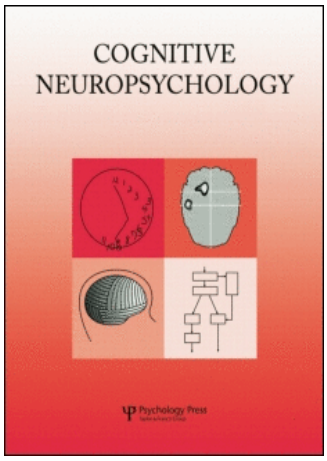

Cognitive Neuropsychology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713659042

Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach

Online Publication Date: 01 June 2007
To cite this Article: Crawford, John R. and Garthwaite, Paul H. , (2007) 'Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach', Cognitive Neuropsychology, 24:4, 343 - 372
To link to this article: DOI: 10.1080/02643290701290146
URL: http://dx.doi.org/10.1080/02643290701290146

COGNITIVE NEUROPSYCHOLOGY, 2007, 24 (4), 343–372

Psychology Press
Taylor & Francis Group

Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach

John R. Crawford
University of Aberdeen, Aberdeen, UK

Paul H. Garthwaite
Department of Statistics, The Open University, Milton Keynes, UK

Frequentist methods are available for comparison of a patient's test score (or score difference) to a control or normative sample; these methods also provide a point estimate of the percentage of the population that would obtain a more extreme score (or score difference) and, for some problems, an accompanying interval estimate (i.e., confidence limits) on this percentage. In the present paper we develop a Bayesian approach to these problems. Despite the very different approaches, the Bayesian and frequentist methods yield equivalent point and interval estimates when (a) a case's score is compared to that of a control sample, and (b) when the raw (i.e., unstandardized) difference between a case's scores on two tasks are compared to the differences in controls. In contrast, the two approaches differ with regard to point estimates of the abnormality of the difference between a case's *standardized* scores. The Bayesian method for standardized differences has the advantages that (a) it can directly evaluate the probability that a control will obtain a more extreme difference score, (b) it appropriately incorporates error in estimating the standard deviations of the tasks from which the patient's difference score is derived, and (c) it provides a credible interval for the abnormality of the difference between an individual's standardized scores; this latter problem has failed to succumb to frequentist methods. Computer programs that implement the Bayesian methods are described and made available.

In the typical single-case study in neuropsychology the performance of a patient on a task or series of tasks is compared to that of a healthy control sample. A fundamental question in such studies is how to infer the presence of a cognitive impairment in the patient. The method used most commonly is to convert the case's score on a given measure (e.g., an ability test) to a z score based on the mean and standard deviation of the controls and then refer this score to a table of areas under the normal curve (Howell, 2002).

Thus, if a researcher has formed a directional hypothesis for a case's score prior to testing (i.e., that the case has an acquired impairment, and the score will therefore be below the control sample mean), then a score that fell below -1.645 would be considered statistically significant ($p < .05$) and would be taken as an indication

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 2UB, UK (E-mail: j.crawford@abdn.ac.uk).

http://www.psypress.com/cogneuropsychology DOI:10.1080/02643290701290146

that the case was not an observation from the control population—that is, it would be concluded that the case exhibits a deficit on the task in question.

One problem with this approach is that it treats the control sample as if it was a population—that is, the mean and standard deviation are used as if they were *parameters* rather than *sample statistics*. In many contexts this is not a problem in practice as the normative or control sample is large and therefore should provide sufficiently accurate estimates of the parameters (as when, for example, a clinical neuropsychologist compares a patient's test scores against norms obtained from a large standardization sample).

However, the control samples in single-case studies in neuropsychology are typically modest in size; *N*s of around 10 are not unusual (Crawford & Howell, 1998). With samples of this size it is not appropriate to treat the mean and standard deviation as though they were parameters.

One solution to this problem is to use a method described by Crawford and Howell (1998) that treats the control sample statistics *as* sample statistics. Their approach is based on a formula for a modified *t* test given by Sokal and Rohlf (1995). It uses the *t*-distribution (with $n - 1$ degrees of freedom), rather than the standard normal distribution, to estimate the abnormality of the individual's scores and to test whether it is significantly lower than the scores of the control sample.

Theory tells us that the effect of using *z* with small control samples will be to exaggerate the rarity/abnormality of an individual's score and to inflate the Type I error rate (in this context a Type I error occurs when a case that is drawn from the control population is incorrectly classified as not being a member of this population; i.e., they are incorrectly classified as exhibiting a deficit). Monte Carlo simulation studies confirmed this empirically and also demonstrated that, in contrast, Crawford and Howell's (1998) method controls Type I errors (Crawford & Garthwaite, 2005b).

The formula for Crawford and Howell's (1998) method is

$$t = \frac{x^* - \overline{x}}{s\sqrt{(n+1)/n}} \quad (1)$$

where x^* is the patient's score, $\overline{x}$ and s are the mean and standard deviation of scores in the control sample, and n is the size of the control sample. The p value obtained when this test is used to test significance also simultaneously provides a point estimate of the abnormality of the patient's score; for example if the one-tailed probability is .013 then we know that the patient's score is significantly ($p < .05$) below the control mean and that it is estimated that ($p \times 100 =$) 1.3% of the control population would obtain a score lower than the patient's.

As noted by Crawford and Howell (1998), this point estimate of abnormality is a useful complement to the significance test given that the use of an alpha of .05 is essentially an arbitrary convention (albeit one that has, in general, served science well). A formal proof that the p value for the significance test and the point estimate of the abnormality of the score are equal is provided in Crawford and Garthwaite (2006b). In keeping with the contemporary emphasis on the use of confidence intervals, Crawford and Garthwaite (2002) developed a method to provide accompanying confidence limits on the abnormality of the individual's score.

THE BAYESIAN APPROACH TO INFERENCE

The methods outlined above can be classified as "classical" or "frequentist" statistical methods. Bayesian statistics provides an alternative to the classical approach to inference. The essential difference between the classical and Bayesian approaches is that the classical approach treats parameters as *fixed* but unknown whereas, in the Bayesian approach, parameters are treated as random variables and hence have probability distributions. For example, suppose μ is the (unknown) mean IQ in a specified population. Then a Bayesian might say "the probability that

μ is less than 90 is .95" while classical statistics does not permit a probabilistic statement about μ. With the classical approach, μ is either less than 90 or it is not, and there is no random uncertainty, so the probability that μ is less than 90 is either 1 or 0 and nothing in between.

In the Bayesian approach, a *prior* distribution is used to convey the information about model parameters that was available before the sample data were gathered. This is combined with the information supplied by the data, which is contained in the *likelihood*, to yield a *posterior* distribution. Formally,

$$\text{posterior} \propto \text{prior} \times \text{likelihood},$$

where $\propto$ means "is proportional to". Prior distributions are both a strength and a weakness of Bayesian statistics. They are a strength in that they enable background knowledge to be incorporated into a statistical analysis, but they are a weakness because a prior distribution *must* be specified, even if no useful background knowledge is available or if one does not wish to use background knowledge, perhaps to ensure the analysis is transparently impartial, or because it is difficult and time consuming to specify a prior distribution that accurately portrays prior knowledge. For these reasons, in practice a prior distribution is almost always chosen in a mechanical way to yield a flat, noninformative prior distribution.

There has been an explosion of interest in Bayesian statistical methods over the last 15 years. The reason for this is the development of numerical techniques, notably Markov Chain Monte Carlo (MCMC) methods, which have solved many of the remaining computational problems formally associated with practical application of the Bayesian paradigm. With these new techniques, Bayesian methods can now handle complex problems that cannot be tackled by frequentist approaches (Garthwaite, Jolliffe, & Jones, 2002). This has led to a general acceptance of Bayesian methods. (Before the development of MCMC methods and the power they unleashed, a sizeable proportion of statisticians regarded the classical approach as the only valid way of analysing data.) MCMC methods make inferences by generating a large number of observations from the posterior distribution, thus capturing the information in that distribution. The approach of sampling from posterior distributions is adopted in the methods developed here, but the problems addressed are reasonably tractable, and, while we use Monte Carlo methods, we do not need to run any Markov chains.

In the present paper we develop Bayesian methods for the analysis of the typical single-case study design in which a patient is compared to a matched control sample. Although the Bayesian approach is used increasingly in other areas of scientific enquiry, to our knowledge it has not previously been applied to this problem. We therefore incorporate explicit details of how to set up a Bayesian analysis for this type of problem in the hope that this may encourage other researchers or methodologists to develop additional Bayesian single-case methods.

EXPERIMENT 1

Development of a Bayesian analysis for the simple difference between a case and controls

As a starting point, we develop a Bayesian approach for the simple but fundamental problem of detecting an impairment in a single case. The aim is to obtain a Bayesian point estimate of the abnormality of the case's score (i.e., an estimate of the percentage of the control population that would obtain a score lower than the patient's) and an accompanying interval estimate of this quantity. A second aim is to compare the results obtained from a Bayesian analysis of this problem to the corresponding frequentist methods for obtaining a point estimate (Crawford & Howell, 1998) and interval estimate (Crawford & Garthwaite, 2002) of this quantity.

Method

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 μ and unknown variance θ ($\theta = \sigma^2$ in standard notation). A single case has a value of x^*. We want to estimate p, the proportion of controls who have a value of x that is less than x^*; p therefore will provide an estimate of the abnormality of the case's score. If p is small (say less than .05) then the hypothesis that the case is an observation from the control population is unlikely, and there is evidence that the patient has a deficit.

In addition to the point estimate of p we also want an interval estimate (i.e., we want an interval estimate of the abnormality of the case's score). This Bayesian interval estimate is called a credible interval rather than a confidence interval.

We start with a noninformative prior distribution. Specifically, we suppose the prior conditional distribution of μ given θ is $\mu|\theta \sim N(0,\infty)$, and the prior marginal distribution of θ is proportional to θ^{-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 θ, the conditional posterior distribution of μ is a normal distribution with mean $\overline{x}$ and variance θ/n (see for example DeGroot & Schervish, 2001). The following iterative procedure is then followed to obtain estimates of p:

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

$$\hat{\mu} = \overline{x} + z\sqrt{(\hat{\theta}/n)} \qquad (2)$$

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

3. We have estimates of μ and θ. We calculate the value of p conditional on these being the correct values of μ and θ. That is, we put

$$z^* = (x^* - \hat{\mu})/\sqrt{\hat{\theta}} \qquad (3)$$

and find the tail-area of a standard normal distribution that is less than z^*. This tail area is an estimate of p, which we call $\hat{p}_i$ for the ith iteration.
4. We repeat Steps 1 to 3 a large number of times; in the present case we will perform 100,000 iterations. Then the average value of $\hat{p}_1, \ldots, \hat{p}_{100,000}$ is the point estimate of p. (It is the Bayesian posterior mean of p.) To obtain a 95% Bayesian credible interval, we take the 2,500th smallest $\hat{p}_i$ and the 2,500th largest $\hat{p}_i$. Call these values p_l and p_u. Then the 95% Bayesian credible interval is (p_l, p_u). Note that if a *one*-sided 95% credible limit is required then (again assuming 100,000 iterations have been performed) we simply take the 5,000th smallest $\hat{p}_i$ to obtain the lower limit or the 5,000th largest $\hat{p}_i$ for the upper limit.

