This article was downloaded by:[Hull University]
[Hull University]

On: 2 May 2007

Access Details: [subscription number 768495754]

Publisher: Psychology Press

Informa Ltd Registered in England and Wales Registered Number: 1072954

Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Cognitive Neuropsychology

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t713659042>

Maximizing the power of comparing single cases against a control sample: An argument, a program for making comparisons, and a worked example from the Pyramids and Palm Trees Test

First Published on: 21 February 2007

To cite this Article: , 'Maximizing the power of comparing single cases against a control sample: An argument, a program for making comparisons, and a worked example from the Pyramids and Palm Trees Test', Cognitive Neuropsychology, 24:3, 279 - 291

To link to this article: DOI: 10.1080/02643290601181890

URL: <http://dx.doi.org/10.1080/02643290601181890>

PLEASE SCROLL DOWN FOR ARTICLE

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

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

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

© Taylor and Francis 2007

Maximizing the power of comparing single cases against a control sample: An argument, a program for making comparisons, and a worked example from the Pyramids and Palm Trees Test

Johan Hulleman

University of Hull, UK

Glyn W. Humphreys

University of Birmingham, UK

In neuropsychological research, it is frequently necessary to compare the performance of a single case with that of a control sample. Recently, there has been a discussion about whether existing methods are adequate in preventing Type 1 errors due to increased variability in the single-case data (Crawford, Garthwaite, Howell, & Gray, 2004; Mitchell, Mycroft, & Kay, 2004; Mycroft, Mitchell, & Kay, 2002) and about the power of any comparisons. In this paper, we propose an extension of the modified t test introduced by Crawford and Howell (1998). We provide Monte Carlo simulations that show that it is possible to increase the power of the modified t test by retesting the single case. We also make available a computer program that implements the method introduced in this paper and that can be used by neuropsychologists to test for a difference between a single case and a control group. We apply our procedure to an example of a patient examined on the “Pyramids and Palm Trees” test.

In neuropsychological testing, it is frequently important to know whether the performance of a single case is different from that of a control group. The statistical analysis of comparing a single case against a control sample has a long history in the literature (e.g., Huber, 1973; Zubin, 1950). Crawford and Howell (1998) introduced the modified t test, to provide such an analysis. In the modified t test, the single case is considered to be a group of one, and the data of the single case do not contribute to the error

term against which the difference between the single case and the mean of the control group is tested. Crawford and Howell (1998) argued that the modified t test is well suited for this kind of comparison, because it takes into account the fact that control groups are not always very large. Recently, Crawford, Garthwaite, Azzalini, Howell, and Laws (2006) showed that the modified t test is reasonably robust against departures from normality in the data and that with a slight adjustment of the critical t it is possible to

Correspondence should be addressed to Johan Hulleman, Department of Psychology, University of Hull, Hull HU6 7RX, UK (E-mail: j.hulleman@hull.ac.uk).

This work was supported by grants from the MRC and the Stroke Association.

control the Type 1 error level in these situations. However, the modified t test approach of Crawford and Howell was criticized by Mycroft, Mitchell, and Kay (2002) for ignoring the fact that there is often increased variability in single-case data. Mycroft et al. (2002) presented a review of the methods that have been employed for comparing the performance of a single case with that of a control group. They argued that neuropsychologists who use the modified t test of Crawford and Howell (1998) run the risk of finding spuriously significant results, because the modified t test is not robust in situations where the variability in the performance of the single case is much larger than that of the control group. Mycroft et al. (2002) ran Monte Carlo simulations and found that, in certain situations, when the variability in the data from the single case was much larger than that in the control group, there was a 30% chance of finding a significant difference between the single case and the controls, even though there was no difference in reality. To remedy this increase in Type 1 error, Mycroft et al. (2002) proposed that neuropsychologists should adjust (increase) the critical level of their test statistic (i.e., the level where they judge the difference between a single-case mean and the control group mean to be significant). According to Mitchell, Mycroft, and Kay (2004), it is not possible to distinguish whether a single case differs from a control group in terms of their mean level of performance, the variability in their performance, or both, without an adjustment of the critical level. This argument has, however, been refuted by Crawford, Garthwaite, Howell, and Gray (2004). They argued that the situation cited by Mycroft et al. (2002)—larger variability in the single-case data than in the control group data with equal performance for single case and control group—was extremely unlikely to occur in practice. According to Crawford et al. (2004), an increase in variability without a change in the mean level of performance would require some single cases to improve relative to their premorbid scores. Because it is hard to imagine how postmorbid performance could actually be better than premorbid performance, Crawford et al. (2004) argued that any increase in variability has to be accompanied by a decrease in mean level

of performance. Consequently, there is no need to protect against an increase in Type 1 error, because Type 1 errors only occur when there is no real difference between the means. So, according to Crawford et al. (2004), whenever the modified t test yields a significant result, it is safe (at the specified alpha level) to infer a decrease in performance for the single case. This argument against an increased Type 1 error due to increased variability in the single-case data forwarded by Crawford et al. (2004) seems rather compelling, and the remainder of this paper is based on the assumption that increased variability in the single-case data does indeed not increase the number of Type 1 errors.

