Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Research report

Comparing a single case to a control sample: Testing for neuropsychological deficits and dissociations in the presence of covariates

John R. Crawford[a,*], *Paul H. Garthwaite*[b] and *Kevin Ryan*[a]

[a] School of Psychology, University of Aberdeen, UK
[b] Department of Mathematics and Statistics, The Open University, Milton Keynes, UK

ARTICLE INFO

Article history:
Received 7 December 2010
Reviewed 9 February 2011
Revised 16 February 2011
Accepted 22 February 2011
Action editor Sergio Della Sala
Published online 5 March 2011

Keywords:
Single case
Covariates
Statistical methods

ABSTRACT

Existing inferential methods of testing for a deficit or dissociation in the single case are extended to allow researchers to control for the effects of covariates. The new (Bayesian) methods provide a significance test, point and interval estimates of the effect size for the difference between the case and controls, and point and interval estimates of the abnormality of a case's score, or standardized score difference. The methods have a wide range of potential applications, e.g., they can provide a means of increasing the statistical power to detect deficits or dissociations, or can be used to test whether differences between a case and controls survive partialling out the effects of potential confounding variables. The methods are implemented in a series of computer programs for PCs (these can be downloaded from www.abdn.ac.uk/~psy086/dept/Single_Case_Covariates.htm). Illustrative examples of the methods are provided.

1. Introduction

The focus of the present paper is on single-case studies in neuropsychology that employ the case–controls design; i.e., studies in which inferences concerning the cognitive performance of a single-case are made by comparing the case to a sample of healthy controls. Crawford, Garthwaite, Howell and colleagues (Crawford and Howell, 1998a; Crawford and Garthwaite, 2002, 2005, 2007a) have developed a set of classical and Bayesian methods for this design. These methods allow researchers to test for a deficit in the single-case, and to test whether the standardized difference between a case's score on two tasks differs from the differences observed in controls; the latter methods are useful in testing for dissociations (Crawford et al., 2003; see Crawford and Garthwaite, 2007a for further details).

Although these methods are sound and useful they do not currently offer solutions to some of the more complex issues faced by the single-case researcher. The aim of the present paper is to extend upon these existing methods to allow researchers to test for deficits or dissociations while controlling for the effects of covariates. That is, the aim is to develop a Bayesian Test for a Deficit allowing for Covariates (BTD-Cov) and a Bayesian Standardized Difference Test allowing for Covariates (BSDT-Cov).

These new methods can serve the broad purpose of allowing researchers to control for nuisance variables when comparing a case to controls. When a healthy control sample

* *Corresponding author*. School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 2UB, UK.
E-mail address: j.crawford@abdn.ac.uk (J.R. Crawford).

0010-9452/$ – see front matter © 2011 Elsevier Srl. All rights reserved.
doi:10.1016/j.cortex.2011.02.017

is recruited to match a single case, the controls are intended to represent the case minus the lesion. The controls should therefore be closely matched on as many potentially important attribute variables as possible. For example, performance on many neuropsychological tasks is moderately-to-highly correlated with age and educational level and, for some cognitive functions, gender may also exert an influence on performance. In practice it can be difficult and time-consuming to recruit an adequately matched sample, particularly if a researcher wants to match the controls on cognitive variables as well as on demographic/attribute variables. Indeed it is obvious from even a casual inspection of published single-case studies that matching is often sub-optimal. One could adopt a pious attitude to these difficulties: the researcher should simply work harder to find more suitable controls. However, the methods developed in the present paper offer a practical alternative when such attempts have failed. Note also that it is not uncommon for researchers to use a single control group as a reference sample for the comparison of the performance of multiple single cases; the methods set out here can play a particularly useful role in such comparisons (see the Discussion section for a fuller treatment of this issue).

The methods can also be used for two, more targeted, purposes: First, they can be used to increase the power to detect a deficit or dissociation in a single case by controlling for the effects of a suppressor variable. A suppressor variable in this context can be defined as any variable that obscures or attenuates the difference between the case and controls. The issue of the statistical power of inferential methods for the single-case has been largely neglected (Crawford and Garthwaite, 2006b). However, it is clear that statistical power will typically be lower than that found in group studies in neuropsychology: a single case, rather than a clinical sample, is compared to a control group and, moreover, the control groups typically employed are modest in size. As sample size is an important determinant of power, it can be seen that power will be low unless effects are very large (neurological damage can have dramatic consequences on cognition and so of course large effects are often there to be detected). Therefore, anything that can increase statistical power to detect deficits or dissociations should be encouraged (provided that it does not achieve this at the cost of failing to control the Type I error rate).

Furthermore, the methods can be used to test whether a difference in task performance between a single case and controls can be explained by the effects of a third variable. That is, in contrast to the foregoing potential application, the methods can also be used to test whether differences survive controlling for the effects of covariates. For example, a difference between a case and controls on a task of interest may be attributable to a general slowing of information processing rather than impairment of the putative specific function measured by the task. This possibility could be approached by testing for a deficit on the task while controlling for a measure of processing speed.

1.1. *Testing for deficits and dissociations in the presence of covariates: desirable statistics*

The statistical methods developed previously by Crawford, Garthwaite, Howell and colleagues provide a comprehensive set of statistics. For example, when testing for the presence of a deficit the methods provide a significance test – if the *p* value from this test (Crawford and Howell, 1998a) is below the specified value for alpha (normally .05) then the researcher can reject the null hypothesis that the case's score is an observation from the scores in the control population. The *p* value from this test is also the optimal point estimate of the abnormality of the case's score (i.e., it is the estimated proportion of controls that will obtain a score lower than the case; multiplying this figure by 100 gives the percentage of controls expected to obtain a lower score). Thus, if the *p* value is .0240, then only 2.4% of controls are expected to obtain a lower score. For a mathematical proof of this dual role for the *p* value see Crawford and Garthwaite (2006b).

Crawford and Garthwaite (2002) have developed methods that supplement the point estimate of the abnormality of the case's score with an interval estimate of the same quantity. Such an interval estimate is in keeping with the contemporary emphasis in both psychology and statistics on the provision of confidence intervals. Finally, Crawford et al. (2010) have emphasized the importance of reporting point and interval for effect sizes in single-case studies and provided methods for achieving this.

Fortunately it will be possible to provide the direct equivalents of all of these statistics when controlling for the effects of covariates. The meaning of these statistics will remain broadly the same but with some important differences. That is, the significance test will still test if we can reject the null hypothesis that the case's score is an observation from the scores in the control population, but the control population is redefined as controls having the same value(s) on the covariate(s) as the case. Similarly, the point estimate of the abnormality of the case's score on the task of interest is the percentage of controls, with the same value(s) on the covariate(s), that are expected to obtain a lower score than the case. To develop these methods we extend Crawford and Garthwaite's (2007a) Bayesian approach to the analysis of the single case.

2. Method

2.1. *Bayesian method of testing for a deficit or dissociation controlling for covariates*

We assume that, conditional on the values of the covariates, in the control population the task(s) of interest follow either a normal distribution (when there is only one task of interest, i.e., when we wish to test for a deficit) or a bivariate normal distribution (when there are two tasks of interest, i.e., when we wish to test for a dissociation); see later section for a discussion of these assumptions. No assumptions are made about the distribution of the covariates and, indeed, their values need not be random as will happen, for example, with some experimental designs. We have a control sample of n individuals with scores on k tasks and values for m covariates from which to estimate $\mathbf{B} = (\underline{\beta}_1, \ldots, \underline{\beta}_k)$, where $\underline{\beta}_i$ is the vector of $m+1$ regression coefficients that relates the ith task to the covariates in the control population, and Σ, a $k \times k$ matrix of the control population covariances for the scores on tasks, conditional on the covariates (the first component of each $\underline{\beta}_i$ is the constant term). It is assumed that Σ does not vary as the values of the covariates change. When there are two tasks we may write Σ as

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

where ${\sigma_1}^2$ and ${\sigma_2}^2$ are the variances of a control's scores on the two tasks, conditional on the values of the covariates, and ρ is their correlation.

If we wish to test whether a case's performance on a single task differs from that of controls (i.e., if we wish to test for a deficit), while allowing for m covariates, then $k = 1$. If we are comparing a case's standardized difference between two tasks to the standardized differences in controls (i.e., if we are testing for a dissociation) while allowing for m covariates then $k = 2$. As noted, we will identify these two methods as BTD-Cov and BSDT-Cov respectively.