Note that Equation 3 is essentially the familiar formula for a z *score* (i.e., a standard score). As noted, when z is used in frequentist methods to test for a deficit (i.e., the sample mean and standard deviation are plugged directly into the formula for z, and inferences are drawn) the result is a marked inflation of the Type I error rate and an exaggeration of the abnormality of the case's score when the control sample n is small (Crawford & Garthwaite, 2005b). However, in the present approach, values for the mean and standard deviation are repeatedly drawn at random from their posterior distributions and are entered into the formula, thus allowing for the uncertainty of these quantities.

Results and discussion

Comparison of frequentist and Bayesian methods

In order to compare the Bayesian and frequentist methods of comparing a case to a control sample we examined their performance over a range of

N and x^*. With no loss of generality we fixed the mean ($\bar{x}$) and standard deviation (s) of the control sample to be 100 and 10, respectively. We varied the size of the control sample (for Ns ranging from 5 to 100) and varied the individual case's score (x^*) from a score of 95 (i.e., only .5 of a standard deviation below the control mean) to a score of 65, representing an extreme score (i.e., 3.5 standard deviations below the control sample mean).

The results are presented in Table 1. It can be seen from this table that, despite the radically different approaches employed, the two methods yield point estimates and confidence limits/Bayesian credible limits that are, for all practical intents and purposes, indistinguishable at all values of N and x^*. (Any differences that do exist are within the bounds expected solely from Monte Carlo variation.) The equivalence of these point estimates was anticipated from theoretical results developed by Geisser and Cornfield (1963) for comparing the means of two unrelated groups. Equivalent theoretical results for confidence/credible intervals have not been developed. They are more difficult to derive because noncentral t distributions are involved rather than standard t distributions (Crawford & Garthwaite, 2002). However, the agreement between confidence intervals and credible intervals observed in the present problem is not uncommon when uninformative prior distributions are used (Garthwaite et al., 2002).

Point estimates

It can be seen from Table 1 that, for both methods, when scores are extreme, the point estimates of the rarity of a given score vary markedly as a function of the size of the control sample. For example, with an N of 5, the estimated percentage of the population that would obtain a score lower than

Table 1. *Frequentist and Bayesian inferential methods for comparing a case to a control sample*

	N	*Bayesian*			*Frequentist*		
		Point	*95% LL*	*95% UL*	*Point*	*95% LL*	*95% UL*
x^*=95	5	33.61	7.90	67.89	33.59	7.89	67.87
	10	32.22	12.61	56.82	32.25	12.54	56.84
	20	31.55	16.90	48.87	31.56	16.85	48.87
	50	31.14	21.43	41.87	31.14	21.42	41.94
	100	31.00	24.04	38.58	31.00	23.98	38.57
x^*=85	5	12.13	0.26	44.40	12.14	0.26	44.55
	10	9.33	0.82	28.76	9.32	0.81	28.76
	20	7.97	1.64	19.79	7.98	1.63	19.88
	50	7.19	2.87	13.77	7.20	2.86	13.77
	100	6.94	3.73	11.29	6.94	3.72	11.28
x^*=75	5	4.23	0.001	27.38	4.23	0.001	27.23
	10	2.05	0.008	11.66	2.05	0.008	11.65
	20	1.23	0.035	5.61	1.23	0.034	5.59
	50	0.84	0.110	2.68	0.84	0.110	2.68
	100	0.73	0.187	1.79	0.73	0.189	1.79
x^*=65	5	1.64	0.000	15.33	1.65	0.000	15.38
	10	0.44	0.000	3.75	0.44	0.000	3.71
	20	0.15	0.001	1.07	0.15	0.000	1.06
	50	0.06	0.001	0.30	0.06	0.001	0.30
	100	0.04	0.003	0.15	0.04	0.003	0.15

Note: Comparison of point estimates of abnormality (percentage of control population obtaining lower scores) and accompanying 95% Bayesian credible limits/frequentist confidence limits. In these examples the mean of the control sample was set at 100 and the *SD* at 10. LL = lower limit. UL = upper limit. x^* = individual case's score. N = size of control sample.

75 is approximately 4.23%; this is more than four times the point estimate (0.73%) obtained with a control sample of 100 and reflects the greater uncertainty when the control sample is small.

If, as is commonly done in single-case studies, the patient's score was simply converted to a z score and evaluated using a table of the normal curve (i.e., the statistics from the normative samples were treated as parameters), the point estimate of the abnormality of this score would be 0.62% regardless of the size of the control sample. Thus, in the majority of cases, the commonly used (z score) method will exaggerate the abnormality of an observed score; when the normative or control sample is small this effect can be substantial. Because the Bayesian method and Crawford and Howell's (1998) frequentist method both treat the control sample statistics *as* statistics, they are not subject to this systematic bias.

Credible limits and confidence limits

The Bayesian credible limits and frequentist confidence limits quantify the degree of uncertainty surrounding the estimate of the abnormality (i.e., rarity) of a given test score. It can be seen from Table 1 that the limits are wide with small sample sizes. However, even with more moderate *N*s the limits are not insubstantial and thus serve as a useful reminder of the fallibility of control or normative data.

Table 1 illustrates another feature of these limits: They are nonsymmetrical around the point estimate. The lower limit (which must exceed 0; zeros appear in Table 1 only because the lower limits in some cases are less than 0.001%) is nearer to the point estimate than is the upper limit. Furthermore, the degree of asymmetry increases the further the case's score is from the mean of the controls.

These asymmetries occur because p follows a noncentral t-distribution. The frequentist approach (Crawford & Garthwaite, 2002) exploits the fact that p follows this distribution in order to determine confidence limits. However, the frequentist method is technically complex and is only possible at all because the distribution of p is tractable. In contrast, the Bayesian approach does not need to know the distribution of p.

The fact that the two methods yield equivalent limits means that we can apply a Bayesian interpretation to either set of limits. Thus, even if a researcher (or clinician) chose to use the frequentist method for this problem, they can legitimately avoid the convoluted frequentist interpretation of these limits. As Antelman (1997) notes, the frequentist (classical) conception of a confidence interval is that, "It is one interval generated by a procedure that will give correct intervals 95% of the time. Whether or not the one (and only) interval you happened to get is correct or not is unknown" (p. 375).

Thus, in the present context, the frequentist interpretation is as follows, "if we could compute a confidence interval for a large number of control samples collected in the same way as the present control sample, about 95% of them would contain the true percentage of the population with scores lower than the patient's". In contrast, the Bayesian analysis shows that it is legitimate to state, "there is a 95% probability that the true level of abnormality of the case's score lies within the stated limits". This statement is not only less convoluted but, we suggest, it also captures what a single-case researcher or clinician would wish to conclude from an interval estimate. It is likely that most psychologists who use frequentist confidence limits in fact construe these in what are essentially Bayesian terms (Howell, 2002).

The limits presented in Table 1 are two-sided limits. However, as noted, one-sided limits are readily obtained from the Bayesian analysis. For example, if a researcher is concerned about how abnormal a case's score might be, but uninterested in whether it may be even more unusual than the point estimate indicates, then a one-sided upper limit is more appropriate. Take the example from Table 1 of the case with a x^* of 75 and a control sample N of 10. Rather than state "There is 95% confidence that the percentage of people who have a score lower than the patient's *is between 0.008% and 11.66%*", we might prefer to state "There is 95% confidence that the percentage of

people who have a score lower than the patient's *is less than 8.45%*". The latter one-sided upper limit was obtained by finding the 5,000th largest $\hat{p}_i$ (rather than the 2,500th largest as required for two-sided limits; see Method section).

EXPERIMENT 2

Development of a Bayesian analysis for comparing differences observed in a case to differences observed in controls

Up to this point we have been concerned with the simple case of comparing a single test score obtained from a patient with a control or normative sample. However, in the assessment of acquired neuropsychological deficits, simple normative comparison standards have limitations because of the large individual differences in premorbid competencies. For example, an average score on a test of mental arithmetic is liable to represent a marked decline from the premorbid level in a patient who was a qualified accountant. Conversely, a score that fell well below the normative mean does not necessarily represent an acquired deficit in an individual who had modest premorbid abilities (Crawford, 2004; Franzen, Burgess, & Smith-Seemiller, 1997; O'Carroll, 1995).

Because of the foregoing, considerable emphasis is placed on *individual* comparison standards when attempting to detect and quantify the severity of acquired deficits (Crawford, 1992; Lezak, Howieson, Loring, Hannay, & Fischer, 2004; Vanderploeg, 1994). In the simplest case, the neuropsychologist may wish to compare an individual's score on two tests; a fundamental consideration in assessing the importance of any discrepancy between scores on the two tests is the extent to which it is rare or abnormal.

The need for a sound method to conduct such comparisons is even more acute in single-case research than it is in clinical neuropsychological practice. Although the detection of impairments 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. That is, the aim in many single-case studies is to fractionate the cognitive system into its constituent parts, and it proceeds by attempting to establish the presence of dissociations of function (Caramazza & McCloskey, 1988; Crawford, Garthwaite, & Gray, 2003a; Ellis & Young, 1996; Shallice, 1988).

Dissociations have come to play a central role in the building and testing of theory in cognitive neuroscience. For example, Dunn and Kirsner (2003) note that, "Dissociations play an increasingly crucial role in the methodology of cognitive neuropsychology ... they have provided critical support for several influential, almost paradigmatic, models in the field" (p. 2). Indeed Rossetti and Revonsuo (2000) have gone as far as to state that "dissociation is the key word of neuropsychology" (p. 1).

In the typical definition of what is termed a classical dissociation (Shallice, 1988), the requirement is that a patient is "impaired" or shows a "deficit" on task *X*, but is "not impaired", "normal", or "within normal limits" on task *Y*. For example, Ellis and Young (1996) state, "If patient X is impaired on task 1 but performs normally on task 2, then we may claim to have a dissociation between tasks" (p. 5). Crawford, Garthwaite, and Gray (2003a) have argued that these conventional criteria for a classical dissociation should be supplemented by a comparison of the difference between a patient's scores on the two tasks of interest to the differences on these tasks observed in the control sample (principally because, otherwise, researchers have to rely on the null result that the patient is not different from controls on one of the tasks).

Striking evidence in favour of this view has been provided by Monte Carlo simulations studies (Crawford & Garthwaite, 2005a, 2006a). These studies indicated that high percentages of the healthy control population would be incorrectly classified as exhibiting a dissociation if a test on the case's difference is not incorporated into the criteria. Even higher percentages (over 50% in some scenarios) of patients with *strictly*

equivalent impairments on two tasks would be misclassified as exhibiting a dissociation (such patients are impaired on both tasks but do not have a dissociation).

Having established the need to compare a case's difference score with that of controls, we turn now to the question of how we should conduct such a comparison. A complication arises because the tasks being compared commonly have different means and standard deviations. The patient's scores therefore have to be standardized as a comparison of the raw scores would not be meaningful. Obtaining a sound inferential method of examining the difference between an individual's standardized scores has proved to be much more difficult than might have been anticipated.

One obvious candidate is the method developed by Payne and Jones (1957): The patient's scores on the two tasks are converted to z scores based on the mean and standard deviations of controls, and the difference between them is divided by the standard deviation of the difference. The resultant quantity is treated as a standard normal deviate (z_D) and is referred to a table of areas under the normal curve to estimate the percentage of the control population that would exhibit a discrepancy that exceeds the discrepancy observed for a patient.

A number of authors have commented on the usefulness of this formula in neuropsychology (Crawford, 1996; Ley, 1972; Miller, 1993; Shallice, 1979; Silverstein, 1981), and it has been applied to the analysis of differences on a variety of tests (Atkinson, 1991; Grossman, Herman, & Matarazzo, 1985; Mittenberg, Thompson, & Schwartz, 1991). However, just as was the case with the use of z to compare a single score with a control sample, the Payne and Jones (1957) formula treats the statistics of the normative or control sample as if they were population parameters.

