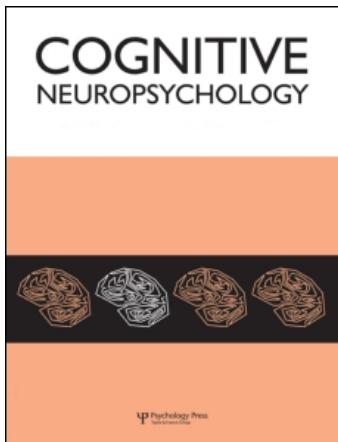


This article was downloaded by: [University of Aberdeen]  
On: 5 November 2010  
Access details: Access Details: [subscription number 917198493]  
Publisher Psychology Press  
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Cognitive Neuropsychology

Publication details, including instructions for authors and subscription information:  
<http://www.informaworld.com/smpp/title~content=t713659042>

### Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards

John R. Crawford<sup>a</sup>, Paul H. Garthwaite<sup>b</sup>, Sara Porter<sup>a</sup>

<sup>a</sup> University of Aberdeen, Aberdeen, UK <sup>b</sup> Department of Mathematics and Statistics, The Open University, Milton Keynes, UK

First published on: 09 October 2010

**To cite this Article** Crawford, John R. , Garthwaite, Paul H. and Porter, Sara(2010) 'Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards', *Cognitive Neuropsychology*, 27: 3, 245 — 260, First published on: 09 October 2010 (iFirst)

**To link to this Article:** DOI: 10.1080/02643294.2010.513967

URL: <http://dx.doi.org/10.1080/02643294.2010.513967>

## PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards

John R. Crawford

*University of Aberdeen, Aberdeen, UK*

Paul H. Garthwaite

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

Sara Porter

*University of Aberdeen, Aberdeen, UK*

It is increasingly common for group studies in neuropsychology to report effect sizes. In contrast this is rarely done in single-case studies (at least in those studies that employ a case-controls design). The present paper sets out the advantages of reporting effect sizes, derives suitable effect size indexes for use in single-case studies, and develops methods of supplementing point estimates of effect sizes with interval estimates. Computer programs that implement existing classical and Bayesian inferential methods for the single case (as developed by Crawford, Garthwaite, Howell, and colleagues) are upgraded to provide these point and interval estimates. The upgraded programs can be downloaded from [www.abdn.ac.uk/~psy086/dept/Single\\_Case\\_Effect\\_Sizes.htm](http://www.abdn.ac.uk/~psy086/dept/Single_Case_Effect_Sizes.htm)

**Keywords:** Neuropsychology; Effect sizes; Single-case studies; Single-case methods; Statistical reform; Confidence intervals; Credible intervals.

It is now generally recognized that, although significance testing may have served science well in many respects, it has been overemphasized in psychology at the expense of attention to the size of effects. In an attempt to redress this imbalance a number of authorities in statistics and psychology have made strenuous calls for research papers to include indexes of effect size. 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 & The APA Task Force on Statistical Inference, 1999, p. 599).

Advice aimed specifically at neuropsychologists has also been offered (e.g., Bezeau & Graves, 2001; Crawford & Henry, 2004; Zakzanis, 2001), and editorial policies requiring the reporting of effect sizes in neuropsychology journals

---

Correspondence should be addressed to John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 3HN, UK (E-mail: [j.crawford@abdn.ac.uk](mailto:j.crawford@abdn.ac.uk)).

(Becker, Knowlton, & Anderson, 2005) have provided a further impetus. Although it is true to say that the take-up of such advice has been relatively slow, reporting of effect sizes in group-based neuropsychological research is now fairly common.

The focus of the present paper is on single-case studies employing 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 matched sample of healthy controls). The reporting of effect sizes in such studies is still relatively uncommon. This is unfortunate as the methodological standards in single-case studies should, where possible, be as high as those demanded in group research. Moreover, in contrast to the neglect of effect sizes in the case-controls design, the issue of effect sizes has been tackled for single-case research based solely on intraindividual comparisons (such as studies evaluating the effects of an intervention in the single case by comparing baseline and postintervention scores; e.g., see Parker et al., 2005).

In the present paper we: (a) identify suitable indexes of effect size for the case-controls design; (b) briefly set out the advantages of using effect sizes; (c) derive methods of supplementing point estimates of effect size with interval estimates; (d) describe and make available computer programs that calculate these point and interval estimates; and (e) propose a set of standards for the reporting of statistical results in the case-controls design.

### **Effect size index for comparison of a single case to controls**

In *group* studies the most commonly used effect size index is Cohen's *d*. This index expresses the difference between the means of a patient sample and control sample in standardized units by dividing the difference by the pooled standard deviation of the two samples (an alternative index is Hedges's *g*, which differs only in the method used to estimate the pooled standard deviation; Cohen's *d* and Hedges's *g* typically have similar

values, especially when the number of controls greatly exceeds the number of patients).

The obvious, direct, analogue of Cohen's *d* when comparing a single-case's score to a control sample is *z*, computed using a single-case's score (*x*) and the controls' sample mean ( $\bar{x}$ ) and standard deviation (*s<sub>x</sub>*); we hereafter denote this quantity as *z<sub>CC</sub>* (the subscript *CC* representing "case-controls") to differentiate it from a further index developed later. Thus the proposed index of effect size when comparing a single case to controls is simply

$$z_{CC} = \frac{x - \bar{x}}{s_x}. \quad (1)$$

This index is an estimate of the average difference, measured in standard deviation units so as to be scale independent, between a case's score and the score of a randomly chosen member of the control population. Like Cohen's *d*, *z<sub>CC</sub>* is insensitive to the size of the control sample. This characteristic makes it unsuitable as a significance test (Crawford, Garthwaite, & Howell, 2009), although regrettably it is still widely used for such a purpose. However, in contrast, this is a required characteristic for an index of effect size.

### **The case for effect sizes for cases**

The main advantages of including effect sizes when reporting the result of a single-case study are those that also apply to the reporting of group research. Moreover, there are characteristics of single-case research that make the use of effect sizes particularly compelling. Namely, (a) single-case studies typically employ a large number of neuropsychological measures, (b) more often than not these measures are expressed on different metrics, (c) there is an emphasis in neuropsychological single-case studies on examining the *profile* (i.e., the relative strengths and weaknesses) of a patient's performance, and (d) it is not uncommon for single-case studies to employ more than one control sample (so that the influence of control sample size on *p* values is not necessarily constant across measures).

