Running Head: Comparing Single Cases

Cognitive Neuropsychology, in press

Inferential methods for comparing two single cases

John R. Crawford[1], Paul H. Garthwaite[2], and Liam T. Wood[1]

[1]School of Psychology

University of Aberdeen

Aberdeen, UK

[2]Department of Mathematics and Statistics

The Open University

Milton Keynes, UK

_____

Address for correspondence: Professor John R. Crawford, School of Psychology,

College of Life Sciences and Medicine, King's College, University of Aberdeen,

Aberdeen AB24 2UB, United Kingdom.

E-mail: j.crawford@abdn.ac.uk

Inferential methods for comparing two single cases

ABSTRACT

In neuropsychological single-case studies it is not uncommon for researchers to compare the scores of two single cases. Classical (and Bayesian) statistical methods are developed for such problems which, unlike existing methods, refer the scores of the two single cases to a control sample. These methods allow researchers to compare two cases' scores, with or without allowing for the effects of covariates. The methods provide a hypothesis test (one- or two-tailed), point and interval estimates of the effect size of the difference, and point and interval estimates of the percentage of pairs of controls that will exhibit larger differences than the cases. Monte Carlo simulations demonstrate that the statistical theory underlying the methods is sound and that the methods are robust in the face of departures from normality. The methods have been implemented in computer programs and these are described and made available (to download, go to http://www.abdn.ac.uk/~psy086/dept/Compare_Two_Cases.htm).

Key terms: single-case methods; Bayesian statistics; dissociations; neuropsychological methods; credible limits; multiple indicators

INTRODUCTION

In neuropsychological single-case studies it is not uncommon for researchers to compare the scores of two single cases, either as the primary focus of such studies or as a supplementary feature (Akiyama et al., 2006; Blazely, Coltheart, & Casey, 2005; Forde & Humphreys, 2005; Humphreys & Forde, 2005; Linebarger, 2004; Martin, Miller, & Vu, 2004; Mochizuki-Kawai et al., 2004; Riddoch & Humphreys, 2004; Shallice, Venable, & Rumiati, 2005; Smith & Gilchrist, 2005; Toraldo & Shallice, 2004).

One potentially compelling reason for conducting such comparisons stems from arguments made by Shallice (1988) over how to test for a double dissociation. He is critical of the view that a double dissociation occurs "when patient A performs task I significantly better than task II, but for patient B, the situation is reversed..." because " this pattern of performance is not sufficient grounds for inferring the existence of separate subsystems" (p. 234). Shallice (1988) suggests that a more rigorous approach is to directly compare the performance of the two cases. That is, he argues that "the valid formulation of the double dissociation … is that on task I, patient A performs significantly better than patient B, but on task II, the situation is reversed". (p. 235).

In Shallice's (1988) definition, the use of the word *significant* is a little ambiguous, as it might be taken to mean (and indeed probably was intended to mean) "statistically significant". However, if a task involves a sufficiently extensive test, even tiny differences between two cases can be detected and give a *statistically* significant result. Two tiny differences in opposite directions should not be interpreted as a double dissociation. That is, a significant difference should be interpreted in its colloquial sense, and taken to mean *a large difference.* The criteria of

large that we shall adopt here is that a large difference must be greater than the typical difference that arises between two controls from random variation.

In the next section we consider existing inferential methods of testing for a difference between two cases. In contrast to the methods developed later in this paper, the existing methods typically conduct such comparisons without reference to a control sample.

Comparing two single cases in the absence of a control sample

Probably the most common existing method of comparing two cases is to use a 2 x 2 chi-square test of independence (or related method, such as a Fisher's Exact Test) in which the number of passes and fails on a given task for Case A is compared to the number of passes and fails for Case B. One obvious issue here is that these tests are based on the assumption that the observations (the scores on each of the k items of the task) are independent. At one level this assumption is always violated: each of the two sets of k observations are obtained from the same cases. Moreover, it is often not very realistic and sometimes completely unrealistic to assume that the cases' responses to a particular item are independent of their responses to earlier items. (These same issues arise when chi square tests are used to compare the performance of one single case on two tasks).

Even if the independence problem is ignored there are further serious difficulties with such an approach; these are best illustrated with a concrete example. Suppose that a task (Task X) has 135 pass-fail items and that a case (Case A) passes 68 of these, whilst another case (Case B) passes only 47 (for reasons that will become apparent let us also suppose that both cases are 65 years old). A chi-square test of independence to compare the performance of the two cases yields a chi-square value

of 5.494 on 1 df. This is statistically significant, $p = 0.019$. Thus a researcher may

conclude that there is a significant difference between the two cases and use this as

evidence either for or against some theoretical proposition.

However, the task in question is the Coding subtest from the Wechsler Adult

Intelligence Scale Fourth Edition (WAIS-IV; Wechsler, 2008) and, for 65 year olds,

raw scores of 68 and 47 convert to subtest scaled scores of 12 and 9 respectively[1].

The standard deviation of the difference between a pair of individuals on the same

task ($s_{x_1-x_2}$) is

$$s_{x_1-x_2} = \sqrt{2s_X^2} \, , \qquad\qquad (1)$$

where s_X^2 is the square of the task's standard deviation (i.e. its variance). As

WAIS-IV subtests have a standard deviation of 3, the standard deviation of the

difference between a pair of individual's scaled scores on the Coding subtest (or

indeed any Wechsler subtest) is therefore 4.243. If we divide the difference between

the two cases' scores ($12 - 9 = 3$) by the standard deviation of the difference we

obtain a z of 0.707. Referring to a table of areas under the standard normal curve we

can see that the area below -0.707 is approximately 0.24, as is the area above 0.707

(the areas therefore sum to 0.48). Thus, if we randomly selected pairs of cases from

the general adult population, a very substantial percentage of these pairs (48%) will

exhibit differences that exceed that observed for the two cases. While the chi-squared

test for independence addresses the question *"Do the two cases differ?"* it does not

consider the question *"Do the two cases differ by an unusually large amount?"*

---

[1] As a reviewer pointed out, when raw scores are converted to scaled scores, some information is lost as, commonly, more than one raw score can be converted to the same scaled score. However, the practice is ubiquitous in neuropsychology, not least because it allows the neuropsychologist to refer a case's scores to normative values.

When the items in a particular task are not dichotomous (i.e., pass-fail) an alternative approach used in the single-case literature is to is to analyse the data using a paired samples *t*-test, or ANOVA, in which scores on each of the items are treated as cases and the two cases are treated as (two) samples. As with the use of the chi-square test, such usage is problematic as the methods assume that, for each case, the *k* task items are independent of each other (*t*-tests and ANOVAs are also often used to compare the scores of one single case on two tasks and suffer from the same difficulty).

Just as we saw with the use of chi square, there are further problems in using these methods to compare two cases (if anything the problems are more striking than those encountered when using a chi square test). To illustrate, suppose that a case (Case A) performs a task consisting of 26 items (each scored 0, 1 or 2) and obtains a mean score on these items of 1.192 (SD= 0.694). Suppose also that a further case obtains a mean of 1.423 (SD = 0.703). Finally suppose that the correlation between the two sets of scores is 0.729 (not unreasonable if we assume that the task is of graded difficulty), and that both cases are 19 years old.

A paired sample *t*-test (two-tailed) to compare the scores of Case A and Case B records a statistically significant difference: $t = 2.287$, df = 25, $p = 0.031$. However, this example was constructed using the Spatial Span subtest from the WAIS-IV and the raw scores (i.e., item totals) for Cases A and B (31 and 37) convert to scaled scores of 9 and 11. Using the same procedure as that outlined earlier for the chi-square example it is estimated that 64% of pairs of individuals drawn from the general healthy adult population would exceed the difference observed for the two cases. Therefore, as in the previous example, we should make little if anything of the

difference between the performance of the two cases despite the statistically

significant result.

Testing for a difference between the scores of two single-cases by referencing the

difference to a control sample

The inferential methods outlined in the foregoing section do not refer the difference

between the cases to a control or normative sample. An alternative, more stringent,

approach to this problem is to test whether the difference found between two cases is

greater than would be found in a pair of controls. That is, we can specify the null

hypothesis that the difference between the cases is an observation from the

distribution of differences in the control population and, if we can reject this

hypothesis, we can conclude that there is a statistically significant difference between

the cases.

This approach to the problem can be seen as an extension of methods

developed by Crawford, Garthwaite, Howell and colleagues for comparing a single

case to a control sample (Crawford & Garthwaite, 2002; Crawford & Garthwaite,

2005, 2007; Crawford & Howell, 1998). These methods assume that the scores in the

control population have a normal distribution. If we make the same assumption for

the current problem it transpires that it is remarkably simple to obtain suitable

methods (the robustness of the methods in the face of violations of this assumption is

examined in Study 2).

The methods for comparing one single-case to controls provide a hypothesis

test (Crawford & Howell, 1998), point and interval estimates of the abnormality of a

case's score (Crawford & Garthwaite, 2002), and point and interval estimates for the

effect size of the difference between the case and controls (Crawford, Garthwaite, &

Porter, 2010). The aim here is to provide the equivalent statistics for the difference between two single cases. In Study 1 we develop classical and Bayesian methods for the basic problem of comparing the scores of two single cases. In Study 2 we evaluate these methods using Monte Carlo simulation, and in Study 3 we go on to develop methods for comparing the scores of two cases in the presence of covariates.

Before formally setting out the methods it is worth noting that existing methods of comparing one single case to controls do not provide a satisfactory solution to problems involving the comparison of two cases.   For example, suppose Crawford and Howell's test (1998) was used to compare the scores of each of the two cases to controls and suppose that one of the cases was significantly lower than controls, whereas a null result was obtained for the second case. This should not be taken as an indication that the cases differ: the *difference* between the scores of the two cases in this scenario could be very trivial and could arise very frequently among pairs of controls (as would occur when the *p* value for the first case was just below 0.05 and that for the second case just above).