Increased variability in single-case data and power

Although there is no inflation of Type 1 error rates, this does not mean that the power to detect a difference between single case and control group is immune to differences in variability between single case and control group. On the contrary, it seems highly likely that a difference in variability will influence the likelihood of finding a significant difference. Because a lot of the time comparisons of single cases to control groups are not the most powerful to begin with (i.e., small control group size), it is important to trace the effect of increased variability in the single-case data on the power of the comparison.

To see how we might accommodate for differences in variability and actually increase the power of single-case testing, relative to the modified t test, we first need to go back to the underpinnings of the methods proposed by Crawford and Howell (1998) and Mycroft et al. (2002). Crawford and Howell assume that there is a population of controls, and this population has a true population mean μ_{control} and a population variance $\sigma_{\text{control}}^2$. The question that Crawford and Howell (1998) try to answer is whether the mean of the single case could conceivably come from this population of controls.

The assumptions of Mycroft et al. (2002) about the control population are identical to those of Crawford and Howell (1998). The difference is

in the assumptions that Mycroft et al. (2002) make about the single-case data. They assume that, just as the individual controls are sampled from a population of potential controls, the single case is sampled from a population as well (the single-case population). This single-case population has a true mean $\mu_{\text{single case}}$ and a variance $\sigma_{\text{single case}}^2$. The question that Mycroft et al. (2002) try to answer is whether the observed single-case data could conceivably come from a single-case population with a different mean from that of the control population. In practice, the difference between the methods of Crawford and Howell (1998) and Mycroft et al. (2002) is that Crawford and Howell do not allow for any influence of $\sigma_{\text{single case}}^2$, whereas Mycroft et al. do. Crawford and Howell (1998) use a t test, and Mycroft et al. (2002) use an F test, but because $t^2 = F$, the only real difference between their methods is the fact that Mycroft et al. (2002) adjust their critical value, whereas Crawford and Howell (1998) do not. In essence, Crawford and Howell (1998) assume that the variability for the controls and the single case is the same. Crawford et al. (2004) have pointed out that the single-case population assumed by Mycroft et al. (2002) is hypothetical, and that there is no good way of estimating the actual value of $\sigma_{\text{single case}}^2$. On the other hand, in our experience of single-case testing it does not seem unreasonable to assume that single-case data are frequently more variable than the data of the controls. Consequently, we agree with Mycroft et al. (2002) that we should not assume that the variability is the same for single case and control group.

Both Crawford and Howell (1998) and Mycroft et al. (2002) assume that the single case and the control group members are tested only once. What we propose is to test the single case repeatedly. This will make it possible to estimate $\sigma_{\text{single case}}^2$. Moreover, it will also enable us to increase the power of a comparison between a single case and a control group.

In our conceptualization of the comparison of a single case with a control group we follow Crawford and Howell (1998), and we do not assume that the single case is a member of a

hypothetical single-case population. As pointed out by one of our reviewers, a number of influential theorists have used neuropsychological considerations (e.g., Caramazza, 1986; Caramazza & McCloskey, 1988; Coltheart, 2001) to conclude that the single case is not a member of a patient population. That is, it has been argued that because (a) the functional architecture of cognition is enormously complex, and (b) there is substantial variability in the site and extent of naturally occurring lesions, each single case should be considered to be unique (Vallar, 2000). McCloskey (1993) provides an unequivocal expression of this position when he states, "In the single-patient approach, patients are not identified as members of patient populations" (p. 729).

Unlike Crawford and Howell (1998) however, we allow for retesting of the single case. The performance of the single case on the particular test will vary over repetitions of the test. This, $\sigma_{\text{single case}}^2$, is the source of the variation in the performance of the single case. Typically, this variation will be large, relative to the controls. For instance, Mitchell et al. (2004) refer to reaction time data reported by Moreno, Buchanan, and Van Orden (2002), where there was considerable variability in the reaction times.

The members of the control group belong to a population of potential control group members. The true level of performance will vary from one control group member to the next, so the scores of the control group members will straddle the true mean performance of the control group population. This, $\sigma_{\text{control}}^2$, is a source of variation when a particular set of control group members is tested. A different set of control group members will yield a different estimate of the true control group mean. Note that this source of variance is absent for the single case, because in a repetition of the experiment, we would test exactly the same single case with the same true mean. Please note that σ_{single}^2 and $\sigma_{\text{control}}^2$ are not equivalent: σ_{single}^2 is the within-subject variability of the single case, whereas $\sigma_{\text{control}}^2$ is the between-subject variability of the controls. We will assume that there is only a single test per control group member, because there will not be many instances where the controls are tested more than once.