All these factors conspire to make the use of effect sizes particularly useful in single-case research. By expressing all of a case's score on a common metric, consumers of the research can readily assimilate the pattern of relative strengths and weaknesses in the case's profile of scores. This is much more satisfactory than only presenting  $p$  values or, even worse, simply recording whether  $p$  values are below, say, .05. A case's  $p$  value for one task could be just below .05 and just above .05 for another. In such circumstances the relative difference in the case's performance on the two tasks is entirely trivial—this would be immediately apparent from the effect sizes. Calculation of effect sizes also allows researchers to compare their results with those found in studies of other single cases (even when the tasks used to assess a function of interest in these studies differ from the tasks employed in the study at hand).

### Reporting of effect sizes in contemporary single-case research

It was suggested in an earlier section that effect sizes are not commonly reported in single-case studies that employ the case-controls design. In order to examine this suggestion empirically we reviewed all papers published between 2004 and 2006 in four major neuropsychology journals (*Cognitive Neuropsychology*, *Journal of the International Neuropsychological Society*, *Neuropsychologia*, and *Neuropsychology*).<sup>1</sup> Two of these journals (*Cognitive Neuropsychology* and *Neuropsychologia*) were chosen because they have a history of publishing single-case research. The other two were chosen because although, like the others, they are major journals (i.e., they are official journals of the International Neuropsychological Society and American Psychological Association, respectively), they predominantly publish group studies. That is, we felt

that single-case studies published in journals that do not specialize in this area should also be represented.

Papers were selected for inclusion if they were single-case studies and involved the analysis of cognitive/behavioural processes. Papers were included if more than one patient was featured (i.e., multiple single-case studies), provided that the paper included analysis at the level of the single case. This resulted in the identification of 219 single-case studies, 159 of which included control samples. For present purposes attention was then restricted to studies ( $k = 98$ ) that used inferential methods to *directly* compare a single-case to controls—that is, studies that used the case-controls design.

Of these 98 studies, only 20 (20.4%) reported any effect sizes ( $z$ ), and only 13 (13.3%) presented effect sizes to accompany all of the inferential tests reported. A further 2 studies reported *raw* differences between cases and controls and explicitly described these differences as "effect sizes". However, such differences are not effect sizes as commonly defined because, crucially, they are not *standardized* differences. Thus the vast majority (79.6%) of studies using the case-controls design did not report any effect sizes.

The number of studies credited with reporting effect sizes (20) falls below the number of studies (31) that used  $z$  for inferential purposes (i.e., as a significance test). This disparity occurs because, in many of these studies,  $z$  itself was not directly reported. Rather, either the results were reported in a dichotomous form (i.e., cases were classified as exhibiting or failing to exhibit a statistically significant deficit without reporting the  $z$  value upon which the decisions were based), or reporting was limited to presentation of the  $p$  values.

This pattern of reporting is particularly unfortunate;  $z$  was used for a purpose to which it is unsuited (hypothesis testing), but was not used for a purpose to which it is suited (to record

---

<sup>1</sup>Papers published after 2006 were not included because the figures reported here have been extracted from a more detailed ongoing survey of single-case research. This includes collecting data on citation impact and therefore requires a lag between publication and measurement of impact (another factor is that coding up the attributes of these studies is more complex and time consuming than was anticipated).

effect sizes). Even among the 20 studies that did report  $z$ , in most cases it was clear that it was being used as a significance test (i.e., cases were described as being significantly different from controls), rather than being identified as an index of effect size.

Two of the present authors may have inadvertently contributed to the low rate of reporting of effect sizes. In previous papers (e.g., Crawford & Garthwaite, 2005b) we have shown that  $z$  is inappropriate as a mean of testing for a significant difference between a single case and controls because it is associated with inflated Type I errors (i.e., false positives) when used with the sample sizes that typify single-case research. The appropriate hypothesis test is that proposed by Crawford and Howell (1998) because, under the null hypothesis (that the case's score is an observation from the scores in the control population), the difference between a case and controls follows a  $t$ -distribution rather than a standard normal distribution. The Crawford and Howell method was the second most common method used to compare a single case to controls (26 studies, compared to the 31 using  $z$ ).

We suggest that single-case researchers should employ and report *both* these statistics: the  $t$  value from Crawford and Howell's (1998) procedure to test and record whether the difference between the single case and controls is statistically significant; and  $z_{CC}$ , as an effect size index for the difference between the case and controls. (The  $t$  value obtained from the Crawford and Howell method is not suitable as an index of effect size because it varies as a function of the size of the control sample.)

### Interval estimates for effect sizes in the case-controls design

In group-based research there is an increasing recognition that *point* estimates of effect size should be accompanied by *interval* estimates (i.e., confidence intervals or credible intervals)—for example, see Steiger (2004), Fidler and Thompson (2001), and Thompson (2007). In keeping with the principle alluded to earlier, that

the standards of reporting in single-case studies should be as high as those expected in group research, we suggest that interval estimates for effect sizes should also be reported in single-case research. None of the single-case studies reviewed earlier reported such intervals. This, however, should not be surprising as this problem has not previously been tackled explicitly.

Fortunately, the statistical theory necessary to form these interval estimates already exists. Crawford and Garthwaite (2002) derived a classical method of setting confidence limits on the abnormality of a patient's score using noncentral  $t$ -distributions. An intermediate step in obtaining these limits involves computing two standard normal deviates, and these provide the required upper and lower limits on the effect size index. The derivation of these limits on an effect size and the calculations required to obtain them are set out in Appendix A. For a broader and excellent treatment of noncentral  $t$ -distributions in psychological measurement see Cumming and Finch (2001).

A Bayesian analysis of this problem proceeds differently but, as will be shown, gives the same results as the classical method (save for very minor differences attributable to Monte Carlo variation). The Bayesian approach that is used is based on methods developed by Crawford and Garthwaite (2007). The formal details are set out in Appendix B. Informally, the method consists of repeatedly applying Equation (1) with inputs consisting of the case's score and estimates of the control mean and standard deviation obtained through random draws from the posterior distribution for controls (we denote the value obtained on any one of these iterations as  $\hat{z}_{CC}$ ). If one million iterations are performed, then the 25,000th lowest and 25,000th highest values of  $\hat{z}_{CC}$  provide the lower and upper endpoints of a 95% credible interval for the true effect size.

To illustrate the application of these intervals, take the example of a neuropsychological task for which the mean in a healthy control sample of 10 persons is 50, and the standard deviation is 10. Suppose a case obtains a score of 24. Then the point estimate of the effect size,  $z_{CC}$ , for the

comparison of the case to controls is  $-2.60$ . Using the classical method of obtaining an interval estimate for this quantity, the 95% confidence interval on the effect size ranges from  $-3.919$  to  $-1.253$  (the calculations for this example can be found at the end of Appendix A). This interval captures the uncertainty over the true effect size—if the patient's score had been referred to another control sample the resultant point estimate would differ from that calculated using the sample in hand. It can be seen that, in this example, the confidence interval on the effect size is wide. This largely stems from the fact that the control sample is modest in size, although an  $n$  of this size is not atypical in single-case studies; indeed the sample size used in this example ( $n = 10$ ) corresponds to the median  $n$  for the 98 single-case studies reviewed earlier that directly compared a single case to controls; the mean  $n$  was slightly higher at  $11.69$  ( $SD = 10.66$ ).