Study 1

Method

<u>Classical hypothesis test of whether the difference found between two cases is greater than would be found in a pair of controls (and point estimate of abnormality).</u>

We assume that the value taken by a control, X say, is normally distributed, $N(\mu, \sigma^2)$. Let D be the difference in values between two controls. Then $D \sim N(0, 2\sigma^2)$. A random sample of n controls is taken and the variance of their observed values is calculated. If this sample variance is s^2, then $(n-1)s^2/\sigma^2$ has a chi-square distribution on $(n-1)$ degrees of freedom.

If we denote the difference between two single cases as G, then $G/\sqrt{2s^2}$ has a t-distribution on $n - 1$ df. (Note that this problem is greatly simplified by the fact that, perhaps surprisingly, the control mean does not need to feature when developing the test statistic, nor when developing interval estimates). Denote the observed value of G by g. We compare

$$\frac{g}{\sqrt{2s_X^2}} \qquad (2)$$

with a t-distribution on $n - 1$ degrees of freedom and reject the null hypothesis (that the difference between the values of the cases could have come from a pair of controls) if $g/\sqrt{2s^2}$ is large., e.g., if it exceeds the critical value for t on $n-1$ df for significance at the conventional 5% level. The hypothesis test can be either one-tailed, when the researcher has a clear *a priori* directional hypothesis, or two-tailed when this is not the case.

The rejection probability for the hypothesis test also serves as a point estimate of the proportion of differences between pairs of controls that would be more extreme than the difference observed between the two cases. If we multiply this probability by 100 we have a point estimate of the percentage of pairs of controls that would be more extreme than the difference observed between the cases; that is we have a point estimate of the abnormality of the difference between the two cases.

Classical confidence interval for the percentage of (pairs of) controls with more extreme differences than the difference found in a pair of single cases.

The preceding section provided a point estimate of the abnormality of the difference between two cases. A confidence interval for this quantity can easily be obtained:

We have that $(n-1)s^2/\sigma^2$ is a value from a chi-square distribution on $(n - 1)$ degrees

of freedom. Let *a* be the 0.025 point of a chi-square distribution on $(n-1)$ df and let *b* be the 0.975 point. Then $(n-1)s^2/b$ and $(n-1)s^2/a$ are endpoints of a 95% confidence interval for σ^2. We have that $D \sim N(0, 2\sigma^2)$.

If *g* is the observed difference between the two controls, put

$$z_1 = g\sqrt{a/\{2(n-1)s^2\}}, \qquad (3)$$

and

$$z_2 = g\sqrt{b/\{2(n-1)s^2\}}. \qquad (4)$$

Determine the tail areas of a standard normal distribution to the left/right (depending on whether *g* is less/greater than 0) of z_1 and z_2 i.e. $\Phi^{-1}(z_1)$ and $\Phi^{-1}(z_2)$. These are the 95% confidence limits on the proportion of pairs of controls who would have a greater difference than the two cases. As with the point estimate, these endpoints can be multiplied by 100 so they are expressed as percentages rather than proportions.

Classical point and interval estimate of the effect size for the difference between a pair of cases versus the differences in controls

There is increasing emphasis on the use of point and interval estimates of effect sizes in group-based psychological research. For example, in a report on statistical inference, the American Psychological Association strongly endorsed the reporting of effect sizes. The report recommends that researchers should "always provide some effect-size estimate when reporting a *p* value" and goes on to note that "reporting and interpreting effect sizes… is essential to good research" (Wilkinson and The APA Task Force on Statistical Inference, 1999, p. 599). In keeping with the general principle that the standards of reporting in single-case studies should be as high as those expected in group research, Crawford, Garthwaite and Porter (2010) have

proposed a set of effect size indexes for use in the case-controls design in

neuropsychology, and have developed methods for obtaining point and interval

estimates of these quantities. In this section we extend on that work to provide point

and interval estimates of effect sizes for the present problem.

For this problem the index is very straightforward. The statistic in equation (2)

is evaluated against a *t*-distribution on $n-1$ df (i.e., it is treated as a *t* value for

hypothesis testing purposes). However, it can also be treated as a *z* value: as can be

seen, the difference between the scores of the two cases (*g*) is standardized by

dividing by the standard deviation of the difference in controls. Thus equation (2) can

provide an index of effect size: it is an analogue of Cohen's *d*. We label the index as

z_{PCC} to clearly differentiate it from other effect size indexes proposed by Crawford et

al. (Crawford, Garthwaite, & Porter, 2010); the "P" in the subscript identifies that we

are comparing the difference between a *pair* of single-cases against the differences

between *pairs* of controls, the CC refers to Case-Controls. The formula for the index

is

$$z_{PCC} = \frac{g}{\sqrt{2s_X^2}} , \qquad (5)$$

where all terms have been previously defined. Note that, as should be the case for an

effect size index, the control sample size (*n*) does not feature in this equation (*n* does

of course feature when applying the hypothesis test as it determines the *t*-distribution

against which z_{PCC} is evaluated). Having obtained a formula for the point estimate of

the effect size for this problem we need a corresponding interval estimate. Again this

is very straightforward. An intermediate step in deriving a confidence interval for the

abnormality of the difference between two cases, see equations (2) and (3), involved

obtaining two quantities, z_1 and z_2: these are the endpoints of a 95% confidence interval on the effect size.

<u>A Bayesian approach to comparing two cases</u>

The methods set out in the foregoing sections are classical (frequentist) methods. As will be shown, a Bayesian analysis of this problem gives the same results. This is reassuring regardless of whether one is classical, Bayesian, or eclectic in orientation; a similar convergence between classical and Bayesian methods has been observed for other problems in the analysis of the single case (Crawford & Garthwaite, 2007).

As was the case for the classical approach, the Bayesian approach to the present problem is greatly simplified by the fact that we do not need to be concerned with the control mean. It proceeds by combining the control data with a diffuse, non-informative prior distribution for the control variance (we denote the population variance as σ^2). The posterior distribution states that the distribution of $(n-1)s^2/\sigma^2$ is a chi-squared distribution on $n-1$ degrees of freedom. From this it is almost immediate that there is perfect agreement between the Bayesian and classical interval estimates for the proportion of pairs of controls with a greater difference than the cases' difference. That is, the Bayesian $1-\alpha$ credible interval for this proportion will have precisely the same endpoints as the $1-\alpha$ confidence interval. (The Bayesian interval estimate is called a credible interval.) The result is derived in Appendix 1.

In Appendix 1 we also show that, if D is the difference between two controls, then the posterior distribution of $D/\sqrt{2s_x^2}$ is a t-distribution on $n-1$ degrees of freedom. This corresponds to the result from classical approach, that $g/\sqrt{2s_x^2}$ also

has a t-distribution on $n-1$ degrees of freedom (equation (2)). The consequence is that, for any level of confidence $(1-\alpha)$, the endpoints of the two-tailed $(1-\alpha)$ Bayesian credible interval for $D/\sqrt{2s_x^2}$ are identical to those of the two-tailed $(1-\alpha)$ classical confidence interval for $g/\sqrt{2s_x^2}$. This leads to the result that the Bayesian and classical approaches give identical point estimates for the proportion of differences between pairs of controls that would be more extreme than the difference observed between the two cases. (The result is derived in Appendix 1.)

The Bayesian paradigm aims to judge hypotheses on the basis of Bayes factors rather than p-values, but appropriate methods of calculating Bayes factors are a contentious issue. With classical statistics, rejecting a hypothesis test at significance level p is equivalent to the hypothesised value being outside the $1-p$ confidence interval. Bayes p-values may be defined by relating them to credible intervals in the same way, and then the Bayes p-value and the classical p-value are identical when testing whether the differences found between two cases is greater than would be found in a pair of controls (cf Appendix 1).

<div align="center">Results and Discussion</div>

<u>Examples of these methods</u>

To illustrate the use of the methods, two examples are provided in Table 1. In both these examples the control sample's standard deviation for a task of interest is 10. In the first example (Case A versus Case B) the difference between the cases scores on the task is 30 (this would be obtained, for example, if Case A scored 110 and Case B 80 on the task, or if Case A scored 90 and Case B 60, etc). The sample size for controls in this example is varied from 5, through 20, to 50, and results are presented

for both the classical and Bayesian methods. For the second example (Case C versus

Case D) the control sample standard deviation remains at 10 but the difference in

scores between the cases is 20.

For each case, and at each control sample n, the following results are

presented (in column order): (a) the one-tailed probability used to test whether we can

reject the null hypothesis that the difference between the two cases could be observed

in a pair of controls (two-tailed probabilities are also reported in the computer

programs accompanying this paper but are omitted here for reasons of space); (b) the

point estimate of the effect size for the difference between cases with its

accompanying interval estimate (the point estimate of the effect size does not of

course vary with sample size, whereas the interval estimate does), and (c) point and

interval estimates of the percentage of controls that will exceed the difference

observed between the cases.

Thus, taking a difference of 30 points between the cases scores (i.e., Case A

vs. Case B), a control sample n of 20, and limiting ourselves to the classical results, it

can be seen that the difference between cases is statistically significant (the one-tailed

p-value is 0.0236). The effect size (point estimate) is large in this example (2.12), but

the uncertainty associated with this estimate remains considerable (1.45 to 2.79);

finally, it is estimated that only 2.36% of pairs of controls would exhibit a difference

larger than that observed for the cases. Again, however, there is uncertainty over this

estimate – the uncertainty is quantified with the accompanying interval estimate,

which ranges from 0.26% to 7.32%.

The first point to make about these results is that they illustrate that the

classical and Bayesian methods give the same estimates. The results are identical in

all scenarios. As noted, this equivalence between methods has been obtained when

tackling other problems in the analysis of single cases (Crawford & Garthwaite, 2007) and is reassuring.

