Running head: Equations from summary data

Neuropsychology, in press

(Neuropsychology journal home page)

Using regression equations built from summary data in the neuropsychological assessment of the individual case

John R. Crawford

University of Aberdeen

Paul H. Garthwaite

Department of Statistics

The Open University

Address for correspondence: Professor 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

Abstract

Regression equations have many useful roles in neuropsychological assessment. This paper is based on the premise that there is a large reservoir of published data that could be used to build regression equations; these equations could then be employed to test a wide variety of hypotheses concerning the functioning of individual cases. This resource is currently underused because (a) not all neuropsychologists are aware that equations can be built using only basic summary data for a sample, and (b) the computations involved are tedious and prone to error. To overcome these barriers we set out the steps required to build regression equations from sample summary statistics and the further steps required to compute the associated statistics for drawing inferences concerning an individual case. We also develop, describe and make available computer programs that implement the methods. Although caveats attach to the use of the methods, these need to balanced against pragmatic considerations and against the alternative of either entirely ignoring a pertinent dataset or using it informally to provide a clinical “guesstimate”.

Keywords: neuropsychological assessment; regression equations; single-case methods

INTRODUCTION

Regression equations serve a number of useful functions in the neuropsychological assessment of individuals (Chelune, 2003; Crawford, 2004; Crawford & Howell, 1998; Strauss, Sherman, & Spreen, 2006; Temkin, Heaton, Grant, & Dikmen, 1999). For example, they are widely used to estimate premorbid levels of ability in clinical populations using psychological tests that are relatively resistant to neurological or psychiatric dysfunction (Crawford, 2004; Franzen, Burgess, & Smith-Seemiller, 1997; O'Carroll, 1995).

Another common application of regression is in the assessment of change in neuropsychological functioning in the individual case. Here a regression equation can be built (usually using healthy participants) to predict a patient's level of performance on a cognitive ability measure at retest from their score at initial testing. An obtained retest score that is markedly lower than the predicted score suggests cognitive deterioration (Crawford, 2004; Heaton & Marcotte, 2000; Sherman et al., 2003; Temkin et al., 1999).

Clinical samples can also profitably be used to build regression equations for predicting retest scores. For example, Chelune, Naugle, Lüders, Sedlak, and Awad (1993) built an equation to predict memory scores at retest from baseline scores in a sample of temporal lobe epilepsy cases who had not undergone any surgical intervention in the intervening period. The equation was then used to assess the effects of surgery on memory functioning in further individual patients.

Regardless of whether the equation was built from a healthy or clinical sample, this approach simultaneously factors in the strength of correlation between scores at test and retest (the higher the correlation the smaller the expected discrepancies), the effects of practice (typically scores will be higher on retest) and

regression to the mean (extreme scores on initial testing will, on average, be less extreme at retest).

Yet another common role of regression equations is as an alternative to the use of conventional normative data (Heaton & Marcotte, 2000). For example, if (as is commonly the case) performance on a neuropsychological test is affected by age, then age can be incorporated into a regression equation to obtain an individual's predicted score on the test. This provides what Zachary and Gorsuch (1985) have termed "continuous norms" and can be contrasted with the discrete norms formed by creating arbitrary age bands. With the latter approach, the apparent relative standing of an individual can change dramatically as they move from one age band to another.

It can be seen from the foregoing that regression equations perform many useful roles in the neuropsychological assessment of individuals. However, we suggest that the potential of regression equations is far from being fully realized. There is a large reservoir of published data that could be used to build regression equations; these equations could then be employed to test a wide variety of hypotheses concerning the neuropsychological functioning of individual cases.

For example, there are literally hundreds of published studies that have examined performance at test and retest on a wide variety of commonly used neuropsychological tests. Many of these studies have been tabulated by McCaffrey, Duff and Westervelt (2000) and are classified by the test involved (e.g., verbal fluency, Wisconsin Card Sorting Test, and so forth), the type of sample (e.g. healthy elderly, TBI sample etc), the retest interval, and whether any intervention was applied in the intervening period.

Provided that studies such as those described above record the means and standard deviations for the sample's scores at test and retest, and the correlation

between these sets of scores, a regression equation can be built to predict retest scores from initial scores (these basic summary statistics are commonly reported). The equation built using the summary statistics can then be used with an individual to examine the discrepancy between the predicted score and the actual score obtained at retest. The statistics required to draw inferences concerning this discrepancy can also be calculated from summary data alone; the only information required beyond that already noted is the size of the sample (essentially it can be assumed that this will always be available).

The remainder of this paper has two principal aims. The first is to (a) set out the calculations required to build regression equations from summary data and (b) set out the further calculations required when applying these equations to draw inferences concerning an individual case. The second aim is to describe and make available computer programs that implements all the methods described.

Statistically minded neuropsychologists will be aware that regression equations can be built using summary data alone and that the associated statistics required to apply such equations to an individual case can also be obtained without the sample raw data. However, on the basis of discussions at conferences, workshops and elsewhere, it is clear that many neuropsychologists are unaware, or only vaguely aware, that such a possibility exists.

Moreover, those neuropsychologists who know that summary statistics are sufficient also know that, if the optimal statistical methods are to be employed, the calculations involved are time-consuming, tedious, and prone to error. Therefore there is the temptation to either (a) not use regression equations in situations in which it would be helpful, or (b) use computationally simple methods that only approximate the results that would be obtained from the optimal methods. Alternatively, if the

optimal methods are used, there is the danger that clerical errors will unknowingly be made when carrying out the computations. The provision of a computer program that implements the optimal methods deals with all of these problems.