2.2. *Two forms of non-informative prior distribution*

The control sample data are combined with a non-informative prior distribution (i.e., a prior distribution that assumes no prior knowledge or data) to obtain the posterior distribution of (**B**, Σ). The posterior distribution is used to estimate parameters and make inferences or decisions. Note that a common approach in Bayesian statistics is to employ Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution (Garthwaite et al., 2002). The problems tackled here however are reasonably tractable, and, while we use Monte Carlo methods, we do not need to run any Markov chains.

As the approach used here is Bayesian, in principle, it is not designed to have frequentist properties. However, having such properties seems desirable. For example, a Bayesian interval estimate is termed a credible interval and it is preferable that 95% credible intervals (like 95% confidence intervals) should contain the true value of the quantity of interest 95% of the time in repeated sampling.

In testing for a dissociation we will consider two forms of non-informative prior distribution, one is a standard method while the other has better frequentist properties. We do not consider two different priors when testing for a deficit, as the standard non-informative prior has good frequentist properties. This prior is $f(\mathbf{B}, \Sigma) \propto \Sigma^{-1}$ (the symbol $\propto$ is read as "is proportional to") and we use it when there is a single task ($k = 1$). The posterior distribution of Σ^{-1} takes the form of a Wishart distribution on $n - m - 1$ df (Tiao and Zellner, 1964). The Wishart distribution is a multivariate generalisation of the chi-square distribution. (For sampling from a univariate normal distribution with variance σ^2, the posterior distribution of the inverse of σ^2 is a chi-square distribution.) In all the posterior distributions formed in this paper, the conditional distribution of **B** given Σ follows a matrix normal (MXN) distribution (Tiao and Zellner, 1964).

Choosing a non-informative prior distribution to test for a dissociation is trickier. Here there are two tasks ($k = 2$) and the traditional choice (Jeffreys' prior) is $f(\mathbf{B}, \Sigma^{-1}) \propto |\Sigma|$. This is one of the two prior distributions that we consider for $k = 2$ and we will refer to it as the "standard theory" prior. It is analogous to the prior used in Crawford and Garthwaite (2007a) for comparisons of a case and controls in the *absence* of covariates. With this prior, the posterior distribution of Σ^{-1} takes the form of a Wishart distribution on $n - m$ df (Tiao and Zellner, 1964). This posterior distribution has perfect frequentist properties for estimating σ_1 or σ_2, for example, but is less good for other quantities, such as estimating ρ or the size of a dissociation (Berger and Sun, 2008). This illustrates that a non-informative prior distribution may be perfect for estimating some quantities but less good for others.

Berger and Sun (2008) focused on the current case of interest, $k = 2$. They examined the frequentist performance of a variety of non-informative priors in estimating many different quantities. They did not consider the quantity needed to test for a dissociation, but they recommend using

$$f(\mathbf{B}, \Sigma) \propto \frac{1}{\sigma_1\sigma_2(1-\rho^2)} \tag{1}$$

as a general purpose prior when the appropriate prior for the quantity of interest is unknown. They considered the distribution in (1) because it is a "one-at-a-time reference prior", arguing that reference prior theory (Bernardo, 1979) is probably the most successful technique for deriving non-informative prior distributions and that, based on experience and numerous examples, the best reference priors come from the class of *one-at-a-time* reference priors. The basis of their recommendation, though, was a large simulation study in which this prior gave a good general performance for a broad range of quantities.

In Monte Carlo simulations, we examined the posterior distributions that resulted from the prior in equation (1), and found that credible intervals to test for a dissociation tended to be too narrow, so that hypothesis tests of a dissociation tended to be liberal. However, if we treated the sample estimate of Σ as if it were based on $n - 1$ observations, rather than n, then further simulations showed that the posterior distribution had good frequentist properties. These simulations are reported in Garthwaite and Crawford (in preparation). We adopt this approach for our second method of analysis. That is, we take the prior recommended by Berger and Sun and combine it with the sample estimates of **B** and Σ while treating $n - 1$ as the sample size. There is a prior distribution that would give this posterior distribution without modifying the sample size, and we refer to that as the "calibrated prior". We strongly recommend the use of this latter prior, although the computer programs that accompany this paper also offer the "standard theory" prior as an option, as some users may believe that frequentist properties are irrelevant when performing a Bayesian analysis, and prefer to use the prior distribution given by Jeffreys (1961).

In the next section we set out the procedure for conducting the Bayesian analysis. Part of this procedure yields point and interval estimates of effect size for the difference between a case's score and controls (which we denote as z_{CCC}) and a point estimate of the effect size for the difference between a case's standardized difference and controls (which we denote as z_{DCCC}). The subscripts for these effect size indexes may appear unnecessarily lengthy but they are used to differentiate them from the equivalent indexes of effect size (z_{CC} and z_{DCC}) developed by Crawford et al. (2010) for use when there are no covariates. The *CCC* subscript identifies that the effect size allows for a **C**ovariate and is for use in a **C**ase–**C**ontrols design; the *D* subscript denotes that an index is concerned with a (standardized) **D**ifference between scores.

2.3. *Generation of the required statistics*

The following procedure yields the statistics required:

1. Generate estimates of **B** and Σ. When testing for a deficit ($k = 1$) or using the standard theory prior to test for a dissociation ($k = 2$), the estimates of Σ are obtained by random sampling from an inverse-Wishart distribution on $n - m + k - 2$ degrees-of-freedom (df). For details of how to sample from an inverse-Wishart distribution, including information on suitable algorithms, see Crawford and Garthwaite (2007a). These random draws are then used, in combination with random draws from a multivariate normal (MVN) distribution, to obtain estimates of **B**. The general procedure is the same when using the calibrated prior except that we sample from an inverse-Wishart on $n - m - 2$ df and apply an accept/reject algorithm to the random draws obtained. Details of the accept/reject procedure are provided in Appendix 1. As some readers will be familiar with Bayesian regression (and others will have little interest in the technical details) we have consigned the procedural details (which are the same regardless of which prior is employed) to Appendix 2. Let $\hat{\mathbf{B}}_{(i)}$ and $\hat{\Sigma}_{(i)}$ identify the ith set of estimates.
2. From $\hat{\mathbf{B}}_{(i)}$ and $\hat{\Sigma}_{(i)}$, calculate the conditional means (denote these as $\underline{\hat{\mu}}_{(i)}$) and standard deviations (denote these as $\underline{\hat{s}}_{(i)}$) for the *k* tasks of interest given the values obtained by the case on the *m* covariates. At this point the methods diverge depending on whether the aim is to make inferences concerning (*a*) a case's score on a single task (BTD-Cov), or (*b*) a case's standardized difference between two tasks (BSDT-Cov).

3a. For problem (*a*), in which we only have one task of interest, subtract the conditional mean, $\underline{\hat{\mu}}_{(i)}$, from the case's score (y) and divide by the conditional standard deviation, $\underline{\hat{s}}_{(i)}$ (as there is only one task these vectors are of length one for this problem); denote the resultant quantity as $\hat{z}_{CCC(i)}$. That is

$$\hat{z}_{CCC(i)} = \frac{y - \underline{\hat{\mu}}_{(i)}}{\underline{\hat{s}}_{(i)}} \tag{2}$$

3b. For (*b*) convert the case's scores on Tasks Y_1 and Y_2 to z scores using the estimated conditional means and standard deviations for controls. We will denote these conditional means as $\hat{\mu}_{1(i)}$ and $\hat{\mu}_{2(i)}$, and the standard deviations as $\hat{s}_{1(i)}$ and $\hat{s}_{2(i)}$. Then divide the difference between the z scores by the estimated standard deviation of their difference and denote the result as $\hat{z}_{DCCC(i)}$,

$$\hat{z}_{DCCC(i)} = \frac{\left(\frac{y_1 - \hat{\mu}_{1(i)}}{\hat{s}_{1(i)}}\right) - \left(\frac{y_2 - \hat{\mu}_{2(i)}}{\hat{s}_{2(i)}}\right)}{\sqrt{2 - 2\hat{\rho}_{12(i)}}} \tag{3}$$

Note that the denominator requires the estimated conditional correlation between Tasks Y_1 and Y_2 ($\hat{\rho}_{12(i)}$). This is obtained from the estimated conditional variances and covariances, i.e.,

$$\hat{\rho}_{12(i)} = \frac{\hat{s}_{12(i)}}{\sqrt{\hat{s}^2_{1(i)}\hat{s}^2_{2(i)}}} \tag{4}$$

where the elements on the right-hand side of (4) are obtained from $\hat{\Sigma}_{(i)}$.