A Monte Carlo simulation study (Crawford & Garthwaite, 2005b) has demonstrated that this method can be associated with very high Type I error rates (i.e., it misclassifies a high percentage of the control population as exhibiting an abnormal difference); error rates were as high as 25.7% for a nominal rate of 5% when the control sample was modest in size. These results demonstrate that the method is only suitable for comparison of an individual's standardized difference to the differences obtained from large normative samples.

Crawford and Garthwaite (2005b) proposed two solutions to this problem. First, they noted that in some scenarios it *is* meaningful to compare the raw difference between two tasks observed for a case with the raw differences in a control sample—that is, it is not always necessary to standardize the scores. They presented a modified paired-samples t test (which they termed the unstandardized difference test) for this purpose. A Monte Carlo simulation study showed that the Type I error rate is under control when this test is applied regardless of the size of the control sample and magnitude of the correlation between the tasks.

Although this approach is sound from a statistical point of view, as noted, it can only be used in fairly circumscribed circumstances. It is far more common for neuropsychologists to attempt to demonstrate dissociations between tasks of different cognitive functions in which the two tasks also have different means and standard deviations (i.e., these quantities are essentially often arbitrary). In this latter situation it is necessary to standardize the patient's score against the control's performance in order to conduct a meaningful test on the difference between a patient's performance on the two tasks. Therefore it would be very useful if a method could be found that permits standardization of the patient's scores whilst also maintaining control of the Type I error rate.

Garthwaite and Crawford (2004) determined the asymptotic distribution of the difference between an individual's standardized difference and the standardized differences in controls. They also obtained a test statistic that asymptotically approximates a t-distribution. Monte Carlo simulation studies revealed that the approximation to t was very satisfactory in all scenarios examined (Crawford & Garthwaite, 2005b).

In other words, for the first time a method was available that would control the Type I error rate

when comparing the standardized difference for a case to the standardized differences for controls (a Type I error was defined as misclassifying a member of the control population).[1] Although Crawford and Garthwaite suggested that this test (which they termed the Revised Standardized Difference Test; RSDT) provides an approximate point estimate of the abnormality of a case's score, it did not prove possible using asymptotic methods to obtain an expression that would permit the setting of confidence limits on the abnormality of a score.

The purpose of the present study was to extend the Bayesian method developed in Experiment 1 to cover examination of score differences in the single case. Specifically, (a) we aimed to determine whether or not a Bayesian approach would exhibit convergence with Crawford and Garthwaite's (2005b) frequentist methods of examining unstandardized and standardized differences; (b) should convergence not occur, we aimed to explore the reasons for this result; (c) we hoped, for standardized differences, to obtain an exact (rather than approximate) point estimate of the abnormality of a case's difference; and (d) we hoped to solve the problem of setting credible limits/confidence limits on the abnormality of such differences.

This last aim is in keeping with the increasing importance placed on confidence limits by many scientific bodies including the American Psychological Association (Wilkinson & APA Task Force on Statistical Inference, 1999). Confidence limits or credible limits serve the useful general purpose of reminding us that all results are fallible and serve the specific purpose of quantifying the degree of uncertainty attached to a particular result (Crawford & Garthwaite, 2002).

Method

We measure the values of x and y on a sample of n controls. Let $\bar{x}$ and $\bar{y}$ denote the sample means. We need to form a scale matrix **A**, which consists of the sums-of-squares and cross-products for x and y. In keeping with our aim of requiring that these methods can be used when only basic summary statistics are available for the sample, the elements of this matrix can be obtained from the sample standard deviations of x and y (s_x, s_y) and the sample correlation between x and y (r_{xy}). Let

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = s_x^2(n-1),$$

$$s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = s_y^2(n-1),$$

and

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = s_x s_y r_{xy}(n-1).$$

Put

$$\mathbf{A} = \begin{pmatrix} s_{xx} & s_{xy} \\ s_{xy} & s_{yy} \end{pmatrix}.$$

We assume each (x, y) comes from a bivariate-normal distribution with unknown mean $\mu = (\mu_x, \mu_y)'$ and unknown variance

$$\Sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}.$$

A case has a value (x^*, y^*); without loss of generality we assume that x^* is greater than y^*. We want to estimate p, the proportion of controls whose value of $x - y$ is greater in magnitude than $x^* - y^*$. We also want an interval estimate of p.

We want to start with a standard noninformative prior distribution. For sampling from a bivariate normal distribution, as here, there are two standard forms of noninformative prior

[1] A computer program that implements this test (dissocs.exe), and the unstandardized difference test described earlier, can be downloaded from http://www.abdn.ac.uk~psy086/dept/SingleCaseMethodsComputerPrograms.HTM

distributions for μ and Σ, either $f(\mu, \Sigma^{-1}) \propto |\Sigma|$ or $f(\mu, \Sigma^{-1}) \propto |\Sigma|^{3/2}$. We examined both of these priors but here we only report the results for $f(\mu, \Sigma^{-1}) \propto |\Sigma|$ because (as will be seen) it gives classical and Bayesian interval estimates that are identical for unstandardized differences. The heuristic motivation for this choice is that frequentist methods necessarily ignore prior information, so presumably a prior distribution conveys no information (i.e., is noninformative) if it leads to Bayesian inferences that are the same as those given by a frequentist analysis.

The posterior distribution is obtained by combining the prior distribution with the data, and inferences and estimates are based on the posterior distribution. With our choice of prior, the posterior distribution states that the marginal distribution of Σ^{-1} is a Wishart distribution with parameters n and $\mathbf{A}^{-1}$:

$$f(\Sigma^{-1}) \propto |\Sigma^{-1}|^{(n-3)/2} \exp\left[-\frac{1}{2}\text{trace}(\mathbf{A}\Sigma^{-1})\right].$$

Given Σ, the conditional posterior distribution of μ is a normal distribution with mean $(\bar{x}, \bar{y})$ and variance Σ/n (Geisser & Cornfield, 1963). The following steps are then followed to obtain estimates of p from the posterior distribution:

1. The first step is to generate a random observation (a 2×2 matrix in this case) from an inverse-Wishart distribution on n *df* and scale matrix $\mathbf{A}$. The procedure for generating these random observations is set out in Appendix 1. Let

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_{xx} & \hat{\sigma}_{xy} \\ \hat{\sigma}_{xy} & \hat{\sigma}_{yy} \end{pmatrix}$$

denote the generated value. Then $\hat{\Sigma}$ is the estimate of Σ for this iteration.

2. Generate two observations from a standard normal distribution. Call the generated values (z_1, z_2). Find the Cholesky decomposition of $\hat{\Sigma}$. That is, find the lower triangular matrix $\mathbf{T}$ such that $\mathbf{T}\mathbf{T}' = \hat{\Sigma}$. Put

$$\begin{pmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} + \mathbf{T}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}/\sqrt{n}.$$

Then $(\hat{\mu}_{xx}, \hat{\mu}_y)$ is the estimate of μ for this iteration.

3. We have estimates of μ and Σ. We calculate the value of p conditional on these being the correct values of μ and Σ. At this point the method diverges depending upon whether we want to form inferences concerning an individual's unstandardized or standardized scores on X and Y. Considering first the *unstandardized* case, put

$$z^* = \frac{(x^* - \hat{\mu}_x) - (y^* - \hat{\mu}_y)}{\sqrt{(\hat{\sigma}_{xx} + \hat{\sigma}_{yy} - 2\hat{\sigma}_{xy})}}. \tag{4}$$

When inferences are to be made concerning an individual's *standardized* scores (as will more commonly be the case) put

$$z_x = \frac{x^* - \hat{\mu}_x}{\sqrt{\hat{\sigma}_{xx}}} \tag{5}$$

$$z_y = \frac{y^* - \hat{\mu}_y}{\sqrt{\hat{\sigma}_{yy}}} \tag{6}$$

and

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{yy}}{\sqrt{(\hat{\sigma}_{xx}\hat{\sigma}_{yy})}} \tag{7}$$

Then

$$z^* = \frac{z_x - z_y}{\sqrt{(2 - 2\hat{\rho}_{xy})}}. \tag{8}$$

4. We find the tail-area of a standard normal distribution that is less than z^*. This tail area is an estimate of p, which we call $\hat{p}_i$ for the ith iteration.
5. We repeat Steps 1 to 4 a large number of times; in the present case we perform 100,000

iterations. Then the average value of $\hat{p}_1, \ldots, \hat{p}_{100,000}$ is the point estimate of p. (It is the Bayesian posterior mean of p.) To obtain a 95% Bayesian credible interval, find the 2,500th smallest $\hat{p}_i$ and the 2,500th largest $\hat{p}_i$. Call these values p_l and p_u. Then the 95% Bayesian credible interval is (p_l, p_u).

Recall that in Experiment 1 the Bayesian method used the familiar formula (3) for a z score (i.e., a standard score) in computing the point and interval estimates for the abnormality of a case's score. Analogously, in the present scenario, in which we are concerned with obtaining point and interval estimates for the abnormality of the difference between a case's standardized scores, the Payne and Jones (1957) formula is used (8). As before, the crucial difference between the present approach and the standard use of the Payne and Jones formula is that, in the latter case, the sample statistics are plugged directly into the formula. In contrast, with the Bayesian approach, these quantities are repeatedly drawn at random from their posterior distributions, thus allowing for their uncertainty.

Results and discussion

Unstandardized difference for a case compared to the differences for controls

In order to compare the Bayesian and frequentist methods of comparing the unstandardized difference for a case to those of controls we examined their performance over a range of control sample *Ns* and correlations between tasks (the correlations varied from 0 to .80 but, in the interests of economy, we present results for a correlation of .6 only). Without loss of generality we fixed the means and standard deviations of the control sample to be 100 and 10, respectively.

Table 2 reports results for an individual case's x^* score of 105 (0.5 standard deviations above the control mean) combined with y^* scores of either 95 (0.5 standard deviations below the control mean) or 85 (1.5 standard deviations below the control mean). In addition we report results for an individual case's x^* score of 85 combined with y^* scores of 75 or 65. These combinations were chosen so that the raw difference (D in Table 2) was either relatively modest (10) or large (20) and were such that differences could be obtained by a case whose x^* score either was well within the normal range or was fairly poor.

It can be seen from Table 2 that the Bayesian and frequentist methods yield equivalent results with regard to both the point estimates and the interval estimates of the abnormality of the case's scores (again such differences as do exist are within the range expected solely from Monte Carlo variation). This equivalence was observed for other combinations of x^* and y^* scores and for the different correlations between the tasks, although these results are not reported here. The equivalence of the frequentist and Bayesian point estimates (but not the interval estimates) follows from (Geisser & Cornfield, 1963). Note that, should they wish, readers can easily confirm that equivalence occurs for other values of these variables as the Bayesian unstandardized difference test has been implemented in a computer program (see later section), and the results can be compared with those from the equivalent frequentist program described in Footnote 1.

Difference between standardized scores for a case compared to the differences for controls

A similar procedure to that used to compare the Bayesian and frequentist unstandardized difference tests was followed for the standardized difference tests. However, an interval estimate to accompany the frequentist point estimate is not available so there is no equivalent to the Bayesian credible interval developed in the present paper. With no loss of generality we fixed the means of the control sample to 100 and standard deviations to 10.

