


AUTHOR QUERY FORM

 ELSEVIER	Journal: CORTEX Article Number: 688	Please e-mail or fax your responses and any corrections to: E-mail: corrections.esnl@elsevier.tnq.co.in Fax: +31 2048 52789
--	--	---

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult <http://www.elsevier.com/artworkinstructions>.

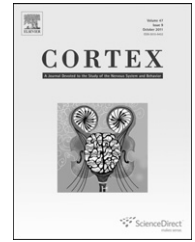
Any queries or remarks that have arisen during the processing of your manuscript are listed below and highlighted by flags in the proof.

Location in article	Query / Remark: Click on the Q link to find the query's location in text Please insert your reply or correction at the corresponding line in the proof
Q1	Kindly check the edit made to the article title.
Q2	Please check the affiliation 'b', if it is two different departments, then split it into two different affiliations.
Q3	Kindly check the phrase '100p%' in the sentence 'The standard error is...' and correct if necessary.
Q4	Kindly update Ref. 'Crawford et al., in press'.

Thank you for your assistance.



ELSEVIER

available at www.sciencedirect.comjournal homepage: www.elsevier.com/locate/cortex

Highlights

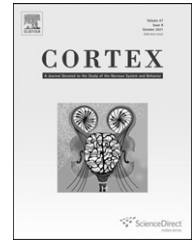
► Five t-tests used to compare a case to controls were examined. ► Three of these methods are shown to be equivalent and not fit for purpose. ► Two of the methods are sound and give equivalent results. ► Crawford & Howell's method is to be preferred as it provides additional information.

UNCORRECTED PROOF



ELSEVIER

available at www.sciencedirect.com

journal homepage: www.elsevier.com/locate/cortex

Research report

Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls

John R. Crawford^{a,*} and Paul H. Garthwaite^b

^a School of Psychology, University of Aberdeen, Aberdeen, UK

^b Department of Mathematics and Statistics, The Open University, Milton Keynes, UK

ARTICLE INFO

Article history:

Received 16 March 2011

Reviewed 31 May 2011

Revised 27 June 2011

Accepted 29 June 2011

Action editor Sergio Della Sala

Published online xxx

Keywords:

Single-case methods

Case-controls design

t-tests

Neuropsychological methods

ABSTRACT

Five inferential methods employed in single-case studies to compare a case to controls are examined; all of these make use of a t-distribution. It is shown that three of these ostensibly different methods are in fact strictly equivalent and are not fit for purpose; they are associated with grossly inflated Type I errors (these exceed even the error rate obtained when a case's score is converted to a z score and the latter used as a test statistic). When used as significance tests, the two remaining methods (Crawford and Howell's method and a prediction interval method first used by Barton and colleagues) are also equivalent and achieve control of the Type I error rate (the two methods do differ however in other important aspects). A number of broader issues also arise from the present findings, namely: (a) they underline the value of accompanying significance test results with the effect size for the difference between a case and controls, (b) they suggest that less care is often taken over statistical methods than over other aspects of single-case studies, and (c) they indicate that some neuropsychologists have a distorted conception of the nature of hypothesis testing in single-case research (it is argued that this may stem from a failure to distinguish between group studies and single-case studies).

© 2011 Elsevier Srl. All rights reserved.

1. Introduction

The focus of the present paper is on single-case studies in neuropsychology that employ the case-controls design; that is, studies in which inferences concerning the cognitive performance of a single case are made by comparing the case to a sample of matched healthy controls. Until relatively recently the standard way of testing for a difference between a case and controls was to convert the case's score to a z score using the control sample mean and standard deviation (SD), and refer the resultant value to a table of areas under the

normal curve or algorithmic equivalent. If z was less than -1.645 (assuming a one-tailed test) then it was concluded that the case was significantly lower than controls.

The problem with this method is that it does not allow for the uncertainty over the control mean and SD; it treats the control sample means and SD as though they were known control population means and SD. The result is that the use of z leads to inflated Type I errors (in this context a Type I error occurs when it is concluded that the case's score is not an observation from the population of controls' scores) and a corresponding exaggeration of the abnormality of a case's scores. As suggested by

* Corresponding author. School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 2UB, UK.

E-mail address: j.crawford@abdn.ac.uk (J.R. Crawford).

0010-9452/\$ – see front matter © 2011 Elsevier Srl. All rights reserved.

doi:10.1016/j.cortex.2011.06.021

Crawford and Howell (1998), the solution is to treat the control mean and SD as what they are, that is, as a sample mean and a sample SD, and use a t-test to test for a difference between the case and controls. However, in the course of reviewing the single-case literature, it became apparent that Crawford and Howell's (1998) method is only one of a variety of forms of t-test used to compare a case to controls. As will be shown, these alternatives are in fact even more problematic than the use of z. The aim of this paper is to evaluate these competing approaches to the analysis of the single case through a mixture of thought experiments/worked examples, and Monte Carlo simulation. In the next section the different methods are set out although, as will be shown, the differences between some of them are more apparent than real.

