Neuropsychology
2007, Vol. 21, No. 4, 419–430

0894-4105/07/$12.00 DOI: 10.1037/0894-4105.21.4.419

Estimating the Percentage of the Population With Abnormally Low Scores (or Abnormally Large Score Differences) on Standardized Neuropsychological Test Batteries: A Generic Method With Applications

John R. Crawford
University of Aberdeen

Paul H. Garthwaite
The Open University

Catherine B. Gault
University of Aberdeen

Information on the rarity or abnormality of an individual's test scores (or test score differences) is fundamental in interpreting the results of a neuropsychological assessment. If a standardized battery of tests is administered, the question arises as to what percentage of the healthy population would be expected to exhibit one or more abnormally low test scores (and, in general, *j* or more abnormally low scores). Similar issues arise when the concern is with the number of abnormal pairwise differences between an individual's scores on the battery, or when an individual's scores on each component of the battery are compared with the individual's mean score. A generic Monte Carlo simulation method for tackling such problems is described (it requires only that the matrix of correlations between tests be available) and is contrasted with the use of binomial probabilities. The method is then applied to Index scores for the Wechsler Adult Intelligence Scale—Third Edition (WAIS–III; D. Wechsler, 1997) and Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV; D. Wechsler, 2003). Three computer programs that implement the methods are made available.

Keywords: neuropsychological assessment, multiple tests, single-case inference, Monte Carlo methods, prevalence of deficits

Information on the rarity or abnormality of test scores (or test score differences) is fundamental in interpreting the results of a neuropsychological assessment (Crawford, 2004; Strauss, Sherman, & Spreen, 2006). When attention is limited to a single test, this information is immediately available; if an abnormally low score is defined as one that falls below the 5th percentile, then, by definition, 5% of the population is expected to obtain a score that is lower (for example, in the case of Wechsler IQs or Indexes, scores of 75 or lower are below the 5th percentile). However, multiple tests are used in neuropsychological assessment. Therefore, the important question arises as to what percentage of the healthy population would be expected to exhibit at least one abnormally low test score. This percentage will be higher than for any single test, and knowledge of it is liable to guard against overinference; that is, concluding impairment is present on the basis of one "abnormally" low score when such a result is not at all uncommon in the general, healthy population. It is also important to know what percentage of the population would be expected to obtain two or more, or three or more, abnormal scores; in general, it is important to know what percentage of the population would be expected to exhibit *j* or more abnormally low scores.

One approach to this issue would be to tabulate the percentages of a test battery's standardization sample exhibiting *j* or more abnormal scores; that is, the question could be tackled empirically. However, to our knowledge, such base rate data have not been provided for test batteries commonly used in neuropsychology. Thus, it would be very useful if the required percentages could be estimated statistically. This could easily be achieved if the tests in the battery were either independent (i.e., uncorrelated) or perfectly correlated.

When tests are independent, the required percentages can be obtained using the binomial distribution (Ingraham & Aiken, 1996). For example, if a battery of six tests is administered, and abnormality is defined as a score falling below the 5th percentile, then the percentage of the population expected to show one or more abnormal scores is 26.49%. This result is obtained by finding the binomial probability for *r* or more "successes" (i.e., $r = 1$ if the interest is in one or more abnormal scores) in $n = 6$ trials (i.e., tests), with *p*, the probability of "success" (i.e., an abnormal score) on each individual trial, set at .05. Multiplying the resultant probability by 100 provides the estimate of the required percentage. Note that the percentage obtained is less than the sum of the percentages for each test (which are all 5%) because a small number of individuals exhibiting an abnormal score on one of the tests will, by chance, also exhibit an abnormal score on one or more of the other tests but will only contribute once to the former percentage.

John R. Crawford and Catherine B. Gault, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen, United Kingdom; Paul H. Garthwaite, Department of Statistics, The Open University, Milton Keynes, United Kingdom.

Correspondence concerning this article should be addressed to John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 3HN, United Kingdom. E-mail: j.crawford@abdn.ac.uk

At the other extreme, if the tests were all perfectly correlated, the percentage of the population exhibiting one or more abnormal scores would be identical to the percentage for any individual test—that is, 5% (if any one test falls below the 5th percentile, then so will all others). The problem with both of these scenarios is that neither is at all realistic for tests commonly used in neuropsychological assessment. As an example, the average correlation between the four Index scores of the Wechsler Adult Intelligence Scale—Third Edition (WAIS–III; Wechsler, 1997) and Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV; Wechsler, 2003) are 0.61 and 0.51 respectively (these figures were obtained from the correlations between Indexes presented in the relevant test manuals).

Ingraham and Aiken (1996) suggested that the use of binomial probabilities is justified even when the tests are not independent, provided that potential users are aware that the estimates provided are upper limits on the percentage of the population exhibiting abnormal scores. That is, they acknowledge that the binomial approach can overestimate the required percentages.

There are two potentially serious problems with Ingraham and Aiken's (1996) suggestion of using binomial probabilities. First, if the correlations between tests are substantial (as they often are for neuropsychological measures), the binomial estimates of the percentage exhibiting at least one abnormally low score will be very inflated. Thus, a patient who exhibits one abnormally low score will look much less unusual than is truly the case. Second, unfortunately, Ingraham and Aiken are wrong in asserting that the binomial estimates invariably provide upper limits on the percentage exhibiting abnormal scores. Although this assertion holds when $j = 1$ (i.e., when it is required to estimate the percentage exhibiting one or more abnormal scores), it will very commonly be the case that the binomial estimates will underestimate the percentages when j is greater than one (i.e., when the interest is in the percentage exhibiting two or more, or three or more, abnormal tests, etc.). Thus, using binomial probabilities, a patient with (say) three abnormally low scores may be estimated to be very unusual when, in fact, an appreciable percentage of the general population will exhibit this number of abnormally low scores.

In view of these problems with the use of binomial probabilities, it is worth exploring alternative approaches. Given that the focus of the present article is on obtaining base rate information for *standardized* test batteries, it is assumed that the tests concerned follow a multivariate normal distribution (standardized batteries are designed to possess this very property). It might be thought that this assumption would allow us to directly calculate the required probabilities when the tests are not independent. However, obtaining a solution to this type of problem by direct analytic means is remarkably difficult (Gentz, 1992; Ingraham & Aiken, 1996). This is because conditional distributions derived from a multivariate normal distribution do not, in general, have a closed form when the condition is an inequality (e.g., the score on Test X is abnormally low) rather than an equality (e.g., the score on Test X is 36). Numerical multiple integration of tail areas from these awkward conditional distributions would be necessary if the required probabilities are to be obtained through calculation.

Given these difficulties, an approach based on Monte Carlo simulation methods has obvious appeal. It has the advantage that it is relatively easy to implement, and—if a large number of Monte Carlo trials are run—it will achieve a high level of accuracy in estimating the quantities required. In the present article, we outline a generic Monte Carlo approach to answering questions of this type and related questions that are also of relevance to neuropsychological assessment. Before setting out the method, we turn first to outlining these related questions.