4. For problem *a*, where *p* is the proportion of controls that will obtain a lower score than the case, we find the tail-area of a standard normal distribution that is less than $\hat{z}_{CCC(i)}$. For problem *b*, where *p* is the proportion of controls that will obtain a larger standardized difference, we find the tail-area of a standard normal distribution that is greater than $\hat{z}_{DCCC(i)}$. For both problems, the tail-area is an estimate of *p*; call this quantity $\hat{p}_{(i)}$.
5. The remainder of the procedure is the same in both problems. (We distinguish between them by referring to the method of testing for a deficit first, then identifying the equivalent step for a standardized difference in brackets.) Steps 1 to 4 are repeated a large number of times; for the present problems we chose to perform one million iterations. Then the average value of $\hat{p}_{(1)}, \ldots, \hat{p}_{(1000000)}$ is the point estimate of *p*. (It is the Bayesian posterior mean of *p*.) This probability is used for significance testing: it tests whether we can reject the null hypothesis that the case's score (or standardized difference) is an observation from the population of scores (or standardized differences) for controls having the same values on the covariate(s) as the case. Multiplying this *p* value by 100 gives the point estimate of the percentage of controls that are expected to obtain a lower score (or larger standardized difference). To obtain a 95% Bayesian credible interval for this quantity, we find the 25000th smallest $\hat{p}_{(i)}$ and the 25000th largest $\hat{p}_{(i)}$. Multiply these values by 100 so that they are percentages rather than proportions and denote them as p_l and p_u. Then the 95% Bayesian credible interval is (p_l, p_u). That is, it is a 95% interval estimate of the percentage of controls obtaining a lower score than the case (or an estimate of the proportion of controls obtaining a larger standardized difference than the case).
6. To obtain a 95% interval estimate for the effect size for the difference between the case and controls we find the 25000th smallest $\hat{z}_{CCC(i)}$ (or $\hat{z}_{DCCC(i)}$) and the 25000th largest $\hat{z}_{CCC(i)}$ (or $\hat{z}_{DCCC(i)}$) and denote them as $\hat{z}_{CCC,l}$ (or $\hat{z}_{DCCC,l}$) and $\hat{z}_{CCC,u}$ (or $\hat{z}_{DCCC,u}$). Then, for problem *a*, the 95% credible interval for the effect size is $(\hat{z}_{CCC,l}, \hat{z}_{CCC,u})$; for problem *b*, the 95% credible interval is $(\hat{z}_{DCCC,l}, \hat{z}_{DCCC,u})$.

By following the procedure set out in these preceding steps we have all the statistics required for both problems with the exception that we do not yet have a *point* estimate of the effect sizes for the difference between a case's score and controls (z_{CCC}) and a point estimate of the effect size for the difference between a case's standardized difference and controls (z_{DCCC}).

These point estimates are easily obtained. For z_{CCC} we use equation (2) but substitute the estimated conditional mean and standard deviation *calculated directly from the control sample statistics* rather than using the estimates of the control population conditional means and standard deviations obtained by making random draws from the posterior distribution. The same applies to the effect size for a standardized difference (z_{DCCC}). That is, equation (3) is used, substituting the observed control sample statistics for the random draws from the posterior distribution.

2.4. Use of a calibrated prior when testing for a standardized difference (i.e., dissociation) in the absence of covariates

As noted above, in developing the method of testing for a standardized difference (i.e., dissociation) between tasks in the *presence* of covariates, we adopted the use of calibrated prior so that the results of the Bayesian hypothesis test had frequentist properties. This would also be a desirable feature when testing for a dissociation in the *absence* of covariates. Therefore we have added this as an option to the BSDT developed by Crawford and Garthwaite (2007a) and Crawford et al. (2010). The implementation of this calibrated prior is very similar to that used when there are covariates; the details are presented in Appendix 1. As was the case for the BSDT-Cov procedure, Monte Carlo simulations of the performance of the BSDT using the calibrated prior showed it yielded very satisfactory results. For example, when the number of controls was only six, the coverage of 95% credible intervals was greater than 94.5% in all the situations we examined (and hence the method was never markedly liberal) and seldom above 97% (Garthwaite and Crawford, in preparation). We therefore recommend it over the use of the standard theory prior used in Crawford and Garthwaite (2007a). Offering this form of prior for the BSDT is also important for pedagogic purposes as the best way to illustrate the effects of controlling for covariates is to directly compare results obtained with and without covariates.

3. Results and discussion

The best way to illustrate the present methods is through the use of concrete examples. As noted, it is particularly informative to contrast the results obtained to those obtained when covariates are ignored. In the following sections some examples of testing for a deficit are provided before moving on to considering tests for dissociations.

3.1. A worked example of testing for a deficit in the presence of covariates: increasing statistical power

Suppose that a single case obtain a score of 78 on a task of interest and that the mean score on this task in a sample of 18 controls is 100, with a standard deviation of 15. In the *absence* of a covariate we can test whether the case's score is significantly lower than controls using either the classical or Bayesian methods developed by Crawford and colleagues since, as noted, these give equivalent results for this problem. Applying Crawford and Howell's (1998a) test, the case's score is not significantly lower than controls: $t = -1.428$, p (one-tailed) $= .0858$. It is estimated that 8.58% of the control population would exhibit lower scores than the case and the 95% confidence interval on this percentage is from 1.66% to 21.66%.

Now suppose that we wish to compare the case's performance to controls allowing for a single covariate, years of education, using the BTD-Cov method. As noted, most neuropsychological tests correlate with years of education so this is a realistic example and will arise commonly in practice. Suppose that the case has 13 years of education. Suppose also that the mean years of education in the control sample is 13.0, with a standard deviation of 3.0, and that the correlation between task performance and years of education in the controls is .65.

Casual inspection of this scenario might suggest that it is not a promising one for the use of the present methods: the case's years of education is exactly equal to the mean years of education of controls. Therefore, as the case appears to be "matched" to the control group for education, allowing for the effects of education in comparing the scores of the case and controls may seem to add nothing. Indeed, given that one df 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 deficit in the case. However, this view would be erroneous.

In the example the expected score on the task for individuals with 13 years of education (i.e., the conditional mean) is, necessarily, equal to the unconditional mean score of controls (100). However, the conditional standard deviation for the task is 11.75, which is smaller than the unconditional standard deviation for the task (15). Thus the uncertainty attached to the expected score has been reduced because education is a predictor of task performance. As a result the case's actual score obtained on testing is more extreme and the null hypothesis that the case's score is an observation from the control population having 13 years of education can be rejected. The one-tailed p value is .0436, compared to the value obtained in the absence of the covariate ($p = .0858$). The effect size for the difference between case and controls controlling for education ($z_{CCC} = -1.872$) is also larger than that obtained without the covariate ($z_{CC} = -1.467$).

As noted, in addition to providing a significance test, the one-tailed p value is also the proportion of controls expected to obtain a lower score than the case. Multiplying this proportion by 100 to express it as a percentage, it can be seen that 4.36% of controls are expected to obtain a lower score. Thus, on allowing for the case's years of education, the case's score is estimated to be more unusual than the estimate provided in the absence of the covariate (8.58%).

In this example the case's years of education was at the mean of controls. If the case was above the control mean for years of education then even greater power will be achieved by inclusion of the covariate: not only will the uncertainty over the expected score be reduced (i.e., as in the original example, the conditional standard deviation will be reduced) but also the conditional mean (i.e., the expected score) will be higher. As a result the (low) score actually obtained by the case will be even more unusual. As a concrete example suppose that all values were the same as those used in the first example except that the case had 14.5 years of education. Then the (one-tailed) p value is .02102 and the effect size (z_{CCC}) is -2.287. Note that, even when the case scores *below* the mean of controls on the covariate, power can potentially still be higher with inclusion of the covariate if the difference is modest. For example, if everything else were the same as in the original example but the case had 12.5 years of education, then the *p* value from the hypothesis test would be .05561 (not significant, but still smaller than the *p* value of .0858 obtained without the covariate, despite the loss of one df).