1.1. Deficit inferred if a significant result is obtained using Crawford and Howell's (1998) method

This method (Crawford and Howell, 1998; see also Crawford and Garthwaite, 2002) is widely used (for recent examples see Berteletti et al., 2010; Tsapkini and Rapp, 2010; Starfelt et al., 2010; Lallier et al., 2010; Herbert and Best, 2010; Ham et al., 2010; d'Honincthun and Pillon, 2008; Dubois et al., 2010; Dalla Barba and Decaix, 2009; Peters et al., 2009). As noted, it differs from the use of z in that the control sample statistics are treated as statistics rather than as population parameters. The formula for the test is

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

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

If the t-value obtained from this test falls below the (negative) one-tailed 5% critical value for t on $n-1$ degrees of freedom (df), then it can be concluded that the case's score is sufficiently low to reject the null hypothesis that it is an observation from the population of scores for controls and the case is considered to exhibit a deficit on the task in question. The one-tailed p-value obtained (unlike the p-value obtained for a z score) is also an unbiased point estimate of the abnormality of a case's score. Thus if the p-value is .023, then it is estimated that only 2.3% of the control population will obtain lower scores (i.e., in this example the case's score is abnormally low). For a formal proof of this dual role for the p-value see Crawford and Garthwaite (2006).

1.2. Deficit inferred if a case's score is outside of prediction interval (PI) on controls versus an additional case

To our knowledge this method was first employed by Barton and colleagues (e.g., Barton et al., 2002) but it has since been employed in a number of other single-case studies (Barton et al., 2003, 2005; Behrmann et al., 2006; Rosenthal and Behrmann, 2006; Ravizza et al., 2005; Barton, 2008). It involves calculating the standard error of the difference between a sample mean and an additional case (this standard error should definitely not be confused with the standard error of the control mean as used in methods covered later).

The standard error is multiplied by t on $n-1$ df to provide a 100p% (two-sided) PI centred on the control mean. To illustrate, suppose that the mean and SD for a sample of 12 controls was 50 and 10 respectively, and that a 95% two-sided PI is required. Then

$$95\% \text{ PI} = \bar{x} \pm t_{n-1, 0.975} \left(s \sqrt{\frac{n+1}{n}} \right) = 50 \pm 22.9, \quad (2)$$

where the quantity in brackets is the standard error of the difference referred to above. Thus the PI is from 27.1 to 72.9. Suppose that a single-case obtains a score of 25, then, as the patient's score lies outside (i.e., below) the PI, it is concluded that the case has a deficit.

In essence this method is equivalent to Crawford and Howell's (1998) method as it will yield exactly the same outcome. That is, if comparison of a case to controls using Crawford and Howell's test yields a significant result ($p < .05$) on a two-tailed test, then the patient's score will be outside the 95% two-sided PI. However, there are a number of differences between the methods; these will be dealt within the Section 4.

1.3. Deficit inferred if a case's score is outside a confidence interval (CI) on the control mean

This method calculates a CI on the control mean and examines whether the case's score lies outside the interval. Although it generates CIs, it relies on t-distributions to obtain these and in practice it is used as a significance test: a deficit is inferred if the patient lies outside of the interval. We therefore classify it as a t-test. The method has been widely used (Butler et al., 2006; Kotz et al., 2005; Smith and Gilchrist, 2005; Rochon et al., 2004; Newport and Jackson, 2006; Woods et al., 2006; Terao et al., 2006; Castelo-Branco et al., 2006; Ferber and Danckert, 2006; Delazer et al., 2006; Steinworth et al., 2005; Larsen et al., 2004; Roy et al., 2004).

The steps involved are to first obtain the standard error of the control mean (s/\sqrt{n}) and then multiply it by $t_{1-\alpha/2}$ on $n-1$ df. For example, if the mean and SD for 15 controls on a task was 50 and 10 respectively then the standard error of the mean is 2.582. If a two-sided 95% CI is required then this standard error is multiplied by the 97.5th percentile point of a t-distribution on 14 df (the t-value is 2.145) and thus the 95% CI is 44.46–55.54 (i.e., 50 ± 5.54). The presence of a deficit is inferred if a case's score lies below the lower confidence limit (or above the upper limit if the score is an error score). Thus, in this example, if a patient obtained a score of 41, it would be concluded they exhibited a deficit on the task in question as the score lies outside the CI.

We could find no explicit rationale or justification for the use of this method but it is clear that the aim is to provide evidence that the patient has "a deficit" or is "impaired" on the task in question. Unfortunately it is far from satisfactory for such a purpose.