Number of Abnormal Pairwise Differences Between Components of a Battery

Comparison of an individual's test scores against normative data is a basic part of the assessment process. However, in neuropsychology, such normative comparison standards should be supplemented with the use of *individual* comparison standards when attempting to detect and quantify the extent of any acquired impairments (Crawford, 2004; Lezak, Howieson, Loring, Hannay, & Fischer, 2004). For example, a patient of high premorbid ability may score at or close to the mean of a normative sample, but this may still represent a serious decline for the individual concerned. Conversely, a patient may score well below the normative mean, but this may be entirely consistent with the individual's premorbid ability. Because of this, emphasis is placed on the use of individual comparison standards: Most tests used in neuropsychology are at least moderately correlated in the general population; thus, large discrepancies in a patient's test profile suggest an acquired impairment on those tasks that are performed relatively poorly (Crawford, 1992).

Base rate data are available for a number of tests used in neuropsychology to assist clinicians to quantify the abnormality of any discrepancies exhibited by their patients. Although these data provide invaluable information, an obvious issue arises: If a patient's profile of strengths and weaknesses are examined, then, by definition, multiple comparisons are involved. Therefore, although the percentage of the population expected to exhibit a given size of discrepancy between each possible test pair is readily available for many test batteries used in neuropsychology (e.g., for Index scores on the WAIS–III and WISC–IV), it would be useful to know what percentage would exhibit j or more abnormal pairwise differences overall. To our knowledge, base rate data of this latter form are not available for batteries currently in use in neuropsychology. Fortunately, however, it is only a little more complicated to estimate the required percentages using Monte Carlo simulation methods than it is to estimate the percentage expected to exhibit a given number of abnormally low scores.

Number of Abnormal Differences Between Components of a Battery and an Individual's Mean Score

An alternative to pairwise comparisons of an individual's scores is to obtain the individual's mean score on the components of the battery and compare each component with this mean. This approach, developed by Silverstein (1984), has a number of advantages over the former approach (Crawford & Allan, 1996). These are perhaps most evident with large batteries, in which Silverstein's approach serves to reduce the number of comparisons involved to manageable proportions. For example, if there are 10 tests in a battery, there are 45 possible pairwise comparisons.

Even with a smaller number of components, Silverstein's (1984) method, by providing a common individualized comparison stan-

dard for each test, has advantages. In arriving at a formulation, neuropsychologists have to integrate the information from a profile analysis of a given battery with other test data and information from a host of other sources (i.e., the medical history, the clinical interview, behavioral observations, etc.). Anything that eases this burden is to be encouraged. For example, we agree with Longman's (2004) suggestion that Silverstein's approach should also be preferred over the pairwise approach when analyzing WAIS–III Index scores. Longman prepared a table to allow neuropsychologists to estimate the abnormality of the deviations of Index scores from patients' mean Index scores. An equivalent table for the WISC–IV has been provided by Flanagan and Kaufman (2004). These tables provide very useful information to aid test interpretation; as part of the present study, we supplement them by examining the percentage of the population expected to exhibit *j* or more deviations (see next section).

Estimating Base Rates for the Number of Abnormal Scores and Score Differences for WAIS–III and WISC–IV Indexes

Analysis of the WAIS–III can be conducted at the level of the subtests, Index scores, or IQs. A strong case can be made for basing the primary analysis of these scales at the level of the Index scores. Index scores have the advantage that they are more reliable than the individual subtests (and only marginally less reliable than the IQs). In addition, because they reflect the factor structure of the instrument, they have superior construct validity to the IQs. Moreover, empirical evidence indicates that such factor-based composites are better able to differentiate between healthy and impaired performance than either the IQs or measures of subtest scatter (Crawford, Johnson, Mychalkiw, & Moore, 1997).

In view of the foregoing points, in the present study we focus on WAIS–III Index scores rather than subtests. The percentage of the population expected to exhibit *j* or more abnormally low Index scores is quantified, as is the percentage expected to exhibit *j* or more abnormal pairwise differences between Index scores; finally, the percentage expected to exhibit *j* or more abnormal deviations from an individual's mean Index score is quantified. The same data are generated for the WISC–IV.

Method

The Generic Simulation Method

To conduct the simulations, neuropsychologists need to have access to **R**, the $k \times k$ matrix of correlations between the k components (i.e., subtests or Indexes, etc.) of the test battery. Fortunately, however, this is the only information that is required. For most standardized batteries of tests, this correlation matrix will be available in the user manual or an accompanying technical manual.

The starting point for the simulation is to obtain the Choleski decomposition of **R**. The Choleski decomposition, which can be seen as the square root of **R**, takes the form of a lower triangular matrix. Macros for widely used spreadsheet or statistical packages are available to perform this decomposition, as are algorithms for most computer languages. This step is only performed once for a given simulation.

The next step (Step 2) is to generate a random vector of k independent, standard normal variates, where k is the number of tests in the battery of interest (all standard statistical packages and spreadsheet packages can generate random variates of this form, either directly or with the use of widely available macros). The vector of independent standard normal variates is then postmultiplied by the lower triangular Choleski decomposition matrix (Step 3) and is then an observation from the desired multivariate normal distribution with mean vector 0 and covariance matrix **R**. Steps 2 and 3 are repeated a large number of times; in the examples used in the present article, we draw one million vectors (i.e., to represent the scores of one million cases) for each problem studied. An example of generating an observation using this method is provided in the Appendix.

Note that the observations obtained by this process have means and standard deviations of 0 and 1, respectively. It would be easy to obtain observations that had the same means and standard deviations as tests in the battery of interest: One would simply multiply the observations by the desired standard deviation (say 15 as in the Wechsler scales) and add the desired mean (say 100). However, there is absolutely no need to do this: The results obtained would be identical because the transformation is linear.

The above steps are employed regardless of the question being posed. We turn now to the procedures adopted to address the three problems identified earlier.

Estimating the Percentage of the Population Exhibiting j or More Abnormally Low Scores

When the aim is to estimate the percentage of the population that would exhibit *j* or more abnormally low scores, the procedure is very straightforward. The first step is to decide the criterion used to define an abnormally low score on a subtest or Index of the battery and to translate this into a standard normal deviate. In the example, we define an abnormally low score as a score that falls below the 5th percentile and so the corresponding standard normal deviate is –1.645. A researcher or clinician may prefer to define an abnormally low score as one that falls below the 10th percentile, in which case the standard normal deviate required is –1.282; alternatively, if an abnormally low score is defined as a score more than one standard deviation below the mean, then the value required is simply –1.0.

To estimate the percentage of the population that would exhibit *j* or more abnormal scores, the number of abnormal scores obtained on each Monte Carlo trial (i.e., for each simulated member of the population) is recorded and a tally kept of the number exhibiting *j* abnormal scores across the course of the simulation. For example, say that only 1,000 trials were run, that 200 cases exhibited one abnormal score each, and a further 100 cases exhibited two abnormal scores. Then the estimate from this simulation would be that 30% of the population will exhibit one or more abnormal scores, and 10% will exhibit two or more abnormal scores.

Estimating the Percentage of the Population Exhibiting j or More Abnormal Pairwise Differences Between Scores

To estimate the percentage of the population exhibiting *j* or more abnormal differences between pairs of scores, it is necessary

to calculate the standard deviation of the difference between each possible pair of tests; if there are k tests, then there are $k(k-1)/2$ possible pairs (e.g., if there are 6 tests, then there are 15 pairs). The formula for the standard deviation of the difference (s_{X-Y}) between standard normal scores is

$$s_{X-Y} = \sqrt{2 - 2r_{XY}}, \quad (1)$$

where r_{XY} is the correlation between any given pair of tests in the battery. The standard deviations of the difference between each pair of tests is then multiplied by the standard normal deviate corresponding to the criterion adopted for an abnormal difference. For example, if an abnormal difference is defined as a difference that would occur in less than 5% of the population regardless of sign (i.e., regardless of the direction of the difference), then the value required is 1.960. (Note that, unlike the first scenario, it is the absolute value of the difference that is evaluated in the Monte Carlo phase of the analysis.)