The equivalence also means that, regardless of which method is used, a Bayesian interpretation can be placed on the interval estimates. As Antelman (1997) notes, the frequentist (classical) conception of a confidence interval is that, "It is one interval generated by a procedure that will give correct intervals 95% of the time. Whether or not the one (and only) interval you happened to get is correct or not is unknown" (p. 375). Thus, in the present context, the classical interpretation of the interval estimate on the percentage of pairs of controls that will exhibit a larger difference than the two cases is as follows, "if we could compute a confidence interval for each of a large number of control samples collected in the same way as the present control sample, about 95% of these intervals would contain the true percentage".

The Bayesian interpretation of such an interval is "there is a 95% probability that the true percentage lies within the stated interval". This statement is not only less convoluted but it also captures what a neuropsychologist would wish to conclude from an interval estimate (Crawford & Garthwaite, 2007). Indeed most psychologists who use frequentist confidence limits probably construe these in what are essentially Bayesian terms (Howell, 2002).

The second point to make is that differences between cases have to be substantial before they can be considered unusual. For example, in comparing Cases C and D, in which there is a difference of 20 points between the cases scores, the results do not even approach significance. Indeed even when comparing cases A and B, where the difference was 30 points, this difference is not sufficiently large to be

significant (although it is very close) when the control sample n is very modest (i.e., n = 5).

It is not of course surprising that a difference between *two cases* has to be considerably larger than the difference between a *case* and a *control sample* before it is statistically significant. This is most easily appreciated if we ignore the uncertainty over the control sample standard deviation and temporarily pretend it is a population standard deviation. If there is a difference of say 20 between a case and the control *sample mean* then dividing this difference by the control SD yields a z score of 2.0; the difference is extreme and on a one-tailed test, statistically significant ($p = 0.0228$). However, the standard deviation of the difference between two cases will be larger than the standard deviation of controls by a factor of $\sqrt{2}$. Thus the standard deviation of the difference in this example (where the control sample SD is 10) is 14.142 and a difference of 20 between the cases converts to a z score of 1.414. Thus a difference of this magnitude is not very unusual in the control population and can never be statistically significant, even if we had an infinitely large sample of controls (z would need to be > 1.645).

The third point to be made concerns the size of control samples. These methods remain valid when used with very modest control sample sizes in that the Type I error rate is controlled, and the confidence intervals have the coverage specified (because the uncertainty over the control sample statistics is allowed for; Study 2 presents empirical evidence for this from simulation). However, just because a method *can* be used with very small control samples does not mean that researchers should use very modest samples. It can be seen from Table 1 that the power to detect a difference is strongly affected by the size of the control sample. Power is inevitably an issue in single case studies: a single case, or in the present scenario, the difference

between two single cases (rather than a patient sample, or differences between patient samples) is compared to a control sample. Typically the control sample is itself relatively modest in size and it makes sense to make the sample as large as is feasible, so as to maximize the power to detect effects; see Crawford et al. (2006) for a more detailed discussion of this topic.

Combining the current methods with methods comparing a single case to controls

The present methods test whether the difference between two cases could be observed in a pair of controls. When the results indicate that we can reject this null hypothesis a researcher may wish to follow up on this: a significant result could be obtained if the score of only one of the cases was not an observation from the distribution of scores in the control population, or if both scores were not. One can test whether the score of each case differs significantly from controls using Crawford and Howell's method (Crawford & Howell, 1998) or its Bayesian equivalent (Crawford & Garthwaite, 2007).

Although Crawford and Howell's method can be used to explore the reasons for a difference between cases it is worth reiterating that it is not an alternative to the methods developed here. Demonstrating a deficit in Case A relative to controls and a null result for Case B is not adequate as a means of comparing cases as it essentially relies on trying to prove the null hypothesis of no difference between Case B and controls. In contrast, the present methods provide a *positive* test for a difference between two cases. Therefore the present methods and the Crawford and Howell (1998) method should be seen as complimentary rather than as alternatives.

Study 2

Evaluation of the proposed methods by Monte Carlo simulation

The statistical theory developed in Study 1 dictates that, provided the normality assumption holds, the classical and Bayesian methods set out in Study 1 will yield estimates that are correct. However, for neuropsychologists with a limited knowledge or interest in statistical matters, an empirical demonstration of the performance of these methods may be more convincing than an appeal to statistical theory alone. Therefore a series of Monte Carlo simulations were performed. In addition, Monte Carlo simulation can also be used to examine the robustness of the methods when the normality assumption is violated.

The first simulation evaluated control of the Type I error rate for the hypothesis test. Given that the probability obtained from the hypothesis test also provides the point estimate of the abnormality of the difference between cases, the simulation also serves to evaluate this latter statistic. The second simulation examined whether the *interval* estimates performed as they should. Both of these simulations evaluated the classical methods but, given that the classical and Bayesian methods are equivalent, the simulations evaluate both approaches.

As noted, an assumption underlying the use of these methods is that the control samples have been drawn from a normal distribution. However, it is not uncommon for the scores of controls on neuropsychological tests to depart from normality (Capitani & Laiacona, 2000; Crawford & Garthwaite, 2005)

Negative skewness is common in control data because the tasks employed often measure abilities that are largely within the competence of most healthy individuals and thus yield ceiling, or near-ceiling, levels of performance. 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, Gale, Leeson, & Crawford, 2005). Ceiling or near-ceiling performance in controls is also a characteristic of most stimulus sets used to test for deficits in the recognition of specific emotions from facial expressions or prosody (Milders, Crawford, Lamb, & Simpson, 2003).

Another potential problem that will arise in the conduct of single-case research is that the distribution of control data will be overly peaked and have heavier tails than a normal distribution. That is, in practice, the distribution of the control data may be leptokurtic. Indeed the effects of ceiling or near-ceiling performance is that the distribution of control data will be both negatively skewed *and* leptokurtic (Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006).

It is therefore important to examine the effects of departures from normality on these methods so as to examine whether they are robust. That is, there is the danger that, for the hypothesis tests, the Type I error rate will be grossly inflated when the assumption of normality is violated. In this context, a Type I error would occur if we wrongly reject the null hypothesis i.e., if we classified the difference between a pair of cases as not having been drawn from the distribution of differences in the control population. Data on the effects of departures from normality would either provide reassurance for researchers, should the effects be relatively mild, or serve as a warning, should the effects be substantial (in the latter scenario strategies for dealing with the problem should also be addressed).

In this study we quantify the control over the Type I error rate for the test on the difference between two cases when the control data possess varying levels of skewness, leptokurtosis, or both skewness and leptokurtosis. We limit ourselves to examination of negative skewness as this will be the more common problem in practice. (Positive skewness can be a feature of control data when performance on a

task is expressed as errors rather than number correct. However, the results of the

present study will be equally applicable in such a scenario because, if the data were

reflected, they would possess an equivalent degree of negative skewness).

Method

Simulations to verify the statistical theory for the hypothesis test and interval

estimates

The first simulation was conducted using control sample sizes of 5, 10, 20, 30, and

100. For each of these five sample sizes, 5 million trials were performed in which

scores for 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 two further observations were drawn from the same distribution: these

represented the scores of two cases. The difference between the scores of the two

cases was computed and entered into equation (2). The percentage of trials in which

the difference between the two cases was significantly larger ($p < 0.05$, one-tailed)

than controls was recorded (that is, on each trial, the quantity obtained from equation

2 was evaluated against a t-distribution on $n-1$ df).

Note that, in this first simulation, both the control sample and the two cases

were drawn from the same population. Therefore, a statistically significant result

constitutes a Type I error (the null hypothesis, that the difference between the scores

of the two cases is an observation from the population of differences among pairs of

controls, has incorrectly been rejected).

As noted, the second simulation examined whether the confidence limits for the

difference between two cases perform as they should. If the theory set out earlier is

correct, then, assuming the control data are drawn from a normal distribution, the

confidence intervals on the effect size for the difference should capture the true effect

sizes 95% of the time. The true effect size, which we denote z^*_{PCC}, is the effect size

that would be obtained using the standard deviation of the control *population* rather

than that of a control *sample*).

For this second simulation, four population values of z^*_{PCC} were selected

ranging from -0.674 (representing a fairly modest difference between cases, i.e. 25%

of pairs of controls would obtain a larger difference) through values of -1.282 (10%

of controls), -1.960 (2.5%), to a value of -2.326 (representing a very substantial

difference; only 1% of pairs from the control population would obtain a larger

difference). For each of these z^*_{PCC}, 100000 Monte Carlo trials were run in which a

sample of controls of size n were drawn from the control distribution (a standard

normal distribution); on each trial the value of z^*_{PCC} was divided by $\sqrt{2s^2}$, where s^2

was the variance of the control sample on that trial. This created an estimated effect

size and the method set out earlier was then applied to calculate a 95% confidence

interval on z^*_{PCC}. The number of trials in which these confidence intervals captured

z^*_{PCC} was recorded. If the method is valid then the percentage of trials on which this

occurred should be 95%, save for Monte Carlo variation.

Five different control sample sizes were used ranging from 5 through 10, 20,

30, to 100. Thus, in total, two million trials were performed: i.e., 100000 trials times

four levels of z^*_{CC}, times five levels of the control sample n.

Simulations examining the robustness of the hypothesis test in the face of skewness

and leptokurtosis

The procedure for the simulations conducted to examine the effects of departures from normality were identical to that for the first simulation, except that the controls and the pairs of cases were sampled from non-normal distributions. To examine the effects of skewness alone, controls and cases were sampled from skew-normal distributions (Azzalini & Dalla Valle, 1996) in which the skewness ranged from absent ($\gamma_1 = 0$), through moderate ($\gamma_1 = -0.31$), severe (-0.70), very severe (-0.93), to extreme negative skewness (-0.99).