Building a regression equation from summary data

A regression equation with one predictor variable (i.e., a bivariate regression equation) takes the following form

$$\hat{Y} = \alpha + \beta X \,, \tag{1}$$

where $\hat{Y}$ is the predicted score on the criterion variable, α is the intercept (the value of *Y* when *X* takes the value of zero), β is the slope of the regression line, and *X* is the score on the predictor variable. Let *a* and *b* denote the estimates of α and β that the complete sample data would give. Then the equation for calculating the slope *b* using summary data from a sample is

$$b = r_{xy} \frac{s_Y}{s_X} \,, \tag{2}$$

where r_{xy} is the correlation between the predictor and criterion variables and s_X and s_Y are the standard deviations of these variables. The equation for the intercept (*a*) is

$$a = \overline{Y} - b\overline{X} \,, \tag{3}$$

where $\overline{Y}$ is the mean score on the criterion or dependent variable and $\overline{X}$ is the mean score on the predictor or independent variable.

Having set out the steps to obtain a regression equation from summary data we now turn to the calculations required to draw inferences concerning the discrepancies between an individual's obtained score and the score predicted by such an equation. The first step is to calculate the standard error of estimate ($s_{Y \cdot X}$). The standard error of estimate is a measure of the variability of observations about the regression line in

the sample used to build the equation. The formula for the standard error of estimate when there is only a single predictor variable is

$$s_{Y \cdot X} = s_Y \sqrt{\left(1 - r_{XY}^2\right) \frac{N-1}{N-2}}, \tag{4}$$

where N is the size of the sample used to build the equation and all other terms have been defined previously. Again it can be seen that this statistic can be calculated when only summary data are available.

Having obtained $s_{Y \cdot X}$, the next step is to calculate the standard error of a predicted score for a new case (denoted as s_{N+1}) (Crawford & Howell, 1998; Howell, 2002). When there is only one predictor variable the required formula is

$$s_{N+1} = s_{Y \cdot X} \sqrt{1 + \frac{1}{N} + \frac{\left(X_0 - \bar{X}\right)^2}{s_X^2 \left(N - 1\right)}}, \tag{5}$$

where X_0 is the individual's score on the predictor, and the remaining terms have been defined previously. Note that, once again, all the necessary quantities are available from summary statistics.

The standard error of a predicted score for a new case (that is, a case not in the sample used to build the equation) captures the uncertainty associated with estimating a population slope from a sample. It can be seen that the s_{N+1} will increase in magnitude the further the patient's score on the predictor variable (X) is from the mean; this is also a consequence of the uncertainty in estimating the slope. Any error in estimating β will exert a more marked effect on extreme values of X; to visualize this effect imagine rotating the regression line around the mean of X (Crawford & Howell, 1998).

Next, consider

$$\frac{Y_0 - \hat{Y}}{s_{N+1}}, \tag{6}$$

where Y_0 is the individuals obtained score and $\hat{Y}$ is the individual's predicted score, to yield a standardized discrepancy between the obtained and predicted score. Under the null hypothesis, that the discrepancy is an observation from the population sampled to build the equation, this quantity has a *t*-distribution on $N - 2$ df (Crawford & Garthwaite, 2006a). Thus, for a specified level of alpha (e.g., 0.05), one can test whether there is a statistically significant difference between the predicted score and the obtained score, using either a one or two-tailed test[1].

Opinions differ on the use of one- versus two-tailed tests in neuropsychological assessment. Our view is that one-tailed tests are acceptable and appropriate if the neuropsychologist wishes to test a directional hypothesis, particularly when a discrepancy in the opposite direction would not be the result of a clinical condition. When working with individual cases, power to reject the null hypothesis (in this context to reject the hypothesis that the patient's discrepancy is an observation from discrepancies in the reference population) is often low (Crawford & Garthwaite, 2006b): use of a one-tailed test provides higher power than a two-tailed test. However, the logic of a one-tailed test is such that the direction of a difference must be specified *a priori* and, if a difference is in an unexpected or counterintuitive direction, then it must be ignored. As such unexpected findings may be clinically informative, some neuropsychologists may prefer to keep their options open and use two-tailed tests. The programs that accompany this paper (see later) provide both one and two-tailed values and thereby allowing clinicians to choose which values to

[1] Note that, as pointed out by a reviewer, this test can be construed in terms of the studentized deleted residual which is equivalent to asking whether, if that individual had been included in the original sample, would a separate dummy code parameter for the individual be necessary.

attend to based on their general preferences or the specific problem in hand.

Note that there is a potential "third way" with respect to this issue. Many of us are used to thinking in terms of making a dichotomous decision between a one or two-tailed test. However, an alternative approach is to use what has been termed a split-tailed test (Harris, 1994). The rationale underlying the split-tailed test is that, having selected a Type I error rate (i.e. alpha) we regard as acceptable (5% by convention) we need not "spend" it by allocating it equally to both tails or solely to one tail. Rather, we can choose to allocate a greater proportion of it to the tail corresponding to our directional hypothesis whilst still leaving a proportion to cover the opposite tail. For example, we may decide to allocate 80% to the former tail and 20% to the latter. The loss of power for such a split-tailed test is only a little lower than a one-tailed test (a *p* value less than 0.04 rather than less than 0.05 is required to be considered significant) but, unlike the latter test, the researcher or clinician need not ignore a large difference in the direction opposite to that expected, provided it is significant beyond the 0.01 level.

To our knowledge, split-tailed tests have not previously been used or advocated for inferences concerning individual cases but they may be worthy of consideration. In practice they are easily applied, that is, the user need simply attend to the one-tailed *p* value from the test. That is, if the difference is in the predicted direction and the one-tailed p is < 0.04 then the difference is considered to be statistically significant at $p < .05$; if the difference is in the direction opposite to that predicted and the one-tailed p is < 0.01 then the difference would also be considered statistically significant at $p < .05$. Just as is the case with use of a one-tailed test, it must be stressed that the user should take a principled approach to such testing – the proportion of alpha allocated to each tail should be decided before the data are

observed and thereafter should be regarded as fixed.