In the Monte Carlo phase of the analysis, the number of abnormal differences is recorded on each trial (i.e., for each case) and summed across all trials (cases) to then express the number of cases exhibiting j or more abnormal differences as percentages.

Estimating the Percentage of the Population Exhibiting j or More Abnormal Scores Relative to Individuals' Mean Scores on the Battery

For this analysis, it is necessary to calculate the standard deviation of the difference between individuals' mean scores on the battery and each of the subtests or Indexes contributing to the mean (Crawford, Allan, McGeorge, & Kelly, 1997; Silverstein, 1984). When, as in the present case, the scores are standard normal scores, the formula is

$$s_{M-X} = \sqrt{1 + \bar{\mathbf{R}} - 2\bar{\mathbf{m}}_X}, \quad (2)$$

where $\bar{\mathbf{R}}$ is the mean of the elements in the full correlation matrix (including the unities in the diagonal), and $\bar{\mathbf{m}}_X$ is the mean of the column (or, equivalently, the row) of the matrix that contains Test X in the leading diagonal (i.e., the correlation of Test X with itself). Formula 2 is applied k times to calculate the standard deviation of the difference between each of the k tests and the mean test score. As was the case for the standard deviation of the difference between a pair of tests, this standard deviation is then multiplied by the standard normal deviate corresponding to the criterion used to define abnormality. In the Monte Carlo phase of the analysis, the number of abnormal deviations from a case's mean score is recorded on each trial and summed across all trials (cases) to then express the numbers of cases exhibiting j or more abnormal deviations as percentages.

Estimating the Percentage of the Population Exhibiting j or More Abnormal Scores and Score Differences on the WAIS–III and WISC–IV

The methods described above were applied to the four Index scores of the WAIS–III and WISC–IV. For each battery, the averaged correlation matrix (i.e., averaged across all age bands) for the Indexes was extracted from the relevant manual and used as the input for the simulations. For both batteries, we tabulated the percentage of the normal population expected to exhibit j or more abnormally low scores, j or more abnormally large pairwise differences, and j or more abnormally large deviations from individuals' mean Index scores.

Results

Illustration of the Role of the Number of Tests in a Battery and Their Intercorrelations in Determining the Percentage of the Population Exhibiting j or More Abnormally Low Scores

Before tabulating results for the WAIS–III and WISC–IV, we first use the Monte Carlo method for a more didactic purpose: namely to illustrate the role played by the number of tests in a battery and the intercorrelations between them in determining the percentage of the population exhibiting j or more abnormally low scores. The number of tests in the battery was set at either 4, 6, or 10, and the average correlation between the tests was set at either 0, 0.3, 0.5, or 0.7 (for simplicity this was achieved by setting all intercorrelations to the same value). As in all simulation results that follow, one million Monte Carlo trials were run (i.e., one million cases were drawn) for each combination of these two factors. The results of this procedure are presented in Table 1.

It can be seen from Table 1 that, as expected, the percentage of the population expected to exhibit j or more abnormal scores increases markedly with the number of tests in the battery. For example, when the average correlation between tests is 0.3 and there are four tests in the battery, it can be seen that 16.38% of the population are expected to exhibit one or more abnormally low test, but this rises to 30.74% when the battery consists of 10 tests. Table 1 illustrates that the correlations between tests in the battery also strongly influence the percentage expected to exhibit abnormally low tests. However, the direction of this effect differs as a function of j. With regard to the percentage exhibiting one or more abnormal tests (i.e., when $j = 1$), it can be seen that the percentages fall as the average correlation rises. For example, if the battery consists of six tests, the percentage expected to exhibit an abnormally low test is 26.49% when the tests are uncorrelated (i.e., average correlation = 0) but falls to 14.85% when the average correlation is 0.7. In contrast, when $j > 1$, the effect is reversed. Using the same example of a battery of six tests, 3.29% are expected to exhibit two or more abnormal tests when the average correlation is zero, but this rises to 7.26% for an average correlation of 0.7.

When the average correlation between tests is set at 0 (i.e., the tests are independent), the results from the Monte Carlo simulations are the same, except for trivial differences stemming from Monte Carlo variation, to those obtained by calculating binomial probabilities (e.g., for a battery of six tests, the Monte Carlo estimate of 26.49% for $j = 1$ is identical to two decimal places to the binomial estimate). Therefore, by comparing the results for different values of $\bar{\mathbf{R}}$ in Table 1, it can be seen that using binomial probabilities to estimate the percentage of the population expected

Table 1
Percentage of Population Expected to Exhibit j or More Abnormally Low Scores (<5th Percentile) as a Function of the Number of Tests (k) in the Battery and the Average Correlation ($\bar{r}$) Between Tests

		Percentage exhibiting *j* or more abnormally low scores									
$\bar{r}$	*k*	1	2	3	4	5	6	7	8	9	10
0	4	18.53	1.40	0.05	0.00						
0	6	26.49	3.29	0.22	0.01	0.00	0.00				
0	10	40.10	8.64	1.15	0.10	0.01	0.00	0.00	0.00	0.00	0.00
.3	4	16.38	3.09	0.52	0.06						
.3	6	22.07	5.89	1.64	0.41	0.09	0.01				
.3	10	30.74	11.54	4.67	1.94	0.79	0.31	0.11	0.04	0.01	0.00
.5	4	14.42	4.14	1.22	0.26						
.5	6	18.64	6.89	2.88	1.19	0.42	0.10				
.5	10	24.67	11.55	6.21	3.52	2.02	1.13	0.61	0.29	0.12	0.03
.7	4	12.07	4.99	2.22	0.80						
.7	6	14.85	7.26	4.09	2.31	1.21	0.49				
.7	10	18.55	10.54	6.94	4.83	3.40	2.38	1.62	1.06	0.60	0.26

to exhibit *j* abnormal test scores will give poor approximations to the correct values if the tests are moderately to highly correlated.

The reason for this can be illustrated with a simple thought experiment. Take a battery consisting of a moderate number (e.g., 6) of highly correlated tests and suppose a case's score on one of these tests (Test *X*) is well within the normal range. Then, for each of the other tests, the chances that they are abnormal will be well below the 5% probability assigned under the assumption of independence. As a result, the binomial approach overestimates the percentage of such cases that will exhibit one or more abnormal scores. Note that, in the converse situation, that is, when a case's score on Test *X* is in the abnormal range, the probability that the other tests will also be abnormal will be well above the 5% probability assigned when independence is assumed. This will not affect the percentage of cases exhibiting one or more abnormal test scores: If a case has already contributed to this percentage by exhibiting an abnormal score on Test *X*, it is irrelevant how many of the other tests are also abnormal. However, this latter feature does explain why (when, as in the present example, there is a small or moderate number of tests) the binomial estimates flip from providing an overestimate of the required percentage to providing an underestimate of the percentage of the population that will exhibit two or more abnormal test scores (and three or more, etc.).