The most common approach to modelling the effects of leptokurtic distributions on test statistics is to sample from t-distributions (Lange, Little, & Taylor, 1989). This is potentially confusing as the present methods use t-distributions to test for a significant difference between the case and controls. However, as noted, the assumption in applying the test statistic is that the controls were drawn from a normal distribution; in the present study we examine the effects of violating this assumption by drawing controls from leptokurtic distributions and it so happens that t-distributions have this required characteristic. To examine the effects of leptokurtosis alone, controls and cases were sampled from a t-distribution on either 7 df (moderate leptokurtosis) or 4 df (severe leptokurtosis).

To examine the effects of the combination of skewness and leptokurtosis, the cases and controls were sampled from skew-t distributions (Azzalini & Capitanio, 2003). These distributions were parameterized so that the combination of skewness and leptokurtosis was varied using the five levels of skewness identified earlier (absent to extreme), and the three levels of leptokurtosis (absent to severe). Technical details on sampling from skew-normal, skew-t, and leptokurtic (i.e., t-) distributions are provided in Appendix 2.

In summary, a total of 75 million Monte Carlo trials were performed; i.e., one million trials for each combination of five sample sizes, five levels of skewness (absent through to extreme), and three levels of leptokurtosis (absent, moderate, and severe). Note that, when there is neither skewness nor leptokurtosis, sampling is from a normal distribution and constitutes a replication of the first simulation.

For illustrative purposes a graphical representation of two of the distributions employed in the present study are presented as Figure 1. The unshaded distribution is the skew-normal distribution used to represent extreme skewness. It can be seen that this is indeed a very extreme example: compared to a normal distribution, the entire right hand side of the distribution is essentially absent. Also presented is the skew-t distribution used to represent the combination of extreme skewness and severe leptokurtosis. Again it can be seen that this also clearly qualifies as an extreme example: compared to the skew-normal (or standard normal distribution) the distribution is very peaked and provides a good representation of a task in which most members of the control population would be at or near ceiling.

Results

Results from Monte Carlo simulations

The results of the first Monte Carlo simulation are presented in Table 2. Recall that this simulation was designed to examine whether the Type I error rate for the hypothesis test was under control. It can be seen that the observed Type I error rate closely matches the specified error rate of 5% (minor differences are attributable to Monte Carlo variation). Therefore the error rate is under control; it also follows that the point estimate of the abnormality of the difference between cases has no systematic bias (recall that the p value for the hypothesis test is also used to provide

this point estimate). Finally, it was the classical method that was evaluated in the

foregoing simulation. However, as these classical and Bayesian methods are

equivalent, the results apply equally to both.

The results of the second simulation are set out in Table 3; this simulation

evaluated whether the confidence intervals on the effect size for the difference

between cases performed as they should. It can be seen that the percentage of trials in

which the limits captured the true effect size is very close to 95%, regardless of both

the size of the control sample and the extremity of the true effect size (the small

departures from 95% are within the range expected from Monte Carlo variation).

Thus, these results provide empirical confirmation of the statistical theory set out in

the present paper. In other words, these results confirm the veracity of the methods

derived here for confidence limits on an effect size for the difference between a pair

of cases and, because the confidence limits on the abnormality of the difference

between cases depends on the same theory, they also confirm the veracity of these

latter limits.

Effects of departures from normality

The results of the Monte Carlo simulations examining the effects of skewness and

leptokurtosis on the Type I error rate are reported in Table 4. Examining the effects

of leptokurtosis alone first (see the first column of results in Table 4 in which the

skewness was zero), it can be seen that the effects of leptokurtosis, even when severe,

are modest: for small control sample sizes the Type I error rate is inflated above the

nominal rate of 5% but not dramatically; i.e., for a control sample of size 10 the Type

I error rate is 5.86%. For large sample sizes the Type I error rate falls *below* the

specified rate (but again only marginally). The explanation for this latter (perhaps

counterintuitive) result is that a probability distribution is leptokurtic if it has thicker

tails and is more peaked than a normal distribution. It follows that, between the tails

and the peak, the distribution must be lower (the area under the distribution must

equal one); that is, a leptokurtic distribution has thinner shoulders than a normal

distribution. In the present scenario the effects of the thin shoulders are in evidence;

the statistical test is applied in a region of the distribution that is not sufficiently far

out in the tails to produce inflated Type I errors. Inflation of the error rate would be

expected to occur even for larger samples were a more extreme (i.e., more

conservative) value of alpha employed rather than the conventional value of 0.05.

Turning to the effects of skewness alone on Type I errors, it can be seen from

the first row of results in Table 4 that again the results are very encouraging. Whilst

the error rate is above the nominal rate at all sample sizes and at all levels of

skewness, the effects are relatively modest, particularly for larger sample sizes and

with less extreme degrees of skewness. Even when skewness is *extreme* (i.e., see the

unshaded distribution in Figure 1) and sample size is modest ($N = 10$) the error rate is

only 5.54%.

As noted, neuropsychological task data will not infrequently be both skew and

leptokurtic. The effects of the combination of both skewness and leptokurtosis are

reported in the cells of Table 4 not already discussed above. It can be seen that,

although the effects are typically larger than those observed for skewness or

leptokurtosis alone, again the results suggest that the test is surprisingly robust given

the very extreme departures from normality that are represented by some of these

combinations. For example, even when the control sample size is modest ($N = 10$)

and is combined with severe leptokurtosis and extreme skewness, the Type I error rate

is 6.45%; the error rate is inflated but not dramatically even when departures from

normality are very extreme (see Fig. 1). In conclusion, the results suggest that the test

for comparing two cases is robust in the face of violations of the normality

assumption. However, it does remain the case that the Type I error rate is inflated and

so some guidance might be useful.

The first obvious means of dealing with skewness is for researchers to select or

design measures that are not subject to ceiling effects in controls in the first place (i.e.,

by upping the item difficulty). Secondly, as the inflating effects of departures from

normality are attenuated with increasing control sample size, researchers should use

as large a sample of controls as is practical (this would have the additional positive

effect of increasing statistical power). Thirdly, researchers could attempt to apply a

normalizing transformation to their control distribution using, for example, a Box-Cox

transformation (note that, although the search for a normalizing transformation should

be conducted using the control data only, it is crucial that the same transformation is

then applied to the scores of the cases). Finally, if researchers are concerned about

departures from normality then simply adopting a slightly more conservative value of

alpha would provide protection over inflation of the error rate even when the

underlying control distributions were severely skew and leptokurtic. For example, an

additional simulation revealed that for control sample sizes of 10 or more, and a

combination of severe leptokurtosis and very severe skewness, the maximum

observed error rate was 5.21% (for $N = 10$) when using a nominal alpha of 3.5% and

was well below 5% for larger sample sizes.

The robustness of the test on the difference between two cases may come as a

surprise to some researchers. However, it is in keeping with results from Type I error

studies of other parametric tests used in single case studies, such as Crawford and

Howell's (1998) test for comparing one single case to a control sample (Crawford &

Garthwaite, 2005; Crawford et al., 2006), in which the effects of non-normal data were modest. That said, the simulation data indicates that the present methods are even more robust than those existing single case methods.

Study 3

Comparing two cases in the case-controls design allowing for covariates

Two of the present authors have recently developed methods to permit the comparison of a single case to a control sample allowing for the effects of covariates (Crawford & Garthwaite, submitted). This work prompted us to examine whether the methods presented in Study 1 of the present paper could be extended to allow comparison of two cases while controlling for covariates. In some respects the ability to control for the effects of covariates may be even more useful when comparing two cases than when comparing a case to a control sample. For example, although the process may be difficult and time-consuming, it is, in principle, always possible to obtain a control sample that closely matches a case on demographic variables known to effect neuropsychological test performance (e.g., age, or years of education, and for some tasks, gender). However, when comparing two cases it will be largely a matter of chance whether the cases are matched on such variables (they will normally have been selected for comparison because they allow testing of some theoretical proposition). It also follows that, if the two cases differ markedly on demographic variables, the control sample against which the two cases scores are referred cannot simultaneously be a good match for both cases.

Fortunately it is possible to compare two cases allowing for the effects of covariates. The problem can be specified as testing whether the difference found between two cases (Case A and Case B) is greater than would be found in a pair of

controls, one of whom has the same values on the covariates as Case A, the other the same values on the covariates as Case B. That is, as before, we can specify the null hypothesis that the difference between the cases is an observation from the distribution of differences in a control population and, if we can reject this hypothesis, we can conclude that there is a difference between the cases that is larger (by a statistically significant amount) than would be observed in two randomly chosen controls whose covariate values are similar to the cases. The difference in the present scenario is that the control population is more narrowly defined to consist of only those pairs of controls that match the pair of cases' values on the covariates. It is possible to provide the same full range of statistics for this problem as was provided when covariates were ignored. Namely: (a) a one- or two-tailed hypothesis test, (b) point and interval estimates of the effect size for the difference between cases, and (c) point and interval estimates of the abnormality of the difference between the cases.

A Bayesian analysis of this problem gives the same results as a classical approach. However, having already illustrated equivalence between classical and Bayesian methods for these types of problems in Study 1, and in the interests of brevity, we present only the classical methods here. The method is set out formally in the next section (and in Appendix 3), this is then followed by the presentation of illustrative worked examples.

Method

A classical method for comparing two cases in the presence of covariates

As noted, the aim is to develop a hypothesis test and confidence interval for whether the difference found between two cases is greater than would be found in a pair of controls when there are covariates. The basic method is developed in this section for

the simpler case when there is only one covariate. Appendix 3 extends the method

developed to cover the vector case, i.e., to compare the scores of the two cases when

there are two or more covariates.

We have two cases whose observed values are y_1^* and y_2^*, and the values of a

covariate (X) that they take are x_1 and x_2. There is also a population of controls. We

suppose that we pick one control from the set of controls whose covariate value is x_1

and a second control from the set of controls whose covariate value is x_2. The Y

values of these controls are Y_1 and Y_2. We want $P(Y_1 - Y_2 > y_1^* - y_2^*)$.