So, according to our analysis, there are two sources of variation when a single case is compared with a control group: $\sigma_{\text{control}}^2$, reflecting the differences between the members of the control group, and $\sigma_{\text{single case}}^2$, reflecting the varying performance of the single case on a repetition of the same test. When a neuropsychologist tries to establish whether the single case has a mean score that is different from the controls, these are the sources of variation that should be taken into account.

So, the only difference between the approach we propose and the approaches proposed by Crawford and Howell (1998) and Mycroft et al. (2002) is the way we take into account $\sigma_{\text{single case}}^2$. Because we take it into account in the first place, our approach is necessarily different from that of Crawford and Howell. However, it also differs from Mycroft et al. (2002), because we do not assume that $\sigma_{\text{single case}}^2$ reflects the variability of a single-case population. Instead, we assume that it reflects the variability within the single case, from one test to the next. Just as Crawford and Howell (1998) is a special case of Mycroft et al. (2002) with equal variability for the control group members and the single case, Mycroft et al. (2002) can be considered a special case in our approach where the single case is tested only once.

After these theoretical preliminaries, we are now ready to look at the influence of increased single-case variability on the power of a comparison with a control group.

Increasing power

In a recent paper, Crawford and Garthwaite (2006) have compared the power of the modified t test with that of the method proposed by Mycroft et al. (2002) in a series of Monte Carlo simulations. They found that the modified t test had more power for all scenarios tested. Crawford and Garthwaite also reported that a difference in variability between single case and control group only has a marginal effect on the power to detect a 2-standard-deviation (SD) difference. We have repeated the simulation for three effect sizes (1 SD , 2 SD , and 3 SD). We used 100,000 simulations, with the random values generated by the algorithm *ran2* to draw from the standard normal distribution using the algorithm *gasdev* (Press, Teukolsky, Vetterling, & Flannery, 1992). Table 1 shows the power of the modified t test as a function of the variability in the single-case data, the size of the effect, and the size of the control group. From Table 1 it becomes clear that the power actually increases with increased variability in the single-case data for an effect size of 1 SD , stays reasonably constant for an effect size of 2 SD (as previously reported by Crawford & Garthwaite, 2006), and decreases with increased variability when the effect size becomes 3 SD . The power increase for the smallest effect size is a reflection of the increased likelihood of finding a significant result due to the higher

Table 1. Power estimates for the modified t test, as a function of control group size, effect size, and variance ratio, when allowing for increased single-case variability

Group size	Variance ratio								
	1			2			3		
	1 SD	2 SD	3 SD	1 SD	2 SD	3 SD	1 SD	2 SD	3 SD
5	.19	.45	.73	.24	.46	.68	.27	.46	.66
10	.22	.55	.84	.28	.53	.77	.32	.53	.73
20	.24	.59	.88	.30	.57	.81	.33	.56	.76
50	.25	.62	.90	.32	.59	.82	.35	.57	.77
100	.26	.63	.91	.32	.59	.83	.35	.58	.78

Note: Variance ratio is $s_{\text{single case}}^2/s_{\text{control}}^2$. Effect size (1 SD , 2 SD , 3 SD) relative to s_{control}^2 . Note that the column with variance ratio of 1 is equivalent to Table 1 in Crawford and Garthwaite (2006).

variability in the single-case data: the higher the variability, the larger the possibility of finding an extreme result. If, as Crawford et al. (2004) have argued, this increase in variability has to be accompanied by a drop in the mean level of performance, then there is no need to adjust the critical value to protect against a Type 1 error, and increased variability in the single-case data actually increases the likelihood of finding a significant result for small effects. An increase in power with an increase in variability of the single-case data for small effect sizes was reported earlier by Crawford and Garthwaite (2006). They argued that this result will be observed whenever the baseline power of a test is below .5.

Although the increase in power is reassuring, it still remains the case that the power to detect differences between single cases and controls generally remains quite low. We think it is possible to increase the power by retesting the single case. To see how, we have to go back to the F -adjustment proposed by Mycroft et al. (2002). Mycroft et al. (2002) provided a table with adjustments in F for various combinations of control group size and relative variability for the single-case data, based on Monte Carlo simulations. However, Crawford and Garthwaite (2006) found an analytical expression for these adjusted F -values:

$$\frac{(\sigma_{\text{single case}}^2 / \sigma_{\text{control}}^2) * n_{\text{control}} + 1}{(n_{\text{control}} + 1)} * F_{\text{original}} \quad (1)$$