Results for the Bayesian and frequentist methods are presented in Table 3. This table compares results over a range of control sample *Ns* and a range of differences between standardized scores. We also compared results over a range of correlations between tasks (from 0 to .80) but again, in the interests of economy, we present results for a correlation of .6 only (with

Table 2. *Frequentist and Bayesian inferential methods for comparing the unstandardized difference between two measures for a case to the differences in a control sample*

		Bayesian			*Frequentist*		
	N	*Point*	*95% LL*	*95% UL*	*Point*	*95% LL*	*95% UL*
$x^* = 105, y^* = 95; D = -10$	5	18.28	1.26	52.92	18.26	1.28	52.76
	10	15.73	2.89	38.22	15.71	2.86	38.28
	20	14.48	4.72	29.25	14.44	4.73	29.24
	50	13.70	7.12	22.34	13.69	7.09	22.36
	100	13.45	8.61	19.34	13.43	8.59	19.32
$x^* = 105, y^* = 85; D = -20$	5	5.57	0.00	31.59	5.54	0.00	31.24
	10	3.10	0.03	15.19	3.09	0.03	15.12
	20	2.11	0.11	8.18	2.09	0.11	8.12
	50	1.58	0.30	4.34	1.58	0.30	4.35
	100	1.42	0.46	3.11	1.42	0.46	3.10
$x^* = 85, y^* = 75; D = -10$	5	18.28	1.26	52.92	18.26	1.28	52.76
	10	15.73	2.89	38.22	15.71	2.86	38.28
	20	14.48	4.72	29.25	14.44	4.73	29.24
	50	13.70	7.12	22.34	13.69	7.09	22.36
	100	13.45	8.61	19.34	13.43	8.59	19.32
$x^* = 85, y^* = 65; D = -20$	5	5.57	0.00	31.59	5.54	0.00	31.24
	10	3.10	0.03	15.19	3.09	0.03	15.12
	20	2.11	0.11	8.18	2.09	0.11	8.12
	50	1.58	0.30	4.34	1.58	0.30	4.35
	100	1.42	0.46	3.11	1.42	0.46	3.10

Note: Comparison of point estimates of abnormality and accompanying 95% Bayesian credible limits/frequentist confidence limits (D = raw difference between x^* and y^*). In these examples the mean of the control sample was set at 100 and the *SD* at 10 for both tasks, and the sample correlation between tasks was set at .6. LL = lower limit. UL = upper limit. x^*, y^* = individual case's scores. N = size of control sample.

one exception, see later). Note also that we compare results for a control sample N of 1,000 (rather than the N of 100 used to compare the unstandardized difference tests) in order to examine whether the Bayesian and frequentist methods converge when N is very large.

The most important point illustrated by Table 3 is that, unlike the previous problems studied, the frequentist and Bayesian methods for examining the difference between a patient's standardized scores do not yield equivalent results. It can be seen that the point estimates exhibit substantial divergence when control samples are small to moderate in size; it will be appreciated that it is control samples of this size that are of most relevance for the single-case researcher. With very large N there is convergence between the results from the two methods.

Considering the results in more detail, to illustrate the factors influencing the results we use Case A as a reference case and contrast the results for this case with other cases in which various features of their data differ. In Case B the difference between standardized scores is smaller (−1.0) and, in Case C, larger (−3.0). It can be seen that, across all three values of the difference between standardized scores, the Bayesian and frequentist methods yield very different results. Case D is designed to be compared to Case A to illustrate the influence of the estimated correlation between tasks on the results; unlike all other example cases the correlation between tasks for Case D is .3. For both Case A and Case D the difference between standardized scores is −2.0 but, because the tasks are more highly correlated in Case A ($r_{xy} = .6$), it can

Table 3. *Frequentist and Bayesian inferential methods for comparing the difference between standardized scores for a case to the differences observed in a control sample*

			Bayesian			
		N	*Point*	*95% Lower*	*95% Upper*	*Frequentist point estimate*
(A)	$r_{xy} = .6$; $x^* = 110$, $y^* = 90$; $z_x = 1.0$, $z_y = -1.0$; $D_z = -2.0$	5	3.74	0.001	24.21	9.32
		10	2.47	0.021	12.70	4.36
		20	1.86	0.091	7.42	2.55
		50	1.50	0.276	4.15	1.72
		1,000	1.28	0.932	1.70	1.29
(B)	$r_{xy} = .6$; $x^* = 105$, $y^* = 95$; $z_x = 0.5$, $z_y = -0.5$; $D_z = -1.0$	5	15.47	0.884	48.33	23.03
		10	14.45	2.51	36.12	17.92
		20	13.87	4.46	28.38	15.48
		50	13.47	6.97	22.00	14.08
		1,000	13.19	11.56	14.93	13.22
(C)	$r_{xy} = .6$; $x^* = 115$, $y^* = 85$; $z_x = 1.5$, $z_y = -1.5$; $D_z = -3.0$	5	1.05	0.000	10.25	4.09
		10	0.37	0.000	3.13	0.97
		20	0.16	0.000	1.15	0.29
		50	0.08	0.002	0.38	0.10
		1,000	0.04	0.022	0.07	0.04
(D)	$r_{xy} = .3$; $x^* = 110$, $y^* = 90$; $z_x = 1.0$, $z_y = -1.0$; $D_z = -2.0$	5	7.72	0.05	35.71	13.64
		10	6.17	0.31	22.40	8.58
		20	5.39	0.79	15.31	6.42
		50	4.89	1.64	10.21	5.26
		1,000	4.57	3.69	5.54	4.58
(E)	$r_{xy} = .6$; $x^* = 70$, $y^* = 50$; $z_x = -3.0$, $z_y = -5.0$; $D_z = -2.0$	5	13.81	0.000	87.66	9.32
		10	8.09	0.000	56.92	4.36
		20	4.69	0.002	30.42	2.55
		50	2.55	0.034	12.66	1.72
		1,000	1.33	0.641	2.36	1.29

Note: Comparison of point estimates of abnormality (95% Bayesian credible limits are also presented). Note that the control sample means and standard deviations for x and y are 100 and 10, respectively, in all examples. x^*, y^* = individual case's scores. D_z = difference between standardized x^* and y^* scores.

be seen that the estimates of the abnormality of this difference are more extreme than those in Case D. This holds for both the frequentist and Bayesian methods.

The most important comparison is that between Case A and Case E. Note that the difference in standardized scores (D_z) in these two cases is identical (-2.0) but this difference is arrived at by different means. In Case A the scores on x^* and y^* are not extreme (i.e., x^* expressed as a z score is $+1.0$, and y^* is -1.0. In Case E both scores are extreme (i.e., x^* expressed as a z score is -3.0, and y^* is -5.0). It can be seen that the frequentist RSDT yields identical results for both cases. In contrast, the Bayesian method provides much more conservative results for Case E than for Case A.

It can also be seen that the Bayesian method flips from providing more extreme estimates of the abnormality of the difference than does the frequentist method in Case A, to providing more conservative estimates of the level of abnormality of the difference in Case E. For example, for a control sample size of 10, the Bayesian estimate of abnormality for the difference (i.e., the percentage of controls estimated to exhibit a larger difference) is 2.47% for Case A and 8.09%

for Case E; the frequentist estimate is 4.36% for both cases (Table 3).

In order to explain the marked differences in the results for the two methods we need to return to the fundamentals of the problem: We wish to test whether the difference between standardized scores for a case differs from the differences for controls (i.e., we require a hypothesis test). We also wish a point estimate of the abnormality of the case's difference—that is, we want to estimate the proportion (or equivalently the percentage) of the control population that will obtain a more extreme difference in scores. Formally the aim is to examine whether

$$\left|\frac{x^* - \mu_x}{\sigma_x} - \frac{y^* - \mu_y}{\sigma_y}\right| \tag{9}$$

is sufficiently large that scores x^*, y^* are unlikely to be the scores of a control. The proportion of the control population with a difference larger than that in (9) is

$$\Pr\left(\left|\frac{x - \mu_x}{\sigma_x} - \frac{y - \mu_y}{\sigma_y}\right| > \left|\frac{x^* - \mu_x}{\sigma_x} - \frac{y^* - \mu_y}{\sigma_y}\right|\right). \tag{10}$$

Using classical frequentist statistics we cannot determine this probability. However, it is reasonable to base a significance test on

$$\left|\frac{x^* - \bar{x}}{s_x} - \frac{y^* - \bar{y}}{s_y}\right|. \tag{11}$$

That is, the sample means and standard deviations are substituted for the population means and standard deviations; note that (11) represents the (absolute) difference between two t-variates. This was the strategy used by Garthwaite and Crawford (2004) in developing the frequentist RSDT. Specifically, they used asymptotic expansion to obtain a test statistic for this difference that had a distribution that closely approximated a t-distribution and controlled the Type I error rate.

In the case of the Bayesian approach to this problem we also want to evaluate the probability in (9). However, unlike the frequentist approach, with the Bayesian approach we can *directly* evaluate this probability. Moreover, in the previous problems tackled in the present paper the one-tailed p value and the point estimate of abnormality coincided for both the Bayesian and the frequentist approach. In the present case (where we are examining standardized differences), this does not hold for the frequentist approach but *does* hold for the Bayesian approach. That is, for the Bayesian approach, the proportion of the population exhibiting a more extreme difference and the Bayesian p value are one and the same thing.

It follows from the above analysis that we should not expect the frequentist and Bayesian approaches to give the same result, and, given that the latter approach does not need the pragmatic step of substituting (9) with (11), the Bayesian approach is to be preferred. (With standardized differences the problem is too complex for classical methods to handle easily but, as previously noted, Bayesian methods can handle more complex problems than can classical methods.)

Further consideration of the nature of the differences in results for the two approaches provides a further insight. This is best illustrated by making further use of the examples provided by Cases A and E in Table 3. Because the difference between standardized scores is the same in these two cases, the frequentist standardized difference test gives identical p-values. However, error in estimating a standard deviation will have a greater effect in Case E than in Case A because the standard deviations divide larger quantities in Case E than in Case A. To illustrate, imagine that the standard deviation for y 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 E it would be $(-3.0) - (-5.556) = -2.556$. It can be seen that the modest change in the standard deviation of y has had a substantial effect on Case E's z score for y^* and hence a substantial effect on the difference between standardized scores; the effect on Case A is much less

marked. If the standard deviation of y was 11 rather than 10 then again we see that the effect on Case E is much more pronounced—that is, the difference between z scores for this case is then -1.545 whereas it is -1.909 for Case A.

The frequentist standardized difference test looks only at the estimated difference between the standardized scores and so ignores the individual differences in x and y. In contrast, the Bayesian test factors in the greater uncertainty in Case E and so will not yield the same answer for the two cases; for Case E, the Bayesian estimate of the abnormality of the difference will, appropriately, be less extreme.

These features of the results can also be appreciated by referring to Figure 1. This plots the Bayesian and frequentist point estimates of the abnormality of the difference between standardized scores as a function of the extremity of the case's x^* and y^* scores; the example is based on a control sample N of 10 and a correlation between tasks of .6. The crucial point in this figure is that the difference between the case's standardized scores is the same for all the examples; it was fixed at 2.0. Although the figure presents the results as a function of the extremity of the y^* score (in z score form) it will be appreciated that the value of x^* also varies (it decreases with y^* as we move from left to right). For example, the z score for x^* on the extreme left of this figure is $+1.0$ when y^* is -1.0 (so that the difference between standardized scores is 2.0) but is -4.0 on the extreme right (for the same reason, i.e., to maintain the difference between standardized scores at 2.0). Note that the results obtained from applying the Payne and Jones (1957) method are also plotted in this figure for comparison purposes.

