COGNITIVE NEUROPSYCHOLOGY, 2004, 21 (7), 750–755

INFERENTIAL METHODS FOR COMPARING A SINGLE CASE WITH A CONTROL SAMPLE: MODIFIED *t*-TESTS VERSUS MYCROFT ET AL.'S (2002) MODIFIED ANOVA

John R. Crawford
University of Aberdeen, UK

Paul H. Garthwaite
The Open University, Milton Keynes, UK

David C. Howell
University of Vermont, Burlington, USA

Colin D. Gray
University of Aberdeen, UK

Mycroft, Mitchell, and Kay (2002) have criticised existing inferential methods (e.g., Crawford & Howell, 1998) for comparing a single case with a control sample and propose that such comparisons be made using a modified ANOVA. It is argued that the assumptions made by Mycroft et al. are questionable and, even if they held, would not invalidate Crawford and Howell's method. Crawford and Howell's null hypothesis is that the patient is an observation from the control population whereas Mycroft et al.'s null hypothesis is that the control population and a notional population of patients have a common mean. Even if one accepts Mycroft et al.'s conceptualisation, their arguments only have force if (1) the variance of a notional population of patients was larger than that of the control population, and (2) patients with impaired performance were balanced exactly by patients whose performance had been enhanced relative to controls. Furthermore, the modified ANOVA would have the undesirable consequence of reducing statistical power unnecessarily and it requires users to provide some estimate of the variance of a hypothetical population.

BACKGROUND

In single-case studies in cognitive neuropsychology it is very common to compare a patient's performance with a modestly sized control sample. Mycroft, Mitchell, and Kay (2002) have recently reviewed potential methods of conducting such a comparison. There is much that we agree with in this review and indeed we have made many of their points ourselves elsewhere (Crawford & Garthwaite, 2002; Crawford, Garthwaite, & Gray, 2003a; Crawford & Howell, 1998; Crawford, Howell, & Garthwaite, 1998). In particular we agree with Mycroft et al. that limiting a comparison of a patient with controls to descriptive statistics alone is unsatisfactory. They also endorse our argument (Crawford & Howell, 1998) that, in typical single-case studies, the use of *z* for inferential purposes is not appropriate. When *z* is used to make inferences, the control sample statistics are treated as population parameters, rather than as sample statistics. The upshot of this is that, with

Correspondence should be addressed to Professor John R. Crawford, School of Psychology, King's College, University of Aberdeen, Aberdeen AB24 3HN, UK (Email: j.crawford@abdn.ac.uk).

http://www.tandf.co.uk/journals/pp/02643294.html DOI:10.1080/02643290342000276

modest *N*s, there will be an increase in the Type I error rate and the abnormality of the patient's score will be exaggerated.

However, we have serious areas of disagreement with Mycroft et al. (2002). First, we disagree with their view that the method we have proposed (Crawford & Howell, 1998) for comparison of a single case with controls is invalid. Second, we believe that the conservative test that they advocate is based upon a questionable rationale. Such a test, moreover, would be difficult or impossible to carry out; and even if it could be implemented, there would be undesirable consequences.

VARIABILITY OF A HYPOTHETICAL PATIENT POPULATION

The basis for both Mycroft et al.'s criticism of our approach and their development of an alternative is (1) their view that the patient in a single-case study should be considered to have been drawn from a notional population of patients, and (2) that such a notional population is liable to have markedly greater between-subject variability than the population of controls. They claim that we ". . . fail to note the consequences of unequal variance" and that our method ". . . encounters all the problems described below and cannot be considered reliable where there are differences in variability between patients and controls" (p. 294).

We are of course aware of the effects of unequal variances when comparing *group means* using *t*-tests and ANOVAS. Indeed, Mycroft et al. (2002) repeatedly refer to the work of one of us (Howell, 1997; see also Howell, 2002) in support of many of their statements. However, this consideration does not invalidate the use of the modified *t*-test (Sokal & Rohlf, 1995) advocated by Crawford and Howell (1998), nor subsequent extensions based on the same rationale (Crawford & Garthwaite, 2002, 2004; Crawford, Garthwaite, Howell, & Venneri, 2003b; Crawford et al., 1998).