Significance tests have a role to play in neuropsychological assessment of the individual: when a discrepancy achieves statistical significance the neuropsychologist can be confident that is unlikely to be a chance finding (in other words it is unlikely that the observed discrepancy stems from random variation in an individual or error in estimating the population regression equation from sample data). However, it should be borne in mind that significance levels (whether they be one-, two-, or split-tailed) are largely arbitrary conventions; the conclusion drawn when a patient's discrepancy is just above the significance level threshold should be similar to the conclusion when it is just below that threshold. Thus we suggest that the neuropsychologist should be primarily concerned with the more general issue of the degree of abnormality of the patient's discrepancy.

Fortunately, an estimate of the abnormality of the discrepancy is an inherent feature of the proposed method: the *p* value used to test significance is also a point estimate of the proportion of the population with the same value on the predictor variable (i.e., *X*) as the patient that would obtain a discrepancy more extreme than that observed for the patient (Crawford & Garthwaite, 2006a). As noted, the population referred to here is that sampled to build the regression equation; i.e., if the equation was built using healthy adults then the population is the healthy adult population. Alternatively if, for example, the equation was built in a sample of patients who had suffered a severe traumatic brain injury (TBI) six months earlier, then the population is all patients with a severe TBI six months post injury.

In Crawford and Garthwaite (2006a) the fact that the *p* value from the significance test also equals the estimated proportion of the population exhibiting a more extreme discrepancy than the case was stated without proof. In the present

paper we provide a short formal proof in Appendix 1. It is more convenient (and more in line with convention) to multiply the *p* value referred to above by 100 so that we have a point estimate of the *percentage* (rather than proportion) of the population exhibiting a larger discrepancy. This latter index of abnormality is used in the examples that follow and in the outputs from the computer programs that accompany this paper.

The above quantity is a *point* estimate of the abnormality of the discrepancy between an individual's obtained and predicted score. Recently, Crawford and Garthwaite (2006a), building on earlier statistical results from Crawford and Garthwaite (2002), have provided a method of obtaining an *interval* estimate for this quantity. That is, the method provides 95% confidence limits on the percentage of the population that would obtain a more extreme discrepancy than that observed for the individual.

The provision of these confidence limits is in keeping with the contemporary emphasis in psychological assessment and statistics on the utility of confidence limits (APA, 2001; Wilkinson & APA Task Force on Statistical Inference, 1999). Confidence limits serve the useful general purpose of reminding the user that there is always uncertainty attached to an individual's results (i.e., they counter any tendency to reify the observed scores) but they also serve the specific purpose of quantifying this uncertainty (Crawford & Garthwaite, 2002). The calculations involved in obtaining these limits involve non-central *t*-distributions and are complex but the important point for present purposes is that they can be calculated without requiring the sample's raw data.

Confidence limits on the abnormality of an individual's discrepancy are implemented in the computer program that accompanies this paper and an example of

their use is provided in a later section. For full details of the derivation and evaluation of these limits, and of the calculations required to obtain them, see Crawford and Garthwaite (2006a).

Recovering a correlation from summary data when it is not reported

In the foregoing section it was assumed that the summary data for a sample included the correlation between the predictor and criterion variables. It might be thought that, if the correlation is not reported then all is lost, i.e., the regression equation cannot be built. However, it is possible to recover this correlation if the study has reported the results of a paired samples *t*-test to compare mean scores on the two variables. For example, in a test-retest study, even if the correlation between test and retest was not reported, it will often be the case that a *t*-test was conducted to test whether there had been a significant decline or improvement over the intervening period. In this situation the correlation can be recovered using simple algebra.

The formula for a paired-samples *t*-test is

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2 + s_Y^2 - 2 s_X s_Y r_{XY}}{N}}}, \tag{7}$$

where all terms have previously been defined. Rearranging this formula to solve for r_{XY} gives

$$r_{XY} = \frac{s_X^2 + s_Y^2 - N\left(\frac{\bar{X} - \bar{Y}}{t}\right)^2}{2 s_X s_Y} \tag{8}$$

and the correlation can then be used to obtain the regression equation and its associated statistics as set out in the previous section. Note that this formula should not be confused with formulae used to convert a *t* value to a point biserial correlation;

the latter provide an index of effect size and represent the correlation between occasion (before versus after) or group membership (e.g., control versus clinical) and test scores).

The principal computer program that accompanies this paper (regbuild.exe) builds the regression equation directly from the correlation coefficient. However, a companion program (regbuild_t.exe) is available to build the regression equation from a *t*-value.

If a study has used an ANOVA rather than a *t*-test to compare the two means, it is still possible to build an equation either by hand or by using this latter program. $F = t^2$, so all that need be done is to take the square root of *F* and enter it into formula (8) in place of *t*.

The precision of regression equations and their associated statistics when computed from summary data

As noted, the main purpose of this paper is to set out a procedure for building and using regression equations when only basic summary statistics are available for a suitable sample. This raises the question of whether rounding errors contained in these summary statistics pose a threat to the validity of the results obtained. The main threat to precision comes from calculating the slope of the regression line from the reported correlation coefficient for the sample (note however that the correlation also features in the formula for $s_{Y \cdot X}$ and hence also influences s_{N+1}). This is because the correlation is intrinsically more influential in the calculations but also because the other statistics required (i.e., means and standard deviations) are normally reported with three or more significant digits.

In contrast, the correlation coefficient is normally reported with only two

significant digits, i.e., .72. Thus, because of rounding error, the actual sample correlation coefficient in this example could be as low as 0.715 and would have its upper limit at .725. Supposing, for convenience, that the standard deviations of *X* and *Y* were both 1 (so that the slope is identical to the correlation coefficient) then the slope calculated from the sample raw data could be as low as .715 or as high as .725 rather than 0.72.