This is best illustrated with an extreme example: suppose that a very large sample of controls (1000) had been recruited and that, as in the previous example, the mean and SD in this sample on a task was 50 and 10 respectively. Further suppose that a case obtained a score of 49. The 95% (two-sided) CI on the mean for these data is from 49.38 to 50.62. Thus the case's

score lies outside the CI, but of course so will close to 100% of the healthy control population (or close to 50% if we limit consideration to those below the lower limit). The fatal flaw in the method seems obvious using this example, but is not nearly so obvious with the modest control sample sizes typically found in single-case studies: with a modest sample the uncertainty attached to the mean will be considerable and thus the CI will be wide.

It is likely that another factor contributing to the use of this method is a carryover from group studies. In group studies the interest is often in testing for a difference between population means: the danger is that the single-case researcher inadvertently slips into thinking in terms of whether the single-case is below the control mean, rather than being concerned with whether the case's score is sufficiently low to make it unlikely that it is an observation from the control population.

In some studies using this method a 99% CI was used, rather than a 95% interval, as featured in the examples. This will serve to reduce the very high Type I error rates associated with the method but the effect will be relatively modest. To illustrate, in the first worked example, if we substitute a 99% interval for the 95% interval, then the lower limit is 42.31 and the patient is still classified as exhibiting a deficit. In conclusion, we suggest that this method is entirely inappropriate as a mean of testing for a deficit and should be abandoned.

1.4. Deficit inferred if a control mean was significantly different from case's score using a one-sample t-test

The standard use of one-sample t-test is to test if a sample mean differs significantly from a known population mean, or a mean predicted by theory (Howell, 2002), when the population SD is unknown. Such usage is entirely legitimate. However, this test has been used in a number of studies (e.g., Reinhold and Markowitsch, 2007; Vecera and Rizzo, 2004; Brand et al., 2004) to draw inferences concerning the difference between a single case and a control sample: the single-case's score was designated as the hypothesised population mean and the control sample mean compared to it. That is, the formula used was

$$t_{n-1} = \frac{\bar{x} - x^*}{s/\sqrt{n}}, \quad (3)$$

where all terms have previously been defined. Reinhold and Markowitsch (2007) describe the procedure in this way, "Separately for each patient, all comparisons were computed by t-tests for one group with the score of each patient as the tested value" (p. 60). Similarly, Vecera and Rizzo (2004), in describing their study of Case EVR state that, "We compared the control participants' cuing effect against EVR's cuing effect (i.e., we used EVR's cuing effect as the hypothesized mean)" (p. 1662).

This is a highly unorthodox use of a one-sample t-test. As previously noted, the more common problem in single-case studies is that the control sample statistics are treated as population parameters (i.e., z is used as a significance test). In the present procedure the control sample statistics are treated as statistics, but the score of the case is treated as a parameter. This is even more problematic. We suggest that the

appropriate test in this situation is to treat all data as random variables and thus apply Crawford and Howell's (1998) test.

The effect of using a one-sample t-test is a high Type I error rate. This is easily shown with a thought experiment: Suppose that a control sample, consisting of 10 persons, obtains a mean of 50 and SD of 10 on a task, and that a patient obtains a score of 40; note that the patient is exactly one SD below the control mean. Applying Crawford and Howell's (1998) test yields a t-value on 9 df of .954 and a one-tailed probability of .365. That is, the patient's score is not low enough for us to reject the null hypothesis that the score is an observation from the scores in the control population. In contrast, if we apply a one-sample t-test to this problem, we obtain a t on 9 df of 3.162 and the one-tailed probability is .012. We would thus incorrectly conclude that the case differs significantly from controls when this is far from being the case.

This method is directly equivalent to the method set out in Section 1.3, in which a deficit is inferred if a case's score is outside the 95% CI on the control mean. That is, if the case's score is outside the 95% CI on the control mean, then the two-tailed p-value for the present method will be less than .05. To illustrate, in the first example for the CI method (in which the control mean and SD for 15 controls was 50 and 10 respectively), a case obtaining a score of 41 was outside the 95% CI; using these data the present t-test therefore necessarily yields a two-tailed p-value that is less than .05 (the t-value is 3.486 and the obtained p-value is .0364).

1.5. Use of a t-test employing the standard error of the control mean in the denominator

This method involves subtracting the control mean from the single-case's score (just as would be done in Crawford and Howell's test), then dividing this quantity by the standard error of the control mean (rather than the standard error of the difference between a case and controls, as in Crawford and Howell's method). The result is treated as t and evaluated on $n - 1$ df. The formula is therefore

$$t_{n-1} = \frac{x^* - \bar{x}}{s/\sqrt{n}}, \quad (4)$$

where all terms have previously been defined. This method has been used in a number of single-case studies (Gardiner et al., 2006, 2008; Brandt et al., 2009; Brandt et al., 2006).

The rationale underlying this test is not clear but it does not yield satisfactory results. In fact it can easily be seen that it is directly equivalent to two of the methods previously covered. That is, it will give identical results to those obtained using a CI on the control mean (Section 1.3) and those obtained using a one-sample t-test (Section 1.4). By comparing equations (3) and (4) it can be seen that the t values from the tests will be identical in magnitude and will differ only in sign.