Our methods are designed to test the hypothesis that an *individual* patient did not come from a population of controls (under the null hypothesis, the individual is an observation from a distribution with the same mean and variance as for the controls). In contrast, the hypothesis that Mycroft et al. attempt to test is that the mean of a notional population of patients (from which they have a sample of one) is different from the mean of a population of controls. In response to the perceived problem of unequal variances, Mycroft et al. (2002) propose that comparison of a single case with controls should be conducted by running a one-way ANOVA in which one of the groups (the patient) has an *N* of 1. Up to this point their method is, in practice, indistinguishable from our proposal of using a modified *t*-test because the tests are directly equivalent (since $F = t^2$, the *p*-values for *F* and *t* are identical). However, their method then requires that the user estimate the extent to which the between-subject variability of the hypothetical population from which the patient has been drawn is larger than the between-subject variability of controls, while, for ours, no such estimate is required or appropriate. Having done this, their method requires the researcher to refer to a table (their Table 2) of modified critical values for *F*. These values were obtained from Monte Carlo simulations for which cases were sampled from a hypothetical population of patients in which the between-subjects variance was set at various multiples of the between-subjects variance of the control samples. The critical values are the 95th percentiles of the empirical distributions so generated.

DEFINING A HYPOTHETICAL PATIENT POPULATION

For our method it is not necessary or appropriate to be concerned with a hypothetical patient population, but for Mycroft et al. it is crucial how this hypothetical population should be defined. If the notional population were to be defined as all patients who have suffered neurological damage, then it would be generally agreed that a population so defined would have greater between-subject variability than the healthy population. On a measure of any given cognitive function, the scores of some patients will remain in the normal range (their neurological damage has spared the cognitive function), some will exhibit mild impairments, and others will exhibit severe impairment.

However, a defining characteristic of the rise of cognitive neuropsychology, and its associated emphasis on single-case research, is the move from an interest in the *general* to the *particular*. When cognitive neuropsychologists test hypotheses concerning their individual patients they do not wish to generalise to the population of all patients with neurological damage; indeed such an aim would be considered the antithesis of their intentions.

Mycroft et al. (2002) do not attempt to define the notional population beyond suggesting, in a footnote, that the notional population consists of patients that are "equivalent" to the patient of interest (p. 295). On the basis of this statement, one potential definition of their notional population would be that it consists of those patients whose premorbid competencies and current cognitive architectures are identical to the patient under consideration. That is, they would have the same pattern of spared and compromised cognitive subsystems and the same pattern of spared and compromised connections between these sub-systems. If a notional population were to be defined in this way, then Mycroft et al.'s assumption that such a population would have markedly greater between-subject variability than controls is questionable. At the very least it highlights that arriving at an estimate of the variance of a hypothetical patient population (which is necessary if their method is actually to be used in practice) would be an exercise in attaching a precise value to an entirely hypothetical quantity.

DIFFERENT VARIANCES IN THE ABSENCE OF A DIFFERENCE IN POPULATION MEANS

Suppose a researcher uses the modified *t*-test (or its ANOVA equivalent) and finds that a patient's score is significantly lower than controls ($p < .05$, one-tailed). We would conclude that the patient has a deficit: The patient's score is low enough that we can reject (at the 5% level) the null hypothesis that the patient is an observation from a distribution having the same mean and variance as for the controls. However, let us, purely for the sake of argument, go along with Mycroft et al.'s suggestions that we should invoke a population of patients from which the individual patient was drawn and that the between-subject variance in this notional patient population will be markedly larger than that of controls. The modified *t*-test (or ANOVA equivalent) is now conceptualised as a test for a difference in population means.

Mycroft et al.'s argument is that a significant test result (such as that above) could occur in the absence of a difference in population means because the variance of the notional patient population is larger than that of controls. In many other contexts this argument might have some force but, in the present context, it is untenable. In order for there to be a difference in variances in the absence of a difference in population means, one of the populations must have more extreme values *in both tails*. Hence, we are required to conclude that, in Mycroft et al.'s scenario, the neurological damage that differentiates patients from controls will have *enhanced* the performance of some of the patients.[1] Indeed, the extent to which the performance of some patients in this population has been impaired would have to be *balanced exactly* by other patients whose performance has been enhanced.

Furthermore, to refer back to our earlier point on defining the notional patient population, in this scenario, patients who have exhibited dramatically different reactions to neurological damage (i.e., some showing impairment, some enhancement of cognitive performance) are defined as being members of the same notional population.

BALANCING TYPE I AND TYPE II ERRORS

The intention of Mycroft et al.'s approach is to reduce the number of Type I errors; in the present

[1] We would not deny that, albeit in some very atypical situations, performance on a task may be enhanced relative to control performance following neurological damage. For example, it is possible that an alexic patient would be faster on colour naming in the Stroop task than healthy controls because they were free from the interfering effects of the colour words. However, examples such as this will be very rare.

context a Type I error occurs when (using a one-tailed test) we erroneously conclude that a patient's score is significantly lower than those of the controls. However, they say little about the effect of their procedure on the rate of Type II errors (a Type II error occurs when we erroneously conclude that a patient does not differ from controls). This is unfortunate as potential users need such information if they are to make informed decisions. Contemporary opinion in statistics (e.g., Cohen, 1988; Howell, 2002; Zar, 1999) is that our historical obsession with reducing Type I errors has led to a relative neglect of Type II errors.