At the margins this rounding error may make a difference to the results obtained for an individual. If there were grounds for concern then, given that the range of error introduced by rounding is known, the calculations could be repeated substituting the upper and lower limits on the value of the correlation coefficient. The results could then be compared with those obtained originally to examine whether they make a difference to the substantive conclusions.

Admittedly this is not a particularly appealing prospect were the calculations to be done by hand or calculator but, using the computer program developed to accompany this paper, it would take seconds to run this check (i.e., by substituting the upper or lower limit on the correlation attributable to rounding error for the correlation as reported in the relevant data field). Before leaving this topic, it should be borne in mind that errors arising from a lack of precision are not limited to equations built from summary data. Such errors can also arise when using published regression equations if the slope and standard error of estimate etc are reported with a limited number of significant digits.

Computer programs implementing these methods

A computer program for PCs is available that implements the methods covered in the present paper. The program (regbuild.exe) prompts the user for the sample

means and standard deviations of the predictor and criterion variables, the correlation between these variables, and the *N* for the sample.

The output consists of the regression equation, i.e., the slope (*b*) and intercept (*a*); and the standard error of estimate. These outputs are followed by the results obtained from analyzing the individual case's data. These consist of: the individual's predicted score; the standardized discrepancy between the predicted and obtained scores; the results of the significance test (one and two-tailed probabilities are provided); the point estimate of the percentage of the population that would obtain a larger discrepancy; and (by default) the 95% confidence limits on this percentage (alternatively, a one-sided upper or lower 95% limit can be requested). The results can be viewed on screen, printed, or saved to a file. There is also the option of entering user notes (e.g., to keep a record of the source of the summary data or further details of the sample or individual case); these are reproduced in the output from the program.

For convenience, the summary statistics for the sample used to build the equation are saved to a file and reloaded when the program is re-run. Therefore, when the program is used with a subsequent individual, the required data for the new case can be entered, and results obtained, in a few seconds. The program has the option of clearing the sample data to allow the user to build a new equation if required. As noted previously, a further program, regbuild_t.exe, is designed for use when a study fails to report the correlation between predictor and criterion variable but does report a *t* value obtained from performing a paired samples *t*-test to compare the means of the two variables. The functionality, analysis and output from this program are identical to regbuild.exe except that it also reports the estimated correlation between the predictor and criterion variables obtained using formula (8).

Compiled version of these programs can be downloaded (as a zip file) from the online supplemental page that accompanies this paper or from the following website address: http://www.abdn.ac.uk/~psy086/dept/regbuild.htm.

Examples of the use of the methods and accompanying programs

In this section we illustrate some ways in which the methods and accompanying computer programs can harness summary data from published studies in order to assist the neuropsychologist draw inferences concerning the cognitive status of individual patients. Suppose that a neuropsychologist has seen a 65 year old patient with suspected early Alzheimer's disease and has administered a semantic (category) fluency test at the initial assessment and again after five months; further suppose that an initial letter fluency test (e.g., FAS) was also administered at the first assessment. The patient's score on the semantic fluency test at initial testing was 21 and the score at retest was 11; the patient's FAS score at initial testing was 25.

Table 1 sets out details of four hypothetical studies: for each study it lists the summary data required to build regression equations and to calculate the associated statistics for drawing inferences concerning an individual case. The resultant regression equations and their associated statistics, calculated using either the formulas presented in the text or using the accompanying computer programs, are also presented in Table 1 (for clarity a blank row separates these statistics from the preceding statistics required for their computation). Although the accompanying computer programs are designed to be intuitive, the provision of the summary data in Table 1 and the worked examples below allows users to run these examples themselves. This will help users become familiar with the mechanics of the process prior to using the methods with other data.

Study 1 was a study conducted on a sample of healthy participants (age range 50 to 80) on the effects of ageing on neuropsychological test performance; in the course of this study the correlation between age and performance on the semantic fluency test (-0.58) was reported. It can be seen that age exerts a substantial effect on performance on the semantic fluency task.

Study 2 was also a study of cognitive ageing in healthy elderly participants and included among its results the correlation between the semantic fluency test and the FAS test. Study 3 was a longitudinal (test-retest) study of neuropsychological tests including the semantic fluency test, again the sample consisted of healthy elderly participants. Finally, Study 4 was another longitudinal study that included the semantic fluency test but was concerned with cognitive change in a sample of patients with early Alzheimer's disease.

Suppose, as is the case for many neuropsychological instruments, that the normative data for the elderly on the semantic fluency test are modest. These conventional normative data could be supplemented by using the data from Study 1 to build an equation for prediction of individuals' semantic fluency scores from their age. If the predicted score is substantially higher than the patient's obtained score this suggests impaired performance. This is an example of the use of regression to provide continuous norms (Zachary & Gorsuch, 1985) as referred to in the Introduction.

Entering the summary statistics for the sample in Study 1 (i.e., the means and SDs for age and semantic fluency together with the correlation between these variables) into equations (2) and (3) the slope (*b*) and intercept (*a*) of the regression equation are calculated as –0.909 and 99.31. Thus based on the patient's age, his *predicted* semantic fluency score is 40.21. Using equations (4) and (5), the standard

error of estimate for this equation ($s_{Y \cdot X}$) is 10.787 and the standard error for an additional individual (s_{N+1}) is 10.821. The difference between these two quantities is modest in this example because the patient's value on the predictor variable (i.e., their age) is not extreme relative to the sample mean and also because the sample used to build the equation is large; it will be appreciated that this will not always be so.

The raw discrepancy between the patient's obtained semantic fluency score of 21 and predicted score of 40.21 is –19.21. Dividing this discrepancy by s_{N+1} (i.e., applying formula 6) yields a standardized difference of –1.775. Under the null hypothesis this standardized difference is distributed as *t* on $N - 2 = 158$ df (in this case the null hypothesis is that the individual's discrepancy is an observation from the population of discrepancies found in the healthy elderly). Evaluating this *t*-value reveals that the patient's obtained score is significantly below the score predicted from her/his baseline score ($p = 0.0389$, one-tailed).