Given that the method is equivalent to these two foregoing methods, the arguments and examples used earlier to highlight the problems with these other methods also apply here. The method, therefore, will be associated with grossly inflated p values and will therefore also exaggerate the abnormality of a case's score.

1.6. Empirical examination of the methods through Monte Carlo simulation

In the foregoing discussion of methods of testing for a deficit we have tried to avoid a mathematical treatment of the issues, instead relying on concrete examples to highlight issues. Single examples however have limitations. In particular, at some points we have used *extreme* examples as they clearly expose the underlying statistical problems with a method. The potential downside of such an approach is that single-case researchers may remain unconvinced; i.e., it may be thought that with the type of data they typically work with the problems exposed are of little concern. Monte Carlo simulation provides a means of supplementing these examples as it can provide a systematic evaluation of the competing methods through examining their ability to control the Type I error rate. It also provides a convenient means of quantifying the severity of these problems as a function of control sample size. In conducting the simulation we include the use of z as a method for inferring a deficit for comparison purposes.

2. Method

2.1. Monte Carlo simulation

The approach employed is based on previous simulations conducted by Crawford and Garthwaite (2005b). Further technical information can be found in that paper: in the present paper we provide only basic details.

To perform the simulation, n controls plus an additional case were sampled from a standard normal distribution. Five values of n were used: 5, 10, 20, 50 and 100. For each of these values of n , 2 million trials were performed. On each trial the score of the case was compared to that of the control sample using each of the five methods referred to in the Introduction (plus the z -value approach), and a tally kept of whether a significant result was obtained. For example, z was computed for the case, based on the control sample mean and SD for that trial, and a significant result recorded if this z was ≤ -1.645 . Similarly, Crawford and Howell's method was applied and a significant result recorded if t was negative (i.e., the case's score was below controls) and its magnitude exceeded the critical value for a one-tailed test on $n - 1$ df with alpha set at .05. For each method, at each sample size the percentage of significant results was obtained.

Note that in this simulation, because the case is an observation from the same population as controls, if a method records a significant result then a Type I error has been committed – it is wrongly concluded that the case exhibits a deficit (i.e., it is erroneously concluded that the score is not an observation from the population of control scores).

3. Results

The results of the simulation are presented in Table 1. This table reports the Type I error rates as percentages for each of the six methods (the five methods using a t -distribution, plus z) at each of the five different sample sizes examined. Most

Table 1 – Results of a Monte Carlo simulation examining control of the Type I error rate for six methods of testing for a deficit by comparing a single-case to a control sample; the nominal error rate is 5% (one-tailed).

Control n	Inferential method					
	z	CH_t	PI_{n+1}	$CI_{\bar{x}}$	SE_t	OS_t
5	10.39	5.00	5.00	21.67	21.67	21.67
10	7.58	5.01	5.01	29.71	29.71	29.71
20	6.25	5.00	5.00	35.49	35.49	35.49
50	5.49	4.99	4.99	40.81	40.81	40.81
100	5.24	5.00	5.00	43.46	43.46	43.46

Note: z = "standard" use of z (one-tailed) i.e., deficit inferred if $z < -1.645$; CH_t = Crawford & Howell's (1998) method; PI_{n+1} = Barton et al's (2003) method; $CI_{\bar{x}}$ = case below lower limit of 95% CI on control mean; SE_t = t -test based on dividing difference between case and controls by SE of control mean; OS_t = one-sample t -test in which case's score is entered as the population mean.

aspects of the data are immediately apparent. As expected from theory and from previous simulation results (Crawford and Garthwaite, 2005b), the Type I error rate obtained closely matches the specified error rate of (5%) for Crawford and Howell's method (the small deviations are within the bounds expected from Monte Carlo variation).

In addition, it was noted earlier that Barton et al's (2003) PI method is equivalent to Crawford and Howell's method. Thus, as expected, identical results were obtained for the two methods and thus Barton's method also maintains the Type I error rate at the level specified. Note however that, in all of the examples we encountered the use of Barton's PI method, it was used to provide a two-sided test. A one-sided version was used in the simulation in order to confirm that the two methods are equivalent. Were a two-sided test applied then the error rate would be below the specified 5% rate (i.e., it would be $\approx 2.5\%$). Thus the test would be more conservative and consequently would have lower power to detect a deficit.

It was suggested that a further three of the inferential methods are directly equivalent to each other: use of a $CI_{\bar{x}}$ on the control mean (denoted as $CI_{\bar{x}}$ in Table 1), use of a one-sample t -test in which the case's score is entered as the population mean (denoted as OS_t), and use of a t -test in which the standard error of the mean is used rather than the standard error of the difference between case and controls (denoted as SE_t). The simulation results confirm that these three methods are directly equivalent: the error rates are identical for all three methods.