Note that if there are a large number of moderately to highly correlated tests, the binomial estimates of the percentage exhibiting *j* or more abnormal tests will tend to be overestimates even for $j = 2$. For example, if there are 20 tests in a battery in which the average correlation was 0.7, the binomial estimate of the percentage exhibiting two or more abnormal tests is 26.41%, whereas the Monte Carlo estimate is 15.66%. However, the binomial estimates will still flip to providing underestimates for larger values of *j*. To continue with the example, the binomial estimate of the percentage exhibiting three or more abnormal tests is 7.54%, whereas the Monte Carlo estimate is 11.65%.

Estimated Percentages of the Population Exhibiting at Least j Abnormally Low Index Scores on the WAIS–III and WISC–IV

The results of estimating the percentage of the population exhibiting at least *j* abnormally low Index scores on the WAIS–III are presented in Table 2. The equivalent results for the WISC–IV are presented in Table 3. In both cases a range of (increasingly stringent) definitions of abnormality were applied, ranging from a score that was more than one standard deviation below the mean (i.e., below the 15.9th percentile) to a score that fell below the 1st percentile. Our own preference is to define abnormality as a score falling below the 5th percentile, and for this reason we have presented these results in bold. These different definitions of abnormality should cover most requirements.

It can be seen from Table 2 that although, by definition, 5% of the population would be expected to exhibit an abnormally low

Table 2
Percentage of the Normal Population Expected to Exhibit at Least j Abnormally Low Index Scores on the Wechsler Adult Intelligence Scale—Third Edition (WAIS–III)

Criterion for abnormality	Percentage exhibiting *j* or more abnormally low WAIS–III Index scores			
	1	2	3	4
<15.9th	34.43	17.40	8.57	3.20
<10th	23.74	10.34	4.53	1.49
<5th	**13.21**	**4.64**	**1.74**	**0.48**
<2nd	5.88	1.58	0.49	0.11
<1st	3.12	0.69	0.19	0.04

Note. Increasingly stringent definitions of abnormality are used ranging from <15.9th percentile (i.e., more than 1 *SD* below the mean) to below the 1st percentile. We define abnormality as a score falling below the 5th percentile, and for this reason we have presented these results in bold.

Table 3
Percentage of the Normal Population Expected to Exhibit at Least j Abnormally Low Index Scores on the Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV)

Criterion for abnormality	Percentage exhibiting *j* or more abnormally low WISC–IV Index scores			
	1	2	3	4
<15.9th	37.07	17.01	7.31	2.20
<10th	25.67	9.81	3.66	0.93
<5th	**14.29**	**4.21**	**1.28**	**0.26**
<2nd	6.33	1.35	0.33	0.05
<1st	3.34	0.56	0.11	0.01

Note. Increasingly stringent definitions of abnormality are used ranging from <15.9th percentile (i.e., more than 1 *SD* below the mean) to below the 1st percentile. We define abnormality as a score falling below the 5th percentile, and for this reason we have presented these results in bold.

score on any single WAIS–III Index, a sizeable percentage of the population (13.21%) would be expected to exhibit one or more abnormally low Index scores out of the possible four Index scores. However, it can also be seen that the percentages fall off fairly steeply when moving to two or more abnormal scores.

Turning to the results for the WISC–IV, again it can be seen that it will not be very unusual for a member of the normal population to exhibit at least one abnormal Index score (defined for present purposes as below the 5th percentile) but that the percentages fall off fairly steeply when moving to two or more abnormal scores and beyond. The results for the WISC–IV are similar to those for the WAIS–III. This is to be expected because both the pattern of correlations between Indexes and the absolute magnitude of the correlations are broadly similar for both batteries. Having said that, the intercorrelations between the WISC–IV Indexes are somewhat lower than those for the WAIS–III. As a result, the percentages exhibiting *j* or more abnormally low scores on the WISC–IV are somewhat higher than those for the WAIS–III when $j = 1$ (14.29% vs. 13.21%), and somewhat lower for $j > 2$ (i.e., the figures for $j = 2$ are 4.21% vs. 4.64%).

Estimated Percentages of the Population Exhibiting j or More Abnormally Large Pairwise Differences Between Index Scores on the WAIS–III and WISC–IV

The results of estimating the percentage of the population exhibiting *j* or more abnormally large pairwise differences between WAIS–III Index scores are presented in Table 4; the equivalent results for the WISC–IV appear in Table 5. It can be seen that for both batteries, a reasonable percentage of the population is expected to exhibit one or more abnormal pairwise differences (20.20% for the WAIS–III, 20.16% for the WISC–IV).

Although Ingraham and Aiken (1996) only considered use of binomial probabilities when estimating the number of abnormally low scores on a battery, it is worth considering how well binomial probabilities can model the percentages of the population expected to exhibit *j* or more abnormal score *differences*. Defining an abnormal pairwise difference as one exhibited by less than 5% of the healthy population, the required binomial probabilities, multiplied by 100 to express them as percentages, for $j = 1$ to 3 are 26.49%, 3.28%, and 0.22% (the percentages for $j > 3$ are very small and so are not reported). Comparing these estimates with those obtained by Monte Carlo simulation reveals that, for both the WAIS–III and WISC–IV, the binomial approach has fairly poor accuracy. Just as was the case when estimating the percentages exhibiting abnormally low scores, the binomial approach overestimates the percentages exhibiting one or more abnormal pairwise differences when $j = 1$ and underestimates the percentages when $j > 1$. For example, for the WAIS–III, the binomial estimate for $j = 1$ is 26.49% compared with the Monte Carlo estimate of 20.20%; for $j = 2$ the binomial estimate is 3.28% compared with 7.69%.

Table 4
Percentage of the Normal Population Expected to Exhibit j or More Abnormal Pairwise Differences, Regardless of Sign, Between Index Scores on the Wechsler Adult Intelligence Scale—Third Edition (WAIS–III)

Criterion for abnormality	Percentage exhibiting *j* or more abnormal pairwise differences (regardless of sign) between WAIS–III Indexes					
	1	2	3	4	5	6
<25%	65.65	47.68	28.10	7.52	1.02	0.01
<15%	47.28	28.03	12.42	2.14	0.14	0.00
<10%	35.15	17.68	6.34	0.80	0.03	0.00
<5%	**20.20**	**7.69**	**1.97**	**0.16**	**0.00**	**0.00**
<2%	9.15	2.41	0.43	0.02	0.00	0.00
<1%	4.90	0.98	0.14	0.00	0.00	0.00

Note. Increasingly stringent definitions of abnormality are used ranging from a difference exhibited by less than 25% of the population to a difference exhibited by less than 1%. Our preference is to define an abnormal difference as one exhibited by less than 5% of the population, and for this reason we have presented these results in bold.

Table 5
Percentage of the Normal Population Expected to Exhibit j or More Abnormal Pairwise Differences, Regardless of Sign, Between Index Scores on the Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV)

Criterion for abnormality	Percentage exhibiting *j* or more abnormal pairwise differences (regardless of sign) between WISC–IV Indexes					
	1	2	3	4	5	6
<25%	65.56	47.61	28.03	7.61	1.06	0.01
<15%	47.21	28.01	12.45	2.19	0.15	0.00
<10%	35.01	17.71	6.35	0.83	0.03	0.00
<5%	**20.16**	**7.71**	**2.00**	**0.17**	**0.00**	**0.00**
<2%	9.12	2.43	0.44	0.02	0.00	0.00
<1%	4.88	0.99	0.14	0.00	0.00	0.00

Note. Increasingly stringent definitions of abnormality are used ranging from a difference exhibited by less than 25% of the population to a difference exhibited by less than 1%. Our preference is to define an abnormal difference as one exhibited by less than 5% of the population, and for this reason we have presented these results in bold.