The first example serves as a general illustration of the use of the present methods to increase power through controlling for the effects of a suppressor variable. However, it also highlights a specific issue surrounding the matching of controls to cases. It is rare in single-case studies for each control to be exactly matched to the case in terms of demographic variables. Rather, as in the example, the emphasis is usually on matching the *mean* of controls to the case. The example demonstrates that a researcher should consider using demographic variables as covariates even when the case's values are at, or close to, the means of controls.

Finally, consider the effect of a researcher recruiting controls that all individually and exactly matched the case for years of education. In these circumstances there would be no advantage to using education as a covariate (indeed it would be impossible as years of education would have zero variance). Application of Crawford & Howell's test with this second control sample would be expected to yield results that were similar to those obtained for the original control sample when the data were analyzed using years of education as a covariate. The control standard deviation for the task of interest would be expected to be smaller than that obtained in the original, more heterogeneous, control sample, thus rendering the case's score more extreme (indeed it would be expected that the standard deviation would be close to the *conditional* standard deviation obtained for the original control sample when controlling for education).

3.2. *A second worked example: testing whether effects survive controlling for a covariate*

In the foregoing scenario the inclusion of a covariate served to increase the power to detect a difference between a case and controls. We turn now to the second situation in which researchers may wish to control for a covariate, namely when the aim is to examine whether observed differences in task performance between a case and controls could be attributed to the effects of a confounding variable.

Suppose that a case obtains a score of 72 on a task of interest and that the mean and standard deviation in a sample of 16 controls is 100 and 15 respectively. Applying Crawford & Howell's (1998a) test reveals that the case's score is significantly below that of controls: $t = -1.940$, *p* (one-tailed) = .0357. Now suppose that both the case and controls had been administered a measure of general processing speed and that the case obtained a score of 32 on this task whilst the mean and standard deviation of controls was 50 and 10. Finally suppose that the correlation between the task of interest and the processing speed task in controls is .70. It can be seen that performance on the task of interest is not independent of processing speed and that the case is substantially below controls on processing speed.

Application of the present methods reveals that the case's score on the task of interest is no longer significantly poorer than controls when controlling for differences in general processing speed, *p* (one-tailed) = .23996. Moreover, the effect size for this difference (z_{CCC}) is only −.821 (95% credible interval = −1.885 to .270). This is considerably smaller than the effect size obtained in the absence of the covariate ($z_{CC} = -1.940$). Finally, the case's score is substantially less extreme after controlling for processing speed: it is estimated that 24.00% of the control population with processing speed scores equal to those of the case would obtain a lower score on the task of interest (95% CI = 2.97 to 60.64%), compared to 3.57% of the overall control population.

3.3. *Comparing a case's standardized difference to those of controls (i.e., testing for a dissociation)*

When the aim is to test for a dissociation the situation is more complex than those outlined earlier. There are now two tasks of interest and they may differ substantially in their relationships with the covariate(s). It can be hard to predict the influence of covariates on the outcome of testing for a dissociation and this underlines the need for formal methods such as those developed here.

To illustrate the method of testing for a dissociation in the presence of covariates (BSDT-Cov), suppose that a researcher wants to examine whether a case exhibits a dissociation in their performance on two tasks, Tasks Y_1 and Y_2, and that a sample of 20 controls obtain a mean of 100 with a standard deviation of 15 on Task Y_1, and a mean of 80 with standard deviation of 10 on Task Y_2. Suppose also that the correlation between tasks in controls is .72 and that the case obtains scores of 70 and 78 on Tasks Y_1 and Y_2 respectively. In this example it can be seen that, as usually holds, the two tasks have different means and standard deviations in the control population and therefore the case's scores need to be standardized in order to compare them meaningfully (Crawford et al., 2009a). In this example the case's scores on Tasks Y_1 and Y_2 are −2.000 and −.200 when expressed as z scores (performance on Task Y_1 is very poor whereas performance on Task Y_2 is only marginally below the control mean).

As in the previous examples, it is useful to analyze these data without allowing for covariates before examining the effects of including the covariate. For this problem the appropriate test is the BSDT (Crawford & Garthwaite, 2007a). Here we apply this test but use the calibrated prior version developed in the present paper because, (a) we consider it desirable that the test has frequentist properties, and (b) the calibrated prior is used when controlling for covariates and so results can be compared directly.

Applying the BSDT the difference between the case's scores is statistically significant (two-tailed $p = .03807$) and it is estimated that only 1.903% of the control population would exhibit a discrepancy between the tasks that was more extreme and in the same direction as that observed for the case (95% CI = .039 to 8.827%).

Now suppose that a researcher is concerned that the case's poorer score on Task Y_1 may be attributable to the task placing higher demands on speeded processing than Task Y_2. Further suppose that the mean score for controls on a speed-of-processing task is 50 with a standard deviation of 10 and that this task has a correlation of .52 with Task Y_1 and .12 with Task Y_2. Finally, suppose the case's processing speed score was 30.

Applying the BSDT-Cov procedure (using the calibrated prior) for testing for a standardized difference in the presence of covariates the two-tailed probability is .1625. That is, the difference is not statistically significant: we cannot reject the null hypothesis that the case's standardized difference is an observation from the population of differences in controls. It is estimated that 8.124% of controls with the same value on the covariate would exhibit a larger difference in the same direction as the case (95% CI = .251 to 32.114%). The effect size for the difference between the case' scores after allowing for the covariate (z_{DCCC}) is −1.737 (95% CI = −2.805 to −.465) compared to the effect size (z_{DCC}) of −2.405 (95% CI = −3.359 to −1.351) without the covariate.

In this first example, application of the BSDT indicated a dissociation between the case's performance on Tasks Y_1 and Y_2. However, the conclusion that the case exhibits a dissociation has to be strongly tempered by the finding that the difference between scores is not significant or very unusual when the covariate is taken into account.

For our second example, suppose that the values for controls and the case on Tasks Y_1 and Y_2 are the same as the first example. However, in this new example suppose that a researcher wishes to test for a dissociation controlling for years of education and suppose that (as in the example of testing for a deficit) the mean years of education for the controls is 13.0, with a standard deviation of 3.0, and the case has 13 years of education (they are at the mean of controls). Finally, suppose that years of education has a correlation of .55 with Task Y_1 (the task on which the case obtained a very poor score) and a correlation of .12 with Task Y_2 (the task on which performance was only marginally below the control mean).

Analyzing these data using BSDT-Cov reveals that the dissociation is now much more striking than it appeared when education was ignored. The two-tailed probability is .00788 (compared to .03807 with no covariates) and the effect size for the difference (z_{DCCC}) is −3.375 (compared to a z_{DCC} of −2.405 with no covariates). It is estimated that only .394% of the control population having the same number of years of education as the case would exhibit a discrepancy between the tasks that was more extreme and in the same direction as that observed for the case, 95% CI = .000 to 2.825% (the lower limit appears as zero because of rounding to three decimal places).

For this present example it is easy to see why there is a substantial difference in the results obtained with and without the covariate: (a) education is moderately correlated with performance on the impaired task, (b) the case has an average years of education (and therefore higher than many in the control population), (c) therefore the case's score is in fact more impaired than it appears when education is ignored, and as a result (d) the difference in the case's relative standing on the two tasks is more extreme.

A less immediately intuitive result is obtained if we analyze this example with one change only: let us reverse the correlations of education with Tasks Y_1 and Y_2 so that the moderate correlation is observed for Task Y_2, and the low correlation for Task Y_1. For this analysis we also find the difference between the case's scores is more extreme than without the covariate: the two-tailed probability is .02504, despite the loss of one df, and the effect size (z_{DCCC}) is −2.731. Clearly the dissociation is not nearly as striking in this example compared to the last example but is still more marked than in the absence of the covariate. This occurs because when one task (either Y_1 or Y_2) has a high correlation with the covariate while the other task has a low correlation with it, then adjusting for the covariate increases the correlation between Y_1 and Y_2. This results in a smaller denominator in equation (3).

The only situation in which the results obtained with inclusion of a covariate will be very similar to those obtained without covariates is when the covariate's correlations with the tasks of interest are very similar. Indeed, if there is only one covariate and its correlations with each of the two tasks are the same, then adjusting for the covariate will have almost no affect on inferences about the dissociation: the only difference is the loss of one df to allow for the covariate. This is proved in Appendix 4. For example, using the same data as the last examples, if the correlations between the tasks and education were both .55, then the effect size (z_{DCCC}) is identical to that obtained without the covariate (−2.405) and the two-tailed probability is similar (if a little higher because of the loss of one df; $p = .04466$ *vs* .03807).