The larger the ratio between $\sigma_{\text{single case}}^2$ and $\sigma_{\text{control}}^2$, the larger the adjustment in F -value. Because $\sigma_{\text{single case}}^2$ will in all likelihood be larger than $\sigma_{\text{control}}^2$, the adjustment will always be upward, resulting in an increased critical F -value and a decrease in power. In our conceptual model, we are able to incorporate retesting of the single case. However, we will have to take into account that $s_{\text{single case}}^2$ measures within-subject variability. The consequence of testing a single case repeatedly is that the test results will be correlated. This means that $s_{\text{single case}}^2$ will actually underestimate $\sigma_{\text{single case}}^2$. The higher the

correlation, the higher the underestimation. Without a correction for this underestimation, adjusted F -levels would become too low, and Type 1 error would increase. With correction for correlation, the adjustment in F -value becomes:

$$\frac{\{[1 + (n_{\text{single case}} - 1) * r_{\text{single case}}] * s_{\text{single case}}^2\}}{[(1 - r_{\text{single case}}) * n_{\text{single case}} * s_{\text{control}}^2] * n_{\text{control}} + 1} \\ n_{\text{control}} + 1 \\ * F_{\text{original}} \quad (2)$$

This change in the formula opens up the possibility of a downward adjustment for the F -value: by increasing the $n_{\text{single case}}$, it becomes possible to make the ratio

$$\frac{[1 + (n_{\text{single case}} - 1) * r_{\text{single case}}] * s_{\text{single case}}^2}{(1 - r_{\text{single case}}) * n_{\text{single case}} * s_{\text{control}}^2}$$

smaller than 1. Whenever this is the case, the adjusted F -value will be smaller than the original F -value. This will result in an increase of the power of the test. The reason for this increase in power is the following: The outcome of the F -test is determined by s_{control}^2 , because this is used to estimate the variability of the difference between the performance of the single case and the performance of the control group. There will be an overestimation of the real variability of this difference whenever

$$\frac{[1 + (n_{\text{single case}} - 1) * r_{\text{single case}}] * s_{\text{single case}}^2}{(1 - r_{\text{single case}}) * n_{\text{single case}} * s_{\text{control}}^2} < 1.$$

In turn this means that the critical F -value becomes overly conservative, reducing the chance of a significant result.

There are two caveats, however. For larger values of $r_{\text{single case}}$, the number of retests needed to achieve a decrease in the critical F -value becomes prohibitive, even for modest ratios of $s_{\text{single case}}^2 / s_{\text{control}}^2$. This is illustrated in Figure 1. Second, power is also determined by the effect size. Whenever the data are more variable, extreme

Figure 1. The number of retests needed to achieve a decrease in the critical F -value as a function of the variance ratio $s^2_{\text{single case}}/s^2_{\text{control}}$ and correlation $r_{\text{single case}}$. Note that the number of retests needed increases steeply with increased correlation. For large correlations combined with large variance ratios, a decrease in critical F -value will never be achieved.

results become more frequent. Whenever retesting takes place, the estimate of an effect becomes more stable. For small effects, this might actually have a detrimental effect on the power of the test. As we have seen in Table 1, the smallest effect sizes pocket the largest increase in power due to increased variability in the single-case data. This increase can only be explained by the increased frequency of extreme values for the single case, because the variability for the control group has remained constant. By retesting the single case, we are reducing the frequency of extreme values. This in turn will reduce the power of the comparison.

Monte Carlo simulations

Determining the power analytically becomes difficult when there is a difference in variability between the single case and the control group. We therefore used Monte Carlo simulations to determine the influence of retesting the single case for a combination of correlations, different effect sizes, and number of retests. We compared the power of the modified t test and our adjusted F -test, which takes retesting into account. In Tables 2–4, we have listed the power yielded by the most powerful method. To make the comparison fair, we used an alpha of .1 for our F -test. This is the equivalent of the one-sided modified t test that has been advocated by Howell and Crawford

(1998). The one-sidedness of the significance test has the consequence that any result where the single case has significantly better performance than the control group does not count towards the observed power. Rather than concluding that the postmorbid performance of the single case has improved relative to the controls, we would conclude that postmorbid performance has not deteriorated relative to the controls.

Tables 2–4 show that whenever

$$\frac{[1 + (n_{\text{single case}} - 1) * r_{\text{single case}}] * s^2_{\text{single case}}}{(1 - r_{\text{single case}}) * n_{\text{single case}} * s^2_{\text{control}}} \geq 1,$$

most power is afforded by the modified t test. This is by virtue of the fact that the modified t test is using a critical value that is lower than the equivalent critical value used by our adjusted F -test. Consequently, the adjusted F -test obtains larger power than the modified t test for those entries where

$$\frac{[1 + (n_{\text{single case}} - 1) * r_{\text{single case}}] * s^2_{\text{single case}}}{(1 - r_{\text{single case}}) * n_{\text{single case}} * s^2_{\text{control}}} < 1.$$