Estimated Percentages of the Population Exhibiting j or More Abnormally Large Deviation Scores Relative to Their Mean Index Score on the WAIS–III and WISC–IV

The estimated percentages of the population exhibiting *j* or more abnormally large deviation scores relative to their mean Index score on the WAIS–III are presented in Table 6; the equivalent results for the WISC–IV appear in Table 7. As was the case for pairwise differences, when an abnormal deviation score for each individual Index is defined as one exhibited by less than 5% of the population, a sizeable percentage of the healthy population (16.74% for the WAIS–III, 16.71% for the WISC–IV) is expected to exhibit one or more abnormal deviation. Again, however, the percentages fall off rapidly as *j* increases. For example, for the WAIS–III, only 3.21% of the population is expected to exhibit two or more abnormal deviations (the corresponding figure for the WISC–IV is 3.22%).

Discussion

An Example of the Application of the Present Approach Using the Results for WAIS–III Index Scores

To illustrate the potential applications of the present approach, we take the hypothetical case of a patient who had suffered a severe traumatic brain injury and had been administered the WAIS–III as part of a more comprehensive neuropsychological assessment. The patient's scores on the WAIS–III Indexes were as follows: Verbal Comprehension (VC) = 118; Perceptual Organization (PO) = 106; Working Memory (WM) = 78; and Processing Speed (PS) = 68. Suppose we define an abnormally low score as one that falls below the 5th percentile (Index scores of 75 or lower are below the 5th percentile). Then one of the patient's scores (PS) would be classified as abnormally low. Referring to Table 2, it can be seen that 13.21% of the population is expected to exhibit one or more abnormally low scores. Thus, although the patient's PS score is unusually low, it is by no means highly unusual for a member of the general population to exhibit an abnormally low score on one of the four Indexes.

Table 6
Percentage of the Normal Population Expected to Exhibit j or More Abnormal Wechsler Adult Intelligence Scale—Third Edition (WAIS–III) Index Scores Relative to Individuals' Mean Index Scores (Regardless of Sign)

Criterion for abnormality	Percentage exhibiting *j* or more abnormal deviation scores (regardless of sign) on the WAIS–III			
	1	2	3	4
<25%	61.19	32.86	4.91	0.95
<15%	42.27	16.21	1.28	0.20
<10%	30.51	8.96	0.43	0.06
<5%	**16.74**	**3.21**	**0.07**	**0.01**
<2%	7.21	0.81	0.01	0.00
<1%	3.73	0.28	0.00	0.00

Note. Increasingly stringent definitions of abnormality are used ranging from a difference exhibited by less than 25% of the population to a difference exhibited by less than 1%. Our preference is to define an abnormal deviation as one exhibited by less than 5% of the population, and for this reason we have presented these results in bold.

Table 7
Percentage of the Normal Population Expected to Exhibit j or More Abnormal Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV) Index Scores Relative to Individuals' Mean Index Scores (Regardless of Sign)

Criterion for abnormality	Percentage exhibiting *j* or more abnormal deviation scores (regardless of sign) on the WISC–IV			
	1	2	3	4
<25%	61.16	32.83	5.01	0.94
<15%	42.22	16.20	1.33	0.20
<10%	30.49	8.97	0.45	0.06
<5%	**16.71**	**3.22**	**0.07**	**0.01**
<2%	7.21	0.82	0.01	0.00
<1%	3.72	0.29	0.00	0.00

Note. Increasingly stringent definitions of abnormality are used ranging from a difference exhibited by less than 25% of the population to a difference exhibited by less than 1%. Our preference is to define an abnormal deviation as one exhibited by less than 5% of the population, and for this reason we have presented these results in bold.

Turning to pairwise comparisons of the patients' Index scores, suppose we define an abnormal pairwise difference as one that is exhibited by less than 5% of the population. By consulting Table B.2 of the WAIS–III manual, it can be concluded that four of the patients' six pairwise comparisons are abnormal (VC vs. WM, VC vs. PS, PO vs. WM, and PO vs. PS). For example, a discrepancy of 26 or more points between VC and WM is required to meet our chosen criterion for abnormality. That is, from Table B.2 of the manual it can be seen that 5.4% of the standardization sample exhibited a difference of 25 or more points between VC and WM, whereas only 4.3% exhibited a difference of 26 or more points. The discrepancy between the patient's scores on VC and WM is 40 points and therefore easily meets the criterion for abnormality. Referring to Table 4, we can see that only 0.16% of the population is expected to exhibit four or more abnormal pairwise differences.

Turning to examination of the deviations from the patient's mean Index score, suppose that, as previously, we use a deviation that is expected to be exhibited by less than 5% of the population as our criterion. The mean Index score is 92.50; therefore, the deviations from this mean are 25.50 for VC, 13.50 for PO, –14.50 for WM, and –24.50 for PS. Referring to Longman's (2004) estimated base rate data for deviations, two of the Index scores qualify as abnormally large according to our chosen criterion (VC is abnormally high, PS abnormally low). For example, for the PS Index, the (absolute) observed difference must exceed 17.73 to meet the chosen criterion for abnormality. Referring to Table 6, it can be seen that only 3.21% of the population is expected to exhibit two or more abnormal deviations from their mean Index scores.

In this example, the indications of abnormally large pairwise discrepancies and deviation scores survived the further scrutiny applied; that is, it is estimated that few healthy individuals would

exhibit this number of abnormal score differences. Note also that the client's pattern of strengths and weaknesses is consistent with a head injury. This combination of profile and degree of abnormality gives a high degree of confidence in a conclusion that the client has suffered significant acquired impairment.

It will be appreciated, however, that this need not be the case. For example, suppose an individual exhibited only one abnormally large pairwise difference or one abnormally large deviation. Such results will not be uncommon in the healthy population; 20.2% are expected to exhibit one or more abnormally large pairwise differences (see Table 4), and 16.74% are expected to exhibit one or more abnormally large deviation scores (see Table 6). Even when the weakness in such a case is in line with clinical expectations (e.g., PS or WM in a client with head injury), much more in the way of converging evidence from other sources is required before one could be confident in inferring the presence of acquired impairment. In cases when there is little in the way of theory or prior empirical evidence to specify a likely pattern of impairment, a knowledge of the base rates for the number of abnormal scores or score differences assumes particular importance.

Before concluding this example, it should be noted that making use of the data provided in the present article adds little to the time taken to analyze a patient's profile. That is, the time consuming aspects are those that precede use of the tables presented here, and these will already form a part of many neuropsychologists' practice (i.e., most neuropsychologists are aware that they should be concerned with the degree of abnormality of any scores or score differences present in a patient's profile). Moreover, we are not suggesting that neuropsychologists use both the pairwise and deviation approach to examining discrepancies. Our own preference is for the approach of Longman (2004), see also Flanagan and Kaufman (2004), but we recognize that many may prefer, or at least be more familiar with, the pairwise method adopted in the WAIS–III and WISC–IV manuals.