By assumption,

$$Y = \mu + \beta x + \varepsilon,$$

where ε is the random error which we assume is normally distributed with a mean of

0 and an unknown variance σ^2. Then

$$Y_1 - Y_2 = \beta(x_1 - x_2) + \varepsilon_1 - \varepsilon_2.$$

Let $D = Y_1 - Y_2$. If β and σ^2 were known, then

$$D \sim N(\beta(x_1 - x_2), 2\sigma^2). \tag{6}$$

Data from n controls give paired values $(x_1. y_1)$, … , $(x_n. y_n)$, from which we calculate

corrected sums of squares and products:

$$S_{xx} = \sum (x - \bar{x})^2$$

$$S_{yy} = \sum (y - \bar{y})^2$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}).$$

We put

$$\hat{\beta} = S_{xy} / S_{xx} \tag{7}$$

and

$$\hat{\sigma}^2 = (S_{yy} - \hat{\beta} S_{xy}) / (n-2). \tag{8}$$

Then

$$\hat{\beta} \sim N(\beta, \sigma^2 / S_{xx}). \tag{9}$$

Also, $(n-2)\hat{\sigma}^2 / \sigma^2$ has a chi-square distribution on $(n-2)$ degrees of freedom.

Hence, $P(Y_1 - Y_2 > y_1^* - y_2^*)$ is equal to

$$P(t > \frac{(y_1^* - y_2^*) - \hat{\beta}(x_1 - x_2)}{\sqrt{2\hat{\sigma}^2 + \{(x_1 - x_2)^2 \hat{\sigma}^2 / S_{xx}\}}}) \tag{10}$$

where t has a t-distribution on $(n-2)$ df. This is the point estimate of the required

probability.

The confidence interval is derived from a non-central t-distribution. This

distribution is defined by

$$T_\nu(\delta) = \frac{(Z + \delta)}{\sqrt{\psi / \nu}},$$

where Z has a standard normal distribution with a mean of zero and variance 1, and ψ

is independent of Z with a Chi-square distribution on ν degrees-of-freedom. δ is

referred to as the non-centrality parameter.

For specified values y_1^* and y_2^*, let $P^* = \Pr(Y_1 - Y_2 < y_1^* - y_2^*) \times 100$. Let

$$c = \frac{(y_1^* - y_2^*) - \hat{\beta}(x_1 - x_2)}{\hat{\sigma}\sqrt{2}}$$

And let $c^* = \{(y_1^* - y_2^*) - \beta(x_1 - x_2)\} / \{\sigma\sqrt{2}\}$. Then c is an estimate of c^*. Also,

$\Pr(Y_1 - Y_2 < y_1^* - y_2^*) = \Pr(Z < c^*)$, so percentage points of a normal distribution

determine P^* from c^*. Thus a $100(1-\alpha)\%$ confidence interval for c^* will yield the

required confidence interval for P^*. Now,

$$\frac{c\sqrt{2S_{xx}}}{x_1 - x_2} = \frac{(\beta - \hat{\beta})\sqrt{S_{xx}}/\sigma + \{y_1^* - y_2^* - \beta(x_1 - x_2)\}\sqrt{S_{xx}}/\{\sigma(x_1 - x_2)\}}{\sqrt{\hat{\sigma}^2/\sigma^2}}$$

so $c\sqrt{2S_{xx}}/(x_1 - x_2)$ has a non-central t-distribution with non-centrality parameter

$\delta = c^*\sqrt{2S_{xx}}/(x_1 - x_2)$ and $n - 2$ df. The $100(\alpha/2)\%$ and $100(1 - \alpha/2)\%$ points of

this distribution will depend on the value of δ. Let δ_L denote the value of δ for

which the $100(1 - \alpha/2)\%$ point is $c\sqrt{2S_{xx}}/(x_2 - x_1)$. Similarly, let δ_U denote the

value of δ for which the $100(\alpha/2)\%$ point is $c\sqrt{2S_{xx}}/(x_1 - x_2)$. Then

$(\delta_L(x_1 - x_2)/\sqrt{2S_{xx}}, \ \delta_U(x_1 - x_2)/\sqrt{2S_{xx}})$ is a $100(1 - \alpha)\%$ confidence interval for c^*.

Hence a $100(1 - \alpha)\%$ confidence interval for P^* is

$$(\Pr(Z < \delta_L(x_1 - x_2)/\sqrt{2S_{xx}}), \ \Pr(Z < \delta_U(x_1 - x_2)/\sqrt{2S_{xx}}))$$

where $Z \sim N(0,1)$.

A special case arises when the two cases have the same score on the covariate. When this occurs the problem is simplified: the method reduces to that set out in Study 1 when there are no covariates except that (a) the *conditional* standard deviation ($\hat{\sigma}$) is substituted for the unconditional standard deviation (s) in all equations, and (b) the degrees-of-freedom become $n - 2$ rather than $n - 1$ when evaluating t and in setting the confidence limits (one degree-of-freedom is lost because of the covariate).

As noted, the method set out in this section can be extended to allow for the effects of more than one covariate. The details are provided in Appendix 3. Note that the same special case identified above can occur when there are multiple covariates. If the two cases obtain the same scores on all covariates, then again the problem is simplified; the only difference between this scenario and that in which there is only

one covariate is that the degrees-of-freedom become $n - m - 1$ when applying the methods set out in Study 1, where *m* is the number of covariates.

Results and Discussion

These methods can be used for a number of purposes. For example they can be used to increase the power to detect a difference between two cases. For example, if premorbid IQ is related to task performance on a task of interest, and the case with the poorer score has a higher estimated premorbid IQ than the other case, then controlling for premorbid IQ will more clearly expose the difference between the two cases. Perhaps less intuitively, power can be increased even when the two cases have the same score on the covariate; an example of this outcome is provided later.

Statistical power is inevitably an issue in single case research (Crawford & Garthwaite, 2006) and is even more so when comparing two cases. As noted, a difference between cases has to be substantial before we can reject the null hypothesis that it is an observation from the control population. Thus, by allowing for the possibility of an increase in power, the ability to control for covariates is particularly useful.

The methods can also be used to test whether a difference between cases survives the effect of controlling for a covariate. For example, to continue with the example of premorbid IQ as a covariate: if the scenario just outlined was reversed, in that the case with the superior task score had the higher premorbid IQ, then there is the possibility that the difference observed may be an artefact of pre-existing differences rather than the result of contrasting lesions / acquired impairments.

Some fully worked examples are provided in the next section. It is particularly informative to contrast the results obtained when covariates are included to those

obtained when covariates are ignored. For clarity, we limit these examples to the use of one covariate.

<u>Worked example of testing for a difference between two cases in the presence of</u>

<u>covariates: increasing statistical power</u>

Suppose that a case (Case A) obtains a score of 64 on a task of interest while Case B obtains a score of 96. Suppose also that the standard deviation for this task is 15 in a sample of 18 controls. In the *absence* of a covariate we can test whether the two cases differ significantly using the classical or Bayesian methods developed in Study 1. Applying either method, the case's scores do not differ significantly: $t = 1.509$, p (one-tailed) $= 0.0749$. It is estimated that 7.49% of pairs drawn from the control population would exhibit a larger difference and the 95% interval estimate for this percentage ranges from 2.22% to 15.72%.

Now suppose that we wish to compare the two cases allowing for a single covariate, years of education. Most neuropsychological tests correlate with years of education so this is a realistic example and will arise commonly in practice. Suppose that both cases have 13 years of education. Further suppose that the mean number of years of education in controls is also 13 with a standard deviation of 3.0, and that the correlation between task performance and years of education in the controls is 0.65.

Casual inspection of this scenario might suggest that it is not a promising one for the use of the present methods: the cases have the same years of education (and this is exactly equal to the mean years of education of controls). Therefore, as the cases appear to be "matched" for education, allowing for the effects of education in comparing the scores of the two cases may seem to add nothing. Indeed, given that one degree-of-freedom is lost by incorporating the covariate into the analysis, it might

even be thought that such a procedure will *lower* power, thereby lessening the chances

of detecting a difference between the cases. However, this view would be erroneous

because the conditional standard deviation for the task is 11.75, which is smaller than

the unconditional standard deviation for the task (15). The uncertainty attached to the

cases' expected scores has been reduced because education is a predictor of task

performance. As a result the difference between the cases' scores obtained on testing

is more extreme (the raw difference is divided by $11.75 \times \sqrt{2}$, rather than $15 \times \sqrt{2}$),

and the null hypothesis that the difference between the cases is an observation from

the distribution of differences between pairs of controls with 13 years of education

can be rejected. The one-tailed p value is 0.036 on 16 df, compared to the value

obtained in the absence of the covariate ($p = 0.075$). The effect size for the difference

controlling for education ($z_{PCCC} = 1.926$) is also larger than that obtained without the

covariate ($z_{PCC} = 1.508$) and the percentage of controls expected to exhibit a larger

difference is smaller (3.60%; 95% CI $= 0.49\%$ to 10.28%).

In the foregoing example the cases had the same years of education. If the case

with the poorer score (Case A) had a higher number of years of education than Case

B, then even greater power will be achieved by inclusion of the covariate: not only

will the uncertainty over the cases' expected scores be reduced (i.e., as in the original

example, the conditional standard deviation will be reduced) but also the conditional

mean for Case A (i.e., the expected score for Case A given their years of education)

will be higher. As a result the difference between the cases becomes even more

unusual. As a concrete example, suppose that all values were the same as those used

in the first example except that Case A had 14.5 years of education. Then the

(one-tailed) p value is 0.021and the effect size (z_{PCCC}) is 2.10.

Testing whether a difference between cases survives controlling for covariates

In the foregoing examples it was demonstrated that allowing for covariates can increase power to detect differences between cases. In contrast, another potential application of these methods is to test whether differences survive after controlling for covariates. Suppose that a significant difference between cases had been observed prior to allowing for covariates but that, in contrast to the previous example, it was the case with the superior score who had the higher number of years of education. This raises the question of whether the difference in task scores may be largely an artefact of the differences in education: we would therefore want to test whether the difference in task scores survives when allowing for the differences in education.