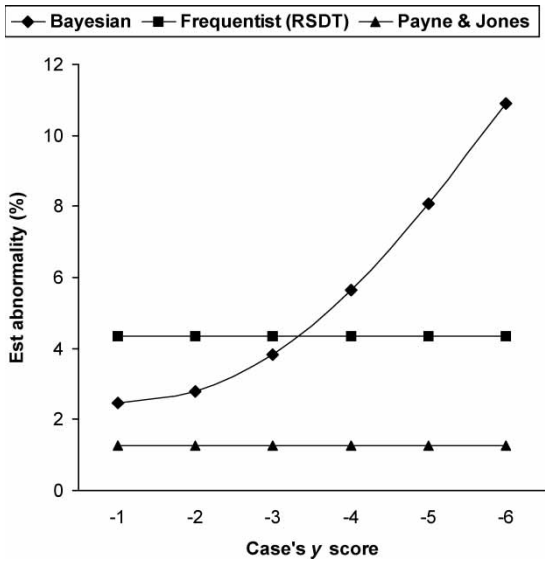

Figure 1. *Demonstration of the divergence between the point estimates of abnormality (i.e., estimated percentage of controls exhibiting a larger difference than the single case) for the Bayesian (BSDT) and frequentist (RSDT) standardized difference tests as a function of the extremity of the single case's* x *and* y *scores; crucially, the difference between the case's standardized scores is the same (2.0) in all examples (the example is based on a control sample of 20 and correlation between tasks of .6). Point estimates obtained from the method of Payne and Jones (1957) are also presented.*

It can be seen from Figure 1 that the frequentist point estimates of the abnormality of the difference between a patient's standardized scores (i.e., the point estimates obtained from the RSDT and Payne and Jones, 1957, method) are constant regardless of the extremity of the patient's x^* and y^* scores. In contrast, the Bayesian point estimate is lower than the frequentist estimates (i.e., the case's difference is estimated to be more extreme) when the x^* and y^* scores themselves are not extreme but is higher (i.e., more conservative) than the frequentist estimates when the x^* and y^* scores are extreme. As noted, this occurs because of the ability of the Bayesian method to allow for the greater uncertainty attached to the standardized values z_x and z_y in these circumstances.

The point at which the Bayesian method will produce more conservative point estimates of the abnormality of the difference than the frequentist methods is determined by a complex interaction of factors including the size of the difference between standardized scores, the extremity of the scores, the correlation between tasks and the control sample size. However, the fundamental underlying reasons for the differences in results stem from (a) the ability of the Bayesian method to directly evaluate the probability in (9), and (b)

the fact that it differentiates between a case for which both x^* and y^* are extreme and a case in which x^* and y^* are less extreme, even though the magnitude of the difference between their standardized scores is the same.

This latter feature of the Bayesian method makes it particularly suitable for use in single-case research in neuropsychology. As noted, if the aim of a single-case study is to demonstrate a (classical) dissociation, the Monte Carlo simulation studies provide striking evidence that a test on the difference between a single-case's scores should be performed (Crawford & Garthwaite, 2005a). Without such a test—that is, if a researcher demonstrates only that the patient's performance on task y is significantly below controls whereas performance on task x is not significantly lower (i.e., is "within normal limits")—the chances are high that the apparent dissociation will be spurious.

Furthermore, in neuropsychological single-case research, evidence of what have been termed "strong dissociations" (Shallice, 1988) is also used as evidence for modularity (Coltheart, 2001). In such cases it is necessarily the case that the scores of the single case are extreme on both tasks (i.e., their performance on both tasks was sufficiently low to conclude that they were impaired on both tasks). Thus, the Bayesian test will provide a more rigorous means of testing for a strong dissociation than any of the available frequentist methods. This is an important practical advantage for the Bayesian method as cases with strong dissociations are more likely to be encountered than cases with classical dissociations (Crawford & Garthwaite, 2006a).

In conclusion, the frequentist and Bayesian approaches to the analysis of a case's standardized difference will not yield equivalent results. The Bayesian analysis has a number of advantages over its frequentist alternative. First, it can directly evaluate the probability required, rather than estimate it second hand. Second, the Bayesian test will, appropriately, yield a more conservative result when the x^* or y^* scores are extreme (because extreme scores will be more sensitive to error in estimating the control standard deviations). In contrast, the frequentist approach will not capture the differing level of uncertainty attached to estimating the standard deviations. As noted, this feature makes the Bayesian method particularly useful in neuropsychological applications. Third, a very appealing feature of the Bayesian method is that the p value simultaneously provides an exact point estimate of the abnormality of the difference. In contrast, the frequentist p value cannot serve this dual function (although, as noted, frequentist p values do serve this dual function for the earlier problems studied in the present paper). Fourth, in previous work by the present authors (Garthwaite & Crawford, 2004), it did not prove possible to obtain frequentist *interval* estimate of the abnormality of a standardized difference whereas this is readily achieved with the Bayesian method.

EXPERIMENT 3

Monte Carlo simulation of Type I errors for frequentist and Bayesian methods of testing for a standardized difference

In this study a Monte Carlo simulation is performed to further evaluate the frequentist and Bayesian methods for examining standardized differences. The aim was to study the Type I error rate for these methods; in this context a Type I error occurs when we falsely conclude that the difference between a case's standardized scores is not an observation from the corresponding differences in the control population—that is, we claim that the case's difference is abnormal when it is not. Note that, when the interest in a single-case study is in the difference between an individual's performance on two tasks, one can define two forms of Type I error. The most fundamental occurs when a healthy (i.e., cognitively intact) control is misclassified. However, a second form of error can also occur—namely, misclassifying a patient with an equivalent level of acquired impairment on two tasks as exhibiting an abnormal standardized difference. In practice, the latter form of Type I error will be much

more of a threat to validity than the former (Crawford & Garthwaite, 2006a) and is the primary focus of the present study. The method adopted is based on an approach used by Crawford and Garthwaite (2005a) in which controls and a single case are sampled from the same control distribution but the case is then "lesioned" to impose strictly equal impairments on the two tasks.

It should be stressed that this simulation is based on the frequentist approach to statistics. With this approach, the question of interest is "If these are the values of my population parameters, what data values would be observed?" In the simulations the parameter values are indeed fixed, and data are being generated and observed. The Bayesian approach asks the question "If these are the values of the data, what values might the parameters take?" That is, in the Bayesian approach the data values are fixed, and the population parameters are the variable quantity. Hopefully, good methods of inference will generally perform reasonably well under most sensible criteria, rather than just the criteria for which they were designed. This simulation examines how the Bayesian method of testing for a dissociation performs under a frequentist criterion.

Method

The Monte Carlo simulation was run on a PC and implemented in Borland Delphi (Version 4). The algorithm ran3.pas (Press, Flannery, Teukolsky, & Vetterling, 1989) was used to generate uniform random numbers (between 0 and 1), and these were transformed by the polar variant of the Box–Muller method (Box & Muller, 1958). The Box–Muller transformation generates pairs of normally distributed observations, and, by further transforming the second of these, it is possible to generate observations from a bivariate normal distribution with a specified correlation (e.g., see Kennedy & Gentle, 1980).

The simulation was run with two different values of N (the sample size of the control sample): 10 and 20 (these are fairly typical Ns in single-case research, although even smaller Ns are not uncommon). For each of these values of N, 10,000 samples of $N + 1$ were drawn from one of two bivariate normal distributions in which the population correlation (ρ_{XY}) was set at either .3 or .6.

In each trial, the first N pairs of observations were taken as the control sample's scores on x and y, and the $(N + 1)$th pair was taken as the premorbid scores of the single case. The single case was then "lesioned" by imposing an acquired impairment of a specified number of standard deviations (1, 2, 4, 6, and 8) on both x^* and y^*. (An impairment of 8 standard deviations is clearly extremely severe but not beyond the bounds of possibility given the catastrophic effects of some cerebral lesions.) As the observations are sampled from a standard normal bivariate distribution, the standard deviation is 1.0 for both tasks x and y. Hence the required impairments are achieved simply by subtracting 1, 2, 4, 6, or 8 from the x^* and y^* scores of the single case. These cases are used to represent patients who had suffered *strictly equal* deficits on x and y. Thus, if any of the statistical methods record a difference between x^* and y^* for such cases, this constitutes a Type I error.

Note that, although this procedure is designed to model patients with identical *acquired deficits*, it does not produce cases with identical *scores* on x and y. Rather, the method recognizes that, (a) patients are initially members of the healthy control population until the onset of their lesion, (b) there will be premorbid differences in competencies on x and y, and (c) the magnitude of premorbid differences between x and y will be a function of the population correlation between the two tasks (i.e., the magnitude of such differences will, on average, be smaller when the population correlation is high than when it is low). The standardized scores (z_x and z_y) will also be affected by random variation in estimating the population means and standard deviations.

The simulations also contained a condition in which no (zero) impairments were imposed on the cases' scores; these cases represent healthy control cases.

In total 240,000 Monte Carlo trials were run—that is, 10,000 trials for each combination of the

two sample sizes (10 and 20), the two values for the population correlation (.3 and .6), and the six levels of acquired impairment (this includes acquired impairments of zero as noted above). The number of Monte Carlo trials were limited to 10,000 per condition because of the computationally intensive nature of the simulations; that is, on each trial the Bayesian Standardized Difference Test (BSDT) was applied to the scores of the single case, and (as noted in Experiment 2) this required a further 100,000 Monte Carlo trials per case.

On each Monte Carlo trial the three statistical methods (the BSDT, the frequentist RSDT, and the Payne and Jones, 1957, method) were applied to the scores of the single case using a specified Type I error rate of 5% for all three methods. A Type I error was recorded for the Payne and Jones method when z for the difference between the case's standardized scores exceeded 1.946; for the frequentist RSDT a Type I error was recorded when t exceeded the critical value for t on $n - 2$ *df*; a Type I error was recorded for the Bayesian test, when the mean p value from the 100,000 iterations was below .05. The number of Type I errors for each method were then expressed as a percentage of the total number of trials in each condition.

Results and discussion

The full results from the simulation are presented in Table 4. In addition, the results are presented graphically in Figure 2 but are limited to those results obtained for a control sample size of 20 and population correlation of .6.

The first thing to note is that, despite its widespread use, the Payne and Jones (1957) method achieves very poor control over Type I errors in all scenarios. Compared to the specified error rate of 5%, the minimum Type I error rate was 8.32% (when no impairments were present, the control sample size was 20, and correlation between tasks was .6), and the maximum error rate was 50.15% (for impairments of 8 standard deviations, the control sample size was 10, and correlation was .6).

Turning to the two more serious contenders, it can be seen that the RSDT yields Type I error rates that are nearer to 5% than the Bayesian test when impairments on x^* and y^* are either absent (i.e., impairment = 0) or mild (i.e., 1 *SD*). To illustrate the differences in results for the two methods: The error rate for a control sample size of 20, correlation between tasks of .3, and acquired impairments of 1 standard deviation is 5.70% for the RSDT but is 6.52% for the Bayesian test. Thereafter, however, the error rates for the RSDT increasingly exceed those of the Bayesian test as the cases' impairments on x^* and y^* become more extreme; as can be seen from Table 4, the error rate was as high as 37.57% for the RSDT (for a control sample size of 10, correlation between tasks of .6, and impairments of 8 *SD*s). In contrast, the Bayesian error rates are much lower; the maximum error rate is 8.67% (for a control sample size of 10, correlation of .6, and zero impairments). It can also be seen that the Type I error rate for the Bayesian test is fairly consistent across the varying levels of the extremity of the x^* and y^* scores (in keeping with the fact that it allows for the extremity of these scores thereby requiring a larger standardized difference before p falls below .05 than that required when the scores are less extreme).