To further ease use of these data, we have written two computer programs, one for the WAIS–III (WAISIII_Percent_Abnorm.exe) and one for the WISC–IV (WISCIV_Percent_Abnorm.exe). The programs require that users enter an individual's scores on the four Indices and select a criterion for abnormality. For example, they might choose to define an abnormally low score as a score falling below the 5th percentile. The programs automate the process of determining whether each of an individual's Index scores are abnormally low and whether they exhibit abnormally large pairwise differences or abnormally large deviations from the mean Index score. The programs report the percentage of the population expected to exhibit lower scores for each Index and either the percentage of the population expected to obtain larger pairwise differences or larger deviations (the user selects between the latter of these two options).

It then reports the number of scores for the individual that meet the chosen criterion for abnormality and reports the percentage of the population expected to exhibit this number, or more, of abnormally low scores. The reporting of abnormally large pairwise differences or deviations follows a similar format. That is, if the user has chosen to analyze pairwise differences, the program records the number of the individual's pairwise differences that meet the criterion for abnormality and reports the percentage of the population expected to exhibit this number, or more, of abnormal differences. In summary, these programs perform all the necessary clerical work required to examine the abnormality of Index scores as well as Index score differences and deviations. See a later section for details of where to download these programs.

It should be noted that, to estimate the abnormality of each individual pairwise difference, the programs use Formula 1; that is, abnormality is estimated statistically rather than by using empirical base rates as found in the relevant test manuals (e.g., Table B.2 of the WAIS–III manual). The former approach is in keeping with the use of statistical, rather than empirical, methods to estimate the abnormality of deviations from individuals' mean Index scores in the present article (Formula 2) and in the tables provided by Longman (2004) and Flanagan and Kaufman (2004).

The result is that, on occasion, the number of pairwise differences or deviations estimated to be abnormal will differ depending on whether the program is used or the empirical base rate data in the relevant manual. For example, take the earlier example in which a traumatic brain injury case exhibited four abnormal pairwise differences between the case's Index scores. Suppose, however, that the case scored 105 rather than 106 on the PO Index. Using the statistical method (as implemented in the WAIS and WISC programs), the case is still classified as exhibiting four abnormal pairwise differences—the difference (of 27 points) between the PO Index and WM remains abnormal (it is estimated that 3.64% of the healthy population will exhibit this size of a difference or larger; i.e., less extreme than in the original example [3.00%] but nevertheless still meets the selected criterion for abnormality). However, using the empirical base rates, this difference is not classified as abnormal (from Table B.2, 5.2% of the normative sample exhibited a difference of this magnitude or larger compared with 4.6% for the difference of 28 points in the original example).

A number of factors contribute to differences between empirical and statistical base rates; for example, statistical base rates are unaffected by the inevitable small "bumps and wiggles" that will occur in empirical distributions even when normative samples are large. However, the most important factor is that a number of people in the normative sample will obtain the same difference score as that obtained by the individual of interest. When empirical base rates are employed then, typically, the percentage recorded and read off by the user is the percentage equaling or exceeding this difference. In contrast, because the statistical approach treats the data as continuous, it will, in essence, credit half those obtaining the same difference score as having a larger difference and the other half as obtaining a smaller difference; thus, the percentages will normally be a little lower than those obtained from empirical rates. This is akin to the procedure used for forming standard percentiles; moreover, the same holds when estimating the abnormality of deviation scores using existing tables (as these too are derived statistically).

The Distinction Between the Abnormality of Differences and the Reliability of Differences

When examining differences in an individual's score profile, a crucial distinction is that between the abnormality of the differences and the reliability of the differences (Crawford, 2004). If a difference is reliable, then this means it is unlikely to have arisen

as a result of measurement error in the tests. That is, the observed difference is liable to reflect a genuine difference. However, reliable differences can be very common in the cognitively intact population, particularly if the tests have high reliability. Thus reliable differences, on their own, cannot serve as a basis for inferring impairment on the test that is performed more poorly. The present article is solely focused on the abnormality of differences.

Use of These Methods in Group Research

The focus of the foregoing discussion has been on the use of the generic Monte Carlo method as an aid to interpretation of an individual patient's profile. However, the method also has a number of applications in group-based research, either as a means of evaluating the results of an existing study or to inform decision making at the design stage. For example, in studies of the prevalence of neuropsychological deficits in various patient populations (e.g., multiple sclerosis, HIV, diabetes, mild cognitive impairment, patients who have undergone coronary artery bypass grafts, etc.), it is common practice to administer multiple neuropsychological tests to assess performance across a range of cognitive domains.

The prevalence of deficits in these populations is then estimated by calculating the percentage of cases meeting a predefined criterion for impairment. The criteria used in such studies are usually of the form that a patient should exhibit *j* or more abnormally low scores on the battery, where both *j* and the definition of abnormality varies from study to study. For example, the criterion may be that a patient's scores on at least three tests must be one standard deviation below the mean.

A number of authors have stressed the need to be concerned with the base rate of false positives in such studies (Grunseit, Perdices, Dunbar, & Cooper, 1994; Heaton, Miller, Taylor, & Grant, 2004; Iverson, White, & Brooks, 2006; Lewis, Maruff, & Silbert, 2004), and there are a number of vivid empirical demonstrations of the dangers when such base rates are ignored (de Rotrou et al., 2005; Lewis, Maruff, Silbert, Evered, & Scott, 2006a, 2006b; Palmer, Boone, Lesser, & Wohl, 1998).

As noted by most of these studies, one clear solution to this problem is to include a matched healthy control sample to obtain empirical base rates for the criterion used. However, practical constraints mean that many prevalence studies do not include control samples. In such circumstances, it is therefore important to estimate the base rate at which members of the healthy (i.e., cognitively intact) population would be identified as cases as a safeguard against overestimating the prevalence of cognitive deficits.

Grunseit et al. (1994), in an article that raised a number of important methodological issues concerning research in HIV patients, recommended that, when a control sample is unavailable, studies should test whether the percentage of patients classified as impaired exceeds chance. Although we agree with the need to take account of base rates, the approach they recommend to achieve this aim is inappropriate. Grunseit et al. suggest examining whether the percentage of cases identified as impaired exceeds the base rate for impairment as calculated using the binomial distribution. There are three problems with this.

First, it is unlikely that binomial probabilities will yield accurate estimates of the base rate for the reasons outlined earlier. Second, if binomial probabilities are used, then it is crucial that the probability calculated and reported is the sum of probabilities for *j* or more successes, not the probability of exactly *j* successes. So, for example, if there are six items in a battery, and the criterion for impairment is that a patient exhibits four or more abnormal scores, then the binomial probabilities for four, five, and six successes need to be summed to estimate the base rate. The binomial probability for exactly *j* successes may grossly underestimate the base rate. Grunseit et al. (1994) used the example of a battery of 16 tests and a criterion of scores one standard deviation below the mean on two tests. They reported the probability of meeting this criterion by chance as 0.268 (this is the binomial probability for exactly 2 successes out of 16), whereas the correct figure is 0.747 (the probability of 2 or more successes).

The third problem is that, even if it were appropriate to use binomial probabilities to estimate the base rate (i.e., the tests were independent), and even if the correct probabilities were calculated, this procedure would still not test whether the percentage of cases in a sample meeting the criterion significantly exceeds chance. Grunseit et al. (1994) have suggested that it does; that is, they have referred to the procedure as testing whether the number of patients identified is "significantly different from chance" (p. 906), but it can be seen that the clinical sample does not feature in any of the above calculations, and therefore the method is not an inferential method.