However, from a practical point of view, the more important question is whether an increase in the number of retests actually yields larger power than testing the single case only once. As can be seen from Tables 2–4, this depends on the correlation $r_{\text{single case}}$, the effect size, and the variance ratio $s^2_{\text{single case}}/s^2_{\text{control}}$. The larger the variance ratio, and the smaller the effect size, the more retests are needed to improve on the power afforded by a single test of the single case evaluated with the modified t test. This effect is exacerbated for larger values of $r_{\text{single case}}$. For small effect sizes, the increase in power as a result of the downward adjustment of the critical F -value is (partially) offset by the decrease in frequency of large effects. But, whenever the effect size is large, or the variance ratio between single case and control group is small,

Table 2. Power as a function of control group size, variance ratio, effect size, and number of retests of the single case where $r_{\text{single case}} = 0$

Group size	Number of retests	Variance ratio								
		1			2			3		
		1 SD	2 SD	3 SD	1 SD	2 SD	3 SD	1 SD	2 SD	3 SD
5	1	.19	.45	.73	.24	.46	.68	.27	.46	.66
	2	.26*	.63*	.90*	.19	.45	.73	.22	.45	.70
	3	.31*	.73*	.95*	.23*	.55*	.84*	.19	.45	.73
	4	.34*	.79*	.98*	.26*	.63*	.90*	.22*	.52*	.81*
	5	.37*	.82*	.98*	.28*	.68*	.93*	.24*	.58*	.86*
10	1	.22	.55	.84	.29	.53	.77	.32	.53	.73
	2	.33*	.77*	.97*	.22	.55	.84	.26	.54	.80
	3	.41*	.88*	.99*	.28*	.68*	.93*	.22	.55	.84
	4	.47*	.93*	1.00*	.33*	.77*	.97*	.26*	.64*	.91*
	5	.52*	.96*	1.00*	.37*	.83*	.99*	.29*	.71*	.95*
20	1	.24	.59	.88	.30	.57	.80	.34	.55	.76
	2	.37*	.83*	.99*	.24	.59	.88	.28	.58	.84
	3	.46*	.93*	1.00*	.31*	.74*	.96*	.24	.59	.88
	4	.55*	.97*	1.00*	.37*	.83*	.99*	.28*	.70*	.94*
	5	.61*	.99*	1.00*	.42*	.89*	1.00*	.33*	.77*	.97*
50	1	.25	.62	.90	.32	.59	.82	.35	.57	.77
	2	.39*	.86*	.99*	.25	.62	.90	.29	.60	.86
	3	.50*	.95*	1.00*	.32*	.77*	.97*	.25	.62	.90
	4	.60*	.98*	1.00*	.39*	.86*	.99*	.30*	.73*	.96*
	5	.68*	.99*	1.00*	.45*	.92*	1.00*	.35*	.80*	.98*

Note: Variance ratio is $s_{\text{single case}}^2/s_{\text{control}}^2$. Effect size (1 SD, 2 SD, 3 SD) relative to s_{control}^2 . Only the power provided by the most powerful method is shown. Asterisks identify the entries where our adjusted critical F -value as computed in (2) yields most power; entries without asterisks have been obtained by using the modified t test.

Table 3. Power as a function of control group size, variance ratio, effect size, and number of retests of the single case where $r_{\text{single case}} = .1$

Group size	Number of retests	Variance ratio								
		1			2			3		
		1 SD	2 SD	3 SD	1 SD	2 SD	3 SD	1 SD	2 SD	3 SD
5	1	.20	.45	.72	.25	.46	.68	.28	.46	.66
	2	.24*	.58*	.86*	.20	.45	.71	.23	.46	.69
	3	.27*	.66*	.92*	.20*	.48*	.76*	.21	.45	.71
	4	.30*	.71*	.94*	.22*	.53*	.82*	.20	.45	.72
	5	.32*	.74*	.96*	.24*	.57*	.85*	.20*	.47*	.75*
10	1	.23	.54	.83	.29	.53	.76	.32	.53	.72
	2	.29*	.71*	.95*	.24	.54	.82	.27	.54	.78
	3	.35*	.80*	.98*	.24*	.58*	.87*	.25	.54	.81
	4	.39*	.86*	.99*	.27*	.65*	.92*	.23	.55	.83
	5	.42*	.89*	1.00*	.29*	.70*	.95*	.23*	.57*	.86*
20	1	.25	.59	.87	.31	.56	.79	.34	.55	.75
	2	.32*	.77*	.97*	.26	.58	.86	.30	.57	.81
	3	.39*	.86*	.99*	.26*	.64*	.91*	.27	.58	.85
	4	.44*	.91*	1.00*	.29*	.71*	.95*	.25	.59	.87
	5	.48*	.94*	1.00*	.32*	.76*	.97*	.25*	.62*	.90*
50	1	.26	.62	.89	.32	.58	.81	.35	.57	.76
	2	.34*	.80*	.98*	.27	.61	.88	.31	.59	.83
	3	.42*	.89*	1.00*	.27*	.66*	.93*	.28	.61	.87
	4	.48*	.94*	1.00*	.31*	.74*	.96*	.26	.62	.89
	5	.53*	.96*	1.00*	.34*	.79*	.98*	.26*	.65*	.92*