When there are m covariates ($m > 1$) the equivalent result is the following: suppose we adjust for $m - 1$ of the covariates. If the partial (adjusted) correlation between Y_1 and the remaining covariate is the same as the partial correlation between Y_2 and the remaining covariate, then additionally adjusting for the remaining covariate will have no additional effect on the dissociation, other than reducing the df by one.

A complex set of factors determine whether a case will exhibit a dissociation, particularly when allowing for covariates. It is worth stressing that these complexities do not stem from the methods developed here but reflect the real world complexities of such problems. The methods simply provide researchers with the potential to deal with these complexities rather than ignore them.

The computer programs that accompany this paper (see a later section for details) can test for deficits and dissociations using summary data for the controls as input. This means that they can serve as a convenient learning tool: an intuitive appreciation of the factors that influence the results of testing for a dissociation can quickly and easily be gained by entering hypothetical data and observing the effects of modifying the inputs (e.g., by varying the correlations between the tasks and the covariates, varying the extremity of the hypothetical case's score on a covariate etc.).

3.4. *Statistical assumptions*

It is worth noting that the methods do not assume a multivariate normal distribution for the tasks and covariates. Rather all that is required is that the *conditional distributions* for the tasks of interest are normal (when there is only one task of interest, i.e., when testing for a deficit) or bivariate normal (when there are two tasks, i.e., when testing for a dissociation).

These latter assumptions (which also apply to classical regression) are considerably less restrictive and mean that the covariates can, in principle, have any distribution. This is fortunate as the assumption of a normal distribution for a single covariate, or multivariate normal distribution for a set of covariates, is unlikely to be met in some potential practical applications. For example, researchers may wish to control for the effects of dichotomous variables (e.g., gender, handedness, etc.) that clearly do not follow a normal distribution, or they may have evidence that a continuous covariate is skewed. (When a categorical covariate is dichotomous, e.g., gender, then a single dummy variable should be created taking values of 0 and 1; in general if there are j categories, then $j-1$ dummy variables are required to code category membership.)

If the unconditional (marginal distributions) of the task or tasks of interest are not normal, then the conditional distribution(s) will not be normal. It is not uncommon, for example, for control data on neuropsychological tasks to be negatively skewed because the tasks largely lie within the competence of most controls. This could lead to inflation of the Type I error rate (in the present context a Type I error would arise if we incorrectly concluded that a case's score, or score difference, was not an observation from the control population).

Fortunately, empirical examinations of the effect of non-normality on Type I error rates have shown that the case–control methods developed by Crawford, Garthwaite and colleagues, of which the present methods are an extension, are surprisingly robust even in the face of severe departures from normality (Crawford and Garthwaite, 2005; Crawford et al., 2006; Crawford et al., 2011). However, it is clearly best to avoid markedly non-normal data if possible. One solution is to tackle the issue from the outset by selecting or developing tasks that are not subject to ceiling or near ceiling performance in controls (e.g., by upping the task difficulty of items). Second, the effects of departures from normality on Type I errors become attenuated with increasing sample size so the recruitment of as large a control sample as is practical is to be encouraged (as noted, this would have the additional positive effect of increasing statistical power). Finally, when the tasks depart markedly from a normal distribution, researchers could attempt to find a normalizing transformation such as a Box-Cox power transformation (note that, although the search for a normalizing transformation should be conducted using the control data only, it is crucial that the selected transformation is then applied to the score of the case before running the analysis); see Crawford et al. (2006) for further details.

3.5. *Comparison of the present Bayesian methods with classical statistical methods*

The methods developed here are Bayesian and it is worth briefly discussing potential alternative methods based on classical statistics. The Bayesian method of testing for a deficit (i.e., problem *a*) in the presence of covariates has a direct classical equivalent. Crawford and Garthwaite (2006a, 2007b) developed classical methods for drawing inferences concerning the discrepancy between an individual's score and their score predicted by a regression equation built using a control sample; see also Crawford and Howell (1998b). The significance test provided by these methods will give the same results as those obtained here using Bayesian methods, albeit (a) in a much less convenient form for present purposes, and (b) the classical method does not offer the full range of statistics provided by the Bayesian method (e.g., they do not provide interval estimates of effect sizes).

To illustrate this convergence the data used in the first worked example (the case with a score of 78 on a task of interest and 13 years of education etc.) was processed to turn it into a form suitable as the input for Crawford and Garthwaite's (2006a) classical method. The resultant one-tailed classical p value was .0436 which is identical to the Bayesian p value. This convergence between Bayesian and classical methods is reassuring regardless of whether one is Bayesian, classical, or eclectic in outlook. Convergence between Bayesian and classical methods also occurs for a number of other problems involving inference in the single case (Crawford et al., 2009b; Crawford et al., 2011).

The Bayesian method developed here for examining standardized differences in the presence of covariates does *not* have a classical equivalent. This is because classical methods cannot cope with the additional uncertainties introduced by standardizing scores on the two tasks of interest. This limitation does not stem from the need to allow for covariates, it applies equally to classical methods for examining standardized differences in the absence of covariates. Classical methods are restricted to examining the standardized difference score for the two tasks (regardless of whether scores have or have not been conditioned on a covariate) and are therefore blind as to how this standardized difference was obtained. This is best illustrated with a concrete example, which is based on Crawford and Garthwaite (2007a) and, in the interests of simplicity, ignores the effects of covariates.

Suppose we have two tasks Y_1 and Y_2, and that, for both tasks, the sample estimates of their standard deviations are 10.0. Suppose also that the standardized scores (zs) on the two tasks for Case A are $+1$ and -1, and that the standardized scores for Case B are -3 and -5. A classical analysis of these two cases will give an identical result because the difference between the standardized scores is the same for both cases ($y_1 - y_2 = -2$). However, error in estimating the standard deviations for the tasks will have a greater effect in Case B than in Case A because the standard deviations divide larger quantities in Case B than Case A. To illustrate, imagine that the standard deviation for Y_2 was 9 rather than 10. Then the difference between standardized scores for Case A would be $(+1.0) - (-1.111) = -2.111$, while for Case B it would be $(-3.0) - (-5.556) = -2.556$. It can be seen that the modest change in the standard deviation of Y_2 has had a substantial effect on Case B's standardized score on Y_2 and hence a substantial effect on the difference between standardized scores; the effect on Case A is much less marked. The Bayesian method allows for the greater uncertainty over the size of the difference between tasks in Case B because it allows for the uncertainty over the standard deviations of the individual tasks.

Some readers might wonder whether this limitation of classical methods could be sidestepped by not standardizing the scores on the two tasks and simply performing a test on the raw difference between them. However, although this does deal with the former problem, it introduces its own difficulties. Namely, if the standard deviations of the two tasks differ appreciably (as they often will in practice), the comparison of the case's difference score to the differences observed in controls may in effect reduce to a test on whether the case has a lower score than controls on the task with the larger SD. This is an example of what Capitani et al. (1999)

have termed the Cow-Canary problem; see Crawford et al. (2009a) for a more detailed discussion of this issue and examples of how it can seriously mislead.

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

In the interests of simplicity the worked examples provided in earlier sections were limited to the use of one covariate. However, technically the number of covariates is limited only by the df available, which must not be less than k. For example, if testing for deficit there are $n - m - 1$ df so that, if there were say 10 controls, a researcher could potentially include eight covariates as $k = 1$. Similarly, if testing for a standardized difference between two tasks in the presence of covariates and if (as recommended) the calibrated prior was used, then there are $n - m - 2$ df so that a researcher with 10 controls could potentially include six covariates as $k = 2$. However, in practice it could be anticipated that between one and three covariates would be typical (for example, a researcher may wish to simultaneously control for the effects of age and years of education). Moreover, the computer programs that implement these methods (see a later section) are limited to a maximum of five covariates, regardless of the df 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 a case and controls even when the case's value equals the mean of controls. 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 df, there would be a good reason not to include the variable as it will lower power. (Relatedly, in group studies, a popular rule of thumb for choosing between ANOVA *vs* ANCOVA is to use ANCOVA when the correlation of a covariate with the dependent variable exceeds .3.) 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 or tasks of interest.

