSDT (although it does control the overall error rate for a standardized difference) does not allow for the variation in uncertainty over a patient's standardized scores as a function of the extremity of the scores.

Conclusion

It is to be hoped that the interesting position paper by Corballis (2009) will encourage single-case researchers to take a deeper interest in the inferential methods they employ. In addition, as mixed ANOVA designs have been used to draw inferences concerning single-cases, it is very useful to see the rationale for their use spelled out explicitly. However, the issues involved in the analysis of the single-case (particularly those surrounding the detection of dissociations) are much more complex than Corballis (2009) acknowledges and the methods he proposes are unsound. Fortunately, alternative methods provide solutions to the problems identified in earlier sections of the present paper.

References

APA. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington DC: Author.

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative sudies in pediatric neuropsychology. *Child Neuropsychology, 13*, 528-538.

Capitani, E., Laiacona, M., Barbarotto, R., & Cossa, F. M. (1999). How can we evaluate interference in attentional tests? A study based on bi-variate non-parametric tolerance limits. *Journal of Clinical and Experimental Neuropsychology, 21*, 216-228.

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.

Corballis, M. C. (2009). Comparing a single case with a control sample: Refinements and extensions. *Neuropsychologia, in press*.

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. (2006). Methods of testing for a deficit in single case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology, 23*, 877-904.

Crawford, J. R., & Garthwaite, P. H. (2007). Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach. *Cognitive Neuropsychology, 24*, 343-372.

Crawford, J. R., & Garthwaite, P. H. (2008). On the "optimal" size for normative samples in neuropsychology: Capturing the uncertainty associated with the use of normative data to quantify the standing of a neuropsychological test score. *Child Neuropsychology, 14*, 99-117.

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.

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

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.

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.

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, CA: 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.

Table 1. Results of Monte Carlo simulation: Percentage of Type I errors for two methods of comparing a case to controls, in which the null hypothesis is that the case's score is an observation from the population of control scores (results are also shown for a variant of the Corballis method in which the population mean is substituted for the sample mean)

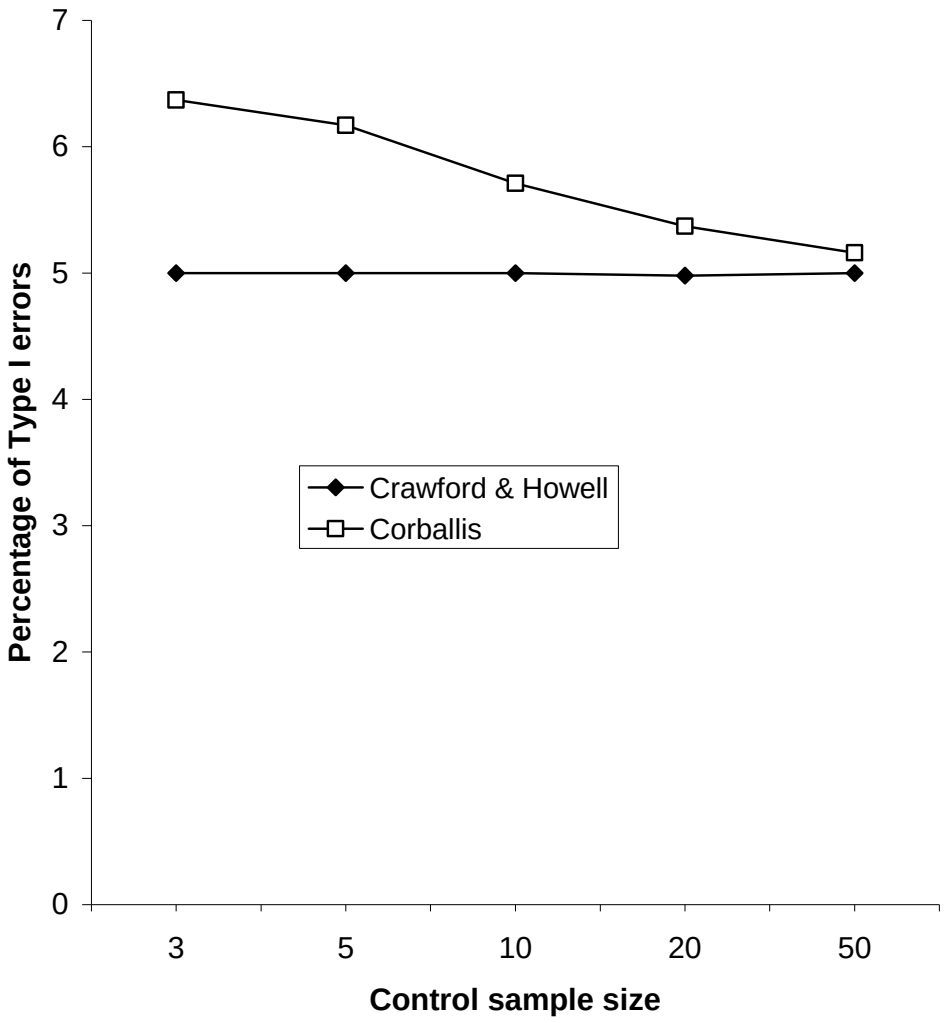
	Control sample size				
	3	5	10	20	50
Crawford & Howell (1998)	5.00	5.00	5.00	4.98	5.00
Corballis (2009)	6.37	6.17	5.71	5.37	5.16
Corballis (2009) – substituting μ for $\bar{x}$	5.02	5.00	5.01	4.99	4.99

Table 2. Demonstration of the problems in using the interaction term from a mixed ANOVA comparing a patient with controls to test for a dissociation in the single-case; in all scenarios the control sample *n* was 20, the correlation (*r*) between the tasks was 0.5, and a two-tailed test was employed.

	Control sample statistics				Case's raw scores		Case's *z*-scores		ANOVA results	
Scenario	Mean *X*	Mean *Y*	SD_X	SD_Y	Task *X*	Task *Y*	Task *X*	Task *Y*	*F*	*p*
A	100	100	10	10	90	70	–1.0	–3.0	3.81	0.066
B	100	1000	10	100	90	700	–1.0	–3.0	8.80	0.008
C	1000	1000	10	100	990	700	–1.0	–3.0	8.80	0.008
D	1000	1000	10	100	970	700	–3.0	–3.0	7.63	0.012

Figure Legend

Fig. 1 Type I error rates (nominal error rate = 5%) estimated by Monte Carlo simulation for the methods of Crawford & Howell (1998) and Corballis (2009)


7
6
5
4
3
2
1
0
Percentage of Type I errors
Crawford & Howell
Corballis
3
5
10
20
50
Control sample size