The point estimate of the abnormality of this discrepancy (i.e., the percentage of the population that would be expected to exhibit a discrepancy larger than that observed) is thus 3.89%. The accompanying 95% confidence interval on the percentage of the population that would exhibit a larger discrepancy than the patient ranges from 2.12% to 6.32%. Thus, in summary: There is a large and statistically significant discrepancy between the patient's predicted and obtained scores. This size of discrepancy is estimated to be unusual in the healthy elderly population and is consistent with impaired performance on the semantic fluency task.

Moving on to Study 2: In neuropsychological assessment much emphasis is placed on the use of intra-individual comparison standards when attempting to detect acquired impairments (Crawford, 2004; Lezak, Howieson, Loring, Hannay, & Fischer, 2004). Comparison of semantic and initial letter fluency performance

provides a good example of such an approach as (a) scores vary widely as a function of an individuals' premorbid verbal ability and thus there are limits to the usefulness of normative comparison standards (Crawford, Moore, & Cameron, 1992), and (b) the two tasks are highly correlated in the general adult population (Henry & Crawford, 2004). Thus, if an individual exhibits a large discrepancy between these two tasks, this suggests an acquired impairment on the more poorly performed task.

In this specific example there is an additional consideration: there is much evidence that semantic fluency performance is more severely disrupted by Alzheimer's disease (AD) than is initial letter fluency. For example, a random effects meta-analysis of a large number of studies of semantic and initial letter fluency in AD versus healthy controls revealed very large effects for semantic fluency coupled with more modest effects on initial letter fluency (Henry, Crawford, & Phillips, 2004). That is, the semantic fluency deficits qualified as differential deficits relative to initial letter fluency. On the basis of such evidence a discrepancy in favor of initial letter fluency over semantic fluency would be consistent with an Alzheimer's process.

One means of examining whether this pattern is observed in the individual case is to use a healthy sample to build an equation to predict semantic fluency from initial letter fluency and to compare the individual's predicted and obtained scores. The regression equation and associated statistics built with the hypothetical data from Study 2 are presented in Table 1. Using this equation, the patient's predicted semantic fluency score is 35.23 (based on his initial letter fluency score of 25) compared to his obtained score of 21. Dividing the raw discrepancy between the obtained score and predicted score (-14.23) by s_{N+1} yields a standardized difference of –1.442. Evaluating this *t*-value reveals that the patient's obtained score is not significantly below the score predicted from his initial letter fluency score ($p = 0.076$,

one-tailed). The point estimate of the abnormality of this discrepancy (i.e., the percentage of the population that would be expected to exhibit a discrepancy larger than that observed) is thus 7.60% and the 95% confidence interval is from 3.94% to 12.63%. In summary, the patient's semantic fluency is considerably lower than expected given his initial letter fluency performance; although not statistically significant at the .05 level, the discrepancy is nevertheless fairly unusual and is consistent with a differential deficit in semantic versus initial letter fluency.

Note that a good case could be made for the use of a two- rather than one-tailed test in this situation. That is, a case may turn out to have a discrepancy favoring semantic fluency over initial letter fluency (a pattern that is liable to be relatively uncommon in AD). Had this occurred in the present case then the logic of hypothesis testing would have precluded testing for the significance of this difference. The two-tailed *p* value in this example is 0.152.

This foregoing analysis also provides a good opportunity to discuss the weight given to null hypothesis significance tests in neuropsychology. When a difference does achieve statistical significance the neuropsychologist can be particularly confident that a problem has been uncovered. However, as noted, we suggest that the principal focus should be on the abnormality of the patient's discrepancy rather than whether it falls below or above the cusp for conventional statistical significance. In this example the discrepancy between initial and semantic fluency tests is fairly uncommon. Thus such evidence should be given weight when arriving at a formulation, particularly when it is consistent with other indications of impairment, such as here, where the patient's semantic fluency score was also lower than expected given his age.

Turning to Study 3: neuropsychologists commonly have to attempt to detect

change in cognitive functioning in the individual case, whether this be to determine whether recovery is progressing following a stroke or TBI, to monitor the positive or negative effects of intervention, or to detect decline in degenerative conditions. In the case of AD, serial assessment plays a particularly important role in diagnosis because the results of testing from a single time period will often be equivocal (Morris, 2004). When test data from two occasions are to be compared, regression provides a useful means of drawing inferences concerning change: the neuropsychologist simply needs to find test-retest data for the measures used in an appropriate sample retested over an interval similar to that of the patient. Study 3 is a hypothetical test-retest study in which a sample of healthy elderly participants (N = 45) were tested on the semantic fluency test and retested after 6 months (this is a slightly longer interval than that for the case but sufficiently close to justify use of the data).

The regression equation and associated statistics obtained from Study 3 are presented in Table 1. Using this equation, the patient's predicted semantic fluency score at retest is 29.29 (based on his initial score of 21) compared to his obtained retest score of 11. Dividing the raw discrepancy between the obtained score and predicted score (-18.29) by s_{N+1} yields a standardized difference of –1.95. Evaluating this t-value reveals that the patient's retest score is significantly below the score predicted from his score on first testing (p = 0.0287, one-tailed). The point estimate of the abnormality of this discrepancy is thus 2.87% and the 95% confidence interval is from 0.27% to 9.88%. In conclusion, the analysis indicates that the patient's performance on semantic fluency has declined over the interval between the two testing occasions. That is, it is unlikely that a member of the cognitively intact elderly population would exhibit this large a decline in performance.

Finally, Study 4 provides an example of the use of an equation built in a

clinical rather than healthy sample. Having found evidence of a decline using the equation built using data from Study 3, the data from Study 4 are used to examine whether or not the change from test to retest is unusual for patients with AD. Note also that, in this last example the study did not report the correlation between test and retest scores for semantic fluency but did report the results form a paired samples *t*-test used to test for a significant difference between mean scores on the two occasions (see Table 1).