Years of education has been used as the covariate in all these examples but it will be appreciated that there may be many other variables that would be potentially important covariates. For example, a researcher (or reviewer) may be concerned that an apparent difference between two cases on a given task is an artefact of differences between the cases in their general psychomotor speed. In this scenario some form of speeded task, such as Part A of the Reitan Trail Making Test (Reitan & Wolfson, 1985), could be employed as a covariate.

Finally, it is worth stressing that the method does not require any assumptions about the distribution of the covariate(s); this is a useful feature as in many scenarios researchers would wish to use covariates that happen to be heavily skewed or leptokurtic, or even dichotomous.

When should an analysis include covariates and how many covariates should be incorporated?

In the interests of simplicity the worked examples provided were limited to the use of

one covariate. However, technically the number of covariates is limited only by the

degrees-of-freedom (df) available. When comparing two cases in the presence of

covariates there are $n - m - 1$ df so that, if there were say 10 controls, a researcher

could, in principle, include nine covariates. However, in practice it could be

anticipated that between one and three covariates would be typical. Moreover, the

computer programs that implement these methods (see a later section) are limited to a

maximum of five covariates, regardless of the number of degrees-of-freedom

available.

Turning now to the question of when a potential covariate should be included

in an analysis: in an earlier section it was shown that including a covariate can

increase the power to detect a difference between cases even when the cases have the

same value on the covariate. Thus, the fact that cases are matched on the covariate

does not mean that the covariate should be excluded. However, when the potential

covariate has a zero, or near zero, correlation with the task or tasks of interest there

would be no point in including it. Indeed, as each covariate leads to the loss of one

degree-of-freedom, there would be every reason not to include the variable as it would

lower power. An exception to this general rule might occur when there are strong

theoretical reasons (or reasons based on previous empirical findings) to incorporate a

variable as a covariate, but even in these circumstances an alternative is simply to

report that the variable is not related to (i.e., is uncorrelated) with the task of interest.

It need hardly be said but researchers should adopt a principled approach to the

use of covariates. While it is sound practice to seek an appropriate model that fits the

available data, it would be very bad science to go on a fishing expedition in which the

effects of all possible combinations of potential covariates were examined until one

was found that gave results closest to the researchers "favoured" outcome. When control sample sizes are small, in particular, such an approach may well lead to spurious results. Interesting questions that can be addressed by the use of covariates will often arise after data collection is completed, but in such circumstances it should be acknowledged that these analyses were post hoc and treated with great caution if sample sizes are modest

General Discussion

Implementation of the present methods in computer programs

The methods developed in the present paper have been implemented in compiled computer programs written in the Delphi programming language for PCs. These programs will also run on a Mac if emulation software is installed (R code may also be available at a future date; if so a note to that effect will be added to the web page listed below).

In keeping with other computer programs made available by the present authors (Crawford & Garthwaite, 2002; Crawford & Garthwaite, 2005, 2007) the programs take summary data from the control sample as inputs. Being able to conduct the analysis from summary data has a number of potential positives: first, the summary data are normally already available from provisional analysis of the study conducted using standard statistical packages or spreadsheets; second, it provides a convenient means of using control data from other researchers; and third it allows users to use large scale normative data as control sample data (such data are normally only available in summary form). All programs reproduce the data entered by the users so that researchers have a record of the inputs.

The program C_CTC.exe (**C**lassical method **C**ompare **T**wo **C**ases) requires only that the researcher enters the standard deviation for the task of interest in the control sample, the control sample *n*, and the scores of the two cases. The output consists of: (a) the results of the hypothesis test, i.e., the *t* value, its df, and associated one- and two-tailed probabilities; (b) the point and 95% interval estimates of the effect size for the difference between cases (z_{PCC}); and (c) the point and interval estimates of the percentage of pairs of controls that will exhibit a larger difference than the cases. The Bayesian version of this test, B_CTC.exe, (**B**ayesian method **C**ompare **T**wo **C**ases) takes the same inputs and provides the same outputs as its classical equivalent.

For the convenience of neuropsychologists three programs were written for the analysis of two cases when allowing for the effects of covariates. The first of these programs, CTC_Cov.exe (**C**ompare **T**wo **C**ases allowing for a **C**ovariate), performs the classical test for a difference between two cases in the presence of a *single* covariate. The inputs are those required for the preceding program but also include the standard deviation for the covariate and the correlation between the covariate and the task of interest in the control sample, and the case's scores on the covariate. The output consists of: (a) the results of the hypothesis test, i.e., the *t* value, its df, and associated one- and two-tailed probabilities; (b) the point and 95% interval estimates of the effect size for the difference between cases (z_{PCCC}); and (c) the point and interval estimates of the percentage of pairs of controls that will exhibit a larger difference than the cases.

When a neuropsychologist wishes to allow for the effects of more than one covariate they should use the program CTC_Vec_Cov.exe (**C**ompare **T**wo **C**ases allowing for a **Vec**tor of **Cov**ariates). This program requires the control sample

means and standard deviations for the task of interest and covariates, and the control

sample correlation matrix (i.e., the correlations between the covariates and the task

and correlations between the covariates). The scores of the two cases on the task and

covariates must also be entered (the program provides explicit guidance on data

entry).

It was noted earlier that there are a number of advantages to allowing the use

of summary data for the controls as data input. However, for this particular problem

it may often be more convenient for users to supply the raw data for controls and let

the program calculate the necessary statistics from scratch. Therefore we have written

a companion program, CTC_Vec_Cov_Raw.exe, which allows that. This latter

program takes a single text file of raw data for the controls and the two cases as input

(guidance on preparation of this text file is provided with the program). The outputs

from this program are the same as its counterpart that uses summary data for the

controls as input. Note that, for those programs that allow for a vector of covariates, a

test is performed to ensure that the correlation matrix is positive definite in case the

user has made an error in entering either the raw data or correlation matrix, or has

inadvertently attempted to use a variable that is linearly dependent on other included

variables. Calculations cannot proceed when the matrix is not positive definite and a

warning to that effect is issued (along with some brief guidance).

The results from all five programs can be viewed on screen, printed, or saved

to a text file. The programs can be downloaded (individually, or in the form of a zip

file containing all five programs) from the following URL:

www.abdn.ac.uk/~psy086/dept/Compare_Two_Cases.htm.

Use of these methods with standardized neuropsychological tests

The emphasis in the present paper has been on the use of these methods when the reference sample against which the difference between the two cases is compared is a sample of matched controls. Given that the control sample sizes in single case studies are often relatively modest in size it was necessary to allow for the uncertainty that stems from using a control sample to estimate the control population parameters. However, the methods can also be used to compare cases when the reference sample is a very large normative sample (so that we can treat the control sample parameters as fixed and known). Thus, the methods could be used to compare two cases' scores on standardized psychological tests such as the Wechsler intelligence scales or memory scales. Readers should be aware that, when the reference sample is a large normative sample and reliability data are available, there are alternatives to the present methods. Most notably Huber (1973) developed a method of comparing the scores of two cases on standardized test batteries based on the difference between their estimated true scores; see Willmes (1985) for an accessible treatment of these methods together with illustrative examples.

Potential extensions to the present methods of comparing two cases in the case-controls design and definitions of double dissociations

Classical (i.e., frequentist) and Bayesian methods have been developed that allow comparison of the standardized difference between two tasks obtained for a case with the standardized differences observed in controls (Crawford & Garthwaite, 2005, 2007). These methods, particularly the Bayesian methods, are useful in testing for dissociations (Crawford, Garthwaite, & Howell, 2009). We believe it would be very useful to attempt to extend these methods so that the standardized differences between two tasks obtained by a pair of cases could be compared directly. This would assist in

the process of testing for double dissociations. It is likely that a satisfactory Bayesian

solution to this problem can be obtained but is also likely that the problem is beyond

classical methods (because they will be unable to capture the additional uncertainties

that arise from standardizing the cases' scores on the two tasks). In principle the

Bayesian methods would also be able to allow for covariates.

Conclusion

Existing methods of comparing two cases in neuropsychological research do not refer

the case's scores to a control sample. It was argued that these methods (i.e., use of chi

square tests of independence, or independent samples t-tests in which the scores

obtained by the two cases on the k items of a task are essentially treated as k cases) do

not address the question of primary interest. Differences in the task performance of

the two cases can be recorded as statistically significant when such differences can be

very common in the healthy control population.

We suggest that the methods developed to address this problem achieve their

intended aims: they provide a comprehensive set of statistics including effect sizes

and interval estimates (and are thus in keeping with contemporary expert opinion on

statistical reporting). The simulations provided in Study 1 show that the methods are

based on sound statistical theory and, reassuringly, the two main schools of statistics

(classical and Bayesian) provide identical results.

The ability to control for the effects of covariates greatly increases the

flexibility of these new methods, particularly as cases will often differ on attribute

variables (e.g., age, years of education, or premorbid IQ) that are known to effect

cognitive performance. Note that this feature constitutes another advantage over

methods that do not refer the cases' scores to a control sample as these latter methods do not control for the effects of covariates.

References

Akiyama, T., Kato, M., Muramatsu, T., Saito, F., Nakachi, R., & Kashima, H. (2006). A deficit in discriminating gaze direction in a case with right superior temporal gyrus lesion. *Neuropsychologia, 44*(2), 161-170.

Antelman, G. (1997). *Elementary Bayesian statistics*. Cheltenam, UK: Elgar.

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society Series B, 61*, 579-602.

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-*t* distribution. *Journal of the Royal Statistical Society Series B, 65*, 367-389.

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika, 83*, 715-726.

Blazely, A. M., Coltheart, M., & Casey, B. J. (2005). Semantic impairment with and without surface dyslexia: Implications for models of reading. *Cognitive Neuropsychology, 22*(6), 695-717.