Note that, when testing for dissociations between tasks, it is crucial to appreciate that the advice on omitting a covariate with a zero or near zero correlation with the tasks only holds when this is true for *both* tasks. Indeed earlier examples illustrates that, when only one of the tasks is highly correlated with the covariate, allowing for the covariate can be particularly useful in arriving at a clearer conception of a case's relative standing on the two tasks.

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 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.

3.7. *Application of the present methods to the study of multiple single cases in the case–controls design*

As alluded to in the Introduction, in contemporary single-case studies it is not uncommon for the investigation to include *multiple* single cases. Typically the performance of each of these single cases is compared against that of a single, common, control sample. This can be problematic unless the single cases are homogenous in terms of attribute variables such as age, years of education, estimated premorbid IQ, etc. If, as is liable to be the rule rather than the exception, the cases differ appreciably on these attribute variables then it follows that a single control sample can, at best, provide a close match to only a few of the cases.

One way round this problem would be to recruit separate control samples for each case – or, more realistically, form these control samples from a common pool of controls (such that any given control can serve as a member of the control sample for more than one of the cases). This is rarely if ever done in practice and it is easy to see that either strategy could be awkward and time-consuming to implement[1].

The methods developed in the present study offer an obvious and convenient alternative solution to the above problem. A single control sample (hopefully larger than is typical in single-case studies) would be used and the performance of the single cases on the tasks of interest compared to this sample with any attribute variables (e.g., age, years of education etc.) entered as covariates.

3.8. *Use of these methods in clinical neuropsychology*

The emphasis thus far has been on the use of these methods in single-case research, specifically for work carried out with the aim of building and testing theory concerning the architecture of human cognition. However, the methods can also play a useful role in clinical neuropsychology. Indeed the resurgence of interest in the analysis of the single case in academic neuropsychology means that the quantitative problems faced by the academic and clinician have never been more similar.

However, although both face the same fundamental problem – how to draw inferences concerning the profile of cognitive performance of an individual – there are some differences. One is that the clinical neuropsychologist predominantly uses standardized tests and therefore can compare a case's performance against a very large normative sample, rather than against the modestly sized control samples typically employed in single-case research.

It is true that computationally simpler methods could provide good approximations to the results obtained using the methods developed here when the reference sample is a large

[1] Note that some single-case studies will report that an individual control was recruited to match each case but this does not address the fundamental problem: these controls are then combined to form a control sample and the cases are then compared against this sample.

normative sample (such as those provided for the Wechsler scales etc.), rather than a modestly sized control sample. These computationally simpler methods treat the reference *sample* as though it were the reference *population*. The results would be close to the correct results because, when the reference sample is very large, sample statistics yield very good estimates of the corresponding population parameters.

However, our view is that the present methods should be used even in this latter situation. There are a number of reasons for this. First, the reality is that normative samples, even when very large, are still *samples* and methods that treat the sample as a sample are always to be preferred over methods that fail to do so. Secondly, there is an emphasis in contemporary psychological measurement on accompanying point estimates of quantities of interest with interval estimates. Methods that treat sample estimates as population parameters cannot provide these interval estimates because they do not acknowledge any uncertainty over the parameters; moreover, it is the case that, even with large samples, the uncertainty is appreciable and is therefore worth quantifying (see Crawford and Garthwaite, 2008). Finally, the computer programs accompanying this paper make the application of the current methods quick and painless so that, if anything, it is easier to apply the technically correct methods than the corresponding approximate methods.

A second difference between the academic single-case researcher and the clinical neuropsychologist is that the latter does not have the luxury of recruiting a “bespoke” matched sample of healthy controls. That is, although the clinician has access to large normative samples, the normative data are typically stratified only by age. The present methods offer a means of controlling for other attribute variables (e.g., years of education) by using them as covariates. Such usage requires that the clinician has access to the correlations between these attribute variables and the tasks of interest but it is not uncommon for such data to be available, either directly in test manuals, or in supplementary publications.

In summary, although the methods developed here are particularly suitable for the single-case researcher (because of the use of modest control samples), they have a wide range of potential applications in clinical neuropsychology (and indeed other applied fields in which there is a need to make inferences concerning an individual).

3.9. *Implementing the present methods in computer programs*

All of the methods developed here have been implemented in a series of compiled computer programs for PCs written in the Delphi programming language. These programs will also run on a Mac if emulation software is installed. (Methods developed previously by the present authors have been implemented in the statistical program R by third parties. It may be that the present methods will also be implemented in R; 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 and Garthwaite, 2002, 2005, 2007a) the programs take summary data from the control sample as inputs. The summary data required are the means and standard deviations for the task(s) and covariates(s) and the correlations between all of the variables involved. 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 (such data are normally only available in summary form).

However, for the specific methods developed in the present paper, it is likely that many researchers would find it more convenient to provide the *raw* data for the controls and have the program calculate the necessary summary statistics. Therefore, alternative, parallel, versions of the programs have been written to cater for this. These programs load the raw data for controls and the single-case from a text file prepared by the user (full details of how to format this text file are provided in the information panel of the programs).

All programs reproduce the data entered by the users so that researchers have a record of the inputs (for the versions of the programs that take raw data as inputs, the full set of raw data is reproduced along with the computed summary statistics for the control sample). The outputs consist of one- and two-tailed *p* values for hypothesis testing, the point and interval estimates of the effect size for the difference between the case and controls, and point and interval estimates of the abnormality of the case’s scores or standardized score differences. Some useful supplementary results are also provided, for example the sample estimates of the conditional means and standard deviations. The results can be viewed on screen, printed and/or saved to a text file.

The program BTD_Cov.exe tests for a deficit controlling for covariate(s) using summary data for the control sample. Its companion program, BTD_Cov_Raw.exe, performs the same analysis but takes raw data for the controls as input. The program BSDT_Cov.exe compares a case’s standardized difference to that of controls controlling for the effects of covariate(s); it takes summary data for the control sample as its inputs. Its companion program, BSDT_Cov_Raw.exe performs the same analysis but takes raw data for the controls as input.

Finally, as noted, we have also updated the computer program (DiffBayes_ES.exe) that performs the BSDT (Crawford & Garthwaite, 2007a) to allow use of a calibrated prior, we identify this test as the BSDT-ES-CP; the ES suffix denotes that the method reports point and interval estimates of effect sizes (Crawford et al., 2010) and the CP suffix denotes that it also now offers a calibrated prior (use of this prior is set as the default but can be overridden by the user should they wish to use the “standard theory” prior). As noted we recommend the use of the calibrated prior over the standard theory prior that was implemented in the original version. The updated computer program is DiffBayes_ES_CP.exe. The BSDT is also used in a further program (DissocsBayes_ES.exe) that tests whether a case’s pattern of performance on two tasks meets Crawford and Garthwaite’s (2007a) criteria for a classical or strong dissociation. Use of the calibrated prior has also been added to this program (and is set as the default prior); the upgraded program is DissocsBayes_ES_CP.exe.

All programs can be downloaded (individually, or in the form of a zip file containing all six programs) from the following URL: www.abdn.ac.uk/~psy086/dept/Single_Case_Covariates.htm

3.10. *Conclusions*

The methods developed in the present paper are a useful extension to existing methods of testing for the presence of a deficits or dissociation in the single case. Being able to control for the effects of covariates allows single-case researchers to tackle more complex problems than hitherto. The methods can be applied quickly and reliably using the computer programs written to accompany this paper.

Appendix 1. Procedure for employing a variant on the Berger and Sun (2008) prior, including a description of the accept/reject method

The calibrated prior we employ, is equivalent to the prior identified by Berger and Sun (2008) as π_{R_ρ}

$$\pi_{R_\rho}(\mathbf{B},\Sigma) \propto \frac{1}{\sigma_1\sigma_2(1-\rho^2)}.$$

Simulations showed that the frequentist properties when using this prior could be improved by reducing the effective control sample size by 1 so that our hypothesis test and credible intervals are a little conservative (Garthwaite and Crawford, in preparation). Let **A** be the matrix of products and cross-products so that $\mathbf{A}/(n-m-1)$ is our unbiased estimate of Σ. To retain an unbiased estimate of Σ when we reduce the sample we treat $\mathbf{A}^*$ as the matrix of products and cross-products where

$$\mathbf{A}^* = (n-m-2)\mathbf{A}/(n-m-1).$$

We generate an observation of Σ from an inverse-Wishart distribution on $(n-m-2)$ df and scale matrix $(\mathbf{A}^*)^{-1}$. Let

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\sigma}_1^2 & \widehat{\sigma}_{12} \\ \widehat{\sigma}_{12} & \widehat{\sigma}_2^2 \end{pmatrix}$$

denote this observation. Put $\widehat{\rho} = \widehat{\sigma}_{12}/(\widehat{\sigma}_1^2\widehat{\sigma}_2^2)^{1/2}$.