With larger  $n$  the limits are narrower but there will still be considerable uncertainty unless the control sample  $n$  is very much larger than the  $n_s$  typically found in single-case studies. For example, if all other quantities were the same as those in the previous example but the control sample  $n$  was  $40$ , the 95% confidence interval on the effect size ranges from  $-3.249$  to  $-1.943$ ; for a sample size of  $150$  the interval is from  $-2.934$  to  $-2.263$ .

Because the intervals follow a noncentral  $t$ -distribution, their width will also be affected by the extremity of the patient's score: The more extreme the score, the wider the interval. In addition, the end points (i.e., the upper and lower limits) of the interval will also be more asymmetrical around the point estimate of the effect size ( $z_{cc}$ ) when scores are extreme: For scores below the mean the lower limit will be further from the point estimate (the converse will occur for scores above the mean). This latter characteristic will be attenuated when the control sample  $n$  is large (noncentral  $t$ -distributions become less skewed as sample size increases).

As noted, the Bayesian credible intervals will give the same results as the classical limits and

therefore possess the same characteristics as those described above. The Bayesian 95% interval for the original worked example ( $n = 10$ ) is  $(-3.918, -1.254)$ , and it can be seen that this corresponds very closely to the interval obtained using the classical method  $(-3.919, -1.253)$ . The difference arises from Monte Carlo variation.

The fact that the two methods yield equivalent intervals means that we can apply a Bayesian interpretation to either set of limits. Thus, even if a single-case researcher chose to use the classical method for this problem, they can legitimately avoid the convoluted classical (i.e., frequentist) interpretation of these limits. As Antelman (1997) notes, the frequentist 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 frequentist interpretation is as follows, "if we could compute confidence intervals from a large number of control samples collected in the same way as the present control sample, about 95% of them would contain the true effect size". In contrast, the Bayesian analysis gives the conclusion, "there is a 95% probability that the effect size lies within the stated limits". This statement is not only less convoluted but, we suggest, it also captures what a single-case researcher would wish to conclude from an interval estimate. It is likely that most psychologists who use frequentist confidence intervals in fact construe these in what are essentially Bayesian terms (Crawford & Garthwaite, 2007).

### An empirical test of the method of setting limits on effect sizes using Monte Carlo simulation

The statistical theory set out in Appendix A implies that, if the control data are drawn from a normal distribution, then the confidence intervals should capture individuals' true effect sizes 95% of the time (the true effect sizes are the  $z$  scores the individuals would obtain were they computed

using the mean and standard deviation of the control *population* rather than a control *sample*). To confirm this empirically a Monte Carlo simulation was performed. This simulation should help inform those neuropsychologists who have either little interest in or little knowledge of the underlying statistical theory.

Four population values of  $z_{CC}$  were selected ranging from  $-0.674$  (representing a score only slightly below the population mean, i.e., 25% of the control population would obtain a lower score) through values of  $-1.282$  (10%),  $-1.960$  (2.5%), to a value of  $-2.326$  (representing a very low score; only 1% of the control population would obtain a lower score). These represent individuals' *true z* scores—that is, the *z* scores they would obtain had we access to the entire population of controls and therefore knew the population mean and standard deviation (hereafter denoted  $z_{CC}^*$  to differentiate it from  $z_{CC}$  computed using sample statistics). For each of these  $z_{CC}^*$ , one hundred thousand 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 sample mean was subtracted from  $z_{CC}^*$  and the result was divided by the sample standard deviation. This created a *z* based on the sample statistics on that trial. The method set out in Appendix A was then applied to calculate a 95% confidence interval on  $z_{CC}^*$ . The number of trials in which these confidence intervals captured  $z_{CC}^*$  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  $n$ s were used ranging from 5 through 10, 20, and 50, to 100. Thus, in total, 2.5 million trials were performed—that is, 100,000 trials times 5 levels of  $z_{CC}^*$ , times 5 levels of the control sample  $n$ .

The results of the simulation are set out in Table 1. 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 individuals' scores (the small departures from 95% are within the range expected from Monte

**Table 1. Monte Carlo simulation**

$z_{CC}^*$ (%)	95% interval estimate of effect size ( $z_{CC}$ )				
	$N = 5$	$N = 10$	$N = 20$	$N = 50$	$N = 100$
-0.253 (40)	94.94	94.93	94.96	94.95	95.00
-0.675 (25)	94.97	94.99	95.00	95.00	95.09
-1.036 (15)	94.99	94.89	94.95	95.01	94.92
-1.282 (10)	94.89	95.00	94.95	94.95	95.02
-1.645 (5)	95.04	94.96	94.92	95.01	95.06
-1.960 (2.5)	94.95	95.03	95.04	95.01	94.98
-2.326 (1)	94.98	94.94	95.02	95.05	95.02

*Note:* Results of a Monte Carlo simulation to verify that the proposed method of forming interval estimates of the effect size for the difference between a single case and control captures the true effect size ( $z_{CC}^*$ ) 95% of the time; the size of the control sample is varied (from 5 to 100) as is the size of the true effect size ( $z_{CC}^*$ ). Percentages shown in parentheses.

Carlo variation). Thus, these results provide empirical support for the statistical theory set out in Appendix A of the present paper. In other words, these results confirm the veracity of the method derived here for confidence limits on an effect size for the difference between a single case and controls. Moreover, as the Bayesian equivalent of these classical methods gives the same interval estimates, the simulation simultaneously confirms the veracity of the Bayesian approach to this problem.

### Recommendations for the analysis and reporting of statistical results involving basic comparisons of a single case to controls

The foregoing analysis and discussion suggests that the following information should be provided in single-case studies when comparing a case to controls: the mean and standard deviation for controls on the task (and control  $n$ ), the raw score of the single case, the point estimate of the effect size ( $z_{CC}$ ) for the difference between the case and controls accompanied by its 95% confidence interval (or credible interval), as set out in the present paper; the *t*-value and its associated probability obtained from application of Crawford and Howell's (1998) test or its Bayesian equivalent

(Crawford & Garthwaite, 2007); and finally, the point estimate and 95% confidence interval (or credible interval) for the abnormality of the case's score; these latter statistics are obtained by the classical methods developed by Crawford and Garthwaite (2002), or in the case of their Bayesian equivalent, by Crawford and Garthwaite (2007). The point estimate of the abnormality of the patient's score is the estimated percentage of the control population that would obtain a score lower than the case and the interval estimate quantifies the uncertainty over this percentage.