There are much published data of this form in neuropsychology (McCaffrey et al., 2000), and it is therefore fortunate that we can still build the equation in the absence of a reported correlation between the predictor and criterion variable. Using formula (8), or using the program that implements it (regbuild_t.exe) we can recover the correlation between the scores at test and retest using the data provided. In this example, the correlation is 0.72 and we now have the information required to build the equation to predict the patient's score at retest score from his initial score (see Table 1 for the regression equation statistics). Had the study used a two-way ANOVA to compare means at test and retest (with each participant as a block) then, as noted earlier, we can still recover the correlation. For this example, the *F* value from the ANOVA for the difference between test and retest would be 22.94 and taking its square root gives us the necessary *t* value for entry into formula (8) or into the program.

Using the equation built using the data from Study 4, the patient's predicted semantic fluency score at retest is 15.44 (based on his initial score of 21) compared to his obtained retest score of 11. Dividing the raw discrepancy between the obtained score and predicted score (-4.44) by s_{N+1} yields a standardized difference of –0.532. Evaluating this *t*-value reveals that the patient's retest score is not significantly below

the score predicted from his score on first testing (p = 0.597, two-tailed). The point estimate of the abnormality of this discrepancy is 29.9% and the 95% confidence interval is from 20.23% to 40.68%. Note that the population referred to on this occasion is the population of patients with early AD, not the general (healthy) elderly population as in the previous examples. Note also that in the present example a two-tailed test is employed: even if the neuropsychologist had independent grounds to believe that the patient's cognitive decline would be atypically rapid for AD, or atypically slow, it is unlikely that they would have sufficient confidence in this to rule out the alternative possibility. In conclusion, the present analysis indicates that, although the patient has shown a decline that is somewhat greater than the average decline found in the AD sample, the indications are that the difference in the patient's performance on the two occasions will nevertheless not be unusual in patients with early AD.

The foregoing example of the use of equations built using data from clinical samples is only one of many potential uses. Indeed, given the vast number of clinical studies in the literature, this process is limited only by the ingenuity of the neuropsychologist and by the time involved in conducting a search for published studies relevant to the question in hand. For example, data such as that in Study 4 could also be used to study the potential effectiveness of a pharmacological (or other form of) intervention in the individual case. That is, in the example, the data were obtained from untreated early AD cases and thus, if a treated early AD patient's score at retest substantially exceeded that predicted by the regression equation (i.e. if the discrepancy was estimated to be unusual among untreated AD cases), this would be consistent with a beneficial effect of the intervention.

Finally, the examples featured have primarily had a clinical focus. However,

the use of regression equations to draw inferences concerning the individual case (whether these equations are provided by a third party, or are built from summary data using the methods set out here) have numerous applications in neuropsychological research. Most obviously there has been a massive resurgence of interest in single-case studies in neuroscience as a means of specifying the functional architecture of cognition (Coltheart, 2001). Although the focus of these studies is on a single-case, the patient's pattern of performance is usually interpreted by referring to control values. The use of regression equations can play a useful role in providing such comparisons (i.e., through analysis of the discrepancy between a patient's obtained scores and those predicted by equations built in the healthy control sample). Note that, as the control samples in single-case studies are normally modest in size (Crawford & Garthwaite, 2005), this is an example of where the exact methods set out here are clearly to be preferred over the approximate method.

As pointed out by a reviewer, the methods can also usefully supplement results obtained in group studies. For example, in a clinical trial of a new drug or other form of intervention, researchers are interested not only in the magnitude of the group differences between the active and placebo groups, but also in the number of individuals that show significant improvement from baseline (or attenuation of decline). The methods set out here could be used to set the criterion for improvement or attenuation in the individual case which in turn could be used as a basis for calculating relative risk (via odds ratios) or numbers-needed-to-treat.

<u>Advantages of the proposed method of drawing inferences concerning an individual's discrepancy over existing methods</u>

As noted, regression equations are already widely used to draw inferences

concerning an individual's neuropsychological status. Currently, the "standard" approach in neuropsychology to analyzing the discrepancy between an individual's predicted and obtained score is simply to divide the discrepancy by the standard error of estimate for the equation, treat the resulting quantity as a standard normal deviate (i.e., *z*), and convert this quantile to a probability using a table of areas under the normal curve (Crawford & Garthwaite, 2006a). Just as is the case with the approach presented here, it is possible to apply this method when only summary data are available; i.e., a pre-existing equation is not required. It is therefore appropriate to briefly contrast the two approaches.

In essence the alternative method treats the sample used to build the equation as if it were the entire population; that is the regression statistics are treated as population parameters rather than sample statistics. Monte Carlo simulations have demonstrated that, in contrast to the method used in the present paper, the Type I error rate is inflated above the nominal rate and, relatedly, the abnormality of discrepancies between individuals' obtained and predicted scores is exaggerated (Crawford & Garthwaite, 2006a). These effects can be marked when the sample used to build the equation is modest in size and when an individual's score on the predictor variable is extreme. This latter phenomenon is a direct consequence of failing to allow for the uncertainty in estimating the slope of the population regression line from a sample: as noted earlier, error in estimating the slope will have a greater effect on extreme values.

It should be acknowledged, however, that the two methods will yield very similar results when the sample used to build the regression equation has a moderately large *n*, particularly when the individual's score on the predictor variable is not extreme (see the worked example for Study 1 in which s_{N+1} and $s_{Y \cdot X}$ are very similar

in magnitude). As noted by a reviewer, in such circumstances the $s_{Y \cdot X}$ method could be used as in place of the technically correct method presented here. On the other hand, if the computer programs that accompany this paper are used, there is little to be gained by such a substitution as the correct calculations can be applied in much less time than would be required for hand calculation of the approximate method.