Note: Variance ratio is $s_{\text{single case}}^2/s_{\text{control}}^2$. Effect size (1 SD, 2 SD, 3 SD) relative to s_{control}^2 . Only the power provided by the most powerful method is shown. Asterisks identify the entries where our adjusted critical F -value as computed in (2) yields most power; entries without asterisks have been obtained by using the modified t test.

Table 4. Power as a function of control group size, variance ratio, effect size, and number of retests of the single case where $r_{\text{single case}} = .2$

Group size	Number of retests	Variance ratio								
		1			2			3		
		1 SD	2 SD	3 SD	1 SD	2 SD	3 SD	1 SD	2 SD	3 SD
5	1	.21	.45	.72	.26	.46	.67	.29	.47	.65
	2	.22*	.52*	.81*	.22	.45	.70	.25	.46	.68
	3	.24*	.59*	.87*	.20	.45	.72	.23	.46	.69
	4	.26*	.63*	.90*	.19	.45	.73	.22	.45	.70
	5	.27*	.65*	.92*	.20*	.47*	.76*	.21	.45	.71
10	1	.24	.54	.82	.30	.53	.75	.33	.53	.71
	2	.26*	.64*	.91*	.26	.54	.80	.29	.53	.76
	3	.30*	.72*	.95*	.24	.54	.82	.27	.54	.78
	4	.33*	.77*	.97*	.22	.55	.84	.26	.54	.80
	5	.35*	.80*	.98*	.24*	.58*	.87*	.25	.54	.81
20	1	.26	.58	.86	.32	.56	.78	.35	.55	.74
	2	.29*	.70*	.94*	.28	.58	.84	.31	.56	.79
	3	.33*	.78*	.98*	.26	.59	.86	.29	.57	.82
	4	.37*	.83*	.99*	.24	.59	.88	.28	.58	.84
	5	.39*	.86*	.99*	.26*	.63*	.91*	.27	.58	.85
50	1	.27	.61	.88	.33	.58	.80	.36	.56	.75
	2	.30*	.73*	.96*	.29	.60	.86	.33	.58	.81
	3	.36*	.81*	.98*	.27	.61	.88	.30	.59	.84
	4	.39*	.86*	.99*	.25	.62	.90	.29	.60	.85
	5	.42*	.89*	1.00*	.27*	.66*	.92*	.28	.60	.87

Note: Variance ratio is $s_{\text{single case}}^2/s_{\text{control}}^2$. Effect size (1 SD, 2 SD, 3 SD) relative to s_{control}^2 . Only the power provided by the most powerful method is shown. Asterisks identify the entries where our adjusted critical F -value as computed in (2) yields most power; entries without asterisks have been obtained by using the modified t test.

retesting once or twice does increase the power of the test, especially with a larger control group. Although the power increase is reduced by an increase in $r_{\text{single case}}$, it does not disappear.

As can be seen from Tables 2–4, a similar pattern can be discerned for the modified t test. It is only for a combination of a large variance ratio $s_{\text{single case}}^2/s_{\text{control}}^2$, a small effect size, and a large value for $r_{\text{single case}}$ that retesting results in a decrease of power for both methods.

Why does the power of the comparison between a single case and a control group depend not only on the number of retests, but also on the correlation $r_{\text{single case}}$, the size of the effect, and the variance ratio $s_{\text{single case}}^2/s_{\text{control}}^2$? Although intuitively it might seem that retesting the single case should always be a good idea, it turns out that there are scenarios where retesting actually results in a reduction of power. This state of affairs is the direct consequence of two factors that help to improve the power to detect small differences. First, there is the assumption that increased variability in the single case has to be accompanied by a decrease in level of performance. Without this assumption (i.e., in a scenario where increased variability could be accompanied by a static level of performance), we would have to increase the critical value to compensate for an increase in Type 1 error. (This is in fact what has been argued by Mycroft et al., 2002.) Crawford and Garthwaite (2006) have shown that countering this increase in Type 1 error results in a massive power reduction. Second, the larger variability in the single-case data increases the probability to find an extreme result. When the single

case is retested, the probability of an extreme result is reduced. So, although retesting actually gives a more stable estimate of the true level of performance in the single case, this is not very helpful for small effect sizes, because testing only once gives a higher chance of finding an extreme value. But because we assume that increased variability is always accompanied by a decrease in level of performance, this does not constitute a Type 1 error. Still, there is a price to pay: Testing only once in the presence of larger variability in the single-case data might result in an overestimation of the actual effect size. So, although a neuropsychologist is more likely to find a significant result, the actual size of the deficit becomes more uncertain.