Provision of this information would, we suggest, provide the consumer of single-case studies with all the pertinent information required to come to an informed judgement concerning a single case's performance on the tasks in question. All of the aforementioned results can be obtained by use of computer programs described in a later section: These programs require only that the researcher enters the control  $n$ , the control mean and standard deviation for the task, and the case's score.

Given the amount of information to be reported, we recommend presenting it in a table along with the equivalent analysis performed on other tasks. An illustration of how this could be laid out is given in Table 2. This table sets out a hypothetical example of a single case administered four tasks where performance was compared to a

matched sample of 16 healthy controls. Note that the means and standard deviations for the controls on these four tasks are very different (as they often are in single-case studies given that fully normed and standardized tasks are rarely available to assess the constructs of interest). By expressing the differences between the case and controls case as effect sizes ( $z_{CC}$ ) the patient's relative standing on the four tasks can readily be assimilated by the reader.

In this example there are clear deficits on Tasks A and B: The case's scores are significantly poorer than those of the controls, the effect sizes are very large, and the case's score are highly abnormal—that is, a very small percentage of the control population are expected to exhibit scores as low as these. It can also be seen from the effect sizes that the deficit on Task B is markedly larger than that exhibited on Task A. In contrast, there is little or no evidence for a deficit on Tasks C and D: The case's scores do not differ significantly from controls, the effect sizes are fairly modest, and it is estimated that a large percentage of the control population would obtain scores lower than that observed for the patient.

The results for Task C were chosen intentionally to illustrate a feature of the interval on the effect size: The value 0 (zero) may be outside the 95% confidence interval for the effect size (in this example the interval is from -1.106 to -0.043) while the two-tailed  $t$  test is not

**Table 2.** An example table of results comparing a single case to controls in which the current proposals are implemented

Task	Control sample			Case's score	Significance test <sup>a</sup>		Estimated percentage of the control population obtaining a lower score than the case <sup>b</sup>		Estimated effect size ( $z_{CC}$ )	
	<i>n</i>	Mean	SD		<i>t</i>	<i>p</i>	Point	(95% CI)	Point	(95% CI)
Task A	16	12.78	3.45	4.0	-2.47	.013	1.30	(0.02 to 6.56)	-2.545	(-3.561 to -1.509)
Task B	16	46.3	8.20	19.0	-3.23	.003	0.28	(0.00 to 2.05)	-3.329	(-4.597 to -2.044)
Task C	16	30.4	14.42	22.0	-0.57	.290	29.0	(13.43 to 48.30)	-0.583	(-1.106 to -0.043)
Task D	16	22.2	6.20	21.0	-0.19	.427	42.7	(24.66 to 61.96)	-0.188	(-0.685 to 0.304)

*Note:* Includes reporting point and interval estimates of the effect size ( $z_{CC}$ ) for the differences between case and controls using the method set out in the present paper.

<sup>a</sup>Crawford & Howell (1998); the results are for a one-tailed test. <sup>b</sup>Crawford & Garthwaite (2002).

significant (Table 2 reports one-tailed  $p$  values; the two-tailed  $p$  value for this example is .580). Consequently, finding that the effect size interval does not contain 0 should not be taken as evidence that the case differs from controls. This situation cannot arise in group studies (see Appendix C for a brief explanation).

Returning to the general issue of how the results of single-case study should be reported: One of the advantages of setting out the results in a table, rather than in the text of a paper, is that the reader can immediately compare the patient's performance across the range of tasks administered. In reviewing the sample of single-case studies referred to earlier we found that many studies did provide at least some of the recommended information in table form for the results of *background* neuropsychological testing (where the need for effect sizes is perhaps not so pressing as the tests are often on a standard metric) but often abandoned this in favour of setting out the results from the experimental sections of the study entirely in the text.

Moreover, many single-case reports contain multiple studies of the single case and controls. In such circumstances the benefits of systematically setting out results in one large master table, or at least in a series of tables, are even more apparent—trying to obtain an overview of a patient's relative strengths and weaknesses across multiple studies is either time-consuming and frustrating (in cases where the results are distributed throughout the text of the relevant results section) or impossible (when the necessary information required to construct such a profile is omitted).

A further serendipitous advantage of setting out results in a table is that it prevents the practice of only reporting effect sizes for results that are statistically significant. In group studies such a practice is surprisingly common but has been rightly criticized because it flies in the face of the rationale underlying the reporting of effect sizes (Thompson, 2007). The statistical significance and the likely practical significance of results (the latter addressed by the effect sizes) are two different issues, and therefore the decision to report the latter should not be contingent on the former.

A decision to report effect sizes only for statistically significant results would also be particularly unfortunate and potentially misleading in single-case research where much emphasis is placed on a case's profile of performance across cognitive tasks. Where a case is found to have a statistically significant difference on one task but not on another, the difference in the case's relative level of performance can still be very trivial (Crawford, Garthwaite, & Gray, 2003); reporting of the effect sizes for both tasks allows an assessment of this possibility (where such a comparison is of central theoretical interest further formal testing of the difference between tasks is indicated; see the next section of the present paper).

Finally, Table 2 may also prove useful to those readers wishing to familiarize themselves with the use of the recommended methods and the accompanying computer programs that implement them: The table includes all the required inputs for the programs (i.e., the control means and standard deviations, the control sample size, and the case's scores), and all outputs. The data provided thus allow researchers to conduct a dry run ahead of applying the methods to their own data.

It may be that some readers consider the amount of information provided in Table 2 to be overkill. In contrast, in the course of reviewing the single-case literature for this paper, we were continually struck by what we see as an imbalance in the coverage of the different aspects of such studies. That is, much space was often devoted to a detailed review of the previous empirical literature and to the theory that motivated a given study, the Method sections also often presented very detailed descriptions of the design, materials, and administration of the tasks of interest, and Discussion sections provided in-depth considerations of the implications of the results obtained. That is all as it should be. However, set against this level of detail, the information provided on the inferential statistical methods employed was often very sparse. For example, it was not unusual for the reporting of results to be limited solely to  $p$  values (that is, the corresponding test statistics were omitted, as were the summary statistics for the control samples). An illustration of

this was provided in an earlier section by the disparity in the number of cases using  $z$  for significance testing and the number actually reporting the  $z$  value itself.

### Testing for differences (dissociations) in a case's performance across tasks

Although the detection of deficits is a fundamental feature of single-case studies, evidence of an impairment on a given task usually only becomes of theoretical interest if it is observed in the context of less impaired or normal performance on other tasks (Crawford & Garthwaite, 2005b). That is, much of the focus in single-case studies is on establishing dissociations of function (Caramazza & McCloskey, 1988; Coltheart, 2001; Crawford et al., 2003; Ellis & Young, 1996; Shallice, 1988).