It can also be seen that these latter methods are associated with particularly high error rates. For example, it is estimated that 29.7% of controls would be incorrectly classified as exhibiting a deficit when compared to a control sample of size 10. It can also be seen that, in contrast to Crawford and Howell's and Barton et al's. method, the Type I error rate varies with sample size. The error rate becomes more inflated as the control sample size is increased; this is to be expected given that the standard error of the mean shrinks with increasing sample size.

Table 1 also shows that, as expected, the Type I error rate is not under control when z is used as an inferential method. However, it can also be seen that the inflation of the error rate

is not nearly as marked as it is for the three foregoing methods. For example, for a sample size of 10, the estimated error rate for z is 7.58%, whereas it is 29.71% for the foregoing methods. Thus, although Crawford and Howell (1998) rightly cautioned against the use of z for inferential purposes and recommended the use of a t -test in its stead, it is very clear that it has to be the right kind of t -test, otherwise things can be made much worse rather than better. Finally, it can also be seen that, in contrast to the pattern observed for the foregoing methods, the inflation of the error rate observed for z when sample size is small becomes attenuated as sample size increases (because z approaches t from Crawford and Howell's test for large n).

4. Discussion

4.1. Convergence conveys confidence in results: an exception to the rule

It may be surprising to some readers that identical yet erroneous results should be obtained from the application of three ostensibly different approaches ($CI_{\bar{x}}$, SE_t , and OS_t) to the same problem. Usually, when a series of different approaches converge on the same solution, one would have confidence that the approaches are sound. For example, classical (i.e., frequentist) statistics and Bayesian statistics often yield identical results for a particular problem, despite their radically different assumptions. Indeed a Bayesian approach to comparing a case's score to that of controls yields results that are identical to those obtained by Crawford and Howell's (1998) classical test (Crawford and Garthwaite, 2007). Such convergence is reassuring, regardless of whether a neuropsychologist is classical, Bayesian, or eclectic in orientation.

Thus the current situation is highly unusual. It suggests that many single-case researchers may be operating on a common underlying set of erroneous assumptions. As was noted in Section 1, when discussing the use of CIs on the control mean, we suspect this stems from a failure to make the cognitive shift between testing for a difference between population means (as is done when testing for a difference between groups) to testing whether an individual's score is sufficiently low to allow rejection of the null hypothesis that it is an observation from the control population.

4.2. Reporting the effect size for the difference between a case and controls

Crawford et al. (2010) have recently argued that single-case research reports should always report effect sizes for the difference between a case and controls. The relevant effect size index is simply z , computed in the standard way, and is a direct analogue of Cohen's d , as used in group studies. Crawford et al. (2010) denote this z as z_{CC} , the subscript serving to identify it as an effect size index for the case-controls design and to help avoid any potential confusion with critical values of z etc. Note that sample size does not feature in calculating a z score; this characteristic (as we have seen in the results of the current simulation) makes z unsuitable as a significance test (because it cannot factor in the uncertainties over the control mean and

SD). However, this very feature makes it eminently suitable as an index of effect size.

There are a variety of arguments in favour of the reporting of effect sizes in single-case studies (Crawford et al., in press, 2010), but the present findings provide a particularly compelling justification. That is, although there may be factors that go some way to explaining the adoption of the faulty methods reviewed above, it remains the case that their use suggests that some single-case researchers may exhibit a worrying lack of engagement with their data. In many instances simply computing an effect size (z_{CC}) for a case would make it very obvious that the case's score is not extreme, and that an appropriate inferential method cannot possibly support such a claim. For example, referring back to the very first worked example for the $CI_{\bar{x}}$ method, the case's score of 41 is less than one SD below the control mean of 50 (i.e., $z_{CC} = -1.9$); for the second (extreme) example the case's score of 49 is only very marginally below the control mean of 50 (i.e., $z_{CC} = -1$) and yet both differences would be recorded as statistically significant under some of the methods discussed here.

4.3. Comparison of Crawford and Howell's method with that of Barton's PI

The Monte Carlo simulation confirmed that, when used as a significance test, Crawford and Howell's (1998) method and the PI used by Barton et al. yield strictly equivalent results. However, some important differences remain. First, although Crawford and Howell's method can be used as a one- or two-tailed test, Crawford and colleagues have recommended use of a one-tailed test when testing for a deficit. In contrast, Barton and colleagues, and all of the other papers employing this CI method, used two-sided intervals. As noted, this will result in a more conservative test and consequently will have lower power to detect a deficit. (Note that, just as Crawford and Howell's method can be used as a one- or two-tailed test, there is nothing inherent in the Barton method to prevent it being used as a one-sided test; thus the difference here is not with the methods themselves but with how they have been applied in practice).