There is therefore a trade-off. Unless the single case is retested, $s_{\text{single case}}^2/s_{\text{control}}^2$ can not be computed.¹ However, depending on the value of $s_{\text{single case}}^2/s_{\text{control}}^2$, retesting might actually reduce the chance of finding a significant result. So, although it might be gratifying to arrive at a more stable estimate of the true level of performance of the single case, it will be grating if an effect that was significant after testing only once, is suddenly no longer. It probably depends on what goal a neuropsychologist is trying to achieve. If it is establishing the presence of a deficit, testing only once might be the best way forward. However, when the size of the deficit is important as well, retesting has to be considered. Our simulations show that most of the time, it is possible to compensate for any initial loss of power, by retesting several times.

¹ Although $s_{\text{single case}}^2$ and s_{control}^2 are the best estimators of $\sigma_{\text{single case}}^2$ and $\sigma_{\text{control}}^2$, they are actually surrounded by a confidence interval. The true value of $\sigma_{\text{single case}}^2$ and $\sigma_{\text{control}}^2$ will lie somewhere in that confidence interval. The width of the confidence interval is determined by the number of degrees of freedom for s^2 . For instance, an estimate of σ^2 based on 100 degrees of freedom will have a narrower confidence interval than an estimate based on only 3 degrees of freedom. From our simulations it has become clear that the relative values of the upper and lower limit of the confidence interval only depend on the number of degrees of freedom of s^2 , rather than on the value of s^2 itself. This means that a neuropsychologist does not need to know σ^2 (as will probably always be the case) to construct a confidence interval around s^2 . The observation that the estimates of the σ^2 -values are surrounded by a confidence interval is important, because we have seen that the size of $\sigma_{\text{single case}}^2$ relative to the size of $\sigma_{\text{control}}^2$ is very important for the power of the F -test. By moving around inside the confidence intervals, we will influence the variance ratio between control group and single case, making our test either more conservative or more liberal. For a conservative estimate of the critical F -value, a neuropsychologist could use the upper limit of the confidence interval for $\sigma_{\text{single case}}^2$ and the lower limit of the confidence intervals for $\sigma_{\text{control}}^2$. For a liberal estimate, a neuropsychologist could use the lower limit of the confidence interval for $\sigma_{\text{single case}}^2$ and the upper limit of the confidence intervals for $\sigma_{\text{control}}^2$.

Whether retesting is possible in practice probably depends on the kind of test that is being administered. At one end of the spectrum there are reaction time tests. In our approach, every single reaction time datum would add to the number of tests performed by the single case. Here, increasing the number of tests should be easy and almost effortless. At the other end of the spectrum there are the tests that after several hours of testing yield a single number (e.g., the derivation of a parameter value in the formal analysis of visual attention given by "TVA"; see Duncan et al., 1999). Here, administering the test twice could be the maximum achievable, and it would be better to test the single case only once and use the modified t test (but note that this will only yield the presence of a deficit).

An example

To provide an example of our approach, we now briefly report data from a case study of a patient tested on the "Pyramids and Palm Trees" (PPT) test (Howard & Patterson, 1992). In the PPT test, a patient is presented with three pictures—one target and two items that can be matched to the target. The task is to choose which of the latter pictures is semantically the closest to the target. For example, for the target "pyramid" the choice alternatives are "palm tree" and "deciduous tree", and the correct match would be pyramid and palm tree. Patient K.S. was reported in detail in Humphreys and Riddoch (2003). Over the period of testing, she was 70–72 years old and had suffered herpes simplex encephalitis some 6 years previously. Following this infection, she sustained damage to the middle and anterior regions of her left temporal lobe. She presented with surface dyslexia and dysgraphia and a category-specific impairment for living things. The deficit for living things was particularly pronounced for naming tasks, but she performed rather better on tasks requiring matching on the basis of semantic knowledge (e.g., the associative matching task from the Birmingham Object Recognition Battery, BORB; Riddoch & Humphreys, 1993). We have assessed K.S.'s performance on the

pictorial version of the PPT on four occasions (respectively, 36, 42, 60, and 68 months after herpes simplex infection), where she scored 49, 45, 48, and 44/52, respectively (mean = 46.5, $SD = 2.38$). Howard and Patterson (1992) report that a score below 90% correct (47/52) is impaired, but they provide no statistics on the reliability of this cut-off. We collected data from 9 controls who were matched for age and background to K.S., and they scored 48, 50, 52, 49, 51, 48, 48, 52, and 50, respectively (mean = 49.8/52, $SD = 1.64$). Using a 2 SD cut-off, K.S. would fall just within the control population (cut-off 46.5), but she would be judged as impaired using Crawford and Howell's (1998) approach, $t(8) = 1.89$, $p < .048$, one-sided. The ratio of K.S.'s variance compared to the variance across the controls was 2.1. If we assume a correlation of .1, our approach, which takes into account our retesting of K.S., finds that the adjusted critical F -value is 2.71, and the adjusted (one-sided) $p < .033$. We suggest that our approach could provide a more sensitive means to testing a single case against a control population: For this example, the power estimates are .54 for the modified t test and .62 for our approach, respectively.