The conventional criteria for a classical dissociation requires a researcher to demonstrate that a case is "impaired" (significantly different from controls) on task  $X$  and "unimpaired" (not significantly different from controls) on task  $Y$  (Crawford et al., 2003). Crawford et al. (2003) have argued that these criteria for a dissociation are insufficiently rigorous: for example, a case could be just below a designated cut-off for impairment on task  $X$  and just above the cut-off on task  $Y$  so that the difference in the case's relative standing on the two tasks is trivial. Subsequent Monte Carlo simulation studies have supported this position. When a Type I error was defined as misclassifying a healthy control as exhibiting a dissociation, the conventional criteria generated very high error rates (Crawford & Garthwaite, 2005a); the rates were even higher (they approached 50% in some scenarios) when a Type I error was defined as misclassifying a patient with strictly equivalent deficits on both tasks as exhibiting a dissociation (Crawford & Garthwaite, 2006).

The solution to these problems is to test for a difference between a case's scores on the two tasks. This deals with the issue of trivial differences referred to earlier and, unlike the conventional criteria (which rely on attempting to establish an

absence of a difference between case and controls on one of the tasks), it also provides us with a positive test for a dissociation. A complication in testing for a difference between a case's scores on two tasks is that the tasks in question are usually expressed on different metrics—that is, they differ in their means and standard deviations. Therefore, there is a need to standardize a case's scores on each of the two tasks before they can be meaningfully compared; see Crawford et al. (2009) for a demonstration of the problems that can ensue when there is a failure to standardize a case's scores.

We consider two methods of testing for a difference between a case's scores: one based on classical statistics, the other on a Bayesian approach. Considering the classical solution first: Following standardization, the patient's scores on the two tasks are both  $t$ -variates. Therefore a method is required that tests for a difference between two  $t$ -variates. This was the approach adopted by Crawford and Garthwaite (2005a) in developing the Revised Standardized Difference Test (RSDT). The test is based on asymptotic expansions performed to obtain the standard error for such differences (Garthwaite & Crawford, 2004). However, an even better solution is to use a Bayesian method developed by Crawford and Garthwaite (2007), the Bayesian Standardized Difference Test (BSDT). Among the advantages of this latter test is that it provides not just a hypothesis test, but also a point and interval estimate of the abnormality of the difference between the case's scores; see Crawford and Garthwaite (2007) and Crawford et al. (2009) for a discussion of its other advantages.

In summary: Methods are available to test for a difference between a case's scores on two tasks, and, in the case of the BSDT, it is also possible to obtain a point and interval estimates of the abnormality of the case's difference. The aim in the present paper is to supplement these methods by providing point and interval estimates of the effect size for the difference between case and controls (that is, we require an index that compares the difference between tasks observed for the single case with the differences observed in controls).

## Point and interval estimates of effect size when testing for dissociations

The point estimate of the effect size for this problem is easily obtained. The case's scores on the two tasks are converted to  $z$  scores based on the control means and standard deviations for the two tasks and their difference divided by the standard deviation of the difference between two nonindependent  $z$  scores; this produces a  $z$  score for the difference. We denote this index as  $z_{DCC}$  (the “ $D$ ” suffix in the subscript differentiates it from the index presented earlier and identifies that here we are concerned with the difference between tasks). The formula for the index is

$$\begin{aligned} z_{DCC} &= \frac{[(x - \bar{x})/s_x] - [(y - \bar{y})/s_y]}{\sqrt{2 - 2r_{xy}}} \\ &= \frac{z_x - z_y}{\sqrt{2 - 2r_{xy}}}, \end{aligned} \quad (2)$$

where  $r_{xy}$  = the correlation between the two tasks in the control sample, and all other terms are obvious. Note that the sign of  $z_{DCC}$  is essentially arbitrary as it depends on which task a researcher designates as task  $X$  and which as task  $Y$ . For example, if the case's standardized score on task  $X$  is higher than the standardized score on task  $Y$ , then  $z_{DCC}$  will be positive. Note also that the mean difference between standardized scores in controls is necessarily zero.

In passing, the statistic is used here as an effect size index but it could also be used as an approximate hypothesis test for a difference between case and controls when the control statistics have been obtained from a large normative sample (Payne & Jones, 1957). However, it is decidedly not suitable for this latter purpose with the modest control samples that typify single-case research. In contrast to the hypothesis tests reviewed earlier (i.e., the RSDT and BSDT), the Type I error rate is very inflated: In simulations conducted by Crawford and Garthwaite (2005) rates as high as 25% were observed for a nominal error rate of 5%.

Having specified a point estimate of effect size for the difference between a case's scores versus

those of controls it only remains to supplement this with an interval estimate for this quantity. This can be achieved using a Bayesian approach. En route to obtaining an interval estimate for the abnormality of a cases difference, the Bayesian Monte Carlo method used in the BSDT repeatedly applies Equation 2 using the case's scores and random draws from the posterior distribution for controls as inputs (we denote the values obtained from these iterations as  $\hat{z}_{DCC}$ ). If, say, one million Monte Carlo iterations are performed, then the 25,000th lowest and 25,000th highest values of  $\hat{z}_{DCC}$  provide the lower and upper end points for a 95% credible interval for the true effect size. The formal procedure for obtaining this interval is set out in Appendix D.

To illustrate, take the scores of the single case recorded in Table 2 and suppose that we are interested in the difference between the case's performance on Tasks B and D. All the data required to calculate the point estimate of the effect size for this difference can be obtained from Table 2 except for the correlation between the two tasks in the control sample; let us suppose that this correlation is .65. Dividing the difference between the case's  $z$  scores on these two tasks,  $(-3.329) - (-0.188) = -3.517$ , by the standard deviation of the difference (0.837) yields the point estimate of the effect size,  $z_{DCC} = -3.748$ .

It can be seen that the effect size for the case's difference is very large—that is, the case's difference is well over three standard deviations from the mean difference in controls (the mean difference in controls is of course zero). The 95% interval for the effect size is  $(-5.496, -2.244)$ . Using the Bayesian interpretation of this interval, we can be 95% confident that the true effect size lies in this interval. This interval is wide (not surprising given that control sample size is modest) but, in this example, it can be seen that even the upper limit is very extreme. It is therefore clear that the case shows a very large and striking dissociation between her/his performance on the two tasks.

## Computer programs that incorporate point and interval estimates of effect sizes for the case-controls design

The point estimates for the effect size indexes presented in the present paper could easily be calculated by hand. However, both the classical and Bayesian methods for obtaining interval estimates of effect sizes require a computer. We have therefore implemented the point and interval estimates in a series of six computer programs. These programs are all upgraded versions of earlier programs developed by two of the present authors and their colleagues. In the interests of clarity and continuity the upgraded programs retain the same names as the originals but are given an “ES” (effect size) suffix. Table 3 lists the upgraded programs and provides a short description of their purpose.