Second, Barton's method gives only a dichotomous decision (deficit vs no deficit, depending on whether the patient's score lies within or outside the interval), whereas Crawford and Howell's method yields a precise probability. Note also that, as described earlier, the probability from the significance test is simultaneously a point estimate of the proportion of the control population that will exhibit a lower score (i.e., it provides a point estimate of the abnormality of the patient's score).

Third, the point estimate of the abnormality of a case's score can be supplemented with 95% CIs (Crawford and Garthwaite, 2002). These latter intervals are entirely different from Barton's PI (which essentially serves as a significance test) and they follow a non-central (rather than central) t -distribution. The development of these intervals is in keeping with contemporary opinion in psychology and statistics that point estimates of any quantity should be supplemented with interval estimates in order to quantify the uncertainty associated with them. To illustrate, these intervals allow neuropsychologists to make statements like the following, "it is estimated that only 1.36% of controls would

exhibit a lower score than the case and the 95% CI on this percentage runs from 1.08% to 5.34%". Lastly, as discussed in the previous section, Crawford and Howell's method now incorporates point (and interval) estimates of the effect size (z_{cc}) for the difference between case and controls (Crawford et al., 2010). Although the provision of these limits was motivated by other considerations, it transpires that these effect sizes can serve as a useful reality check for both researchers and consumers of research.

In concluding this section, it is important that Barton's method (e.g., Barton et al., 2002) is not confused with the use of a CI on a control mean. Unlike the latter method, Barton's approach is statistically sound; that is, it will maintain the Type I error rate at the rate specified by the user, subject to the proviso that the assumption of normality is not violated. It is directly equivalent to Crawford and Howell's (1998) method except that it only provides dichotomous information and, in practice, has consistently been used to provide a two-tailed rather than a one-tailed test.

4.4. The assumption of normality

All of the methods considered here assume that control scores are normally distributed. However, it is not uncommon for the scores of controls on neuropsychological tests to depart from normality (Capitani and Laiacina, 2000; Crawford and Garthwaite, 2005b). For example, negative skewness is common in control data because the tasks employed often measure abilities that may be largely within the competence of most healthy individuals and thus yield ceiling, or near-ceiling, levels of performance in controls. As an extreme example, in a review of single-case studies of the living versus non-living distinction in object naming, it was reported that the accuracy of naming in controls was 95% or greater in the majority of these studies (Laws et al., 2005).

It would have been possible to study the effects of violating the normality assumption through Monte Carlo simulation by sampling control scores from skew and/or leptokurtic distributions. However, we did not pursue this because (1) control over the Type I error rate is very poor for most of the methods examined here even when the normality assumption holds so there is little to be gained by further study of them, and (2) for the methods that do control the Type I error rate (Crawford and Howell's method and Barton's PI equivalent), it has previously been established that the effects of violating the normality assumption are surprisingly modest (Crawford et al., 2006; Crawford and Garthwaite, 2005b).

4.5. Conclusion

Most single-case studies in neuropsychology reveal that the researchers concerned have a sophisticated grasp of the relevant cognitive theory. Moreover, the designs of these studies are often found on a careful, logical analysis of the questions to be addressed, and great care is taken to develop tasks and materials (Crawford et al., 2003). In contrast, it is often the case that much less care and attention is given to the statistical methods employed. At the risk of being glib, some studies could be considered to exhibit a classical dissociation between these different components.

The neglect of statistical methods is also reflected in the fact that the Section 2 of many single-case studies provide detailed clinical histories of the cases together with elaborate descriptions of the materials and procedures, but give only a cursory mention of the statistical methods employed (often with no supporting references or justification), or indeed make no mention of these at all. This was evident in some of the studies referred to in the present paper: descriptions were often vague and back engineering from the results reported (i.e., t values or CI widths, together with control summary statistics) was often necessary to determine what method had been employed.

It is clear from the thought experiments and the results of simulation that the three equivalent methods of testing for a deficit in the case-controls design should be abandoned. There would also be a strong argument for reanalysis of studies that have employed these methods, particularly as, in many instances, these were the only inferential methods upon which conclusions were drawn. Note also that, despite evidence and recommendations to the contrary (Crawford and Garthwaite, 2005a), it is still common for dissociations to be inferred when a case is significantly different from controls on one task, but not significantly different from controls on another. Thus, it is not simply the case that these studies may be replete with Type I errors (although they often will) but also that potentially important dissociations may have been passed over. That is, if impaired performance was observed on a targeted task, the task or tasks that would have provided the other half of the dissociation (because the score was comfortably within normal limits of performance) may have been recorded as impaired. In other words a Type I error in testing for a deficit may have led to a Type II error for a dissociation.

Finally, in summary: five ostensibly different inferential methods used to compare a single case to controls can in fact be reduced to two methods; of these two methods, one is clearly not fit for purpose. The remaining method comes in two different flavours (Crawford and Howell, 1998; Barton et al., 2002) both of which are fundamentally sound. The Crawford and Howell method, however, offers a number of advantages, including a point and interval estimate of the abnormality of a case's score (Crawford and Garthwaite, 2002), and has recently been extended to also provide point and interval estimates of effect sizes (Crawford et al., 2010).