$\widehat{\Sigma}$ is an observation generated from an inverse-Wishart distribution. It is not an observation from our posterior distribution. However, we introduce an appropriate accept/reject step, given by following Berger and Sun (2008), whereby we accept some of the $\widehat{\Sigma}$ that are generated, but reject others. The set of $\widehat{\Sigma}$ values that are accepted form a random sample from our actual posterior distribution.

The accept/reject step is as follows. We generate a value from a uniform(0,1) distribution. Denote this value as u. If $u^2 \le 1-\widehat{\rho}^2$ then we accept the generated value $\widehat{\Sigma}$; otherwise we reject it. In the latter case we repeat the process until we generate a $\widehat{\Sigma}$ that is accepted. We let $\widehat{\Sigma}_{(i)}$ denote the ith $\widehat{\Sigma}$ that is accepted. (After obtaining $\widehat{\Sigma}_{(i)}$, we then generate $\mathbf{B}_{(i)}$ from an MVN distribution with variance determined by $\widehat{\Sigma}_{(i)}$.)

As noted in the main body of the paper, we also implemented a calibrated prior for the BSDT, i.e., for use when there are no covariates. The procedure is identical to that followed when there are covariates except that we denote $\mathbf{A}^*$ as

$$\mathbf{A}^* = (n-2)\mathbf{A}/(n-1),$$

and generate observations from an inverse-Wishart on $(n-2)$ df, rather than $(n-m-2)$ df.

Appendix 2. Formal statement of Bayesian regression procedure for testing for a deficit or dissociation in the presence of covariate(s)

This appendix sets out the procedure when using the raw data for each member of the control sample as inputs. As noted, the methods can also be applied using summary statistics for the control sample as inputs (i.e., the vectors of controls means and standard deviations for tasks and covariates and the full correlation matrix for controls). Appendix 3 sets out the pre-processing required when using summary statistics as inputs.

1. We have n controls for whom we have values of k variables that are the tests of interest and m covariates. Let $\underline{x}_j$ 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 for the jth case ($j=1,\ldots,n$). Let $\underline{y}_i$ be the vector of values for the ith test ($i=1,\ldots, k$). Put $\mathbf{X} = (\underline{x}_1,\ldots,\underline{x}_n)'$ and $\mathbf{Y} = (\underline{y}_1,\ldots,\underline{y}_k)$. Then the data estimate of **B** is

$$\mathbf{B}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{5}$$

and the data estimate of Σ is

$$\Sigma^* = \frac{1}{n-m-1}(\mathbf{Y}-\mathbf{X}\mathbf{B}^*)'\cdot(\mathbf{Y}-\mathbf{X}\mathbf{B}^*). \tag{6}$$

When using the standard theory prior, the posterior distribution for Σ^{-1} is a Wishart distribution with parameters $n-m$ and $[(n-m-1)\Sigma^*]^{-1}$. Generate an estimate of Σ from this distribution, using the procedure given in Crawford and Garthwaite (2007a) and denote the ith generated value by $\widehat{\Sigma}_{(i)}$. When using the calibrated prior, generate an estimate of Σ using the method described in Appendix 1 and again denote the ith generated value by $\widehat{\Sigma}_{(i)}$.

Change $\mathbf{B}^* = (\underline{\beta}_1^*,\ldots,\underline{\beta}_k^*)$ from an $(m+1)\times k$ matrix to a $k(m+1)\times 1$ vector by putting $\mathbf{B}_{\text{vec}}^* = (\underline{\beta}_1^{*\prime},\ldots,\underline{\beta}_k^{*\prime})'$. Correspondingly, put $\mathbf{B}_{\text{vec}} = (\underline{\beta}_1',\ldots,\underline{\beta}_k')'$ and$\mathit{\Lambda}_{(i)} = \widehat{\Sigma}_{(i)}\otimes(\mathbf{X}'\mathbf{X})^{-1}$, where ⊗ is the Kronecker product. The posterior distribution of $\mathbf{B}_{\text{vec}}$, given $\underline{\Sigma}_{(i)}$, is a multivariate normal distribution with mean $\mathbf{B}_{\text{vec}}{}^*$ and variance $\Lambda_{(i)}$. An estimate of $\mathbf{B}_{\text{vec}}$ is generated from this distribution and the estimate denoted as $(\widehat{\underline{\beta}}_{1(i)}',\ldots,\widehat{\underline{\beta}}_{k(i)}')'$, where $\widehat{\underline{\beta}}_{j(i)}$ is an $(m+1)\times 1$ vector. Then $\mathbf{B}_{(i)} = (\widehat{\underline{\beta}}_{1(i)},\ldots,\widehat{\underline{\beta}}_{k(i)})$.

2. Let $\underline{x}$ denote the vector of values of the covariates for the case. Then the conditional expected values of the case on the tests is the vector $\widehat{\underline{\mu}}_{(i)} = \mathbf{B}_{(i)}'\,\underline{x}$.

The variances and correlation needed in equations (2) and (3) in the main body of the text are determined from $\widehat{\Sigma}_{(i)}$. When there is only one task of interest, $\widehat{\Sigma}_{(i)}$ is a 1×1 matrix and $\widehat{s}_{(i)}$ is $(\widehat{\Sigma}_{(i)})^{1/2}$. When there are two tasks of interest, $\widehat{s}_{1(i)} = (\widehat{s}_{1(i)}^2)^{1/2}$, $\widehat{s}_{2(i)} = (\widehat{s}_{2(i)}^2)^{1/2}$ and $\widehat{\rho}_{12(i)} = \widehat{s}_{12(i)}/(\widehat{s}_{1(i)}\widehat{s}_{2(i)})$, where

$$\widehat{\Sigma}_{(i)} = \begin{pmatrix} \widehat{s}_{1(i)}^2 & \widehat{s}_{12(i)} \\ \widehat{s}_{12(i)} & \widehat{s}_{2(i)}^2 \end{pmatrix}.$$

Appendix 3. Pre-processing required when using control summary statistics

(a) From Appendix 2 it can be seen that we require the quantities $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{Y}$, $\mathbf{Y}'\mathbf{Y}$, and $(\mathbf{Y}-\mathbf{XB})'(\mathbf{Y}-\mathbf{XB})$. We have the summary statistics from the control sample in the form of the full correlation matrix (i.e., tasks and covariates) and the full vectors of means and standard deviations (i.e., tasks and covariates). We partition the matrix and vectors to form $\mathbf{R}$ (the $m \times m$ sample correlation matrix for the x-variables), $\mathbf{Q}$ (the $m \times k$ matrix of correlations between the m x-variables and k y-variables), $\mathbf{P}$ (the $k \times k$ sample correlation matrix for the y-variables), $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_m)'$ (the $m \times 1$ vector of means of the explanatory variables), $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_k)'$ (the $k \times 1$ vector of means of the y-variables), $v_1,\ldots,v_m$ (the sample standard deviations of $x_1,\ldots,x_m$), and $w_1,\ldots,w_k$ (the sample standard deviations of $y_1,\ldots,y_k$).