Capitani, E., & Laiacona, M. (2000). Classification and modelling in neuropsychology: from groups to single cases. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (2nd ed., Vol. 1, pp. 53-76). Amsterdam: Elsevier.

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. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of

alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology, 19*, 318-331.

Crawford, J. R., & Garthwaite, P. H. (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. (submitted). Comparing a single case to a control sample: Testing for neuropsychological deficits and dissociations in the presence of covariates.

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

Crawford, J. R., Garthwaite, P. H., & Howell, D. C. (2009). On comparing a single case with a control sample: An alternative perspective. *Neuropsychologia, 47*, 2690-2695.

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes in the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology, 27*, 245-260.

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.

Forde, E. M. E., & Humphreys, G. W. (2005). Is oral spelling recognition dependent on reading or spelling systems? Dissociative evidence from two single case studies. *Cognitive Neuropsychology, 22*(2), 169-181.

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

Huber, W. (1973). *Psychometrische einzelfalldiagnostik*. Weinheim: Beltz.

Humphreys, G. W., & Forde, E. M. E. (2005). Naming a giraffe but not an animal: Base-level but not superordinate naming in a patient with impaired semantics. *Cognitive Neuropsychology, 22*(5), 539-558.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modelling using the *t*-distribution. *Journal of the American Statistical Association, 84*, 881-896.

Laws, K. R., Gale, T. M., Leeson, V. C., & Crawford, J. R. (2005). When is category *specific* in Alzheimer's disease? *Cortex, 41*, 452-463.

Lee, P. M. (1997). *Bayesian statistics: An introduction* (2nd ed.). London: Wiley.

Linebarger, M. C. (2004). The role of processing support in the remediation of aphasic language production disorders. *Cognitive Neuropsychology, 21*(2-4), 267-282.

Martin, R. C., Miller, M., & Vu, H. (2004). Lexical-semantic retention and speech production: Further evidence from normal and brain-damaged participants for a phrasal scope of planning. *Cognitive Neuropsychology, 21*(6), 625-644.

Milders, M., Crawford, J. R., Lamb, A., & Simpson, S. A. (2003). Differential deficits in expression recognition in gene- carriers and patients with Huntington's disease. *Neuropsychologia, 41*(11), 1484-1492.

Mochizuki-Kawai, H., Kawamura, M., Hasegawa, Y., Mochizuki, S., Oeda, R., Yamanaka, K., et al. (2004). Deficits in long-term retention of learned motor skills in patients with cortical or subcortical degeneration. *Neuropsychologia, 42*(13), 1858-1863.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery*. Tuscon: Neuropsychology Press.

Riddoch, M. J., & Humphreys, G. W. (2004). Object identification in simultanagnosia: When wholes are not the sum of their parts. *Cognitive Neuropsychology, 21*(2-4), 423-441.

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

Shallice, T., Venable, N., & Rumiati, R. I. (2005). Dissociable distal and proximal motor components: Evidence from perseverative errors in three apraxic patients. *Cognitive Neuropsychology, 22*(5), 625-639.

Smith, A. D., & Gilchrist, I. D. (2005). Within-object and between-object coding deficits in drawing production. *Cognitive Neuropsychology, 22*(5), 523-537.

Toraldo, A., & Shallice, T. (2004). Error analysis at the level of single moves in block design. *Cognitive Neuropsychology, 21*(6), 645-659.

Wechsler, D. (2008). *WAIS-IV technical and interpretive manual*. San Antonio TX: Pearson.

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

Table 1. Results of applying classical and Bayesian methods for comparing two cases (for both examples the control sample SD was 10, for Case A vs B the difference in scores was 30, for C vs D it was 20); the methods provide one and two-tailed p values but only the one-tailed results are presented here.

n	Hypothesis test (one-tailed)		Effect Size (95% CI)		%age of population (95% CI)	
	Classical	Bayesian	Classical	Bayesian	Classical	Bayesian
A vs B						
5	0.0506	0.0506	2.12 (0.74 to 3.54)	2.12 (0.74 to 3.54)	5.06 (0.02 to 23.02)	5.06 (0.02 to 23.06)
20	0.0236	0.0236	2.12 (1.45 to 2.79)	2.12 (1.45 to 2.79)	2.36 (0.26 to 7.32)	2.36 (0.26 to 7.32)
50	0.0195	0.0195	2.12 (1.70 to 2.54)	2.12 (1.70 to 2.54)	1.95 (0.56 to 4.43)	1.95 (0.55 to 4.44)
C vs D						
5	0.1151	0.1151	1.41 (0.49 to 2.36)	1.41 (0.49 to 2.36)	11.51 (0.91 to 31.13)	11.51 (0.91 to 31.16)
20	0.0867	0.0867	1.41 (0.97 to 1.86)	1.41 (0.97 to 1.86)	8.67 (3.15 to 16.65)	8.67 (3.14 to 16.65)
50	0.0818	0.0818	1.41 (1.14 to 1.69)	1.41 (1.14 to 1.69)	8.18 (4.53 to 12.82)	8.18 (4.52 to 12.83)

Note ES = Point estimate of the effect size for the difference; %age = point estimate of the percentage of pairs of controls exhibiting a larger difference; both quantities are accompanied by interval estimates in brackets.

Table 2. Results of Monte Carlo simulation: Percentage of Type I errors for the classical method of comparing the difference between the scores

of a pair of cases to differences between pairs of controls (the null hypothesis is that the cases' difference is an observation from the distribution

of differences in the control population); the nominal error rate is 5%

|	Control sample size				
	5	10	20	30	100
Percentage of Type I errors	4.998	5.011	5.001	5.001	5.001

Table 3. Results of Monte Carlo simulation: Percentage of trials in which the 95% interval estimates on the effect size for the difference

between two cases captured the true effect size; the size of the true effect is varied, as is the size of the control sample

	Control sample size				
True Effect Size (z^*_{PCC})	5	10	20	50	100
−0.674	95.011	95.002	94.995	95.012	94.995
−1.282	94.977	95.006	95.995	94.986	94.999
−1.960	95.007	94.998	95.021	95.023	95.010
−2.326	95.011	95.002	94.995	95.010	94.995

Table 4. Results of Monte Carlo simulation: Percentage of Type I errors as a function of degree of skewness, leptokurtosis, and sample size; the nominal error rate is 5%

Control n	Skewness				
	None	Moderate	Severe	Very Severe	Extreme
No leptokurtosis					
5	5.00	5.10	5.38	5.64	5.72
10	4.99	5.07	5.34	5.50	5.54
20	5.01	5.02	5.19	5.34	5.38
30	5.02	5.01	5.17	5.29	5.33
100	5.01	5.01	5.04	5.17	5.19
Moderate leptokurtosis					
5	5.63	5.86	6.20	6.44	6.50
10	5.48	5.58	5.85	6.02	6.08
20	5.21	5.26	5.41	5.56	5.61
30	5.11	5.12	5.27	5.34	5.35
100	4.92	4.88	4.86	4.90	4.91
Severe leptokurtosis					
5		6.47	6.84	7.10	7.16
10	5.86	6.04	6.27	6.39	6.45
20	5.41	5.46	5.56	5.69	5.71
30	5.18	5.18	5.25	5.33	5.39
100	4.70	4.61	4.56	4.58	4.57

Figure 1
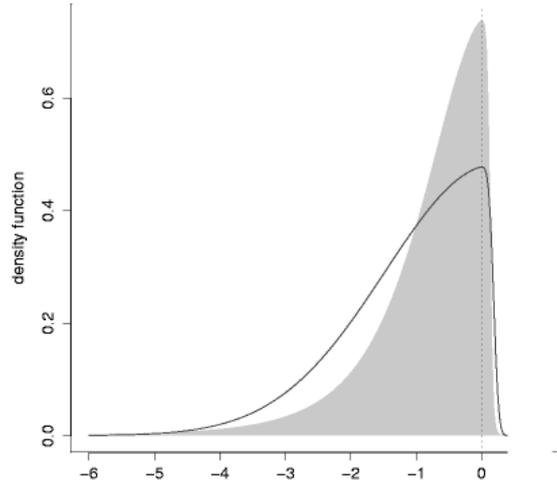
Appendix 1

Bayesian credible intervals for comparing the difference between two cases with differences found in controls

Let $\pi(\sigma^2)$ be the prior distribution for the population variance, σ^2. We adopt Jeffreys' prior, which is the accepted non-informative prior distribution in this context, so $\pi(\sigma^2) \propto 1/\sigma^2$. The data from n controls gives s^2 as the sample variance. The posterior distribution states that the distribution of $(n-1)s^2/\sigma^2$ is a chi-squared distribution on n-1 degrees of freedom; see, for example, Lee (1997), p. 68. We have that a and b are the 0.025 and 0.975 points of a chi-squared distribution on $n-1$ degrees of freedom, so (a,b) is the 95% equal-tailed credible interval for $(n-1)s^2/\sigma^2$. It follows that $(\lambda\sqrt{a}, \lambda\sqrt{b})$ is the 95% equal-tailed credible interval for $\lambda\sqrt{(n-1)s^2/\sigma^2}$, where λ is any positive value. Putting $\lambda = g/\sqrt{2(n-1)s^2}$, we have that (z_1, z_2) is a 95% credible interval for $g/\sqrt{2\sigma^2}$, where z_1 and z_2 are defined in equations (3) and (4). Now $D \sim N(0, 2\sigma^2)$ and $\Pr(D > g)$ is the proportion of pairs of controls whose difference exceeds the difference (g) between the cases. As $\Pr(D > g) = \Pr(Z > g/\sqrt{2\sigma^2})$, where $Z \sim N(0,1)$, we have that $\Phi^{-1}(z_1)$ and $\Phi^{-1}(z_2)$ are the endpoints of the 95% credible interval for that proportion, in agreement with the classical confidence interval.