REFERENCES

- Barton JJS. Structure and function in acquired prosopagnosia: Lessons from a series of 10 patients with brain damage. *Journal of Neuropsychology*, 2(Pt 1): 197–225, 2008.
- Barton JJS, Cherkasova MV, Press DZ, Intriligator JM, and O'Connor M. Developmental prosopagnosia: A study of three patients. *Brain and Cognition*, 51(1): 12–30, 2003.
- Barton JJS, Press DZ, Keenan JP, and O'Connor M. Lesions of the fusiform, face area impair perception of facial configuration in prosopagnosia. *Neurology*, 58(1): 71–78, 2002.
- Behrmann M, Avidan G, Leonard GL, Kimchi R, Luna B, Humphreys K, et al. Configural processing in autism and its relationship to face processing. *Neuropsychologia*, 44(1): 110–129, 2006.
- Behrmann M, Avidan G, Marotta JJ, and Kimchi R. Configural processing in autism and its relationship to face processing. *Journal of Cognitive Neuroscience*, 17(7): 1130–1149, 2005.

- Berteletti I, Hubbard EM, and Zorzi M. Implicit versus explicit interference effects in a number-color synesthete. *Cortex*, 46(2): 170–177, 2010.
- Brand M, Kalbe E, Kracht LW, Riebel U, Munch J, Kessler J, et al. Organic and psychogenic factors leading to executive dysfunctions in a patient suffering from surgery of a colloid cyst of the Foramen of Monro. *Neurocase*, 10(6): 420–425, 2004.
- Brandt KR, Gardiner JM, Vargha-Khadem F, Baddeley AD, and Mishkin M. Using semantic memory to boost 'episodic' recall in a case of developmental amnesia. *NeuroReport*, 17(10): 1057–1060, 2006.
- Brandt KR, Gardiner JM, Vargha-Khadem F, Baddeley AD, and Mishkin M. Impairment of recollection but not familiarity in a case of developmental amnesia. *Neurocase*, 15(1): 60–65, 2009.
- Butler SH, Gilchrist ID, Ludwig CJH, Muir K, and Harvey M. Impairments of oculomotor control in a patient with a right temporo-parietal lesion. *Cognitive Neuropsychology*, 23(6): 990–999, 2006.
- Capitani E and Laiacona M. In Boller F and Grafman J (Eds), *Classification and Modelling in Neuropsychology: From Groups to Single Cases*. Amsterdam: Elsevier, 2000: 53–76.
- Castelo-Branco M, Mendes M, Silva MF, Janeiro C, Machado E, Pinto A, et al. Specific retinotopically based magnocellular impairment in a patient with medial visual dorsal stream damage. *Neuropsychologia*, 44(2): 238–253, 2006.
- 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(8): 1196–1208, 2002.
- Crawford JR and Garthwaite PH. Evaluation of criteria for classical dissociations in single-case studies by Monte Carlo simulation. *Neuropsychology*, 19: 664–678, 2005a.
- Crawford JR and Garthwaite PH. 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(3): 318–331, 2005b.
- Crawford JR and Garthwaite PH. Methods of testing for a deficit in single case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology*, 23(6): 877–904, 2006.
- Crawford JR and Garthwaite PH. Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach. *Cognitive Neuropsychology*, 24(4): 343–372, 2007.
- Crawford JR, Garthwaite PH, Azzalini A, Howell DC, and Laws KR. Testing for a deficit in single case studies: Effects of departures from normality. *Neuropsychologia*, 44(4): 666–676, 2006.
- Crawford JR, Garthwaite PH, and Gray CD. Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex*, 39(2): 357–370, 2003.
- Crawford JR, Garthwaite PH, and Porter S. Point and interval estimates of effect sizes in the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, 27(3): 245–260, 2010.
- Crawford JR, Garthwaite PH, and Wood LT. The case controls design in neuropsychology: Inferential methods for comparing two single cases. *Cognitive Neuropsychology*, in press.
- Q4 Crawford JR and Howell DC. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, 12(4): 482–486, 1998.
- d'Honin P and Pillon A. Verb comprehension and naming in frontotemporal degeneration: The role of the static depiction of actions. *Cortex*, 44(7): 834–847, 2008.
- Dalla Barba G and Decaix C. "Do you remember what you did on March 13, 1985?" A case study of confabulatory hypermnnesia. *Cortex*, 45(5): 566–574, 2009.
- Delazer M, Karner E, Zamarian L, Donnemiller E, and Benke T. Number processing in posterior cortical atrophy – A neuropsychological case study. *Neuropsychologia*, 44(1): 36–51, 2006.