A further difference between the two methods is that, unlike the method advocated here, the alternative method cannot provide confidence limits on the abnormality of an individual's discrepancy: it does not acknowledge any uncertainty in the sample's regression statistics and therefore cannot quantify the effects of this uncertainty on the estimate of the individual's level of abnormality. This can be seen as a significant disadvantage given that (a) confidence limits remind us that our data are fallible and quantify the effects of this fallibility and (b) many bodies, such as the APA (Wilkinson & APA Task Force on Statistical Inference, 1999), stress the importance of incorporating such limits into research and practice.

Multiple regression equations from summary data?

Compared to equations with a single predictor, multiple regression equations provide a more flexible and potentially more sensitive means of testing hypotheses concerning an individual. For example, when testing for change in an individual's neuropsychological functioning, if age is related to the magnitude of practice effects, then age can be incorporated into an equation along with the initial test score to obtain a more precise estimate of an individual's expected change.

Although the computations involved are considerably more complex than those involved in the bivariate case, it is possible to build multiple regression equations from summary data alone. However, there is a major impediment to

actually carrying this out in practice: published studies rarely report the full matrix of correlations between the criterion and predictor variables but this matrix is required to build the equation (the means and standard deviations for all variables are also required but these are usually available).

The studies most likely to include a full correlation matrix are those reporting the results of a multiple regression analysis. For example, the American Psychological Association recommends that such studies report the full matrix (e.g., see APA 2001) but of course there is no need to calculate the coefficients of the regression equation from summary data in these circumstances as they will already be available.

A further practical difficulty is that, even if the correlation matrix is available, the precision with which the correlations are reported would be much more of a problem in using multiple regression than it is in the bivariate case (Sokal & Rohlf, 1995). In view of these practical considerations we have not set out the methods for calculating and analyzing the results of multiple regression equations using summary data, nor have we implemented these methods in a program.

Finally, note that, when a multiple regression *is* available (i.e., from a published study or from a psychologist's analysis of her/his own raw dataset), then it is possible to apply methods for analyzing the discrepancy between an individual's obtained and predicted score that are directly analogous to those covered here for the bivariate case. These methods have been set out by Crawford and Garthwaite (2006a) and have been implemented in an accompanying computer program (this program requires direct entry of the intercept and the regression coefficients for each predictor variable).

Caveats on the use of these methods and some pragmatic considerations

The validity of inferences made using the methods set out here is dependent on the quality of the data used to build the equation; that is, the methods will not provide accurate results if the assumptions underlying regression analysis have been violated (see Tabachnick & Fidell, 1996 for a succinct treatment of this topic). For example, one assumption underlying the use of linear regression is that of homoscedasticity of the residuals. If the size of the residuals increases as scores on the predictor variable increase (as indicated by a fan-like appearance on a scatterplot) then this assumption would be violated. Another assumption is that the relationship between the predictor and criterion variable is linear.

In the case of regression equations published in peer reviewed journals or in test manuals, it is probable (but not guaranteed) that these threats to validity will have been identified (by examination of residual plots and so forth) and rectified or ameliorated (e.g., by transforming the *Y* variable in the case of heteroscedasticity). In the absence of the raw data such strategies are not possible.

These concerns need to be balanced by two pragmatic considerations. First, with many of the combinations of predictor and criterion variables likely to be employed in practice there is little evidence that heteroscedasticity and / or non-linearity is a pervasive problem. For example, if the predictor and criterion variables are both standardized psychological tests (as is the case when attempting to infer change from test to retest or when comparing an estimate of an individual's premorbid functioning with their current functioning) such problems do not appear to be very common.

Second, in an ideal world, neuropsychologists would have access to regression equations that had been built using large samples and had been carefully evaluated.

However, it is clear that the number of such equations is very limited in comparison to the wide variety of hypotheses that neuropsychologists may wish to test. Therefore, in the absence of an existing equation, and when relevant summary data are available, the approach suggested here needs to be contrasted with the alternatives open to the neuropsychologist. These are that the neuropsychologist will either simply ignore the existence of such data despite its relevance to the assessment question, or will attempt to use the data informally to generate a "guesstimate". For example, in the latter case the reasoning might proceed along the following lines: "given that this test appears to have fairly high test-retest reliability and is subject to a moderate practice effect, the difference between this individual's scores looks fairly unusual". It is well known that our subjective estimates of such probabilities are not very accurate and are prone to systematic biases (Beach & Braun, 1994; Tversky & Kahneman, 1971); for example, we typically underestimate the magnitude of the differences expected by chance.

Finally, although it has been demonstrated that it is relatively straightforward to build an equation from summary data (particularly if the computer programs accompanying this paper are used), we also hope this paper will encourage researchers to consider including regression equations in their papers. By doing so they would provide a means whereby their results can be directly applied by other neuropsychologists to draw inferences in the individual case.

References

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

Beach, L. R., & Braun, G. P. (1994). Laboratory studies of subjective probability: A status report. In G. Wright & P. Ayton (Eds.), Subjective probability. Chichester, UK: Wiley.

Chelune, G. J. (2003). Assessing reliable neuropsychological change. In R. D. Franklin (Ed.), Prediction in forensic and neuropsychology: Sound statistical practices (pp. 123-147). Mahwah, NJ: Lawrence Erlbaum.

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base rate information. Neuropsychology, 7, 41-52.

Coltheart, M. (2001). Assumptions and methods in cognitive neuropsychology. In B. Rapp (Ed.), The handbook of cognitive neuropsychology (pp. 3-21). Philadelphia: Psychology Press.

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

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. Neuropsychologia, 40, 1196-1208.

Crawford, J. R., & Garthwaite, P. H. (2005). Evaluation of criteria for classical dissociations in single-case studies by Monte Carlo simulation. Neuropsychology, 19, 664-678.