Turning to the distribution of $D/\sqrt{2s_x^2}$, put $W = D/\sqrt{2s_x^2}$. Then $W \mid \sigma^2 \sim N(0, \sigma^2/s^2)$. As $(n-1)s^2/\sigma^2$ follows a chi-squared distribution on $n-1$ degrees of freedom,

$$f(w) \propto \int_{\sigma^2=0}^{\infty} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{sw}{\sigma}\right)^2 \right\} \right] \cdot \left[\frac{1}{\sigma^{n+1}} \exp\left\{ -\frac{n-1}{2}\left(\frac{s}{\sigma}\right)^2 \right\} \right] d\sigma^2,$$

whence W follows a t-distribution on $n-1$ degrees of freedom (Lee, 1997, pp. 66-67).

The Bayesian point estimate of the proportion of differences between pairs that are more extreme than the cases' difference is

$$\Pr(|D| > |g|) = \Pr(|W| > |g| / \sqrt{2s_x^2}). \tag{11}$$

This gives the same point estimate as classical statistics.

To obtain the Bayesian p-value from testing the hypothesis that the difference between two cases (g) is greater than would be found in a pair of randomly chosen controls, we want $\Pr(|D| > |g|)$. From (11), this gives the same p-value as classical statistics.

Appendix 2.

Sampling from skew-normal and skew-t distributions

The method used to sample from skew-normal distributions is based on work by Azzalini and colleagues (Azzalini & Capitanio, 1999; Azzalini & Dalla Valle, 1996). The starting point for this method is the generation of two independent standard normal variates u_0 and u_1 (u_1 is used to form the X observations and u_0 is used to control the degree of skewness in X). Then u_2 is determined from the formula

$$u_2 = u_0 \rho_{u_0 u_1} + \sqrt{1 - \rho_{u_0 u_1}^2}\, u_1 .$$

The value of $\rho_{u_0 u_1}$ required to introduce the desired degree of skewness, γ_1, can be obtained by algebraic manipulation of Azzalini and Dalla Valle's (1996) formulae for γ_1 to solve for $\rho_{u_0 u_1}$; see Crawford et al., (2006). Then

$$x = \begin{cases} u_2 & \text{if } u_0 \geq 0 \\ -u_2 & \text{otherwise} \end{cases}$$

is an observation from the skew-normal distribution with skewness γ_1.

To sample from the equivalent skew-*t* distribution (i.e., a distribution that is both skew *and* leptokurtic) the above steps are followed by dividing x by $\sqrt{\chi^2 / v}$, where χ^2 is a random draw from a chi-square distribution on v degrees-of-freedom (e.g., $v = 4$ if severe leptokurtosis is required).

To sample from distributions that are leptokurtic only (i.e., not skew) the initial random draws are made from a standard normal distribution, rather than a skew-normal distribution, and the same procedure as just outlined is followed to introduce leptokurtosis; i.e., the observations (*x*s) are divided by $\sqrt{\chi^2 / v}$ (Crawford et al., 2006).

Appendix 3.

Hypothesis test and confidence interval for whether the difference found between two cases is greater than would be found in a pair of controls when there is a vector of covariates.

We have two cases whose observed values are y_1^* and y_2^*, and the values of a $m \times 1$ vector of covariates. Let $\underline{X}$ be an $(m + 1) \times 1$ vector whose first component is 1 (corresponding to the constant term in regression equations) and whose other components give the values of the covariates. Denote the values of $\underline{X}$ for the two cases by $\underline{x}_1^*$ and $\underline{x}_2^*$. There is also a population of controls. We suppose that we pick one control from the set of controls whose covariate vector is $\underline{x}_1^*$ and a second control

from the set of controls whose covariate vector is $\underline{x}_2^*$. The Y values of these controls

are Y_1 and Y_2. We want $P(Y_1 - Y_2 > y_1^* - y_2^*)$.

By assumption,

$$Y = \underline{\beta}'\underline{x} + \varepsilon,$$

where ε is the random error which we assume is normally distributed with a mean of

0 and an unknown variance σ^2. Then

$$Y_1 - Y_2 = \underline{\beta}'(\underline{x}_1^* - \underline{x}_2^*) + \varepsilon_1 - \varepsilon_2.$$

If $\underline{\beta}$ and σ^2 were known, then

$$Y_1 - Y_2 \sim N(\underline{\beta}'(\underline{x}_1^* - \underline{x}_2^*), 2\sigma^2), \tag{12}$$

where ' denotes transpose. This equation forms the basis of the analysis.

Data from n controls give paired values $(\underline{x}_1, y_1), \ldots, (\underline{x}_n, y_n)$, where each $\underline{x}_j$ is

an $(m+1) \times 1$ vector whose first component is 1 and whose other components give the

values of the covariates for the j th control ($j = 1,...,n$). Put $\mathbf{X} = (\underline{x}_1,...,\underline{x}_n)'$ and let $\underline{y}$

be the $n \times 1$ vector of Y values. Then the data estimate of $\underline{\beta}$ is

$$\hat{\underline{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'}\underline{y},$$

and the data estimate of σ is

$$\hat{\sigma}^2 = \frac{1}{n-m-1}(\underline{y} - \mathbf{X}\hat{\underline{\beta}})'(\underline{y} - \mathbf{X}\hat{\underline{\beta}}).$$

We have that

$$\hat{\underline{\beta}}'(\underline{x}_1^* - \underline{x}_2^*) \sim N(\underline{\beta}(\underline{x}_1^* - \underline{x}_2^*), \sigma^2(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*))$$

and

$$Y_1 - Y_2 \sim N(\hat{\underline{\beta}}'(\underline{x}_1^* - \underline{x}_2^*), 2\sigma^2 + \sigma^2(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)).$$

Also, $(n - m - 1)\hat{\sigma}^2 / \sigma^2$ has a chi-square distribution on $(n - m - 1)$ degrees of

freedom. Hence, $P(Y_1 - Y_2 > y_1^* - y_2^*)$ is equal to

$$P(t > \frac{(y_1^* - y_2^*) - \hat{\underline{\beta}}'(\underline{x}_1^* - \underline{x}_2^*)}{\sqrt{2\hat{\sigma}^2 + \hat{\sigma}^2(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)}}) \tag{13}$$

where t has a t-distribution on $(n - m - 1)$ df. This is the point estimate of the required

probability.

The confidence interval is derived from a non-central t-distribution (see

Method section of Study 2). For specified values y_1^* and y_2^*, let

$P^* = \Pr(Y_1 - Y_2 < y_1^* - y_2^*) \times 100$. Let

$$c = \frac{(y_1^* - y_2^*) - \hat{\underline{\beta}}'(\underline{x}_1^* - \underline{x}_2^*)}{\hat{\sigma}\sqrt{2}},$$

and let $c^* = \{(y_1^* - y_2^*) - \underline{\beta}'(\underline{x}_1^* - \underline{x}_2^*)\}/\{\sigma\sqrt{2}\}$. Then c is an estimate of c^*. Also,

$\Pr(Y_1 - Y_2 < y_1^* - y_2^*) = \Pr(Z < c^*)$, so percentage points of a normal distribution

determine P^* from c^*. Thus a $100(1 - \alpha)\%$ confidence interval for c^* will yield the

required confidence interval for P^*. Now,

$$\frac{c\sqrt{2}}{\sqrt{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)}} = \frac{(\underline{\beta} - \hat{\underline{\beta}})'(\underline{x}_1^* - \underline{x}_2^*) + \{y_1^* - y_2^* - \underline{\beta}'(\underline{x}_1^* - \underline{x}_2^*)\}}{\sigma\sqrt{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)}} \cdot \frac{1}{\sqrt{\hat{\sigma}^2 / \sigma^2}},$$

so $c\sqrt{2/\{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)\}}$ has a non-central t-distribution with non-

centrality parameter $\delta = c^*\sqrt{2/\{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)\}}$ and $(n - m - 1)$ df. The

$100(\alpha / 2)\%$ and $100(1 - \alpha / 2)\%$ points of this distribution will depend on the value

of δ. Let δ_L denote the value of δ for which the $100(1 - \alpha / 2)\%$ point is

$c\sqrt{2/\{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X'X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)\}}$. Similarly, let δ_U denote the value of δ for which

the $100(\alpha/2)\%$ point is $c\sqrt{2/\{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X}'\mathbf{X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)\}}$. Then

$(\delta_L\sqrt{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X}'\mathbf{X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)/2},\ \delta_U\sqrt{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X}'\mathbf{X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)/2})$ is a

$100(1-\alpha)\%$ confidence interval for c^*. Hence a $100(1-\alpha)\%$ confidence interval for

P^* is

$(\Pr(Z < \delta_L\sqrt{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X}'\mathbf{X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)/2}),\ \Pr(Z < \delta_U\sqrt{(\underline{x}_1^* - \underline{x}_2^*)'(\mathbf{X}'\mathbf{X})^{-1}(\underline{x}_1^* - \underline{x}_2^*)/2}))$

where $Z \sim N(0,1)$.

As noted in the main body of the text, when the values on all of the covariates

are the same for the two cases this constitutes a special case and the problem is

simplified. The equations set out in Study 1 are applied but the *conditional* standard

deviation ($\hat{\sigma}$) is substituted for the unconditional standard deviation (s), and the

degrees-of-freedom become $n - m - 1$ (rather than $n - 1$, as applies when there are no

covariates) when evaluating t and in setting the confidence limits.

Figure Legend

Figure 1.

Graphical illustration of two of the distributions employed in Study 2; the shaded area

shows the density for a distribution possessing both extreme skewness and severe

leptokurtosis (a skew-*t* distribution), the unshaded line shows the density for the

equivalent distributions with skewness alone (a skew-normal distribution). Note that

the skew-*t* and skew-normal distributions have been scaled to have a common

variance of 1 so they can be meaningfully compared.