It must be stressed that the foregoing simulation is based on the frequentist approach as in each group of simulations the parameter values were fixed at one set of values, and many different sets of data were generated. With the Bayesian approach, in one group of simulations the data would be fixed at one set of values, and many different values of the population parameters would be generated. The results show that the Bayesian method performs well under the frequentist criteria and, indeed, much better than the frequentist method does when the x and y scores of the case are extreme. Whether the frequentist method performs well under the Bayesian criterion would depend upon the data in hand, performing poorly if the case's x and y scores are extreme.

In summary, the simulation, despite being based on frequentist assumptions, supports the

Table 4. *Monte Carlo simulation results*

Deficit	*Population correlation = .3*						*Population correlation = .6*					
	N = 10			*N* = 20			*N* = 10			*N* = 20		
	P&J	*RSDT*	*BSDT*	*P&J*	*RSDT*	*BSDT*	*P&J*	*RSDT*	*BSDT*	*P&J*	*RSDT*	*BSDT*
0	11.92	4.68	7.32	8.57	5.30	6.57	13.42	5.60	8.67	8.32	4.84	6.12
1	12.94	5.36	7.34	9.20	5.70	6.52	14.63	6.62	8.58	8.46	5.37	6.06
2	16.05	7.29	7.02	10.41	6.80	6.36	18.16	9.20	8.26	10.32	6.33	6.07
4	26.13	15.33	7.48	15.85	10.79	5.89	28.93	17.27	7.93	16.98	11.95	6.05
6	37.25	25.05	7.76	23.60	17.82	5.65	40.42	27.90	7.97	25.63	19.80	6.06
8	46.48	34.72	8.26	31.35	25.49	5.80	50.15	37.57	7.82	34.58	28.65	6.03

Note: Type I errors (as percentages) for tests on standardized differences as a function of inferential method, size of control sample (*N*), population correlation between tasks, and size (in *SD* units) of acquired deficits. P&J = Payne and Jones test; RSDT = Revised Standardized Difference Test (frequentist test); BSDT = Bayesian Standardized Difference Test.

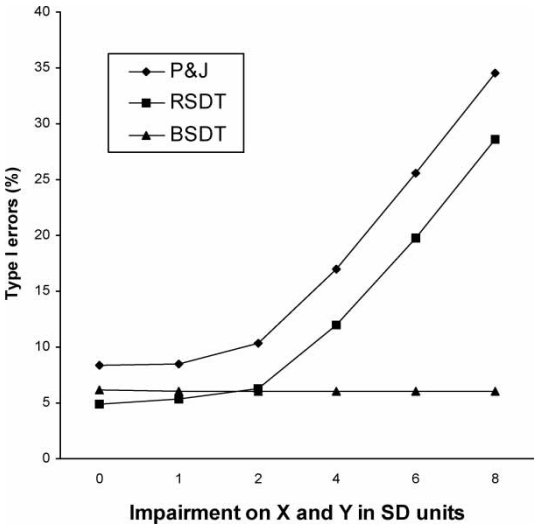

Figure 2. *Percentage of Type I errors for the Bayesian Standardized Difference Test (BSDT), the Revised Standardized Difference Test (RSDT), and the Payne and Jones (1957) method (P&J) as a function of the severity of the (strictly equivalent) impairments imposed on the single case's* x *and* y *scores (results shown are limited to those for a control sample size of 20 and population correlation between tasks of .6).*

use of the BSDT in single-case studies over the frequentist alternatives. That is, the Type I error rate was lower than that observed for the frequentist methods when cases had severe impairments—the error rates were very high for the frequentist methods in these circumstances. Thus, by allowing for the greater uncertainty when the single-case scores are extreme, the BSDT provides better protection against Type I errors. It will be appreciated that extreme scores are very common in single-case research (indeed at least one of the scores will essentially always be extreme), and thus use of the BSDT is in keeping with admonitions to single-case researchers to adopt methods that are as stringent as is practicable (Caramazza & McCloskey, 1988; Willmes, 2004).

GENERAL DISCUSSION

The results from the present series of studies indicate that the Bayesian approach to inference has a significant contribution to make in single-case research. With regard to testing for a deficit (Experiment 1), the Bayesian and frequentist methods agree, despite their radically different approaches. This is reassuring for single-case researchers regardless of whether their statistical orientation is frequentist or Bayesian. Furthermore, the fact that the frequentist confidence intervals and Bayesian credible intervals on the abnormality of a patient's score are equivalent means that the intervals have both good frequentist properties and good Bayesian properties. Also, the intervals can be interpreted in a Bayesian way, and, as noted, we suggest that a Bayesian interpretation is more in keeping with the inferences that most single-case researchers would like to draw from such intervals. When the concern is with examining differences between a patient's scores on two tasks, the Bayesian approach has a number of important advantages over frequentist methods (we will not restate these here).

In the remainder of this paper we consider the use of Bayesian methods in setting criteria for dissociations, describe computer programs that implement the methods, provide a simple example of their use, consider the application of a Bayesian approach to remaining problems in single-case research, and, finally, discuss some caveats.

Using the Bayesian methods to set criteria for dissociations in single-case studies

As previously noted, dissociations observed in single-case studies play a central role in the building and testing of theory in cognitive neuropsychology. Therefore, the criteria used to test for their presence should be as rigorous as practical considerations allow (Caramazza & McCloskey, 1988). Building on a critique of the conventional criteria for dissociations (Crawford et al., 2003a), Crawford and Garthwaite (2005b) proposed formal criteria for classical and strong dissociations. The most important component of these criteria, and the component that most clearly differentiates them from the conventional criteria for a classical dissociation, is the requirement that a patient's standardized difference

Table 5. *Bayesian criteria for dissociations obtained by modifying Crawford and Garthwaite's (2005b) frequentist criteria*

Dissociation		*Criteria*
Dissociation, putatively classical	1.	Patient's score on task *X* is sufficiently low that the probability is <.05 that it is an observation from the control population using Crawford and Howell's (1998) test; i.e., the score meets the criterion for a deficit.
	2.	Probability that the patient's score on task *Y* is an observation from the control population is >.05 using Crawford and Howell's (1998) test; i.e., score fails to meet criterion for a deficit and is therefore considered to be within normal limits.
	3.	Probability that the standardized difference between the patient's scores on tasks *X* and *Y* is an observation from the control population is <.05 on the BSDT. Note that this criterion is two-tailed to allow for the fact that the data are examined before deciding which task is *X* and which is *Y*.
Strong dissociation	1.	Patient's score on task *X* is sufficiently low that the probability is <.05 that it is an observation from the control population using Crawford and Howell's (1998) test; i.e., the score meets the criterion for a deficit.
	2.	Patient's score on Task *Y* is also sufficiently low that the probability is <.05 that it is an observation from the control population using Crawford & Howell's (1998) test; i.e., the score meets the criterion for a deficit
	3.	Probability that the standardized difference between the patient's scores on tasks *X* and *Y* is an observation from the control population is <.05 on the BSDT. Note that this criterion is two-tailed to allow for the fact that the data are examined before deciding which task is *X* and which is *Y*.
Double dissociation, putatively classical	1.	Patient 1 meets the criterion for a deficit on task *X* and meets the criteria for a dissociation (putatively classical) between this task and task *Y*.
	2.	Patient 2 meets the criterion for a deficit on task *Y* and meets the criteria for a dissociation (putatively classical) between this task and task *X*.
Strong double dissociation	1.	Patient 1 meets the criterion for a deficit on task *X* and meets the criteria for a classical or strong dissociation between this task and task *Y*.
	2.	Patient 2 meets the criterion for a deficit on task *Y* and meets the criteria for a classical or strong dissociation between this task and task *X*.
	3.	Only one of the above dissociations is putatively classical; otherwise we have a double dissociation (putatively classical).

Note: These criteria have been implemented in the computer program DissocsBayes.exe (see text for details). BSDT = Bayesian Standardized Difference Test. Criteria that combine Bayesian and frequentist tests are also available (see text for details).

between tasks *x* and *y* should be sufficiently large to render it unlikely ($p < .05$) that it is an observation from the standardized differences in the control population. These criteria were all based on the application of frequentist statistical methods (i.e., Crawford and Howell's, 1998, test was used to test for a deficit on tasks *x* and *y*, and the RSDT was used to test the standardized difference between the two tasks).

The methods developed in the present paper allow us to develop criteria for dissociations based on the results from Bayesian analysis of the scores of the single case. The essential features of these criteria are broadly the same as their frequentist counterparts. Indeed, given that the Bayesian test for a deficit (BTD) has been shown to yield equivalent results to Crawford and Howell's (1998) frequentist test,[2] the only substantive difference is the use of the Bayesian standardized difference test (BSDT) in place of the frequentist Revised Standardized Difference Test. The formal Bayesian criteria for dissociations are presented in Table 5.

Note that in developing these criteria we adopt Crawford and Garthwaite's (2006a) suggested change to nomenclature in that, although we retain the term "strong dissociation", we replace the term "classical dissociation" with "a dissociation, putatively classical". Crawford and Garthwaite (2006a) argued for this change on the basis of the results from Monte Carlo simulations studies. These studies indicated that, although one can have a high degree of confidence that a patient identified as exhibiting a classical dissociation has *some* form of dissociation, one cannot have a similar degree of confidence that it is *classical* in form; many patients who in reality have a strong dissociation will be classified as exhibiting a classical dissociation because the deficit on the less impaired of the two tasks will frequently not be detected. In contrast (provided that Crawford and Garthwaite's criteria are used), a patient with a genuine classical dissociation will rarely be classified as exhibiting a strong dissociation, thereby suggesting that the cautionary term "putative" is not necessary in the case of strong dissociations. This should not be interpreted as suggesting that there is *no* uncertainty with the former form of dissociation; the difference in nomenclature is used to flag that the degree of uncertainty is substantially greater in the case of classical dissociations.

In addition to developing purely Bayesian criteria for a dissociation, it is also possible to use the Bayesian methods to develop criteria that would satisfy those single-case researchers who are firmly frequentist in their orientation. That is, in the presence of extreme acquired impairments on the two tasks, the BSDT performs better than the frequentist tests when evaluated against frequentist assumptions. Thus a "belt and braces" approach to identifying dissociations could be adopted in which the BSDT and the frequentist RSDT are applied to the difference between a patient's standardized scores and a dissociation recorded only when *p* is less than .05 on both tests. The RSDT would provide reassurance for frequentists when the acquired impairments are mild (as Type I errors are lower for the RSDT in these circumstances), and the BSDT provides cover against error when the acquired impairments are extreme. (Note that, for a given single case, we cannot know whether scores that are only moderately lower than the control mean represent a mild impairment or severe impairment because we cannot know the premorbid level of competency.)

Computer programs to implement the Bayesian methods

The methods developed in the present paper could be implemented in Bayesian analysis programs

[2] Crawford and Howell's (1998) test for a deficit is used for computational convenience in setting these criteria. That is, a computer program is available to test whether a case meets these criteria (see General Discussion section for details). Use of Crawford and Howell's test, rather than the equivalent Bayesian test for a deficit, means that only one set of 100,000 observations need be sorted (to obtain credible limits on the abnormality of the standardized difference) rather than three sets (two further sorts would be required to obtain the limits on the abnormality of the case's *x* and *y* scores).

such as BUGS (Spiegelhalter, Thomas, Best, & Gliks, 1996). However, a reasonable level of technical skill would be required. As we are aware that many researchers and practitioners may have neither the skills nor the time to pursue these options, we have written programs (for PCs) that are specifically tailored to implementing the present Bayesian inferential methods.