Moreover, the issue of statistical power (i.e., the ability to avoid making Type II errors) assumes particular importance in single-case studies. This is because a single individual is compared with a control sample and the control sample itself is typically modest in size. As a result, statistical power is inevitably low in single-case studies when compared to group-based studies (Crawford & Howell, 1998; Crawford et al., 1998). Therefore, any procedure that unnecessarily reduces power in single-case studies should be studiously avoided (Crawford et al., 2003a).

Let us suppose that a researcher decides to use Mycroft et al.'s method and further suppose that they decide to estimate the variance of the notional patient population to be 2.5 times greater than controls. Mycroft et al. provide modified *F* values to cover estimated variances that are 5 times larger than control variances so this is a moderate value in their terms. Further, suppose that the *N* of the control sample was 5; not an atypical *N* for single-case studies in cognitive neuropsychology. Referring to the appropriate cell of Mycroft et al.'s Table 2 we find that the modified critical value for *F* at the .05 level is 17.31. Therefore, the patient would have to be more than 3.7 *SD*s below the control mean for the difference to be significant; i.e., the power to detect a difference is very low (if the variance was estimated to be 5 times that of controls then, from Table 2, the *F* value is 33.17 and the patient would have to be more than 6.3 *SD*s below the control mean).

Mycroft et al.'s attempt to reduce Type I errors is misplaced in our view and, in practice, it requires researchers to take a guess at the variance of a hypothesised population. Given that their method also potentially leads to a very substantial increase in Type II errors, we conclude that it should not be employed in single-case research.

UTILITY OF THE TWO METHODS

It will be clear that we reject Mycroft et al.'s (2002) arguments that their method is more valid than our own. However, leaving aside these differences of opinion, the practical utility of the two methods will now be compared. This is of value because some of the differences apply equally to the direct ANOVA equivalent of our method.

To use Mycroft et al.'s method requires (1) setting up and running an ANOVA in a general statistics package, (2) coming up with an estimate of the variance of a hypothetical population from which the patient was drawn, then (3) looking up their table of modified critical values for *F*. If the *N* of the control sample does not exactly match an *N* presented in the table then presumably the researcher will have to estimate a critical value by interpolation from the nearest *N*s offered in the table.

In contrast, if the dedicated computer programs that accompany our methods are employed,[2] then the analysis requires only entry of the score for the patient, the mean and *SD* of the controls, and the control sample *N*. The required data can be entered and the results obtained in less than 10 seconds. As the program uses summary data rather than raw data it is even possible to use the method with control data from third parties.

[2] These programs can be downloaded from the first author's web pages at http://www.abdn.ac.uk/~psy086/dept/SingleCaseMethodology.htm. The program singlims.exe carries out the modified *t*-test, comparing a patient's score on a single test with scores from a control sample. The program difflims.exe extends the approach to testing for a dissociation, i.e., it compares the difference between a patient's scores on two tests with the distribution of differences in controls. The program proflims.exe extends this to comparing the difference between a patient's score on a test with their score on the average of *k* tests. All these programs provide an estimate of the abnormality of the score (or score difference) and 95% CLs on the abnormality of the score (or score difference).

A precise probability for *t* is provided by these programs (rather than only the critical value for *t* or *F* required for a particular significance level as would be the case were Mycroft et al.'s method employed); this also means that there is no need for interpolation. In addition, the *F* values presented by Mycroft et al. (2002) are limited to two-tailed critical values. In contrast our programs report one- and two-tailed probabilities for *t*. As noted, statistical power is inevitably a concern in single-case studies. The single-case researcher is almost invariably interested in testing the null hypothesis of no difference between patient and controls against the alternative directional hypothesis that the patient's score is lower. Therefore, one-tailed tests have much to recommend them in single-case research; they are more powerful and are legitimate given that the possibility of enhanced performance in the patient can be discounted except in very rare and highly specific circumstances.

Another point of departure between our methods and the use of a modified or standard ANOVA is that our methods are not limited to testing the significance of the difference between patient and controls. Rather they also estimate the abnormality or rarity of the patient's score and, in addition, provide confidence limits (95%) on the abnormality of the score using a method developed by Crawford and Garthwaite (2002). That is, the modified *t*-test is used to provide an unbiased *point* estimate of the percentage of the healthy population that would exhibit a score lower than the patient's, and the confidence limits quantify the *uncertainty* associated with this estimate. These limits allow the researcher to make statements of the form "the estimated percentage of the healthy population that would obtain a score lower then the patient is 2.1% and the 95% CI on the percentage is from 0.2% to 6.7%." The uncertainty referred to arises because a sample is being used to estimate a parameter; if another control sample were drawn from the healthy population then the estimate of abnormality would change because of sampling error.