We suggest that problems of this kind should be addressed using a two-stage process. The first stage should be to determine the base rate, that is, to estimate the percentage of individuals that will be classified as impaired by chance. If, as will rarely be the case, the tests in the battery are independent (i.e., uncorrelated), then the binomial distribution can be used to estimate this percentage. In most circumstances, however, the Monte Carlo method outlined here should be used. Note that if estimates of the population correlations between the components of the battery are not available (e.g., if the battery consists of measures derived from diverse sources), then the Monte Carlo method can be used in a more exploratory fashion to examine the base rates under different assumptions for the population correlations.

Regardless of the method used to estimate the base rate, a second analysis should then be performed to determine whether the observed number of cases in the clinical sample that were classified as impaired significantly exceeds the base rate. This second stage should be performed using the binomial distribution with the estimated base rate used to specify *p*, the probability of success on each "trial." Unlike the first stage of the analysis, the use of binomial probabilities is entirely appropriate for this second stage as the observations (trials, in binomial parlance) are independent; that is, they are individual cases rather than tests. (Note that here we assume that the correlations between tests used to estimate the base rate were obtained from a large standardization sample such that we can treat the base rate as fixed. When the correlations are obtained from a more modest sample it would be appropriate to treat the standardization sample as a sample and test for a difference between two independent proportions rather than, as here, compare a sample proportion against a fixed proportion.)

To illustrate the suggested procedure, suppose a study of the prevalence of impairment in newly diagnosed diabetics has administered a battery of 16 tests to a sample of 120 patients and that, for convenience, the correlation between tests in the battery's standardization sample are all 0.5. Suppose also that the criterion used to identify a case as cognitively impaired was performance one standard deviation below the mean on three or more tests and that 48 patients (40% of the sample) met this criterion. Application of the Monte Carlo method reveals that 34.48% of the general (i.e., cognitively intact) population are also expected to meet this criterion. This constitutes Stage 1 (i.e., calculation of the base rate, which we assume is derived from a very large normative sample). Calculation of the binomial probability (Stage 2) for 48 successes (i.e., cases identified) in 120 trials (i.e., the sample *n*) with a probability of success of 0.3448 on each trial (i.e., for each individual) yields a binomial probability of 0.120. Thus, the number of cases identified as impaired does not significantly exceed the base rate. In this hypothetical example, we would conclude that we cannot reject the null hypothesis that the rate is the same as that expected in the general population; that is, there is no evidence for cognitive deficits in the population of newly diagnosed diabetic patients.

In this example, it is clear from the base rate that the criterion for identifying caseness was very lax. To help avoid such problems, the Monte Carlo method developed here can also be used prospectively as a means of selecting criteria for prevalence studies. That is, various candidate criteria can be evaluated to estimate the rate at which they will yield false positives to find a criterion acceptable to the investigator. As referred to earlier, when estimates of the population correlations are unavailable, the method can be used in a more exploratory fashion to model the base rates under different assumptions for the magnitude of these correlations. For example, as Shallice; (1988) and others (Crawford & Garthwaite, 2005) have noted, population correlations among neuropsychological tests typically average around 0.5; this rough and ready rule of thumb could serve as one model for the correlations.

Note that the population correlations referred to here are the correlations in the general population; it would not be appropriate to use correlations obtained from clinical samples to estimate base rates as these will often differ substantially from the latter. The direction of the difference is hard to predict: Correlations are strongly influenced by the degree of dispersion of scores and so correlations will commonly be higher in impaired samples. On the other hand, if dissociations between particular tasks are common in a given clinical population, then correlations will tend to be lower. Before leaving this topic, note that there will be occasions in which the interest is in examining whether a subgroup of patients exceeds a particular base rate in a clinical population; in such circumstances the prescription against using correlations from clinical samples obviously does not hold.

Computer Program for Estimating the Percentage Exhibiting j or More Abnormal Scores and Score Differences on Standardized Test Batteries

A generic computer program for PCs (PercentAbnormK.exe) has been written to accompany this article. It can be downloaded from the following web page: http://www.abdn.ac.uk/~psy086/dept/PercentAbnormKtests.htm.

The program implements the generic methods developed in the present article and can be used to calculate the percentage exhibiting *j* or more abnormally low tests, *j* or more abnormally large pairwise differences between tests, and *j* or more abnormally large deviations from individuals' mean scores. It requires the user to specify the number of tests in the battery (up to a maximum of 20 tests), specify a criterion for abnormality (e.g., below the 5th percentile in the case of abnormally low scores), and enter the correlation between tests in the form of a lower triangular matrix. The outputs consist of the estimated percentage of the population expected to exhibit *j* or more abnormally low scores (e.g., if there are three tests in the battery, it records the percentage of the population expected to exhibit one or more, two or more, and three abnormally low scores). The program also records the percentage of the population expected to exhibit *j* or more abnormally large deviations from individuals' mean scores on the battery, and the percentage of the population expected to exhibit *j* or more abnormally large pairwise differences between tests in the battery.

The program allows the user to select from 1 of 10 criteria for abnormality. In the case of abnormally low scores, these range from a score estimated to be exhibited by less than 25% of the population through to a score exhibited by less than 1% (the intermediate criteria include scores that are 1, 1.5, or 2 standard deviations below the mean; i.e., scores below the 15.8th, 6.6th, or 2.28th percentiles). For simplicity, the criterion selected by the user to define an abnormally low score is also used to define abnormally large pairwise differences and abnormally large deviations from individuals' mean scores. For example, if an abnormally low score is defined as a score falling below the 5th percentile, then an abnormally large pairwise difference (or deviation) is defined as a difference (or deviation) that is exceeded by less than 5% of the normal population, regardless of sign.

Although the program is simple to use and the output resembles that presented in the tables of the present article, users may wish to verify their understanding by running an example drawn from the present article through the program before using it to generate equivalent base rate data for other batteries.

For practical reasons, the program is limited to dealing with batteries consisting of 20 or less components. This should not be a serious imitation as few batteries routinely used in neuropsychology are larger than this. In the case of batteries with more than 20 components, it is likely that components will be combined into composite indexes of some form. Therefore, it should still be possible to generate base rate data for such batteries, albeit only for the composite scores.

As referred to earlier, we have also written programs designed specifically to assist in the analysis of an individual's performance on the WAIS–III and WISC–IV Indexes; these programs can be obtained from the same locations as the generic program.

Some Caveats on the Use of These Methods

The generic simulation method outlined in the present article treats the means, standard deviations, and correlations of the components of a battery as population parameters. The assumption is that the standardization sample was sufficiently large so that the

sample statistics (i.e., the matrix of correlations between tests) provide very good estimates of the population correlations. To avoid any confusion, it can be noted that, although the means and standard deviations are not required as inputs for the simulation, an assumption underlying the simulation is that they are also known and fixed.

Therefore, the method should be used with caution when the population correlations and means and standard deviations have been estimated using data from a modestly sized standardization sample (we suggest caution when the sample *n* is 300 or less). When the concern is with score differences, there will be an overall trend for the level of abnormality for each separate comparison to be overestimated (Crawford & Garthwaite, 2005). However, the effects of error in estimating these parameters for a particular battery is unpredictable: If the population correlations have been underestimated (as may happen if the sample does not adequately cover the full range of ability in the population), then the percentage of the population exhibiting one or more abnormal test scores may be underestimated, and the percentage exhibiting two or more abnormal scores and so forth will be overestimated. The opposite pattern will occur if the population correlations have been overestimated. The interested reader is referred to Crawford and Garthwaite (2005) for a detailed discussion of some of the problems involved in drawing inferences concerning the abnormality of differences between an individual's scores when the tests are standardized on a modestly sized control or normative sample.