A very useful and convenient feature of these programs (and indeed, more generally, of the methods themselves) is that they require only the summary statistics for the control sample rather than the raw data. Armed with these programs a single-case researcher (or clinician) can enter the requisite data and obtain the relevant results in under a minute (given the computer intensive nature of the analysis, this is a testament to the speed of modern bulk standard PCs).

The program SingleBayes.exe implements the Bayesian test for a deficit derived in Experiment 1. The data inputs required are the control sample mean and standard deviation for the task of interest, the control sample *N*, and the patient's score. The output consists of the one- and two-tailed Bayesian *p* value, the point estimate of the abnormality of the individual's score, and the 95% Bayesian credible interval on this percentage. By default the credible interval is two-sided but a one-sided upper or lower limit can be selected if required (see earlier discussion on the use of one-sided limits in Experiment 1).

The program DiffBayes.exe implements the Bayesian methods for comparing a case's difference against the differences in controls. This program prompts the user to select either the Bayesian unstandardized difference test (BUDT) or Bayesian standardized difference test (BSDT). The data inputs required are the means and standard deviations for both tasks and their correlation in the control sample, the *N* for the control sample, and the patient's scores.

The results of applying the selected difference test are reported: namely, the two-tailed probability; the point estimate of the abnormality of the patient's difference (i.e., the estimated percentage of the control population that would exhibit a more extreme discrepancy); and the 95% Bayesian credible interval for this percentage. As is the case for SingleBayes.exe, by default the credible interval is two-sided but a one-sided upper or lower limit can be selected if required.

The program DissocsBayes.exe implements the criteria for dissociations set out in Table 4 and discussed in the previous section. That is, it tests for a deficit on task *x* and task *y* using Crawford and Howell's (1998) test (which, as noted, is equivalent to the Bayesian test for a deficit) and tests the standardized difference using the BSDT. The pattern of results from applying these tests determines whether the patient meets the criteria for a strong dissociation or a dissociation (putatively classical).

It was noted in the previous section that the BSDT could also be combined with use of the frequentist RSDT to provide combined criteria for dissociations that would satisfy those of a firmly frequentist orientation. The DissocsBayes program also applies these combined criteria. That is, when the combined criteria are used, a dissociation is only recorded when the *p* values from both the BSDT and RSDT are below .05. In all other respects these criteria are the same as those set out in Table 5; that is, a strong dissociation is recorded when the patient's performance on both tasks qualify as deficits, and a dissociation (putatively classical) is recorded if only one of the tasks qualifies as deficit. The data inputs required for DissocsBayes.exe are the same as those for DiffBayes.exe: namely, the means and standard deviations for both tasks and their correlation in the control sample, the *N* for the control sample, and the patient's scores.

The results generated by these three programs can be viewed on screen, printed, or saved to a file. In addition, the statistics for the control sample are reloaded when the programs are rerun making it convenient to run analyses on further cases (alternatively, the control data can be cleared). The fact that the methods and programs that implement them require only the summary data for the controls also means that analysis can be run using previously published control data or control data from a third party. The

programs can be downloaded from: www.abdn.ac.uk/~psy086/dept/BayesSingleCase.htm

An examples of the use of the Bayesian methods

To illustrate the use of the Bayesian methods, suppose that a researcher administered a theory of mind (ToM; Baron-Cohen, Leslie, & Frith, 1985) task and a task of executive ability (e.g., set shifting) to a patient. The researcher wants to assess whether there is evidence of deficits on either or both of the two tasks and wishes to determine whether there was an abnormal difference between the two tasks (i.e., they wish to estimate the probability that the difference observed for the patient was drawn from the distribution of differences in the control population). Ultimately, the researcher is interested in assessing whether the patient exhibits a dissociation between ToM and executive ability and, if so, whether this dissociation is putatively classical in form or is best regarded as a strong dissociation.

Suppose also that 20 healthy controls matched to the patient on basic demographic variables have been recruited. The mean score for controls on the ToM task was 60 ($SD = 7.0$), and the mean score on the executive task was 24 ($SD = 4.8$); the correlation between the two tasks in controls was .68. Finally, suppose the patient's raw scores on these two tasks were 33 and 15, respectively.

Applying the Bayesian test for a deficit (BTD) to the patient's ToM score, it is estimated that only 0.067% of the control population would obtain a score lower than the patient's (i.e., this is the point estimate of the abnormality of the score), and the accompanying 95% credible interval on this quantity is from <0.0001% to 0.54%. As the point estimate tells us that less than 5% of the control population are expected to score lower than the patient, the score meets the criteria for a deficit; that is, the null hypothesis, that the score is an observation from the control population, is rejected ($p < .05$, one-tailed). If a researcher considered that the suggested criterion is overly liberal they need only modify it to require a more extreme point estimate of abnormality.

Turning to the patient's score on the executive task, it is estimated that 4.15% of the control population would exhibit a lower score (95% CI = 0.47% to 12.95%), and thus it is concluded that the patient also exhibits a deficit ($p < .05$, one-tailed) on the executive task. The patient's scores on the ToM task and executive task expressed as z scores are -3.857 and -1.875, respectively; the impairment on the ToM task is greater than that on the executive tasks but the question remains of whether this difference between the standardized scores is abnormal. That is, what is the probability that this difference is an observation from the control population? Application of the Bayesian standardized difference test addresses this question: The probability is .043. Note that this probability is two-tailed in that it is based on the absolute difference between the patient's standardized scores (i.e., it estimates the probability that a member of the control population would exhibit a larger difference in either direction). The Bayesian test also provides a point estimate of the percentage of the control population that would exhibit a larger difference *in the same direction as the case's difference*. In this example the point estimate is 2.19% (this is the one-tailed probability multiplied by 100), and the 95% credible interval on this quantity is from 0.003% to 14.17%.

Applying the Bayesian-based criteria for dissociations to this pattern of results, the patient has a deficit on both tasks (ps $< .05$, one-tailed), and the difference between the standardized scores on the two tasks is abnormal ($p < .05$, two-tailed). Thus the patient is classified as exhibiting a strong dissociation. In this example the patient would also be classified as exhibiting a strong dissociation if the combined Bayesian/ frequentist criteria were applied. That is, not only is the p value for the BSDT below .05 but, when the frequentist RSDT is applied to the difference between the patient's standardized scores, the p value for this test is also below .05 ($p = .033$, two-tailed).

Potential future applications of Bayesian methods in investigation of the individual case

Many practically important but difficult statistical problems in single-case studies remain to be tackled. Given the small control samples that typify much single-case research, treating the control sample statistics as population parameters (such as is done in the Payne and Jones, 1957, test) will give misleading results, and deriving frequentist *t* tests like the RSDT will be extremely difficult or impossible in more complex situations. In contrast, we believe the application of a Bayesian approach to inference in single-case studies holds exciting prospects. In this section we limit attention to only a few examples but consider these in some detail.

As noted, dissociations observed in single cases have come to play an important role in building and testing theories concerning the functional architecture of cognition. In the search for dissociations, many neuropsychological single-case studies employ multiple measures of the constructs under investigation. This may simply entail repeat testing using the same pair of tasks (or parallel versions thereof) or may involve administering different but related tasks—for example, x_1 and x_2, and y_1, y_2, and y_3 to measure two putative functions *X* and *Y*. The use of multiple indicators is in keeping with the fact that researchers are ultimately interested in dissociations between functions, not just in dissociations between specific pairs of indirect and impure measures of these functions (Crawford, Garthwaite, Howell, & Venneri, 2003b; Vallar, 2000). Thus, researchers seek converging evidence for a dissociation (Shallice, 1979; Vallar, 2000).

Although the use of multiple indicators in single-case studies is widespread, it is fair to say that there is little consistency across studies in how such data should be analysed (Crawford & Garthwaite, 2005b). One appealing approach to this problem would be to form composites from the two sets of indicators of functions *X* and *Y* and compare the patient's difference on these composites to that of controls. However, just as was the case for comparison of a single pair of tasks, in most cases the tasks will differ in their means and standard deviations; thus standardization of scores would be required to make such an analysis meaningful.

A satisfactory frequentist approach to such an analysis is probably not possible given the need to account for error arising from modestly sized control samples. For example, the asymptotic methods used to derive the frequentist RSDT were performed using a computer algebra package (Maple). Despite the power of this package, it was pushed to its limits to obtain the required results. Given that, in the present scenario, we have multiple rather than single indicators, and all need to be standardized, it is highly unlikely that a satisfactory standard error could be obtained for this analysis using asymptotic methods.

In contrast a Bayesian solution to this problem would be relatively straightforward because sampling error can be entirely ignored (the Bayesian Monte Carlo trials will blindly factor in this error). Thus the control sample statistics can be treated as though they were population parameters, and the problem boils down to specifying the standard deviation of the difference between linear composites in which the components are expressed in *z* score form. This is easily achieved; see Nunnally and Bernstein (1994).

Related to the above problem, frequentist methods are available for estimating the *overall* abnormality of an individual case's profile of performance on *k* standardized tests (Huba, 1985; Willmes, 2004). For example, Huba's (1985) method uses the Mahalanobis Distance statistic (e.g., see Burgess, 1991; Crawford & Allan, 1994; Crawford, Johnson, Mychalkiw, & Moore, 1997; Huba, 1985) but treats the control sample statistics as population parameters. This makes the method inappropriate for use with the modestly sized control samples typically employed in single-case studies. However, just as in the previous example, a Bayesian solution would be fairly straightforward as the problem can be specified in terms of population parameters, and the

sampling error accounted for in the Monte Carlo phase of the analysis.

As a final example, primarily aimed at demonstrating the potential flexibility of Bayesian methods, Crawford and Garthwaite (2004) developed frequentist methods for single-case studies in which a patient's performance is expressed not as simple score but as the slope of a regression line. There are a number of topics in cognitive neuroscience where analysis of a patient's slope is appropriate. For example, estimation of distance, time, musical pitch, or weight can be disrupted by neurological insults. To assess these abilities, the actual distance, elapsed time, pitch, or weight of stimuli are regressed on the patient's estimates of these quantities, and the slope of the regression line compared to the slopes obtained from controls on the same task.

A limitation of Crawford and Garthwaite's (2004) method is that the analysis cannot proceed when there are differences among the controls in their error variances (i.e., differences in the extent to which the observed data deviate from each control's regression line). A Bayesian approach to this problem would have much greater flexibility. For example, the analysis could be run even when the error variances for the controls were heterogeneous; furthermore, the analysis could be run either ignoring or incorporating any differences between the error variances for the patient and controls (the frequentist method incorporates these differences), and, finally, it would be possible to perform a simultaneous test comparing the patient's slope and intercept against controls (the frequentist method only examines the slope).

Some comments and a caveat on the use of these methods

The Bayesian methods developed in the present paper assume that the control sample data are drawn from a normal distribution; all of the corresponding frequentist methods discussed make the same assumption. Given that nonnormal control data are not uncommon in single-case studies it follows that researchers should be aware of this and exercise appropriate caution.

For example, take the simplest case of comparing a single score for a patient to controls: When the ability tested is largely within the competence of healthy controls, the control data will exhibit negative skew. The effects of this on the Bayesian method, and on Crawford and Howell's (1998) frequentist equivalent, will be to exaggerate the abnormality of a patient's score (both the point and interval estimates will be subject to this effect).