As an example, the original program Singlims.exe provides a classical hypothesis test when comparing a case to controls (i.e., it applies Crawford & Howell's, 1998, method) and also a point and interval estimate of the abnormality of the case's test score (Crawford & Garthwaite, 2002). The upgraded version, Singlims\_ES.exe, provides the same results but supplements them with a point estimate and 95% interval estimate of the effect size for the difference between the case and controls. To illustrate the general features of these programs Figure 1 contains screen captures of the input form and output form for Singlims\_ES.exe. The input data (control mean, control  $SD$ , control  $n$ , and score for the case) and the results (hypothesis test results, i.e.,  $t$  and its associated one- and two-tailed probabilities; point and interval estimates of the effect size for

**Table 3.** Computer programs incorporating point and interval estimates of effect sizes

Computer program	Description
Singlims_ES.exe	This program is an upgraded version of the program Singlims.exe (Crawford & Garthwaite, 2002). It implements classical methods for comparison of a single case's score to scores obtained in a control sample. The interval estimate of the effect size for the difference between case and controls is obtained using classical methods.
SingleBayes_ES.exe	This program is an upgraded version of the program SingleBayes.exe (Crawford & Garthwaite, 2007). It implements Bayesian methods for comparison of a single case's score to scores obtained in a control sample. The interval estimate of the effect size for the difference between case and controls is obtained using Bayesian methods.
RSDT_ES.exe	This program is an upgraded version of the program RSDT.exe (Crawford & Garthwaite, 2005). It implements classical methods to test for a difference between a single case's scores on two tasks by comparing the difference against differences observed in a control sample. Note that, although the hypothesis test is a classical test, the interval estimate of the effect size is obtained using Bayesian methods.
DiffBayes_ES.exe	This program is an upgraded version of the program DiffBayes.exe (Crawford & Garthwaite, 2007). It implements Bayesian methods to test for a difference between a single case's scores on two tasks by comparing the difference against differences observed in a control sample. The interval estimate of the effect size is obtained using Bayesian methods.
Dissocs_ES.exe	This program is an upgraded version of Dissocs.exe (Crawford & Garthwaite, 2005). It tests whether a single case meets criteria for a dissociation using classical statistical methods. The interval estimates of the effect size for the difference between the case's score and controls on each of the two tasks is obtained using classical methods; the interval estimate of the effect size for the difference between tasks is obtained using Bayesian methods. Note also that the upgraded version now offers the option of using a one-tailed test when testing for a difference between a case's $X$ and $Y$ scores (a two-tailed test remains as the default).
DissocsBayes_ES.exe	This program is an upgraded version of Bayes_Dissocs.exe (Crawford & Garthwaite, 2007). It tests whether a single case meets criteria for a dissociation using Bayesian statistical methods. All interval estimates of effect size are obtained using Bayesian methods. Note also that the upgraded version now offers the option of using a one-tailed test when testing for a difference between a case's $X$ and $Y$ scores (a two-tailed test remains as the default).

(a)

User's Notes: Illustrative example from Table 1, Task A

Mean of the control or normative sample:

Standard deviation for the normative sample:

Sample size of the normative sample:

Case's test score:

Compute      Clear Data      Exit

(b)

Results viewer: Singlims\_ES.exe: Significance test and point & interval estimates of effect size and abnormality for a patient...

Printer options...

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

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. & Howell, D.C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, 12, 482-486.

**INPUTS:**  
 Mean of the control or normative sample= 12.78  
 Standard deviation for the normative sample= 3.45  
 Sample size of the normative sample= 16  
 Case's test score= 4

**OUTPUTS:**  
 Significance test on difference between case's score and control sample:  
 Significance test: t value = -2.469  
 Significance test: One-tailed probability = 0.01303  
 Significance test: Two-tailed probability = 0.02605  
 Effect Size (Z-CC) for difference between case and controls (plus 95% CI) = -2.545 (-3.561 to -1.509)  
 Estimated percentage of normal population falling below case's score = 1.30264%  
 95% lower confidence limit on the percentage = 0.01846%  
 95% upper confidence limit on the percentage = 6.56209%

Save Output      Clear Results      Return to Worksheet      Exit

Figure 1. Screen captures of input form (a) and results form (b) for the program Singlims\_ES.exe: The data are those recorded in Table 2 for Task A.

the difference between case and controls, and point and interval estimate of the abnormality of the case's score) are those given in Table 2 for Task A.

Some of the original programs use classical statistical methods (Singlims.exe, RSDT.exe, and Dissocs.exe); the remainder use Bayesian methods. In the upgraded versions the methods used to obtain the interval estimates are matched (i.e., Bayesian hypothesis test with Bayesian interval estimate, etc.). The exception to this occurs for programs that test for a difference between a case's scores on two tasks. Classical statistical methods cannot capture the uncertainties involved in this problem, and so the Bayesian method is implemented, regardless of whether the other results reported are based on a classical or Bayesian analysis (the output of these programs clearly identifies the interval estimate as a Bayesian estimate). In all programs that use the Bayesian method one million iterations are performed (hence there is a short delay before the results are obtained). Compiled versions of the modified programs, written for PCs (in the Delphi programming language), can be downloaded individually, or as a single zip file, from the following web page: [www.psyc.abdn.ac.uk/~psy086/dept/Single\\_Case\\_Effect\\_Sizes.htm](http://www.psyc.abdn.ac.uk/~psy086/dept/Single_Case_Effect_Sizes.htm)

## Conclusion

It is to be hoped that the present paper will encourage single-case researchers to report point and interval estimates of effect size in their studies. Such a move would be in keeping with the general principle that standards of reporting in single-case research should be as stringent as those demanded in group-based research. The computer programs written to accompany this paper provide a convenient and reliable means of obtaining the effect size statistics.