Crawford, J. R., & Garthwaite, P. H. (2006a). Comparing an individual's predicted test score from a regression equation with an obtained score: a significance test and point estimate of abnormality with accompanying confidence limits. Neuropsychology, 20, 259-271.

Crawford, J. R., & Garthwaite, P. H. (2006b). Methods of testing for a deficit in single case studies: Evaluation of statistical power by Monte Carlo simulation. Cognitive Neuropsychology, 23, 877-904.

Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. Journal of Clinical and Experimental Neuropsychology, 20, 755-762.

Crawford, J. R., Moore, J. W., & Cameron, I. M. (1992). Verbal fluency: A NART-based equation for the estimation of premorbid performance. British Journal of Clinical Psychology, 31, 327-329.

Franzen, M. D., Burgess, E. J., & Smith-Seemiller, L. (1997). Methods of estimating premorbid functioning. Archives of Clinical Neuropsychology, 12, 711-738.

Heaton, R. K., & Marcotte, T. D. (2000). Clinical neuropsychological tests and assessment techniques. In F. Boller & J. Grafman (Eds.), Handbook of neuropsychology (2nd ed., Vol. 1, pp. 27-52). Amsterdam: Elsevier.

Henry, J. D., & Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance following focal cortical lesions. Neuropsychology, 18, 284-295.

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. Neuropsychologia, 42, 1212-1222.

Howell, D. C. (2002). Statistical methods for psychology (5th ed.). Belmont,

CA: Duxbury Press.

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.

McCaffrey, R. J., Duff, K., & Westervelt, H. J. (2000). Practitioner's guide to evaluating change with neuropsychological assessment instruments. New York: Kluwer.

Morris, R. G. (2004). Neuropsychology of older adults. In L. H. Goldstein & J. E. McNeil (Eds.), Clinical neuropsychology: A practical guide to assessment and management for clinicians (pp. 345-358). Chichester: Wiley.

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

Sherman, E. M. S., Slick, D. J., Connolly, M. B., Steinbok, P., Martin, R., Strauss, E., Chelune, G. J., & Farrell, K. (2003). Reexamining the effects of epilepsy surgery on IQ in children: Use of regression-based change scores. Journal of the International Neuropsychological Society, 9, 879-886.

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

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.

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics (3rd ed.). New York: Harper Collins.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four

models. Journal of the International Neuropsychological Society, 5, 357-369.

Tversky, A., & Kahneman, D. (1971). The belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

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

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. Journal of Clinical Psychology, 41, 86-94.

Acknowledgements

We are grateful to Dr Sytse Knypstra of the Department of Econometrics, University of Groningen, The Netherlands, for providing an algorithm that finds the non-centrality parameter of a non-central *t*-distribution given a quantile, its associated probability, and the degrees of freedom. This algorithm is incorporated in the computer programs that accompany this paper.

Table 1. Illustrative examples of the use of summary data from published studies to draw inferences concerning an individual patient: summary data from four hypothetical studies are presented, together with the statistics for the resultant regression equations built using either the formulas included in the text or accompanying computer programs

	Study 1	Study 2	Study 3	Study 4
Predictor	Age	FAS	SF Time 1	SF Time 1
Predictor mean	63.80	36.60	44.3	22.3
Predictor SD	8.42	12.50	12.8	11.5
SF Mean	41.30	43.4	48.1	16.4
SF SD	13.20	13.14	13.6	11.8
Correlation (*r*)	-0.58	0.67	0.76	-
Paired *t* value	-	-	-	4.79
Sample size (*N*)	160	120	45	50
Slope (*b*)	-0.909	0.708	0.808	0.740
Intercept (*a*)	99.31	17.51	12.33	-0.094
$s_{Y \cdot X}$	10.787	10.110	8.941	8.264
s_{N+1}	10.821	9.872	9.367	8.346

Note: SF = semantic fluency. Note: unlike the other statistics for the regression equation, s_{N+1} is not a fixed quantity as it is partly determined by the extremity of an individual's score on the predictor variable.

Appendix 1

A proof that, for Crawford and Garthwaite's (2006) method, the estimated proportion of the population[2] that exhibit a larger discrepancy in the same direction as a case equals the one-tailed *p* value of the significance test

Let d^* be the discrepancy between the obtained score (Y_0) and predicted score ($\hat{Y}$) for the case. Then the proportion of controls who have a discrepancy that is both in the same direction as the case and larger than that of the case is

$$\Pr\left(d > d^*\right), \tag{9}$$

where *d* is the discrepancy of a control.

Dividing both quantities by s_{N+1} (defined in equation (5)),

$$\Pr\left(d > d^*\right) = \Pr\left(\frac{d}{s_{N+1}} > \frac{d^*}{s_{N+1}}\right). \tag{10}$$

Now

$$\frac{d}{s_{N+1}}$$

has a *t*-distribution on $n-2$ df , so that

$$\Pr\left(d > d^*\right) = \Pr\left(t_{n-2} > \frac{d^*}{s_{N+1}}\right). \tag{11}$$

Also, the test statistic for testing if d^* is from the same normal distribution as the control *d*'s, is

$$\frac{d^*}{s_{N+1}}, \tag{12}$$

and this is compared with a *t*-distribution on $n-2$ df . Comparison of equations

[2] Note that, this proportion is the proportion of the population with the same score on the predictor variable as the case

(11)and (12) shows that $\Pr\left(d > d^*\right)$ is equal to the *p* value for the one-tailed test.

Because the present paper is concerned with the use of regression equations built form summary data and, as noted, this is not practical in the vector case (i.e., when there is more than one predictor variable), the proof is couched in terms of the bivariate case. However, it is easily extended to the vector case by simply substituting $n-k-1$ df for $n-2$ df in the relevant equations, where *k* is the number of predictor variables.