Although beyond the scope of the present paper, it would be useful to examine the robustness of the Bayesian methods in the face of departures from normality—for example, to compare the point estimates when the control data are skewed and/or leptokurtic against the estimates assuming normality. Fortunately, however, such analyses have been conducted for the frequentist equivalents of the Bayesian method for testing for a deficit and the Bayesian test for unstandardized differences (Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006; Garthwaite & Crawford, 2004). In general these methods are more robust than might be expected, even when departures from normality are extreme (Crawford et al., 2006). As the Bayesian methods yield results that match those for the corresponding frequentist methods, the Bayesian methods will be equally robust.

When there is evidence that the control data depart markedly from normality a potential solution would be to transform the scores of controls and the patient in an attempt to normalize the control score distribution (for a brief introduction to transformations see Howell, 2002). For example, to deal with moderate negative skew the scores could be reflected, and a logarithmic transformation applied. A more flexible alternative would be to seek the optimal Box–Cox (Box & Cox, 1964) normalizing power transformation.

A problem here is that, with the small samples typically employed in single-case studies, there are little data with which to assess the true form of the underlying control distribution and thereby select

the appropriate normalizing transformation.[3] Furthermore, the granularity of scores encountered in single-case studies poses a further problem. That is, there may be a limited number of possible scores; no transformation will ever adequately normalize such data (Crawford et al., 2006).

In summary, we recommend that researchers attempt to find a normalizing transformation but recognize that this may not be possible in practice. Note that, although the search for an appropriate normalizing transformation should be conducted using the control data alone, if a reasonable transformation is found, this should then be applied to the data of the controls *and* the patient before running the Bayesian method.

Another potential solution is simply to require that a patient's level of abnormality be more extreme before one concludes that an impairment or dissociation is present (Crawford et al., 2006). For example, if the Bayesian p value is treated as being akin to a classical hypothesis test, then, rather than use the conventional 5% level, a researcher might use a more stringent level (e.g., 2.5%) when there was concern over, say, negative skew; this would protect against inflation of the error rate.

Finally, the emphasis in the present paper has been on evaluating the performance of the Bayesian inferential methods when single-case research is conducted with modestly sized control samples. To avoid any potential confusion it should be noted that the Bayesian methods (and the corresponding frequentist methods) can be used with control samples of any size. Indeed, they remain more valid than commonly used frequentist methods based on z when N is large; in this situation the researcher is still dealing with a sample not a population.

Furthermore, although the methods are suitable when the control sample is modest, this does not mean that researchers should limit themselves to recruiting small control samples. The ability to identify the score of a patient with a genuine impairment as being abnormal will be greater with larger control samples; the present paper focuses on small Ns simply because of the need to reflect the reality of current practice in the majority of single-case studies. It makes sense to increase power by recruiting a large sample of controls when this is practical.

Nevertheless, because new constructs are constantly emerging in neuropsychology, and the collection of large-scale normative samples is a time-consuming and arduous process (Crawford, 2004), the prototypical single-case study is liable to remain one in which a patient is compared to a modestly sized control sample.

In closing, the present study was motivated by the need for rigorous but practical methods for use in single-case studies in neuropsychology and cognitive neuroscience. However, the methods developed could be applied in any areas of research and practice in which there is a need to compare the scores or score differences for an individual (or individual item) with a control or normative sample, particularly when these samples are modest in size.

Manuscript received 5 July 2006
Revised manuscript accepted 7 February 2007
First published online 24 May 2007

REFERENCES

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

Atkinson, L. (1991). Some tables for statistically based interpretation of WAIS–R factor scores. *Psychological Assessment*, *3*, 288–291.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind. *Cognition*, *21*, 37–46.

[3] The public domain software package R (www.cran.r-project.org) provides a very useful routine that finds the optimal Box–Cox normalizing transformation by the method of maximum likelihood. As noted, however, although it will find the best available transformation, this by no means ensures that the data will be normalized successfully.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*, 211–246.

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, *28*, 610–611.

Burgess, A. (1991). Profile analysis of the Wechsler Intelligence Scales: A new index of subtest scatter. *British Journal of Clinical Psychology*, *30*, 257–263.

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. (1992). Current and premorbid intelligence measures in neuropsychological assessment. In J. R. Crawford, D. M. Parker, & W. W. McKinlay (Eds.), *A handbook of neuropsychological assessment* (pp. 21–49). London: Lawrence Erlbaum Associates.

Crawford, J. R. (1996). Assessment. In J. G. Beaumont, P. M. Kenealy, & M. J. Rogers (Eds.), *The Blackwell dictionary of neuropsychology* (pp. 108–116). London: Blackwell.

Crawford, J. R. (2004). Psychometric foundations of neuropsychological assessment. In L. H. Goldstein & J. E. McNeil (Eds.), *Clinical neuropsychology: A practical guide to assessment and management for clinicians* (pp. 121–140). Chichester: Wiley.

Crawford, J. R., & Allan, K. M. (1994). The Mahalanobis distance index of WAIS–R subtest scatter: Psychometric properties in a healthy UK sample. *British Journal of Clinical Psychology*, *33*, 65–69.

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. (2004). Statistical methods for single-case research: Comparing the slope of a patient's regression line with those of a control sample. *Cortex*, *40*, 533–548.

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

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

Crawford, J. R., Garthwaite, P. H., & Gray, C. D. (2003a). Wanted: Fully operational definitions of dissociations in single-case studies. *Cortex*, *39*, 357–370.

Crawford, J. R., Garthwaite, P. H., Howell, D. C., & Venneri, A. (2003b). Intra-individual measures of association in neuropsychology: Inferential methods for comparing a single case with a control or normative sample. *Journal of the International Neuropsychological Society*, *9*, 989–1000.

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.

Crawford, J. R., Johnson, D. A., Mychalkiw, B., & Moore, J. W. (1997). WAIS–R performance following closed head injury: A comparison of the clinical utility of summary IQs, factor scores and subtest scatter indices. *The Clinical Neuropsychologist*, *11*, 345–355.

Daalgard, P. (2002). *Introductory statistics with R*. New York: Springer Verlag.

DeGroot, M. H., & Schervish, M. J. (2001). *Probability and statistics* (3rd ed.). Reading, MA: Addison-Wesley.

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex*, *39*, 1–7.

Ellis, A. W., & Young, A. W. (1996). *Human cognitive neuropsychology: A textbook with readings*. Hove, UK: Psychology Press.

Franzen, M. D., Burgess, E. J., & Smith-Seemiller, L. (1997). Methods of estimating premorbid

functioning. *Archives of Clinical Neuropsychology*, *12*, 711–738.

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

Garthwaite, P. H., Jolliffe, I. T., & Jones, B. (2002). *Statistical inference* (2nd ed.). Oxford, UK: Oxford University Press.

Geisser, S., & Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society Series B*, *25*, 368–376.

Gelfand, A. E., Hills, S. E., Racine-Poone, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, *85*, 972–985.

Grossman, F. M., Herman, D. O., & Matarazzo, J. D. (1985). Statistically inferred vs. empirically observed VIQ-PIQ differences in the WAIS–R. *Journal of Clinical Psychology*, *41*, 268–272.

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

Huba, G. J. (1985). How unusual is a profile of test scores? *Journal of Psychoeducational Assessment*, *4*, 321–325.

Kennedy, W. J., & Gentle, J. E. (1980). *Statistical computing*. New York: Marcel Dekker.

Ley, P. (1972). *Quantitative aspects of psychological assessment*. London: Duckworth.

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Miller, E. (1993). Dissociating single cases in neuropsychology. *British Journal of Clinical Psychology*, *32*, 155–167.

Mittenberg, W., Thompson, G. B., & Schwartz, J. A. (1991). Abnormal and reliable differences among Wechsler Memory Scale–Revised subtests. *Psychological Assessment*, *3*, 492–495.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Carroll, R. (1995). The assessment of premorbid ability: A critical review. *Neurocase*, *1*, 83–89.

Odell, P. L., & Feiveson, A. H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association*, *61*, 199–203.

Payne, R. W., & Jones, G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, *13*, 115–121.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal*. Cambridge, UK: Cambridge University Press.

Rossetti, Y., & Revonsuo, A. (2000). Beyond dissociations. In Y. Rossetti & A. Revonsuo (Eds.), *Beyond dissociation: Interaction between dissociated implicit and explicit processing* (pp. 1–16). Amsterdam: John Benjamins Publishing Company.

Shallice, T. (1979). Case study approach in neuropsychological research. *Journal of Clinical Neuropsychology*, *3*, 183–211.

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

Silverstein, A. B. (1981). Reliability and abnormality of test score differences. *Journal of Clinical Psychology*, *37*, 392–394.

Smith, W. B., & Hocking, R. R. (1972). Algorithm AS 53: Wishart variate generator. *Applied Statistics*, *21*, 341–345.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry* (3rd ed.). San Francisco: W.H. Freeman.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gliks, W. R. (1996). *BUGS: Bayesian inference using Gibbs sampling*. Cambridge, UK: Medical Research Council Biostatistics Unit.

Vallar, G. (2000). The methodological foundations of human neuropsychology: Studies in brain-damaged patients. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (2nd ed., Vol. 1, pp. 53–76). Amsterdam: Elsevier.

Vanderploeg, R. D. (Ed.). (1994). *Clinician's guide to neuropsychological assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Willmes, K. (2004). The methodological and statistical foundations of neuropsychological assessment. In P. W. Halligan, U. Kischka, & J. C. Marshall (Eds.), *Handbook of clinical neuropsychology* (pp. 27–47). Oxford, UK: Oxford University Press.

APPENDIX 1

Generating random observations from an inverse-Wishart distribution

There are various algorithms for generating observations from an inverse-Wishart distribution (e.g., see Gelfand, Hills, Racine-Poone, & Smith, 1990; Odell & Feiveson, 1966; Smith & Hocking, 1972). Alternatively, statistics packages such as BUGS (Spiegelhalter et al., 1996) and R (e.g., Daalgard, 2002) can be used. Below we set out a general algorithm, based on Smith and Hocking (1972), for repeated sampling from an Inverse-Wishart when the starting point is a samples' sum-of-squares and cross-products matrix (denoted **A**) for two variables. This was the method employed in the present study and is implemented in the computer programs for end-users:

1. Calculate $\mathbf{A}^{-1}$, that is, calculate the inverse of the sample sum-of-squares and cross-products matrix, calling the result **C**. Then find the Cholesky decomposition of **C**. That is, find the lower triangular matrix **D** such that $\mathbf{DD}' = \mathbf{C}$. The value of **D** is the same for all iterations.

2a. Generate an observation from a standard normal distribution, N(0,1), and call this observation g_{12}.

2b. Generate an observation from a χ^2 distribution on n df. Call the observation g_{11}.

2c. Generate an observation from a χ^2 distribution on $n - 1$ df. Call the observation g_{22}.

2d. Put

$$\mathbf{G} = \begin{pmatrix} \sqrt{g_{11}} & g_{12} \\ 0 & \sqrt{g_{22}} \end{pmatrix}$$

and calculate $\mathbf{GG}'$, calling the resulting matrix **B**.

2e. Calculate $\mathbf{DBD}'$ calling the resulting matrix **W**. (A useful check can be performed at this stage; over many iterations the average value of **W** should equal $n\mathbf{C}$).

2f. Calculate the inverse of **W**, calling the resultant matrix $\hat{\boldsymbol{\Sigma}}$. $\hat{\boldsymbol{\Sigma}}$ is the estimate of $\boldsymbol{\Sigma}$ for this iteration. [As another check, the average value of $\hat{\boldsymbol{\Sigma}}$ should equal $\mathbf{A}/(n - 3)$.]