When to use adjusted F and when to use modified t

In our opinion, the adjusted F -approach and the modified t test are interchangeable. Although the adjusted F -approach was developed specifically for experiments where single cases have been tested repeatedly, it also encompasses experiments where the single case has been tested only once. (Similarly, the modified t test can be used when the single case has been tested repeatedly.) However, only in the case where $\sigma_{\text{single case}}^2$ is smaller than $\sigma_{\text{control}}^2$ would the adjusted F have more power than the modified t . Moreover, this smaller value of $\sigma_{\text{single case}}^2$ would have to come from the literature, because the experiment itself would not yield an estimate of $\sigma_{\text{single case}}^2$, exactly because the single case has been tested only once.

In general, which method should be preferred for a particular experiment strictly depends on the value of

$$\frac{[1 + (n_{\text{single case}} - 1) * r_{\text{single case}}] * s_{\text{single case}}^2}{(1 - r_{\text{single case}}) * n_{\text{single case}} * s_{\text{control}}^2}$$

Whenever this ratio is larger than 1, the modified t test has most power; whenever it is smaller than 1, the adjusted F -approach is more powerful. When the ratio equals 1, both methods will yield identical outcomes.

Program

We have implemented the procedures used in this paper in a user-friendly program available on <http://www.hull.ac.uk/psychology/staff/pages/johan/adjustedfcalculator>. The program computes p -values for both the modified t test and our adjusted F -approach. It highlights the results of the approach with most power in yellow. Not only can the program be used to evaluate the outcome of a particular experiment; as suggested by one of our reviewers, it should also provide neuropsychologists with a very useful "laboratory" for exploring the potential effects of differing levels of intraindividual and between-subjects variability in patients and controls on inferential methods.

Manuscript received 12 October 2005

Revised manuscript received 11 December 2006

Revised manuscript accepted 18 December 2006

First published online 21 February 2007

REFERENCES

- Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, 5, 41–66.
- 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, PA: Psychology Press.
- Crawford, J. R., & Garthwaite, P. H. (2006). Methods of testing for a deficit in single case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology*, 23, 877–904.
- Crawford, J. R., Garthwaite, P. H., Azzalini, A., Howell, D. C., & Laws, K. R. (2006). Testing for a deficit in single-case studies: Effects of departures from normality. *Neuropsychologia*, 44, 666–677.
- Crawford, J. R., Garthwaite, P. H., Howell, D. C., & Gray, C. D. (2004). Inferential methods for comparing a single case with a control sample: Modified t tests versus Mycroft et al.'s (2002) modified ANOVA. *Cognitive Neuropsychology*, 21, 750–755.
- 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.
- Duncan, J., Bundesen, C., Olson, A., Humphreys, G. W., Chavda, S., & Shibuya, H. (1999). Systematic analysis of deficits in visual attention. *Journal of Experimental Psychology: General*, 128, 450–478.
- Howard, D., & Patterson, K. E. (1992). *The Pyramids and Palm Trees Test*. Bury St. Edmonds, UK: Thames Valley Test Co.
- Huber, H. P. (1973). *Psychometrische Einzelfalldiagnostik* [Psychometric assessment of single cases]. Weinheim, Germany: Beltz.
- Humphreys, G. W., & Riddoch, M. J. (2003). A case series analysis of category-specific deficits of living things: The HIT account. *Cognitive Neuropsychology*, 20, 263–306.
- McCloskey, M. (1993). Theory and evidence in cognitive neuropsychology: A "radical" response to Robertson, Rafal, and Shimamura (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 718–734.
- Mitchell, D. C., Mycroft, R. H., & Kay, J. (2004). Comparing a single case to a control sample: Differences in distribution versus differences in means. *Cognitive Neuropsychology*, 21, 756–760.
- Moreno, M. A., Buchanan, L., & Van Orden, G. C. (2002). Variability in aphasic patients' response times. *Brain and Cognition*, 48, 469–474.
- Mycroft, R., 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.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C. The art of scientific computing* (2nd ed.). Cambridge, MA: Cambridge University Press.

- Riddoch, M. J., & Humphreys, G. W. (1993). *The Birmingham Object Recognition Battery (BORB)*. London: Lawrence Erlbaum Associates Ltd.
- 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.
- Zubin, J. (1950). Symposium on statistics for the clinician. *Journal of Clinical Psychology*, 6, 1–6.