Manuscript received 15 April 2010  
 Revised manuscript received 14 July 2010  
 Revised manuscript accepted 23 July 2010  
 First published online 9 October 2010

## REFERENCES

- Antelman, G. (1997). *Elementary Bayesian statistics*. Cheltenham, UK: Elgar.
- Becker, J. T., Knowlton, B., & Anderson, V. (2005). Editorial. *Neuropsychology*, 19, 3–4.
- Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology*, 23, 399–406.
- Caramazza, A., & McCloskey, M. (1988). The case for single-patient studies. *Cognitive Neuropsychology*, 5, 517–528.
- Coltheart, M. (2001). Assumptions and methods in cognitive neuropsychology. In B. Rapp (Ed.), *The handbook of cognitive neuropsychology* (pp. 3–21). Philadelphia: Psychology Press.
- 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. (2005a). Evaluation of criteria for classical dissociations in single-case studies by Monte Carlo simulation. *Neuropsychology*, 19, 664–678.
- Crawford, J. R., & Garthwaite, P. H. (2005b). 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). Detecting dissociations in single case studies: Type I errors, statistical power and the classical versus strong distinction. *Neuropsychologia*, 44, 2249–2258.
- 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., & Gray, C. D. (2003). Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex*, 39, 357–370.
- Crawford, J. R., Garthwaite, P. H., & Howell, D. C. (2009). On comparing a single case with a control sample: An alternative perspective. *Neuropsychologia*, 47, 2690–2695.
- Crawford, J. R., & Henry, J. D. (2004). Assessment of executive deficits. In P. W. Halligan & N. Wade (Eds.), *The effectiveness of rehabilitation for cognitive deficits* (pp. 233–245). Oxford, UK: Oxford University Press.

- 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.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532–574.
- DeGroot, M. H., & Schervish, M. J. (2001). *Probability and statistics* (3rd ed.). Reading, MA: Addison-Wesley.
- Ellis, A. W., & Young, A. W. (1996). *Human cognitive neuropsychology: A textbook with readings*. Hove, UK: Psychology Press.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, 61, 575–604.
- Garthwaite, P. H., & Crawford, J. R. (2004). The distribution of the difference between two *t*-variates. *Biometrika*, 91, 987–994.
- Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., et al. (2005).
- Effect sizes in single case research: How large is large? *School Psychology Review*, 34(1), 116–132.
- Payne, R. W., & Jones, G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, 13, 115–121.
- Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.
- Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164–182.
- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44(5), 423–432.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Zakzanis, K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology*, 16, 653–667.

## APPENDIX A

### Derivation of classical confidence intervals on the effect size ( $z_{CC}$ ) for the difference between a single-case's score and a control sample

The confidence intervals given in this paper are derived from a noncentral *t*-distribution and are based on theory given in Crawford and Garthwaite (2002). The noncentral *t*-distribution is defined by

$$T_v(\delta) = (Z + \delta)/\sqrt{Y/v},$$

where  $Z$  has a normal distribution with a mean of zero and variance 1, and  $Y$  is independent of  $Z$  with a chi-square distribution on  $v$  degrees of freedom.  $\delta$  is referred to as the noncentrality parameter.

For a specified value  $x_0$ , we require a 100(1 -  $\alpha$ )% confidence interval on the effect size index,  $z_{CC}$ , based on sample data  $\bar{x}$  and  $s^2$ , where  $\bar{x} \sim N(\mu, \sigma^2/n)$  and  $vs^2/\sigma^2 \sim \chi^2(v)$ . (In the present case  $v = n - 1$ ). Put

$$z_{CC} = \frac{(x_0 - \bar{x})}{s}. \quad (3)$$

Let  $z_{CC}^* = (x_0 - \mu)/\sigma$ . Then  $z_{CC}$  is an estimate of  $z_{CC}^*$ . That is,  $z_{CC}$  computed using the *sample* mean and standard deviation

is an estimate of the  $z$ , here denoted  $z_{CC}^*$ , that we would obtain were the mean and standard deviation of the control *population* known. Now

$$z_{CC}\sqrt{n} = \left( \frac{(\mu - \bar{x})\sqrt{n}}{\sigma} + \frac{(x_0 - \mu)\sqrt{n}}{\sigma} \right) / \sqrt{\frac{s^2}{n}},$$

so  $z_{CC}\sqrt{n}$  has a noncentral *t*-distribution with noncentrality parameter  $\delta = z_{CC}^*\sqrt{n}$  and  $v$  degrees of freedom. The 100( $\alpha/2$ )% and 100(1 -  $\alpha/2$ )% points of this distribution will depend on the value of  $\delta$ . Let  $\delta_L$  denote the value of  $\delta$  for which the 100(1 -  $\alpha/2$ )% point is  $z_{CC}\sqrt{n}$ . (The value of  $\delta$  is obtained by a search algorithm that finds the noncentrality parameter of a noncentral *t*-distribution from a quantile, its associated probability, and a specified degree of freedom). Similarly, let  $\delta_U$  denote the value of  $\delta$  for which the 100( $\alpha/2$ )% point is  $z_{CC}\sqrt{n}$ . Then  $(\delta_L/\sqrt{n}, \delta_U/\sqrt{n})$  is a 100(1 -  $\alpha$ )% confidence interval for  $z_{CC}^*$ .

As a fully worked example of finding the 95% two-sided confidence interval on the effect size for the difference between a case and controls, take the first example given in the text, where  $\bar{x}$  was 50,  $s$  was 10,  $n$  was 10, and  $x_0$  was 24. Then  $z_{CC} = -2.600$ , and  $z_{CC}\sqrt{n} = -8.2219$ . Then, to obtain the lower limit, we find the noncentrality parameter for the noncentral *t*-distribution on  $v = n - 1 = 9$  *df* that has -8.2219 as its 0.975th percentile point. The noncentrality parameter is -12.3933, and dividing this by  $\sqrt{n}$  gives

–3.9191. To find the upper limit we find the noncentrality parameter for the noncentral  $t$ -distribution on  $v = n - 1 = 9$   $df$  that has –8.2219 as its 0.025th percentile point. The noncentrality parameter is –3.9632, and dividing this by  $\sqrt{n}$  gives –1.2533. Thus, for this example, the 95% confidence interval on the effect size for the difference between the case and controls is from –3.9191 to –1.2533.

## APPENDIX B

### Obtaining a Bayesian credible interval on the effect size ( $z_{CC}$ ) for the difference between a single-case's score and a control sample

The Bayesian credible interval on the effect size is obtained by a simple extension of methods developed by Crawford and Garthwaite (2007) for comparison of a case to controls. We measure the value of  $x$  on a sample of  $n$  controls. Let  $\bar{x}$  denote the sample mean and  $s^2$  denote the sample variance. We assume each  $x$  comes from a normal distribution with unknown mean  $\mu$  and unknown variance  $\theta$  ( $\theta = \sigma^2$  in standard notation). A single case has a value of  $x^*$ .