- Dubois M, Kyllingsboek S, Prado C, Musca SC, Peiffer E, Lassus-Sangosse D, et al. Fractionating the multi-character processing deficit in developmental dyslexia: Evidence from two case studies. *Cortex*, 46(6): 717–738, 2010.
- Ferber S and Danckert J. Lost in space – The fate of memory representations for non-neglected stimuli. *Neuropsychologia*, 44(2): 320–325, 2006.
- Gardiner JM, Brandt KR, Baddeley AD, Vargha-Khadem F, and Mishkin M. Charting the acquisition of semantic knowledge in a case of developmental amnesia. *Neuropsychologia*, 46(11): 2865–2868, 2008.
- Gardiner JM, Brandt KR, Vargha-Khadem F, Baddeley A, and Mishkin M. Effects of level of processing but not of task enactment on recognition memory in a case of developmental amnesia. *Cognitive Neuropsychology*, 23(6): 930–948, 2006.
- Ham HS, Bartolo A, Corley M, Swanson S, and Rajendran G. Case report: Selective deficit in the production of intransitive gestures in an individual with autism. *Cortex*, 46(3): 407–409, 2010.
- Herbert R and Best W. The role of noun syntax in spoken word production: Evidence from aphasia. *Cortex*, 46(3): 329–342, 2010.
- Howell DC. *Statistical Methods for Psychology*. Belmont, CA: Duxbury Press, 2002.
- Kotz SA, Von Cramon DY, and Friederici AD. On the role of phonological short-term memory in sentence processing: EPP single case evidence on modality-specific effects. *Cognitive Neuropsychology*, 22(8): 931–958, 2005.
- Lallier M, Donnadieu S, Berger C, and Valdois S. A case study of developmental phonological dyslexia: Is the attentional deficit in the perception of rapid stimuli sequences amodal? *Cortex*, 46(2): 231–241, 2010.
- Larsen J, Baynes K, and Swick D. Right hemisphere reading mechanisms in a global alexic patient. *Neuropsychologia*, 42(11): 1459–1476, 2004.
- Laws KR, Gale TM, Leeson VC, and Crawford JR. When is category specific in Alzheimer's disease? *Cortex*, 41: 452–463, 2005.
- Newport R and Jackson SR. Posterior parietal cortex and the dissociable components of prism adaptation. *Neuropsychologia*, 44(13): 2757–2765, 2006.
- Peters J, Thoma P, Koch B, Schwarz M, and Daum I. Impairment of verbal recollection following ischemic damage to the right anterior hippocampus. *Cortex*, 45(5): 592–601, 2009.
- Ravizza SM, Behrmann M, and Fiez JA. Right parietal contributions to verbal working memory: Spatial or executive? *Neuropsychologia*, 43(14): 2057–2067, 2005.
- Reinhold N and Markowitsch HJ. Emotion and consciousness in adolescent psychogenic amnesia. *Journal of Neuropsychology*, 1(1): 53–64, 2007.
- Rochon E, Kave G, Cupit J, Jokel R, and Winocur G. Sentence comprehension in semantic dementia: A longitudinal case study. *Cognitive Neuropsychology*, 21(2–4): 317–330, 2004.
- Rosenthal O and Behrmann M. Acquiring long-term representations of visual classes following extensive extrastriate damage. *Neuropsychologia*, 44(5): 799–815, 2006.
- Roy AC, Stefanini S, Pavesi G, and Gentilucci M. Early movement impairments in a patient recovering from optic ataxia. *Neuropsychologia*, 42(7): 847–854, 2004.
- Smith AD and Gilchrist ID. Within-object and between-object coding deficits in drawing production. *Cognitive Neuropsychology*, 22(5): 523–537, 2005.
- Starrfelt R, Habekost T, and Gerlach C. Visual processing in pure alexia: A case study. *Cortex*, 46(2): 242–255, 2010.
- Steinworth S, Levine B, and Corkin S. Medial temporal lobe structures are needed to re-experience remote autobiographical memories: Evidence from HM and WR. *Neuropsychologia*, 43(4): 479–496, 2005.

- 911 Terao Y, Mizuno T, Shindoh M, Sakurai Y, Ugawa Y, Kobayashi S, 918
912 et al. Vocal amusia in a professional tango singer due to a right 919
913 superior temporal cortex infarction. *Neuropsychologia*, 44(3): 920
914 479–488, 2006. 921
- 915 Tsapkini K and Rapp B. The orthography-specific functions of the 922
916 left fusiform gyrus: Evidence of modality and category 923
917 specificity. *Cortex*, 46(2): 185–205, 2010. 924
- Vecera SP and Rizzo M. What are you looking at? Impaired 'social
attention' following frontal-lobe damage. *Neuropsychologia*,
42(12): 1657–1665, 2004.
- Woods AJ, Mermemeier M, Garcia-Rill E, Meythaler J, Mark VW,
Jewel GR, et al. Bias in magnitude estimation following
left hemisphere injury. *Neuropsychologia*, 44(8): 1406–1412,
2006.

UNCORRECTED PROOF