Finally, note that the further assumption of multivariate normality (which should not be problematic in the case of well standardized tests) may be violated in the case of ad hoc batteries. That is, in addition to any problems stemming from the size and representativeness of the normative sample, sufficient care may not have been taken to normalize the test scores. Moreover, the assumption of multivariate normality includes the assumption that the scores are continuous. Therefore, if the tests in a battery have a limited number of possible raw scores (e.g., 0 to 10) or scaled scores, the accuracy of the estimates will suffer. For example, in contrast to Wechsler Index scores, Wechsler subtest scaled scores have a limited range of scores (the distances between scaled score points represent one third of a standard deviation).

Conclusion

Quantifying the abnormality of an individual's test score or score difference is a standard part of the assessment process in neuropsychology. However, when, as is almost always the case, multiple tests are administered, attention should also be paid to how common it will be to exhibit a given number of abnormal tests or test differences from among the overall set of tests administered. The ability to estimate such percentages helps guard against overinferring the presence of impairment. As was shown, when commonly used criteria for abnormality are used, substantial percentages of the healthy population are expected to exhibit one or more abnormal scores or score differences, even when the overall number of tests is relatively modest. On the other hand, the percentages of the healthy population expected to show *j* or more abnormal scores or score differences drop off rapidly as *j* increases. Being able to demonstrate that the number of abnormal scores or score differences exhibited by a patient occur rarely in the healthy population gives the clinician or researcher confidence in inferring the presence of acquired impairments in such circumstances. Moreover, although the discussion of these methods has mainly focused on the use of these methods as an aid to interpretation of the results for individual patents, the methods can usefully be employed in group based research as outlined in an earlier section.

The Monte Carlo method outlined in the present study has significant advantages over the alternative approach based on binomial probabilities. The downside is that it requires more in the way of computation. However, the provision of the generic computer program that accompanies this article allows neuropsychologists to painlessly apply the methods to other test batteries not covered here.

References

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: Erlbaum.

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, England: Wiley.

Crawford, J. R., & Allan, K. M. (1996). WAIS–R subtest scatter: Base rate data from a healthy UK sample. *British Journal of Clinical Psychology, 35,* 235–247.

Crawford, J. R., Allan, K. M., McGeorge, P., & Kelly, S. M. (1997). Base rate data on the abnormality of subtest scatter for WAIS–R short-forms. *British Journal of Clinical Psychology, 36,* 433–444.

Crawford, J. R., & Garthwaite, P. H. (2005). 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., 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.

de Rotrou, J., Wenisch, E., Chausson, C., Dray, F., Faucounau, V., & Rigaud, A. S. (2005). Accidental MCI in healthy subjects: A prospective longitudinal study. *European Journal of Neurology, 12,* 879–885.

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC–IV assessment.* Hoboken, NJ: Wiley.

Gentz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics, 1,* 141–149.

Grunseit, A. C., Perdices, M., Dunbar, N., & Cooper, D. A. (1994). Neuropsychological function in asymptomatic HIV-1 infection: Methodological issues. *Journal of Clinical and Experimental Neuropsychology, 16,* 898–910.

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead–Reitan battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults (HRB*). Lutz, FL: Psychological Assessment Resources.

Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology, 10,* 120–124.

Iverson, G. L., White, T., & Brooks, B. L. (2006, June). *Base rates of low scores on the Neuropsychological Assessment Battery (NAB).* Paper presented at the annual meeting of the Canadian Psychological Society, Calgary, Alberta, Canada.

Lewis, M., Maruff, P., & Silbert, B. (2004). Statistical and conceptual issues in defining post-operative cognitive dysfunction. *Neuroscience and Biobehavioral Reviews, 28,* 433–440.

Lewis, M. S., Maruff, P., Silbert, B. S., Evered, L. A., & Scott, D. A. (2006a). Detection of postoperative cognitive decline after coronary artery bypass graft surgery is affected by the number of neuropsychological tests in the assessment battery. *Annals of Thoracic Surgery, 81,* 2097–2104.

Lewis, M. S., Maruff, P., Silbert, B. S., Evered, L. A., & Scott, D. A. (2006b). The sensitivity and specificity of three common statistical rules for the classification of post-operative cognitive dysfunction following coronary artery bypass graft surgery. *Acta Anaesthesiologica Scandinavica, 50,* 50–57.

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.

Longman, R. S. (2004). Values for comparison of WAIS–III Index scores with overall means. *Psychological Assessment, 16,* 323–325.

Palmer, B. W., Boone, K. B., Lesser, I. M., & Wohl, M. A. (1998). Base rates of "impaired" neuropsychological test performance among healthy older adults. *Archives of Clinical Neuropsychology, 13,* 503–511.

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

Silverstein, A. B. (1984). Pattern analysis: The question of abnormality. *Journal of Consulting and Clinical Psychology, 52,* 936–939.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms and commentary* (3rd ed.). New York: Oxford University Press.

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale—Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: The Psychological Corporation.

Appendix

Worked Example of Random Sampling of Multivariate Normal Observations

Suppose that a battery consists of three tests and that the correlation matrix between these tests in the standardization sample is

$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.6 & 0.5 \\ 0.6 & 1.0 & 0.4 \\ 0.5 & 0.4 & 1.0 \end{bmatrix}.$$

The Choleski decomposition of this matrix is

$$\mathbf{C} = \begin{bmatrix} 1.00000 & 0.00000 & 0.00000 \\ 0.60000 & 0.80000 & 0.00000 \\ 0.50000 & 0.12500 & 0.85696 \end{bmatrix}.$$

To obtain a random vector of observations (**y**) from the multivariate normal distribution with covariance matrix **R** (i.e., to simulate the scores on the test battery for a member of the general population), we postmultiply **C** by a vector (**z**) (i.e., a one-dimensional array of size 3) of randomly sampled, independent (i.e., uncorrelated) standard normal variates. Suppose that the values for this vector are 0.232, −0.654, and 1.242. Then, as illustrated below, a random vector (i.e., a vector representing scores for an individual on the three tests) from the multivariate normal distribution with covariance matrix **R** is

$$\mathbf{Cz} = \mathbf{y}, \text{ that is } \begin{bmatrix} 1.00000 & 0.00000 & 0.00000 \\ 0.60000 & 0.80000 & 0.00000 \\ 0.50000 & 0.12500 & 0.85696 \end{bmatrix} \times \begin{bmatrix} +0.232 \\ -0.654 \\ +1.242 \end{bmatrix} = \begin{bmatrix} +0.232 \\ -0.384 \\ +1.099 \end{bmatrix}.$$

To simulate further observations (i.e., to simulate the scores of additional individuals), we repeat the process by substituting a new random vector of standard normal deviates. Thereafter, these random vectors are used, as described in the text, to model the percentage of the population that will exhibit a given number of abnormal scores, pairwise differences, or deviations from individuals' mean scores.

Received August 10, 2006
Revision received January 23, 2007
Accepted January 29, 2007 ■