The provision of these confidence limits is in keeping with the contemporary emphasis in statistics, psychometrics, and biometrics on the use of confidence limits in research (American Psychological Association, 2001; Daly, Hand, Jones, Lunn, & McConway, 1995; Gardner & Altman, 1989; Zar, 1999). Gardner and Altman (1989), for example, in discussing the general issue of the error associated with sample estimates, note that "these quantities will be imprecise estimates of the values in the overall population, but fortunately the imprecision itself can be estimated and incorporated into the findings" (p. 3). The specific motivation behind our development of these confidence limits was to provide tools for single-case research that would parallel those taken for granted or even viewed as mandatory in group-based research. For example, the American Psychological Association (2001) take the view that confidence limits or intervals represent "in general the best reporting strategy. The use of confidence intervals is therefore strongly recommended" (p. 22).

It should be noted that these confidence limits (which are obtained from noncentral *t*-distributions), are *entirely* immune to the variance effects discussed by Mycroft et al. (2002) because they do not involve any assumptions about a hypothetical patient population (Crawford & Garthwaite, 2002).

Furthermore, the statistical methods we have developed have been incorporated into a formal system aimed at providing fully specified criteria for the presence of classical and strong dissociations. Monte Carlo simulations indicate that, irrespective of the size of the control sample and the strength of correlation between the tasks involved, very low proportions of the normal population are misclassified as exhibiting either form of dissociation when these criteria are applied (Crawford et al., 2003a). In summary, we believe that the methods we have developed have a number of practical advantages over Mycroft et al.'s modified ANOVA, and also over the direct ANOVA equivalent of our method.

CONCLUSION: A NEGATIVE EFFECT ON SINGLE-CASE RESEARCH?

We are concerned that Mycroft et al.'s (2002) paper unnecessarily muddies the issues surrounding the drawing of inferences from single-case studies. We

do not think we are misrepresenting Mycroft et al.'s views by stating that they, like us, see the most important factors as being whether an analytic approach (1) uses inferential rather than descriptive statistics, and (2) treats the control sample statistics *as* sample statistics. Both our methods and theirs fulfil these criteria. In contrast, when *z* is used for inferential purposes, the control sample statistics are treated as though they were population parameters.

The single-case approach in neuropsychology has made a significant contribution to our understanding of the architecture of human cognition. However, as Caramazza and McCloskey (1988) note, if advances in theory are to be sustainable they "… must be based on unimpeachable methodological foundations" (p. 619). Regrettably the statistical analysis of single-case data is one important aspect of methodology that has been relatively neglected (Crawford, 2004; Crawford & Garthwaite, 2002; Crawford et al., 2003a). Therefore, in principal, any treatment of these issues should be encouraged. Moreover, Mycroft et al.'s paper is certainly of interest to those amongst us who are intrigued by statistical arguments. However, there is the danger that those researchers whose interest is limited to matters neuropsychological will now be confused as to the most appropriate methods. As a reaction to this confusion they may continue to use, or fall back on, the use of *z* to make inferences or, worse still, limit themselves to purely descriptive statistics.

Manuscript received 18 December 2002
Manuscript accepted 25 June 2003

REFERENCES

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Caramazza, A., & McCloskey, M. (1988). The case for single-patient studies. *Cognitive Neuropsychology*, *5*, 517–528.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crawford, J. R. (2003). Psychometric foundations of neuropsychological assessment. In L. H. Goldstein & J. McNeil (Eds.), *Clinical neuropsychology: A practical guide to assessment and management for clinicians* (pp. 133–169). Chichester, UK: Wiley.

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 studies in neuropsychology: Comparing the slope of a patient's regression line with those of a control sample. *Cortex*, *40*, 533.

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., Howell, D. C., & Garthwaite, P. H. (1998). Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples *t*-test. *Journal of Clinical and Experimental Neuropsychology*, *20*, 898–905.

Daly, F., Hand, D. J., Jones, M. C., Lunn, A. D., & McConway, K. J. (1995). *Elements of statistics*. Wokingham, UK: Addison-Wesley.

Gardner, M. J., & Altman, D. G. (1989). *Statistics with confidence—confidence intervals and statistical guidelines*. London: British Medical Journal.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.

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

Mycroft, R. H., Mitchell, D. C., & Kay, J. (2002). An evaluation of statistical procedures for comparing an individual's performance with that of a group of controls. *Cognitive Neuropsychology*, *19*, 291–299.

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

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.