We start with a noninformative prior distribution. Specifically, we suppose the prior conditional distribution of  $\mu$  given  $\theta$  is  $\mu|\theta \sim N(0, \infty)$ , and the prior marginal distribution of  $\theta$  is proportional to  $\theta^{-1}$ . This is the standard noninformative prior distribution when data are from a normal distribution. The posterior distribution is obtained by combining the prior distribution with the data, and inferences and estimates are based on the posterior distribution. The posterior distribution states that the marginal distribution of  $(n - 1)s^2/\theta$  is a chi-squared distribution on  $n - 1$  degrees of freedom, and, given  $\theta$ , the conditional posterior distribution of  $\mu$  is a normal distribution with mean  $\bar{x}$  and variance  $\theta/n$  (see, for example, DeGroot & Schervish, 2001). The following iterative procedure is then followed to obtain an interval estimate of  $z_{CC}$ :

1. Generate a random value from a chi-square distribution on  $n - 1$   $df$ . Let  $\psi$  denote the generated value. Put  $\hat{\theta} = (n - 1)s^2/\psi$ . Then  $\hat{\theta}$  is the estimate of  $\theta$  for this iteration.
2. Generate an observation from a standard normal distribution. Call this generated value  $z$ . Put

$$\hat{\mu} = \bar{x} + z\sqrt{\hat{\theta}/n}. \quad (4)$$

Then  $\hat{\mu}$  is the estimate of  $\mu$  for this iteration.

3. We have estimates of  $\mu$  and  $\theta$ . We calculate the value of  $z_{CC}$  conditional on these being the correct values of  $\mu$  and  $\theta$ . That is, we put

$$\hat{z}_{CC} = (x^* - \hat{\mu})/\sqrt{\hat{\theta}}. \quad (5)$$

4. We repeat Steps 1 to 3 a large number of times; in the present case we will perform one million iterations. To obtain a 95% Bayesian credible interval on the effect size,

we take the 25,000th smallest  $\hat{z}_{CC}$  and the 25,000th largest  $\hat{z}_{CC}$ . Note that if a one-sided 95% credible limit is required then (again assuming one million iterations have been performed) we simply take the 50,000th smallest  $\hat{z}_{CC}$  to obtain the lower limit, or the 50,000th largest  $\hat{z}_{CC}$  for the upper limit.

## APPENDIX C

### Relationship between interval estimate of the effect size and significance test results in single-case studies versus group studies

As noted in the text, the value 0 (zero) will very occasionally be outside the 95% confidence interval for the effect size while the  $t$  test is not significant (two-tailed). This situation cannot arise in group studies. That is, an interval for Cohen's  $d$  or Hedges's  $g$  that excludes zero cannot occur in combination with a non-significant  $t$  test. To see this, suppose that  $\mu_x$  and  $\mu_y$  are the population means for the controls and the patients, respectively. Then, as is well known, a 95% confidence interval for  $\mu_x - \mu_y$  will not contain the value 0 if and only if the hypothesis that  $\mu_x = \mu_y$  is rejected at the .05 significance level in favour of the two-tailed alternative hypothesis. Scaling the end points of a confidence interval by a nonzero quantity will not influence whether the interval contains 0. Hence, a 95% confidence interval for  $(\mu_x - \mu_y)/\sigma$  will have the same property.

In single-case studies, this equivalence between hypothesis tests and confidence intervals does not hold. The reason is that the hypothesis test and confidence interval relate to different random variables. The hypothesis test concerns the question "Could the cases score,  $x_0$ , be the score of a control", and in this test both  $x_0$  and the mean of the controls,  $\bar{x}$ , are treated as random variables. In contrast, the confidence interval for  $(x_0 - \mu_x)/\sigma$  treats  $\mu_x$  as a quantity that must be estimated ( $\bar{x}$  is the estimate) but it treats  $x_0$  as a fixed, known quantity.

## APPENDIX D

### Obtaining a Bayesian credible interval on the effect size ( $z_{DCC}$ ) for a single-case's difference between two tasks

We assume the two tasks ( $x$  and  $y$ ) follow a bivariate normal distribution in the control population. We have a control sample of  $n$  individuals from whom to estimate  $\underline{\mu}$ , the vector of population means, and  $\underline{\Sigma}$ , the variance-covariance matrix of the control population. The control sample data are combined with a noninformative prior distribution (i.e., a prior distribution that assumes no prior knowledge or data) to obtain the posterior distribution of  $\underline{\Sigma}$ . For this problem we use  $f(\underline{\mu}, \underline{\Sigma}^{-1}) \propto |\Sigma|$  as the prior. The posterior distribution of  $\underline{\Sigma}$

takes the form of a Wishart distribution, and the conditional distribution of  $\underline{\mu}$  given  $\Sigma$  follows a bivariate normal distribution. The following procedure yields the statistics required:

1. Generate estimates of  $\underline{\mu}$  and  $\Sigma$ . The estimates of  $\Sigma$  are obtained by random sampling from an inverse Wishart distribution, and in turn these estimates (in combination with random draws from a standard normal distribution) are then used to obtain estimates of  $\underline{\mu}$ ; for procedural details see Crawford and Garthwaite (2007). Let  $\hat{\underline{\mu}}_i$  and  $\hat{\Sigma}_i$  identify these estimates.
2. Convert the case's scores ( $x^*$  and  $y^*$ ) on the two tasks to  $z$  scores using the estimated means ( $\hat{\mu}_{x_i}$  and  $\hat{\mu}_{y_i}$ ) and standard deviations ( $\hat{s}_{x_i}$  and  $\hat{s}_{y_i}$ ). Then divide the difference between these  $z$  scores by the estimated standard deviation of their difference, and denote the result as  $\hat{z}_{DCC_i}$ . That is, put

$$\hat{z}_{DCC_i} = \frac{[(x^* - \hat{\mu}_{x_i})/\hat{s}_{x_i}] - [(y^* - \hat{\mu}_{y_i})/\hat{s}_{y_i}]}{\sqrt{2 - 2\hat{\rho}_{xy_i}}}. \quad (6)$$

Note that the denominator requires the estimated correlation between tasks  $X$  and  $Y$  ( $\hat{\rho}_{xy_i}$ ). This is obtained from the estimated variances and covariances—that is,

$$\hat{\rho}_{xy_i} = \frac{\hat{s}_{xy_i}}{\sqrt{\hat{s}_{x_i}^2 \hat{s}_{y_i}^2}} = \frac{\hat{\Sigma}_{12_i}}{\sqrt{\hat{\Sigma}_{11_i} \hat{\Sigma}_{22_i}}} \quad (7)$$

3.

Repeat Steps 1 to 2 a large number of times; for the present problem we chose to perform one million iterations. To obtain a 95% Bayesian credible interval on the effect size, we take the 25,000th smallest  $\hat{z}_{DCC_i}$  and the 25,000th largest  $\hat{z}_{DCC_i}$ . If a one-sided 95% credible limit is required then (again assuming one million iterations have been performed) we take the 50,000th smallest  $\hat{z}_{DCC_i}$  to obtain the lower limit, or the 50,000th largest  $\hat{z}_{DCC_i}$  for the upper limit.