(b) Multiply the ith row of $\mathbf{R}$ by $(n-1) \times v_i$ for $i = 1,\ldots, m$. Multiply the jth column of the resulting matrix by v_j for $j = 1,\ldots, m$. This gives an $m \times m$ matrix, $\mathbf{R}^*$ say, that needs the mean-correction to be cancelled in order to give the main part of $\mathbf{X}'\mathbf{X}$. Add $n \times \bar{x}_i \times \bar{x}_j$ to the (i, j) element of $\mathbf{R}^*$ (do this for each element of $\mathbf{R}^*$). Call the resulting matrix $\mathbf{R}^{\#}$. Then we can get $\mathbf{X}'\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & n\bar{\mathbf{x}}' \\ n\bar{\mathbf{x}} & \mathbf{R}^{\#} \end{bmatrix}.$$

(c) Multiply the ith row of $\mathbf{Q}$ by $(n-1) \times v_i$ for $i = 1,\ldots, m$. Multiply the jth column of the resulting matrix by w_j for $j = 1,\ldots, k$. This gives an $m \times k$ matrix, $\mathbf{Q}^*$ say, that needs the mean-correction to be cancelled in order to give $\mathbf{X}'\mathbf{Y}$. Add $n \times \bar{x}_i \times \bar{y}_j$ to the (i, j) element of $\mathbf{Q}^*$ (do this for each element of $\mathbf{Q}^*$), calling the resulting matrix $\mathbf{Q}^{\#}$. Then put

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} n\bar{y}_1 & n\bar{y}_2 & \cdots & n\bar{y}_k \\ & \mathbf{Q}^{\#} & & \end{pmatrix}.$$

(d) We need the $k \times k$ matrix $\mathbf{Y}'\mathbf{Y}$. Multiply the ith row of $\mathbf{P}$ by $(n-1) \times w_i$ for $i = 1,\ldots, k$. Multiply the jth column of the resulting matrix by w_j for $j = 1,\ldots, k$. This gives an $m \times k$ matrix, $\mathbf{P}^*$ say, that needs the mean-correction to be cancelled in order to give $\mathbf{Y}'\mathbf{Y}$. Add $n \times \bar{y}_i \times \bar{y}_j$ to the (i, j) element of P^* (do this for each element of $\mathbf{P}^*$). The resulting matrix is $\mathbf{Y}'\mathbf{Y}$.

(e) We can expand $(\mathbf{Y}-\mathbf{XB}^*)'(\mathbf{Y}-\mathbf{XB}^*)$ as

$$(\mathbf{Y}-\mathbf{XB}^*)'(\mathbf{Y}-\mathbf{XB}^*) = \mathbf{Y}'\mathbf{Y} - 2(\mathbf{B}^*)'(\mathbf{X}'\mathbf{Y}) + (\mathbf{B}^*)'(\mathbf{X}'\mathbf{X})\,(\mathbf{B}^*).$$

We have the matrices $\mathbf{Y}'\mathbf{Y}$, $\mathbf{X}'\mathbf{X}$, and $\mathbf{X}'\mathbf{Y}$, and, from Appendix 2 equation (5) we have $\mathbf{B}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Hence we have all the parts we need in the above expression to obtain $(\mathbf{Y}-\mathbf{XB}^*)'(\mathbf{Y}-\mathbf{XB}^*)$ from matrix multiplication and addition/subtraction.

Appendix 4. The effect on a dissociation of a covariate when its correlations with the two tasks are equal

Interest focuses on scores from two tests, Y_1 and Y_2, and a covariate X. If the covariate is ignored, inferences about the dissociation are based on the quantity,

$$\hat{z}_{DCC} = \frac{\left(\dfrac{y_1 - \hat{\mu}_1}{s_1}\right) - \left(\dfrac{y_2 - \hat{\mu}_2}{s_2}\right)}{\sqrt{2 - 2\hat{\rho}_{12}}}, \tag{7}$$

where $\hat{\mu}_1$, $\hat{\mu}_2$, s_1 and s_2 are the means and standard deviations of Y_1 and Y_2 in the control population, $\hat{\rho}_{12}$ is the sample correlation between Y_1 and Y_2, and y_1 and y_2 are the scores of the case.

Suppose now, that the sample correlation between Y_1 and X is r and the sample correlation between Y_2 and X is also r. Let $\hat{\mu}_x$ and s_x denote the mean and standard deviation of X in the control sample and let x be the case's value of X. If $\hat{\mu}_{1(x)}$, $\hat{\mu}_{2(x)}$, $\hat{s}_{1(x)}$, $\hat{s}_{2(x)}$ and $\hat{\rho}_{12(x)}$ denote the estimated means, standard deviations and correlation of Y_1 and Y_2, *conditional on* $X = x$, then

$$\hat{\mu}_{1(x)} = \hat{\mu}_1 - r(s_1/s_x)\,(\hat{\mu}_x - x) \quad \hat{\mu}_{2(x)} = \hat{\mu}_2 - r(s_2/s_x)\,(\hat{\mu}_x - x), \tag{8}$$

$$\hat{s}_{1(x)} = s_1\left(1 - r^2\right)^{1/2} \quad \hat{s}_{2(x)} = s_2\left(1 - r^2\right)^{1/2}, \tag{9}$$

and

$$\hat{\rho}_{12(x)} = \left(\hat{\rho}_{12} - r^2\right)/\left(1 - r^2\right). \tag{10}$$

Conditional on $X = x$, equation (3) states that inferences about the dissociation are based on

$$\hat{z}_{DCCC} = \frac{\left(\dfrac{y_1 - \hat{\mu}_{1(x)}}{\hat{s}_{1(x)}}\right) - \left(\dfrac{y_2 - \hat{\mu}_{2(x)}}{\hat{s}_{2(x)}}\right)}{\sqrt{2 - 2\hat{\rho}_{12(x)}}}. \tag{11}$$

Substitution for $\hat{\mu}_{1(x)}$, $\hat{\mu}_{2(x)}$, $\hat{s}_{1(x)}$, $\hat{s}_{2(x)}$ and $\hat{\rho}_{12(x)}$ using equations (8)–(10), followed by straightforward simplification, shows that $\hat{z}_{DCCC}$ equals $\hat{z}_{DCC}$ in equation (7).

REFERENCES

Berger JO and Sun D. Objective priors for the bivariate normal model. *Annals of Statistics*, 36(2): 963–982, 2008.

Bernardo JM. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society Series B*, 41(2): 113–147, 1979.

Capitani E, Laiacona M, Barbarotto R, and Cossa FM. How can we evaluate interference in attentional tests? A study based on bi-variate non-parametric tolerance limits. *Journal of Clinical and Experimental Neuropsychology*, 21(2): 216–228, 1999.

Crawford JR and Garthwaite PH. Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, 40(8): 1196–1208, 2002.

Crawford JR and Garthwaite PH. Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology*, 19(3): 318–331, 2005.

Crawford JR and Garthwaite PH. Comparing an individual's predicted test score from a regression equation with an obtained score: A significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology*, 20(3): 259–271, 2006a.

Crawford JR and Garthwaite PH. Methods of testing for a deficit in single case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology*, 23(6): 877–904, 2006b.

Crawford JR and Garthwaite PH. Comparison of a single case to a control or normative sample in neuropsychology:

Development of a Bayesian approach. *Cognitive Neuropsychology*, 24(4): 343–372, 2007a.

Crawford JR and Garthwaite PH. Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology*, 21(5): 611–620, 2007b.

Crawford JR and Garthwaite PH. On the "optimal" size for normative samples in neuropsychology: Capturing the uncertainty associated with the use of normative data to quantify the standing of a neuropsychological test score. *Child Neuropsychology*, 14(2): 99–117, 2008.

Crawford JR, Garthwaite PH, Azzalini A, Howell DC, and Laws KR. Testing for a deficit in single case studies: Effects of departures from normality. *Neuropsychologia*, 44(4): 666–676, 2006.

Crawford JR, Garthwaite PH, and Gray CD. Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex*, 39(2): 357–370, 2003.

Crawford JR, Garthwaite PH, and Howell DC. On comparing a single case with a control sample: An alternative perspective. *Neuropsychologia*, 47(13): 2690–2695, 2009a.

Crawford JR, Garthwaite PH, and Porter S. Point and interval estimates of effect sizes in the case–controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, 27(3): 245–260, 2010.

Crawford JR, Garthwaite PH, and Slick DJ. On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist*, 23(7): 1173–1195, 2009b.

Crawford JR, Garthwaite PH, and Wood LT. The case controls design in neuropsychology: Inferential methods for comparing two single cases. *Cognitive Neuropsychology*, 27(5): 377–400, 2011.

Crawford JR and Howell DC. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, 12(4): 482–486, 1998a.

Crawford JR and Howell DC. Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology*, 20(5): 755–762, 1998b.

Garthwaite PH and Crawford JR. Calibrated Bayesian analysis for single-case studies, in preparation.

Garthwaite PH, Jolliffe IT, and Jones B. *Statistical Inference*. Oxford: Oxford University Press, 2002.

Jeffreys H. *Theory of Probability*. Oxford Clarendon Press, 2002.

Tiao GC and Zellner A. On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society, Series B*, 26(2): 277–285, 1964.